Elephant Genomes Reveal Insights into Differences in Disease Defense Mechanisms between Species

Marc Tollis[1,2*], Elliott Ferris[3*], Michael S. Campbell[4], Valerie K. Harris[2,5], Shawn M. Rupp[2,5], Tara M. Harrison[2,6], Wendy K. Kiso[7], Dennis L. Schmitt[7,8], Michael M. Garner[9], C. Athena Aktipis[2,10], Carlo C. Maley[2,5], Amy M. Boddy[2,11], Mark Yandell[4], Christopher Gregg[3], Joshua D. Schiffman[2,12,13], Lisa M. Abegglen[2,12]

[1]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ; [2]Arizona Cancer Evolution Center, Arizona State University, Tempe, AZ; [3]Department of Neurobiology and Anatomy, University of Utah, Salt Lake City, UT 84132-3401, USA; [4]Department of Genetics, University of Utah, Salt Lake City, UT, USA; [5]Center for Biocomputing, Security and Society, Biodesign Institute, Arizona State University, Tempe, USA; [6]Department of Clinical Sciences, North Carolina State University, Raleigh, NC, 27607 USA; [7]Ringling Bros Center for Elephant Conservation, Polk City, Florida USA; [8]William H. Darr College of Agriculture, Missouri State University, Springfield, Missouri, USA; [9]Northwest ZooPath, Monroe, WA 98272; [10]Arizona State University, Department of Psychology, Tempe, AZ; [11]Department of Anthropology, University of California, Santa Barbara CA, USA; [12]Department of Pediatrics & Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA; [13]PEEL Therapeutics, Inc., Salt Lake City, UT, USA & Haifa, IL

*Authors contributed equally

Authors for correspondence: Email: marc.tollis@nau.edu and lisa.abegglen@hci.utah.edu

Abstract

Disease susceptibility and defense are important factors in conservation, particularly for elephants. We report that in addition to endotheliotropic herpesvirus and tuberculosis, Asian elephants are also more susceptible to cancer than African elephants. To determine mechanisms underlying elephant traits including disease resistance, we analyzed genomic datasets from multiple individuals and species. We report a draft genome assembly for the Asian elephant and an improved African elephant assembly. We found 862 and 1,017 potential regulatory elements in Asian and African elephants, respectively, that are enriched near 5,034 differentially expressed genes in peripheral blood mononuclear cells between the two species. These genes are enriched in immunity pathways, including tumor-necrosis factor which plays a role in the elephant response to endotheliotropic herpesvirus. Some elephant *TP53* retrogenes are being maintained by purifying selection and may contribute to cancer resistance in elephants. Positive selection scans revealed genes that may control tusk development, memory, and somatic maintenance. Our study provides an example of how genomics can inform functional immunological studies, which may improve conservation and medical care for elephants and translate into human therapies.

Introduction

Elephants are charismatic megafauna recognized by their prehensile trunks, ivory tusks, intelligence with long-term memory, and large body sizes[1]. The three extant species, Asian (*Elephas maximus*), African bush (*Loxodonta africana*), and African forest elephant (*L. cyclotis*) are the only surviving members of the proboscidean clade of afrotherian mammals. These three species range in body size from African bush elephants (3,000–6,000kg), Asian elephants (2,700–4,000kg), to the smallest, African forest elephants (~2,000kg)[2]. The evolutionary history of elephants included speciation, population size changes, and extensive gene flow between species[3]. Mammoths (genus *Mammuthus*) went extinct between ~11,000 and ~4,300 years ago[4]. Straight-tusked elephants (genus *Palaeoloxodon*) roamed Eurasia until ~34,000 years ago[5], and at an estimated average body mass of 13,000kg many individuals of *P. antiquus* may have been the largest ever land mammals[2].

All extant elephant species are threatened with extinction due to poaching and habitat loss. Asian elephants are recognized as endangered by the International Union for Conservation Red List, with only ~200 wild individuals in some countries[6], and African elephants are considered vulnerable with only about 400,000 wild individuals after a decrease of ~100,000 individuals between 2007 and 2016[7]. DNA-based studies of elephants have shed light on their deeper evolutionary histories[3] and helped focus elephant conservation priorities[8]. Many of these investigations have focused on a relatively small number of neutral genetic markers. Sampling across the entire genome using next generation sequencing can help conservation efforts by identifying adaptive alleles and phenotype-associated variants[9]. While genomic approaches are useful for elephant conservation, the few elephant functional genomic studies currently available are limited to a small number of individuals and species[10–13] and the etiologies of many elephant traits with profound fitness (and therefore conservation-oriented) consequences have not yet been discovered. For instance, increased frequencies of tuskless elephants may be a response

to selective pressures from ivory poaching[14–16]. The genes controlling tusk development, however, remain unknown.

Disease defense is extremely important for elephants and has important ramifications for species conservation[9]. Asian elephants are threatened by their susceptibility to acute hemorrhagic disease resulting from infection with elephant endotheliotropic herpesvirus (EEHV)[17]. Despite recent reports of African elephant calf fatalities from EEHV, mortality rates are higher for Asian elephants, suggesting a genetic component for increased EEHV lethality. Furthermore, elephant cancer mortality rates are low compared to humans, despite their large size and long lifespans, and the fact that cancer is a body size- and age-related disease[18]. This suggests that elephants evolved more efficient cancer suppression mechanisms relative to smaller mammals. Cancer suppression in elephants has profound ramifications for elephant health, and mechanisms of cancer resistance in elephants could inform human cancer research[18–20]. Recent studies have largely focused on elephant-specific duplications of the tumor suppressor gene *TP53* (or *EP53*). However, the evolution of cancer suppression in elephants likely involved many pathways including additional DNA repair genes[13].

Here, we generated the largest elephant genomic dataset to date. We present a draft genome assembly for an Asian elephant and an updated African bush elephant assembly. Using a comparative genomic approach, we discovered elephant genomic regulatory regions and identified differential gene expression between Asian and African elephants. We also leveraged genome sequences from 16 elephants across three living and two extinct species to estimate genetic changes during elephant evolution, including selective sweeps in living populations. The goals of this study are: (1) to provide novel resources for the mammalian comparative genomics and elephant conservation communities; (2) to place genomic changes underlying elephant traits in the contexts of deeper mammalian and elephant evolution; (3) to understand what drives different disease outcomes between elephant species; and (4) to characterize the evolutionary history of cancer suppression in the elephant lineage.

Results

New reference genomes for elephant genomics

All genome sequence data and assemblies generated for this study has been shared under NCBI Bioproject PRJNA622303. We sequenced an estimated 94.4x coverage of the Asian elephant genome using multiple Illumina sequence libraries (Supplementary Table 1), which were assembled into a 3,126,981,281 bp draft reference genome organized on 6,954 scaffolds, with a scaffold N50 of 2.78 Gb and containing 4.88% gaps (Supplementary Table 2). The assembly is estimated to contain >97% of mammalian single-copy orthologs. We annotated 23,277 protein coding genes on 2,541 scaffolds, with an average gene length of 43,098 bp and an average of 8.54 exons per gene. Up to 53% of the Asian elephant genome assembly was comprised of repeats, including a large proportion of Long Interspersed Nuclear Elements (Supplementary Table 4). We used Hi-C sequencing libraries to improve the African bush elephant genome assembly, increasing the length of the longest scaffold 225 Mb to 240 Mb (Supplementary Table 5). A total of 528 previously unjoined scaffolds were linked and nine scaffolds/chromosomes were broken (Supplementary Fig. 1). We suggest future improvements to these assemblies such as high-resolution genomic profiling with long read and/or optical mapping technologies to further resolve structural differences between elephant genomes.

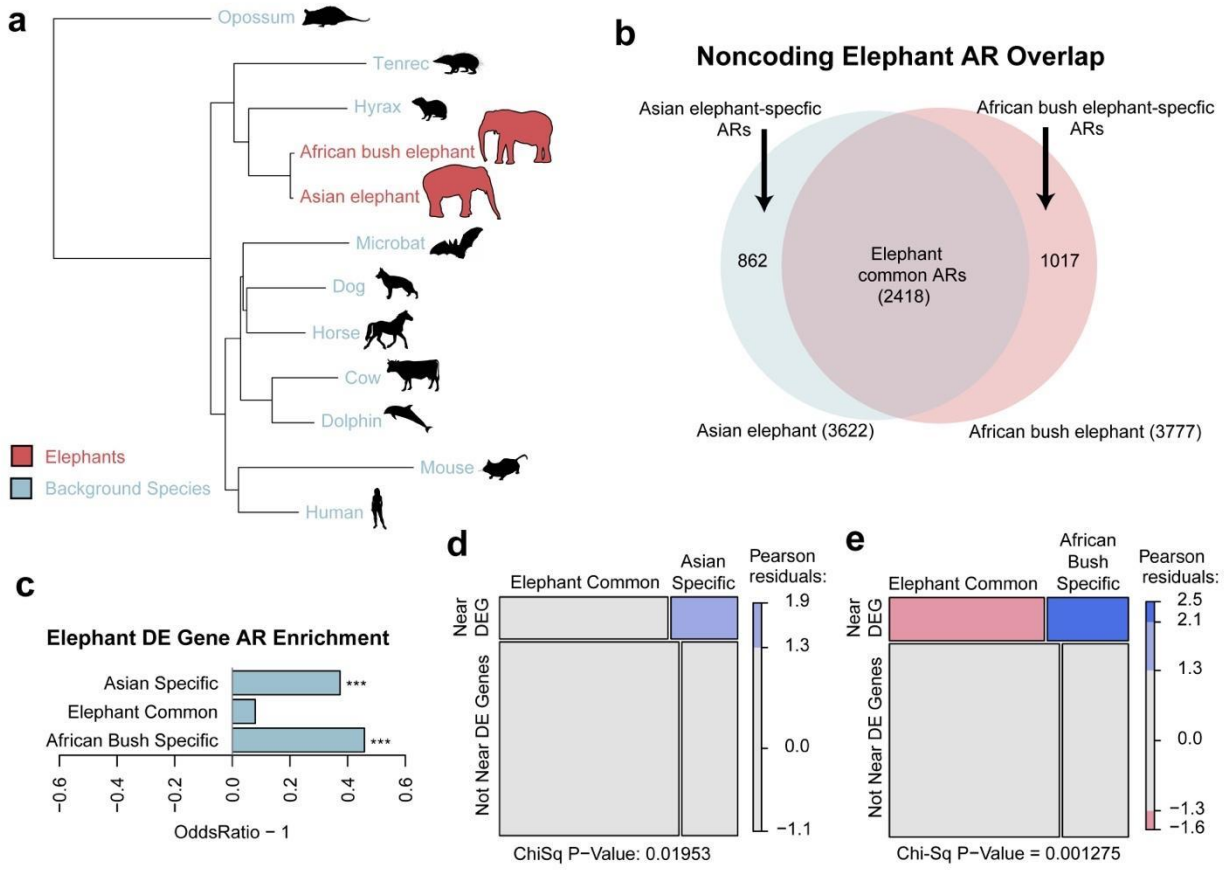
Elephant-specific accelerated regions correlate with species-specific gene expression patterns

Differences between closely related mammals are most often due to changes in non-coding regulatory regions that alter gene expression patterns and transcription factor binding networks[21,22]. To understand how non-coding regulatory regions differ between Asian and African bush elephants, we defined accelerated genomic regions for each species[22]. We first defined 376,899 non-coding regions present in Asian and African elephants and conserved in

10 background mammal species (conserved regions or CRs, Fig. 1a). Of these, 3,622 regions had significantly increased nucleotide substitution rates (accelerated regions or ARs) in the Asian elephant, and 3,777 regions were accelerated in the African bush elephant (q-value < 0.10). We found 2,418 ARs in both species, while 862 are Asian elephant-specific and 1,017 are African bush elephant-specific (Fig. 1b).

Accelerated regions common to Asian and African bush elephants were likely driven by changes pre-dating the evolutionary divergence of the two elephants. In contrast, Asian elephant- and African bush elephant-specific ARs may point to enhancers driving gene expression level changes that impact phenotypes distinguishing the two elephant species. We hypothesized that a disproportionate number of genes near these ARs would be differentially expressed between the two species. Using available African bush elephant and Asian elephant peripheral blood mononuclear cell (PBMC) RNA-Seq data[13,23], we defined 5,034 differentially expressed (DE) elephant genes (false discovery rate or FDR < 0.01). Both Asian elephant- and African bush elephant-specific ARs are significantly enriched near DE genes relative to conserved regions (q-value = 2.05e-4, q-value = 8.30e-7, respectively). Meanwhile, the 2,418 ARs common to both elephants were not significantly enriched near DE genes (Fig. 1c). We used a Chi-squared test to determine if Asian elephant- and African bush elephant-specific ARs disproportionately overlap DE gene regulatory regions relative to the common ARs. Both tests yielded significant p-values, 0.019 and 0.001 (Fig. 1d, Fig. 1e), suggesting that some ARs reflect changes in regulatory regions that alter gene expression patterns in elephant PBMCs.

Figure 1. Elephant accelerated regions. Using a whole genome alignment of 12 mammals (a), we defined genomic regions that were accelerated (ARs) in two elephant species (red), yet conserved in the set of background species (blue). Branch lengths are given in terms of mean substitutions per site at fourfold degenerate sites (neutral model). Among the ARs detected in elephants, we found ARs common to both elephant species as well as ARs specific to either Asian or African bush elephants (b). Differentially expressed (DE) genes were much more likely to be found in species-specific ARs than in common ARs (c). Species-specific ARs disproportionately overlap DE gene regulatory regions relative to the common ARs (d, e).

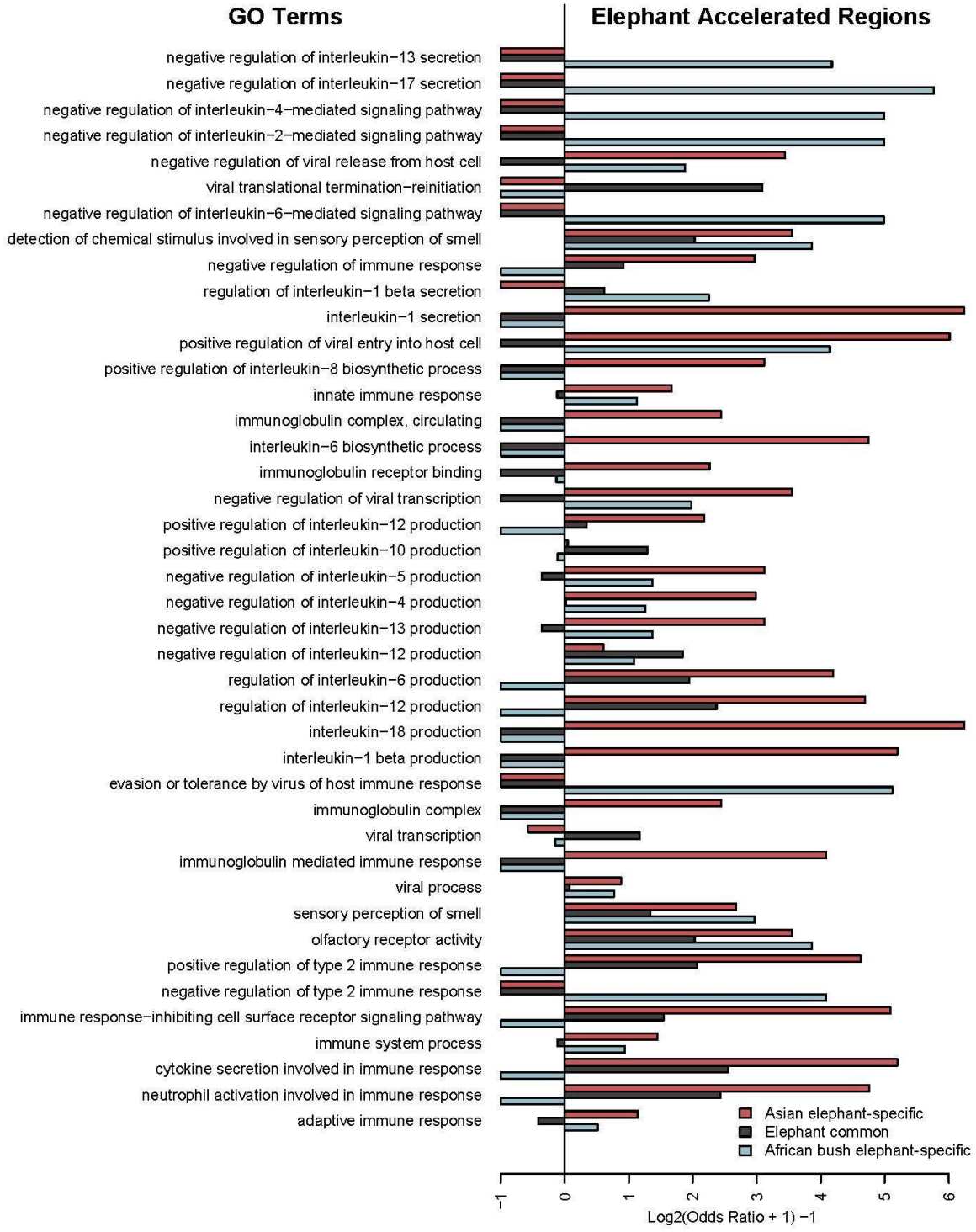


We sought to further explore AR contributions to African and Asian elephant species differences with a gene ontology (GO) enrichment analysis. We tested the elephant species specific and common ARs for Biological Process (BP) GO term enrichments. Of 18,056 terms, 252 were significantly enriched in Asian elephant specific ARs and 275 were enriched in African elephant specific ARs (q-value < 0.05). Many of the GO terms related to the immune system in both species (Fig. 2; Supplementary Fig. 2; Supplementary Fig. 3; Supplementary Fig. 4). The broad term ‘immune system process’, for example, was 4.5 and 2.8 fold enriched with Asian and African elephant-specific ARs (q-value = 4.87e-12, q-value = 7.84e-06, respectively), but not

significantly enriched with elephant common ARs. We found 109 GO terms significantly enriched with elephant common ARs (q-value < 0.05) including a 3.6 fold enrichment of elephant common ARs for the term ‘positive regulation of tumor necrosis factor production’ (q-value = 1.75e-03). In Figure 2 we present the immune and cancer related GO terms significantly enriched with any of the three AR sets. All enrichments are shared as a resource as supplementary data.

The number of African elephant-specific ARs assigned to each gene is well correlated between this study and a previous study (Ferris et al. 2018) (R = 0.82). The gene most enriched with previously defined non-coding African elephant ARs was the DNA repair gene *FANCL* with a 7.3 fold enrichment (q-Value: 2.16e-56). We found that *FANCL* was the third most significantly enriched gene in both African and Asian elephant ARs relative to CRs with 4.6 and 4.9 fold enrichments (Q-Values: 1.27e-14, 4.46e-16). Of 50 African elephant ARs and 51 Asian elephant ARs assigned to *FANLC*, 43 are common to both elephant species suggesting their acceleration predates African-Asian elephant divergence. The similarities between the current and previous studies are consistent with non-coding cis regulatory elements regulating *FANCL* expression for DNA repair as part of elephant cancer resistance adaptation.

Figure 2. Gene ontology terms for enriched biological processes associated with accelerated regions in elephant genomes.



Cancer in elephants

Previous datasets suggest that elephants develop cancer at a lower rate than expected given their long lifespan and large body mass, a phenomenon known as Peto's Paradox[18]. We collected and analyzed pathology data from 26 zoos in the USA that included diagnoses from 76 different elephants (n=35 African and n=41 Asian). We found that 5.71% of the African elephants and 41.46% of the Asian elephants were diagnosed with neoplasia (e.g. benign and malignant tumors), which is a significant difference (Fisher's Exact Test, P=0.0003783; Chi-square test, P=0.0008947) (Table 1, Supplementary Table 5). Overall, the majority of these tumors are benign, with no reports of malignancy in African elephants and 14.63% of Asian elephants diagnosed with malignant tumors (Fisher's Exact Test, P=0.02799; Chi-squared test, P=0.05343). In contrast, the lifetime risk of developing malignant cancer is 39.5% for humans (Surveillance, Epidemiology, and End Results, https://seer.cancer.gov/) and the lifetime risk of developing benign tumors is even higher, with 70%-80% of women developing uterine fibroids (leiomyomas) alone[24]. Asian elephants are reported to have a high prevalence of uterine leiomyomas[25], including seven of the 17 Asian elephants with confirmed neoplasia. Overall, these results confirm that malignant cancer rates in elephants are low. The benign nature of the majority of tumors in elephants further suggests that elephants evolved strong cancer defense mechanisms, and these cancer defense mechanisms may be enhanced in African elephants compared to Asian elephants.

Table 1: Cancer diagnoses and prevalence in African and Asian elephants.

Neoplasia and Malignant Prevalence					
Species	Total Individuals	Neoplasia Cases	Neoplasia %	Malignant Cases	Malignant %
Asian elephant	41	17	41.46	6	14.63
African elephant	35	2	5.71	0	0
Total elephants	76	19	25	6	7.89

Neoplasia Diagnoses in African/Asian Elephants					
Species	Sex	Age	Lesion Type	Lesion Site	Malignant
African elephant	Female	28	Polyp	Vagina	No
African elephant	NA	NA	Adenoma	Parathyroid	No
Asian elephant	Female	45	Polyp	Vulva	No
Asian elephant	Female	50;50	Polyp; Leiomyoma	Uterus;Uterus	No;No
Asian elephant	Female	30;40	Polyp;	Vagina;	No
			Spindle Cell Tumor	Leg	
Asian elephant	Female	39	Leiomyoma	Uterus	No

Asian elephant	Female	39	Mast Cell Tumor	Abdomen	No
Asian elephant	Male	35	Papilloma	Trunk	No
Asian elephant	Female	50	Papilloma	Skin	No
Asian elephant	Female	36	Papilloma	Skin	No
Asian elephant	Female	50	Adenocarcinoma	Breast	Yes
Asian elephant	Female	59;59	Adenocarcinoma; Leiomyoma	Uterus; Uterus	Yes; No
Asian elephant (7)	NA	NA	Adenocarcinoma (2);	Uterus (2);	Yes
			Undifferentiated Malignant Neoplasm (1);	Uterus (1);	
			Leiomyosarcoma (1);	Lung (1);	
			Sarcoma (1)	Liver (1)	
			Leiomyoma (4);	Uterus (4);	No
			Leiomyoma (1);	Stomach (1);	
			Microadenoma (1)	Pituitary Gland (1)	

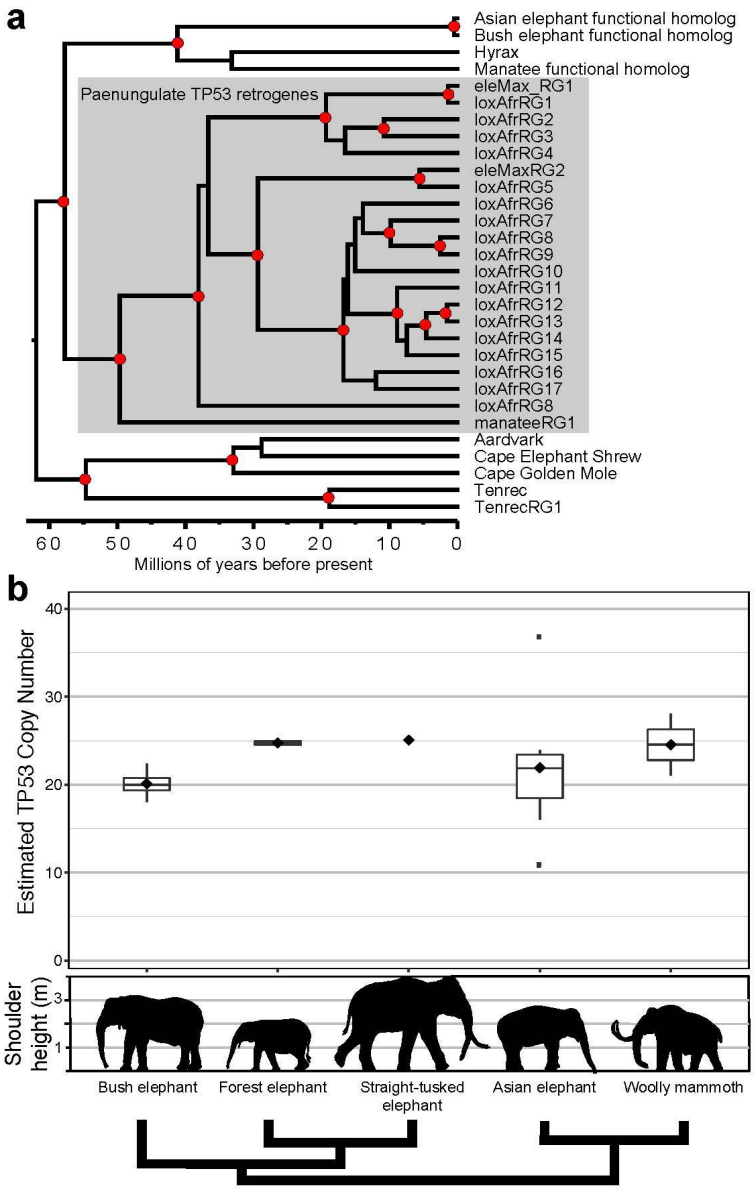
Evolution of TP53 *and its retrogenes (EP53) in elephant genomes*

A potential mechanism for cancer resistance in elephants is an enhanced apoptotic response to DNA damage associated with extensive amplification of elephant *TP53 (EP53)* retrogenes[18,19,26]. By incorporating annotated *TP53* homologs from Asian elephant, African bush elephant, and 41 additional mammals in a molecular clock analysis, we estimated that the *EP53* retrogene families diverged ~35 million years ago (23.51–49.4 95% highest posterior density) (Fig. 3a). The genome of the West Indian manatee (*Trichechus manatus*), a paenungulate afrotherian closely related to proboscideans, contains a retrogene copy of *TP53* as well. Thus, *TP53* retrogenes were likely present in paenungulate ancestors of manatees and elephants, and further expanded in copy number during elephantid evolution.

We obtained whole genome shotgun data from three living and two extinct elephant species from the public databases, and sequenced three additional elephants at ~32–38X coverage (Supplementary Table 6). We mapped these sequences to the African bush elephant genome assembly (loxAfr3.0) with bwa-mem v077[27] (see section on methods). We utilized the loxAfr3.0 version due to its extensive representation in public repositories, including Ensembl (www.ensembl.org). Using normalized read counts to estimate *EP53* copy numbers in the genomes of three living and two extinct elephant species, (Figure 3b; Supplementary Table 7), we found that African bush elephants have on average ~19–23 *EP53* homologs in their genomes, which is similar to previous estimates[18,19,26]. The Asian elephant genomes contain as few as 10 *EP53* copy number variants, or as many as 37, but on average contain ~19–22 *EP53* copies in their genomes. This variation across Asian elephant genomes could be due to factors such as the evolutionary distance between African and Asian elephants which may affect mapping quality, or differences between individual Asian elephants. Regardless, these estimates are similar to previous estimates of *EP53* copy number for Asian elephants based on smaller numbers of individuals[18,19]. We estimated ~21–24 *EP53* copy number variants in forest elephant genomes. The woolly mammoths (*Mammuthus primigenius*) were estimated to have

between 19 and 28 *EP53* copies in their genomes, which was slightly higher than previous estimates[19]. Meanwhile, the straight-tusked elephant genome contained ~22-25 *EP53* copies.

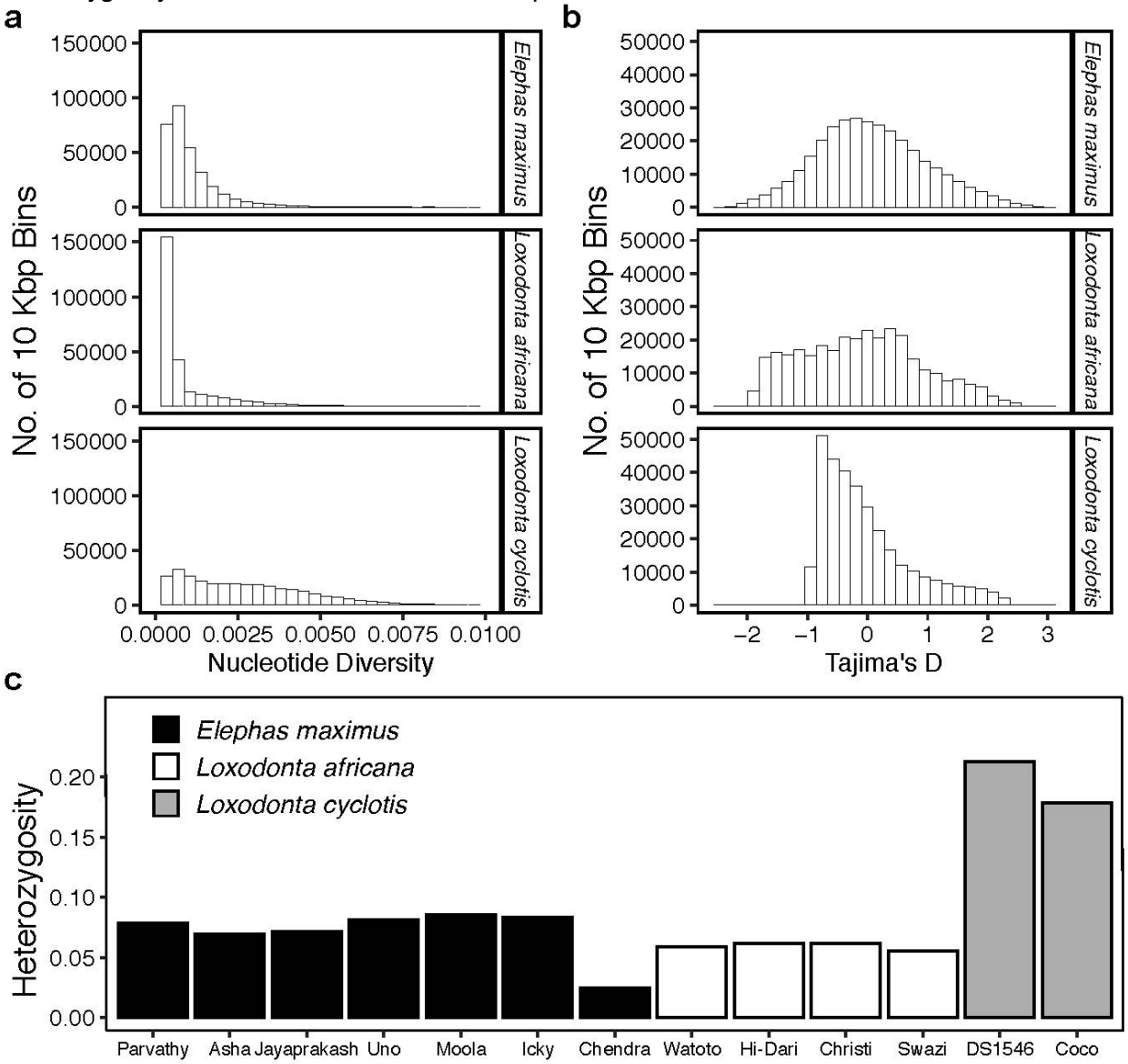
Figure 3. Evolution of TP53 in elephants. (a) Phylogeny of TP53 sequences from afrotherian genomes. TP53 retrogenes appeared early in the evolution of paenungulates ~49 million years ago (MYA), followed by subsequent amplification in the elephant lineage ~35MYA. Red dots indicate estimated nodes with posterior probability ≥90%. (b) TP53 copy number estimates based on read counts from three living elephants (African bush elephant (n=4, minimum 18, maximum 22.4, median 20.1, 25th percentile 19.3, 75th percentile 20.8), forest elephant n=2, minimum 24.3, maximum 25.2, median 24.7, 25th percentile 24.5, 75th percentile 25) and Asian elephant (n=7, minimum 10.9, maximum 36.8, median 21.9, 25th percentile 18.5, 75th percentile 23.4)) and two extinct (straight-tusked (n=1, 25.1) and woolly mammoth (n=2, minimum 21, maximum 28, median 24.51, 25th percentile 22.8, 75th percentile 26.3)) elephant species. Shoulder height estimates from Larramendi et al. (2015). Phylogeny is schematic only and represents relationships from Palkopuolou et al. (2018).



Population genomics of modern elephant lineages

We utilized the aligned genomic sequences from African bush elephants (n=4), African forest elephants (n=2), and Asian elephants (n=7) to call variants using freebayes v1.3.1-12[28] (see section on methods) and genotyped 41,296,555 biallelic single nucleotide polymorphisms (SNPs), averaging one SNP every 77 bases and a genome-wide transition-transversion ratio of 2.46. Altogether, we annotated 290,965 exonic, 11,245,343 intronic, and 32,512,650 intergenic SNPs across the 13 elephant genomes. Among the three living species of elephants, bush elephants averaged the lowest nucleotide diversity (0.0008, s.d. 0.001), followed by Asian elephants (0.0011, s.d. 0.001) and forest elephants (0.002, s.d. 0.002) (Fig. 4a). The distribution of Tajima's D in 10 kb genomic bins calculated for bush elephant revealed an excess of negative values relative to other elephant species (Fig. 4b), indicative of excess low frequency polymorphisms potentially related to a population bottleneck[3]. We also found a larger proportion of heterozygous sites (0.18 and 0.21) in forest elephant genomes compared to all other elephants (Fig. 4c). Asian elephants had an intermediate level of heterozygosity among the three species (mean 0.07, standard deviation 0.007). However, the individual elephant from Borneo contained the smallest number of heterozygous genotyped sites among all elephants (0.03). These results are consistent with recent analyses showing that the Borneo subpopulation of *E. maximus* is genetically isolated, and that forest elephants contain deep genomic divergence[3,29].

Figure 4. Summary population genomics statistics for three living elephant species. We analyzed 10 Kbp bins for (a) nucleotide diversity and (b) Tajima's D for Asian (*Elephas maximus*), bush (*Loxodonta africana*), and forest (*L. cyclotis*) elephants. (c) Per-individual heterozygosity for Asian, bush, and forest elephants.



Genetic variation in EP53 *suggests maintenance of some paralogs by purifying selection*

We analyzed biallelic SNPs called within loxAfr3.0 *EP53* paralogous gene annotations to determine patterns of genetic variation in *EP53* and its retrogenes in modern populations of elephants. Overall, we found little genetic variation in the *EP53* paralogs both within and between the three living elephants (Supplementary Table 8). For instance, the proportion of

polymorphic sites in putatively neutrally evolving ancestral repeats was 0.013, while the average proportion of polymorphic sites across 12 *EP53* paralogous loci was 0.004. We found the most genetic variation of all *EP53* loci in Asian elephants. Meanwhile, despite the deep genomic divergence of forest elephants, we found that *EP53* is conserved in forest elephants, with no SNPs except within three retrogenes which contained one segregating site each. We collected zero nonsynonymous SNPs for the functional homolog (ENSLAFG00000007483). In addition, ENSLAFG00000028299 or ”retrogene 9”, whose protein expression is stabilized by DNA damage in human cells[18], showed no variation across the three modern elephant species.

We annotated variants in 12 *EP53* paralogs based on the bush elephant genome annotation and found few variants affecting gene function (Supplementary Table 9), consisting of mostly missense mutations. There were no variants of high or moderate impact on gene function annotated in the functional homolog (ENSLAFG00000007483), with the majority of variants occurring downstream, in introns, or upstream of the gene. The high degree of sequence conservation across three species of elephant, and in particular the lack of variants with functional effect, especially in “retrogene 9,” suggests that at least some *EP53* retrogenes are being maintained by purifying selection.

An elephant never forgets: positive selection in elephant genomes

To assess the importance of natural selection in elephants, we scanned the genomes of three species (*Elephas maximus*, *Loxodonta africana*, and *L. cyclotis*) for positive selection using SweeD v3.3.1[30,31], hypothesizing that genetic pathways controlling elephant traits would be subjected to selective sweeps. This yielded 24,394 selectively swept outlier regions meeting our statistical thresholds (see Materials and Methods) in Asian elephants, which comprise ~0.07% of the genome and overlapped with 1,611 gene annotations. Out of the 41,204 regions experiencing putative selective sweeps in bush elephants (~1.3% of the genome), we detected

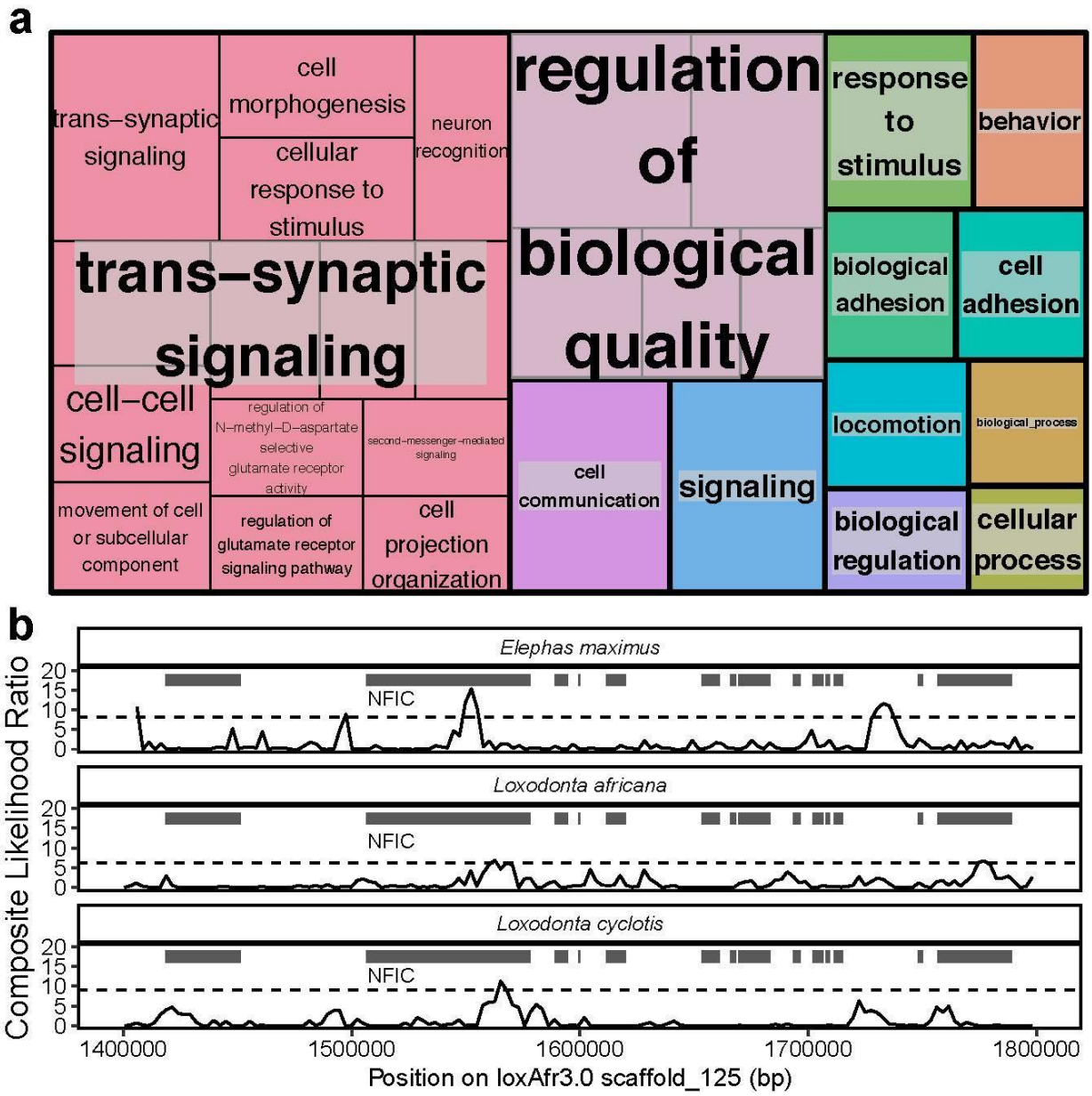
2,882 protein-coding genes. We estimated 4,099 protein-coding genes involved in the 51,249 regions involved in putative selective sweeps in forest elephants (~1.6% of the genome).

We found 242 protein-coding genes that overlapped regions with evidence of positive selection and are shared by all three living elephant species. The most significantly enriched biological processes (in terms of fold enrichment) for these genes were dendrite self-avoidance (27-fold enrichment, FDR=0.02), ionotropic glutamate receptor signaling pathway (17-fold enrichment, FDR=0.02), and regulation of NMDA receptor activity (14-fold enrichment, FDR=0.03), and many significantly enriched GO terms clustered semantically with "trans-synaptic signaling" (Figure 5a). We also found significant protein interactions among outlier genes (75 observed edges versus 45 expected; enrichment p-value=3.68E-5), including an enrichment of genes associated with the glutamatergic synapse KEGG pathway (7 of 98 genes, FDR=0.015). These results suggest strong selection in elephants on pathways involved in memory, learning, and the formation of neural networks. The recognition, storing and retrieving of information in the human brain occurs in the temporal lobe, and the temporal lobes of elephants' brains are relatively larger than those of humans, as well as denser and more convoluted[32]. Retrieving information is likely crucial for elephants to find resources across vast and complicated landscapes[1].

We found that the most significantly enriched mouse phenotypes among genes overlapping selective sweeps were "abnormal upper incisor morphology" (FDR=0.001) and "long upper incisors" (FDR=0.003) represented by two genes: *ANTXR1,* a tumor-specific endothelial marker implicated in colorectal cancer; and *NFIC* (Figure 5b)*,* a dimeric DNA-binding protein involved in both skeletal muscle neoplasms and neuroblastoma, as well as odontogenesis of dentin-containing tooth (GO:0042475). Unlike most tusked mammals which feature elongated canines, the elephant tusk is in fact a highly modified upper incisor[1]. Therefore, our genomic results suggest that positive selection acted on genes involved in elephant tusk formation.

Other significant gene ontology terms from outlier regions in our selective sweep analysis were related to cancer, including cell adhesion (9-fold enrichment, FDR=0.007), cell-cell signaling (3-fold enrichment, FDR=0.01), and cell communication (2-fold enrichment, FDR=0.0001). Significant protein-protein interactions were found associated with EGF-like domain (UniProt keyword enrichment, 13 out of 209 genes, FDR=0.00042; and INTERPRO protein domain enrichment, 13 out of 191 genes, FDR=0.00026). The EGF-like domain contains repeats which bind to apoptotic cells and play a key role in their clearance[33]. Thus, our results suggest ongoing selection in pathways involved with somatic maintenance and in particular apoptosis, a possible key mechanism for cancer suppression in elephants[18].

Figure 5. Selective sweeps in three living elephant species. (a) TreeMap from REVIGO representing semantic clustering of gene ontology biological process terms with a Benjamini-Hochberg false discovery rate of 5% that are associated with genes overlapping selective sweeps common to *Elephas maximus*, *Loxodonta africana*, and *L. cyclotis.* Rectangles represent clusters, and larger rectangles indicate semantically related clusters. Rectangle size reflects corrected p-values. (b) Composite likelihood ratio values in the NFIC region of a genomic scaffold (loxAfr3.0) calculated with SweeD in three elephant species. Gene annotations are represented by dark rectangles; the NFIC gene is indicated. Dashed lines represent p-value threshold of 0.0001.



Discussion

We have leveraged available and *de novo* genomic resources to investigate functional genomic evolution in elephants, presenting a draft assembly for the Asian elephant and an improved assembly for the African bush elephant. By estimating conserved genomic regions in a set of 10 background mammals, we identified elephant-specific accelerated genomic regions associated with significant gene expression differences between Asian and African elephants. These putative regulatory regions are associated with immune function. This may explain observed differential vulnerability to infection and cancer across the two species. We estimated *EP53* copy number and SNP variation in living and extinct elephant species to determine that both neutral and selective forces likely control the segregation of *EP53* alleles. By modeling selective sweeps in living elephant populations, we identified genomic regions under positive selection that may control tusk development, memory and learning, and somatic maintenance related to cancer resistance.

Cancer mortality rates in elephants have been reported using limited elephant necropsy data[18]. Here we investigated cancer prevalence in elephants using elephant biopsy and necropsy data from 26 zoos. We confirm that elephants develop cancer at a lower rate than predicted based on their size and lifespan, i.e. Peto's Paradox, with reports of neoplasia in 19 of the 76 individuals (25%) and malignant cancer reported in 6 of the 76 individuals (7.89%). Tumors that develop in elephants tend to be benign, which may further suggest that elephants have strong cancer defense mechanisms to not only suppress cancer from developing, but also to prevent malignant transformation. Asian elephants develop benign tumors and malignant cancer at higher rates than African elephants; therefore, Asian elephants may benefit from increased monitoring for tumors when under human care as well as in the wild. The collection of additional cancer prevalence data and the genomic analysis of benign and malignant tumors in elephants will be important to continue to understand the evolutionary basis of differences in cancer risk between elephant species. This information could assist with selecting the best

treatment interventions when the rare elephant tumor is diagnosed and benefit the survival of individual elephants. More than half of the tumors reported here were found in reproductive organs. Because even benign reproductive tumors can affect reproduction and conservation, future studies to understand their impact and to develop preventative and treatment measures are needed.

While previous studies suggest that *EP53* copy number increased with body mass during proboscidean evolution as a response to increased cancer risk[19], we estimate some of the highest *EP53* copy numbers in some of the smallest elephants. For instance, while we estimated ~19–21 EP53 copy number variants in the ~44,800 year old woolly mammoth genome from Oimyakon, Russia, we estimated that the much more recent ~4,300-year-old Wrangel Island mammoth had ~1.3X this number of *EP53* copies in its genome. This is remarkably consistent with previous results that estimated an increase in the retention of retrogenes in the genome by 1.3X associated with the demographic decline of the last woolly mammoths on Wrangel Island[12], which by 12,000 years before present shrunk in body size by ~30% relative to more ancient mammoths elsewhere[4]. Therefore, the estimated *EP53* copy number increase in the Wrangel Island mammoth may be related to the random fixation of retrogenes in the population rather than selection acting on body size.

We estimated ~21–24 *EP53* copy number variants in forest elephant genomes, greater than our overall estimates for bush elephants despite the smaller body size of forest elephants. Meanwhile, we estimated ~23–25 *EP53* copy number variants in the genome of the straight-tusked elephant, which was considerably more massive than any extant elephant[2]. Recent genomic evidence suggests that forest elephants are more closely related to straight-tusked elephants than to bush elephants (Fig. 3b)[3,29], and further evidence suggests extensive gene flow may have occurred between forest and straight-tusked elephants, as well as between straight-tusked elephants and other elephantids including mammoths[3]. Palkopoulou et al. (2018) also found that despite hybridization in the modern contact zones between forest and

bush elephants, there is little evidence of historical gene flow between these two lineages. Thus, the possibility exists that, as in the Wrangel Island mammoth, the higher estimated *EP53* copy number in forest elephants relative to bush elephants may not be related to modern differences in body mass between species (and possible increased cancer risk), but instead may be due to complicated evolutionary and demographic histories. However, we also found that genetic variation at some *EP53* retrogenes is conserved in elephant populations, which adds evidence to the potential functionality of at least some *EP53* retrogenes. This suggests that there may be a core set of *EP53* retrogenes that confer the bulk of cancer suppression in elephants, with important implications for research on the translational and therapeutic potential of *EP53*.

We found a significant enrichment for the GO term 'tumor necrosis factor-mediated signaling pathway' (q-value=0.000245) in ARs in common to both Asian and African elephants. Tumor necrosis factor (TNF) is a cytokine involved in cell differentiation and death that can induce the necrosis of cancer cells[34]. The common ancestor of Asian and African elephants was a large-bodied elephantid[1], and large and long-lived species must evolve cancer suppression mechanisms to offset the increased cancer risk associated with a large number of dividing somatic cells and long lifespans[35]. Therefore, it makes sense that there would be accelerated evolution in cancer-related pathways along the elephant lineage after divergence from the small-bodied common ancestor with the rock hyrax (*Procavia capensis*). We noted that, in addition, many Asian elephant- or African elephant-specific ARs were found near genes associated with immunological GO terms. Furthermore, genes near Asian elephant-specific ARs were enriched for the GO term 'negative regulation of tumor necrosis factor-mediated signaling pathway' (q-value = 0.0448). While an enhanced apoptotic response to DNA damage can be a potential mechanism for cancer resistance in elephants[18,19] , increased apoptosis also can be associated with the upregulation of TNF-alpha especially in the setting of EEHV-infected Asian elephant PBMCs[36]. This suggests that regulatory elements may govern immunological differences between the two elephant species. In support of these results, Asian elephant

calves are much more susceptible than African elephant calves to cytokine storm triggered by EEHV infection[17]. This is consistent with our observation that Asian elephants are also more susceptible to both malignancies and neoplasia in general. We also note that, assuming both species have equal rates of exposure, rates of tuberculosis infection are higher in Asian elephants under human care relative to African elephants[37] (Fisher's Exact Test, P=0.0002528; Chi-squared test, P=0.0006843; Supplementary Fig. 5).

Our data for enriched biological processes associated with AR (Fig. 2) support the possibility that innate immunity and enhanced cytokine storm may play a role in the increased infectious susceptibility with inflammatory response of Asian versus African elephants. Compared to African elephants, Asian elephants show increased interleukin-1 beta, interleukin-18, and neutrophil activation. These 3 factors, combined with TNF-alpha are potent mediators of innate immunity. Uncontrolled activation of these factors leads to immune induced disease pathogenesis through excessive inflammation. In humans and other mammals, neutrophil activation directly contributes to tissue damage through the release of neutrophil extracellular traps (NETs)[38,39]. Functional studies to compare cytokine secretion and NET release in response to infectious agents could confirm that genetic differences in innate immunity contribute to differences in disease susceptibility and outcomes between Asian and African elephants. Our study provides an example of how genomics can inform functional immunological and molecular studies, which may lead to improved conservation and medical care for elephants, and could provide important evolutionary insights to translate into human patients with infection or cancer.

Materials and Methods

De novo assembly and annotation of the Asian elephant genome

The DNA sequencing libraries from the Asian elephant ("Icky") from the Ringling Bros. Center for Elephant Conservation were constructed and sequenced at the University of Utah Genomics Core Facility. Paired-end DNA libraries were constructed with the TruSeq Library Prep kit for a target insert size of 200 bp, and mate-paired libraries were constructed for target sizes of 3 kb, 5 kb, 8 kb, and 10 kb using the Nextera Mate Pair Library kit. Genomic DNA was sequenced on an Illumina HiSeq2500. Raw reads were trimmed to remove nucleotide biases, adapters and a quality score cutoff of 30 with Trimmomatic v0.33[40] and SeqClean[41]. Genome assembly was carried out using ALLPATHS-LG[42,43]. The expected gene content was assessed by searching for 4,104 mammalian single-copy orthologs (mammalia_odb9) using BUSCO v3.0.2[44]. We annotated and masked repeats in the resulting assembly using both the *de novo* method implemented in RepeatModeler v1.0.11[45] and a database method using RepeatMasker v4.07[46] with a library of known mammalian repeats from RepBase[47]. Modeled repeats were used in a blast search against Swiss-Prot[48] to identify and remove false positives. We then generated gene models for the Asian elephant assembly using MAKER2[49], which incorporated (1) homology to publicly available proteomes of cow, human, mouse, and all mammalian entries in UniProtKB/Swiss-Prot, and (2) ab initio gene predictions based on species-specific gene models in SNAP (11/29/2013 release)[50], species-specific and human gene models in Augustus v3.0.2[51], and EvidenceModeler[52]. Final gene calls were functionally annotated by using InterProScan[53] to identify protein domains and a blastp search of all annotated proteins to UniProt proteins to assign putative orthologs with an e-value cutoff of 1e-6.

Tissue collection, DNA extraction, and genome sequencing of African bush elephants

The African bush elephant assembly was improved with the addition of Hi-C sequencing libraries. First, a whole blood sample was drawn (in an EDTA tube) from a wild-born female named Swazi (animal ID: KB13542, North American studbook number 532) at the San Diego Zoo Safari Park in Escondido, CA. We selected this individual because her genome was originally sequenced by the Broad Institute[3]. Three Hi-C libraries were constructed and sequenced to ~38X genome coverage and used for scaffolding with HiRise[54] at Dovetail Genomics in Santa Cruz, CA, with the most recent version of the African bush elephant assembly (loxAfr4.0) as an input. DNA was extracted from fresh frozen subcutaneous skin necropsy tissue samples from an African bush elephant named Hi-Dari (animal ID 00003, North American studbook number 33) at the Hogle Zoo in Salt Lake City, UT using a ThermoFisher PureLink Genomic Mini DNA Kit at the University of Utah. Two pieces of tissue were digested and extracted separately and pooled followed extraction. DNA concentration was measured by picogreen (8.66ng/ul) with a total volume of 200ul in 10mM pH8.0. DNA sequencing libraries were generated using the Illumina TruSeq Library Prep Kit for at 350 bp mean insert size, and sequenced on two lanes the Illumina HiSeq2500 platform at Huntsman Cancer Institute's High-Throughput Genomics Core.

Cancer Data Collection

Pathology and necropsy records were collected with consent from 26 zoological institutions across the United States over the span of 26 years. All participating institutions were de-identified and anonymized. Data was collected with permissions from Buffalo Zoo, Dallas Zoo, El Paso Zoo, Fort Worth Zoo, Gladys Porter Zoo, Greenville Zoo, Jacksonville Zoo and Gardens, Louisville Zoological Garden, Oakland Zoo, Oklahoma City Zoo and Botanical Garden, Omaha's Henry Doorly Zoo and Aquarium, The Phoenix Zoo, Point Defiance Zoo and Aquarium, San Antonio Zoological Society, Santa Barbara Zoological Gardens, Sedgwick

County Zoo, Seneca Park Zoo, Toledo Zoological Gardens, Utah's Hogle Zoo, Woodland Park Zoo, Zoo Atlanta, Zoo Miami and three other anonymous zoos. Neoplasia was diagnosed by board certified veterinary pathologists and cancers were identified from pathological services at Northwest ZooPath, Monroe, WA. Published pathology data from San Diego Zoo was also includedbo[55]. Neoplasia prevalence was estimated by the number of elephants that were diagnosed with tumors (benign or malignant) in respect to all elephants documented within our database.

EP53 evolution in African and Asian elephants

To determine *EP53* copy number and evolutionary patterns across placental mammals, we exported the *TP53* human peptide from Ensembl (July 2019), and used it as a query in BLAT searches[56] of 43 mammalian genome assemblies hosted on the UCSC Genome Browser[57]. Within each mammalian genome, we collected putative BLAT hits, filtered out sequences <33.3% of the length of the human peptide sequence, and used the highest scoring BLAT hit in a second BLAT of the target genome to determine copy number. Putative homologs were validated with a BLASTX query of the human peptide database on NCBI in order to ensure the top hit was human TP53 with ≥70% protein identity. Accepted nucleotide sequences were aligned with MAFFT[58], and we weighted and filtered out unreliable columns in the alignment with GUIDANCE2[59] using 100 bootstrap replicates. We reconstructed the phylogeny of all aligned mammalian TP53 homologs and estimated their divergence times in a Bayesian framework with BEAST 2.5[60] using the HKY substitution model, a relaxed lognormal molecular clock model, and a Yule Model tree prior. We used a normal prior distribution for the age of Eutheria (offset to 105 million years or MYA with the 2.5% quantile at 101 MYA and the 97.5% quantile at 109 MYA) and lognormal prior distributions for the following node calibrations from the fossil record[61]: Boreoeutheria (offset the minimum age to 61.6 MYA –164 MYA and the 97.5% quantile to 170 MYA), Glires (56 MYA – 164 MYA), Primates (56 MYA – 66 MYA),

Laurasiatheria (61.6 MYA – 164 MYA) and Afrotheria (56 MYA – 164 MYA). We monitored proper MCMC mixing with Tracer v1.7.1 to ensure an effective sampling size of at least 200 from the posterior distributions of each parameter and ran two independent chains. The final MCMC chain was run for 100,000,000 generations, and we logged parameter samples every 10,000 generations to collect a total of 10,000 samples from the posterior distribution. We then collected 10,000 of the resulting trees, ignored the first 10% as burn-in, and calculated the maximum clade credibility tree using TreeAnnotator.

Detection of accelerated regions in African and Asian elephant genomes

We generated a multiple alignment (whole genome alignment or WGA) of 12 mammalian genome assemblies. First, we downloaded publicly available pairwise syntenic alignments of opossum (monDom5), mouse (mm10), dolphin (turTru1), cow (bosTau7), dog (canFam3), horse (equCab2), microbat (myoLuc1), tenrec (echTel2), and hyrax (proCap1) to the human reference (hg19) from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/downloads.html#human). We also computed two additional *de novo* pairwise syntenic alignments with the human genome as a target and the two elephant genome assemblies reported here queries using local alignments from LASTZ v1.02[62] using the following options: --hspthresh 2200 --inner 2000 --ydrop 3400 --gappedthresh 10000 --scores HOXD70, followed by chaining to form gapless blocks and netting to rank the highest scoring chains[63]. We then constructed a multiple sequence alignment with MULTIZ v11.2[64] with human as the reference species.

To define elephant accelerated regions (ARs), we used functions from the R package rphast v1.6[22]. We used phyloFit with the substitution model 'REV' to estimate a neutral model based on autosomal fourfold degenerate sites from the WGA. We then used phastCons to define 60 bp autosomal regions conserved in the 10 non-elephant species in the WGA with the following options: expected.length = 45, target.coverage = 0.3, rho = 0.31. We further selected

regions with aligned sequence for both African and Asian elephants that have aligned sequence present for at least 9 of the 10 non-elephant species. We tested the resulting 676,509 regions for acceleration in each elephant species relative to the 10 non-elephant species with phyloP using the following options: mode = 'ACC'. We used the Q-Value method[65] to adjust for multiple testing. Statistically significant ARs were defined with a false discovery rate threshold of 10%. We defined regions significantly accelerated in the Asian elephant, but not the African bush elephant as Asian elephant specific ARs and conversely defined African bush elephant specific ARs.

To define genes differentially expressed between Asian and African elephants we took advantage of closeness of the two species. The Asian elephant is more closely related to the African elephant than humans are to chimpanzees (0.01186 substitutions / 100 bp vs 0.01277 substitutions / 100 bp, fourfold degenerate sites). For the purpose of defining differentially expressed genes, chimpanzee RNA-Seq reads have been aligned to human transcriptome indices[66]. We aligned African bush elephant PBMC reads from a previous study[13] and publicly available Asian elephant PBMC data[23] to an African elephant (loxAfr3) transcriptome index with the STAR aligner. After counting reads for each elephant gene with featureCounts, we normalized counts with the TMM method and defined significant DE genes with edgeR[67] (FDR < 0.01) correcting for multiple testing with the Benjamini-Hochberg method. The DE gene list was minimally affected by modest FDR cutoff changes. We note differences in the cell preps, RNA purification methods and sex of the Asian and African elephants as potential confounding factors in defining DE genes. The African elephant PBMC RNA was purified with a Ribo-Zero depletion step while the Asian elephant RNA was purified by Poly-A selection. A study comparing the two RNA purification methods shows a high gene expression correlation (0.931) between the two methods and detects 410 genes as differentially expressed when contrasting these purification methods[68].

Potential regulatory regions for elephant DE genes were defined with custom R scripts implementing logic detailed by McLean et al. (2010)[69] based on transcription start site (TSS) locations obtained for protein coding genes with the R package biomaRt[70] for the African bush elephant genome (loxAfr3) with basal distances of 5 kb upstream and 1 kb downstream an extension distance of 100 kb. We chose this extension distance because the majority of conserved enhancers are located within 100 kb of a TSS[71]. We used the R package LOLA[72] to test for enrichment of ARs relative to CRs in the potential regulatory regions of DE genes in the loxAfr3 genome. Biological processes (BP) and associated elephant orthologs of human genes were obtained with biomaRt. The resulting p-values were Q-Value corrected for multiple testing. We used the same potential regulatory regions and LOLA to test for GO enrichments.
To select GO terms related to traits that distinguish elephants from the background species and Asian and African elephants from one another for Figure 2, we selected significantly enriched GO terms containing any of the words immune, viral, pathogen, interleukin, T cell, B Cell, innate, apoptosis, P53, tumor, or TNF.

Whole genome sequence analysis of living elephants

To understand genetic changes underlying adaptations in *Loxodonta africana*, *L. cyclotis*, and *Elephas maximus*, we obtained ~15–40x whole-genome sequencing data from multiple individuals from across the modern range of living elephants from public databases[3,10,18,23], and the WGS libraries for “Hi-Dari” and “Icky” as well as a third African elephant named “Christie” (Supplementary Table 6). We also obtained sequence data from a straight-tusked elephant[3] and two woolly mammoths[73]. Sequences were quality checked using FastQC and trimmed to remove nucleotide biases and adapter sequences with Trimmomatic where necessary. Reads from each individual were mapped to the *L. africana* genome (loxAfr3.0, Ensembl version) using bwa-mem v077. Alignments were filtered to include only mapped reads and sorted by position using samtools v0.0.19[74], and we removed PCR duplicates using MarkDuplicates in picard

v1.125[75]. Single-end reads from the ancient samples were mapped to loxAfr3.0 with bwa-aln following Palkolpoulou et al. (2018).

We estimated the number of *TP53* paralogs present in the genome of each elephant by calculating the average read depth of annotated sites in Ensembl *TP53* exons and whole genes with samtools, dividing the total average genome coverage, multiplied by the number of EP53 annotations (n=12). We called variants in the living elephant species (n=13) by incorporating the .bam files using freebayes v1.3.1-12, with extensive filtering to avoid false positives as follows with vcffilter from vcflib (https://github.com/vcflib/vcflib, last accessed July 2019): Phred-scale probability that a REF/ALT polymorphism exists at a given site (QUAL) > 20, the additional contribution of each observation should be 10 log units (QUAL/AO>10), read depth (DP>5), reads must exist on both strands (SAF>0 & SAR>0), and at least two reads must be balanced to each side of the site (RPR>1 & RPL>1). We then removed indels from the .vcf file and filtered to only include biallelic SNPs that were genotyped in every individual using vcftools v0.1.17[76] (--remove-indels --min-alleles 2 --max-alleles 2 --max-missing-count 0) and bcftools v1.9[76] (-v snps -m 1). We annotated the biallelic SNPs using SnpEff v4.3[77] based on loxAfr3 (Ensembl), and calculated diversity statistics including per-individual heterozygosity, nucleotide diversity and Tajima's D in 10kb windows with vcftools, and the fixation index F_{ST} with PopGenome v2.7.1[78].

Selective sweep analysis

To detect loci that have been putatively subjected to positive selection within each living elephant species, we used SweeD v3.3.1. SweeD scans the genome for selective sweeps by calculating the composite likelihood ratio (CLR) test from the folded site frequency spectrum in 1 kb grids across each scaffold. We used the folded site frequency spectrum because we lacked a suitable closely related outgroup with genomic resources that would enable us to establish ancestral alleles. For this analysis, we studied each species individually. Following Nielsen et al.

(2005), we established statistical thresholds for this outlier analysis. First, we generated a null model by simulating 1,000 data sets with 100,000 SNPs under neutral demographic models following the Pairwise Sequential Markovian Coalescent trajectories[79] for each species from Palkopoulou et al. (2018), which we implemented with Hudson's ms (October 2007 release)[80] (Supplementary Fig. 4). Then, we categorized regions as outlier regions in the observed SNP data if their CLR was greater than the 99.99th percentile of the distribution of the highest CLRs from the simulated SNP data. For the neutral simulations, we assumed a per-year mutation rate of 0.406E-9 and a generation time of 31 years, following Palkopoulou et al. (2018). We then calculated the CLR with the simulated neutral SNP datasets. SweeD output files were changed to BED format using the namedCapture (https://github.com/tdhock/namedCapture-article/) and data.table[81] R packages, and we used bedtools intersect[82] to collect elephant gene annotations (loxAfr3.0, Ensembl) overlapping putative selective sweeps.

Genomic scans for selection may be complicated by several factors that can increase false positive rates, and false negative rates potentially stem from variable mutation and recombination landscapes[30,83]. We therefore established statistical thresholds using null demographic models. However, the estimated split times within living elephant species differ widely, ranging from 609,000 to 463,000 years before present for forest elephants[84], to as recent as 38,000 to 30,000 years before present for bush elephants[73,85]. Estimated split times between the sampled Asian elephants are highly variable, ranging from 190,000 to 24,000 years before present[3], indicating continental-wide population structure not accounted for here[86,87]. Still, Palkopoulou et al. (2018) found little evidence for gene flow between the three modern species of elephant, which supports our choice of analyzing them separately for selective sweeps. Therefore, we focused on shared outlier regions, which show consistent evidence of being targeted by positive selection across all three elephant species.

Genes overlapping outlier regions of putative selective sweeps were functionally annotated using a variety of methods. First, we tested for the enrichment of Gene Ontology

terms for biological processes[88] in the outlier gene list, using default parameters and the Benjamini-Hochberg correction for multiple testing with an adjusted p-value < 0.05. We used REVIGO[89] to semantically cluster and visualize the most significant GO terms according to their adjusted p-values using default parameters. We also created annotation clusters from the outlier genes using DAVID v6.8[90,91] and constructed protein interaction networks with STRING v11.0[92]. Finally, we tested for enriched mouse phenotypes using ModPhea[93].

Acknowledgements

We acknowledge Leigh Duke for data coordination and Trent Fowler and Rosann Robinson for assistance with sample collection. We acknowledge the collections and veterinary staff at the San Diego Zoo Safari Park and Utah's Hogle Zoo for sample collection. We acknowledge the following institutions for sharing data and/or resources: Buffalo Zoo, Dallas Zoo, El Paso Zoo, Fort Worth Zoo, Gladys Porter Zoo, Greenville Zoo, Jacksonville Zoo and Gardens, Louisville Zoological Garden, Oakland Zoo, Oklahoma City Zoo and Botanical Garden, Omaha's Henry Doorly Zoo and Aquarium, The Phoenix Zoo, Point Defiance Zoo and Aquarium, San Antonio Zoological Society, Santa Barbara Zoological Gardens, Sedgwick County Zoo, Seneca Park Zoo, Toledo Zoological Gardens, Utah's Hogle Zoo, Woodland Park Zoo, Zoo Atlanta, Zoo Miami and three other anonymous zoos. We acknowledge Huntsman Cancer Institute's High-Throughput Genomics Core and the Monsoon computing cluster at Northern Arizona University (https://nau.edu/high-performance-computing/). Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number U54CA217376. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

Dr. Schiffman is co-founder, shareholder, and employed by PEEL Therapeutics, Inc., a company developing evolution-inspired medicines based on cancer resistance in elephants. Dr. Abegglen is share-holder and consultant to PEEL Therapeutics, Inc.

References

1. Shoshani, J. Understanding proboscidean evolution: a formidable task. *Trends Ecol. Evol.* **13**, 480–487 (1998).
2. Larramendi, A. Shoulder height, body mass, and shape of proboscideans. *Acta Palaeontol. Pol.* **61**, 537–574 (2015).
3. Palkopoulou, E. *et al.* A comprehensive genomic history of extinct and living elephants. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2566–E2574 (2018).
4. Vartanyan, S. L., Garutt, V. E. & Sher, A. V. Holocene dwarf mammoths from Wrangel Island in the Siberian Arctic. *Nature* **362**, 337–340 (1993).
5. Stuart, A. J. The extinction of woolly mammoth (*Mammuthus primigenius*) and straight-tusked elephant (*Palaeoloxodon antiquus*) in Europe. *Quat. Int.* **126–128**, 171–177 (2005).
6. Fernando, P. & Pastorini, J. Range-wide status of Asian elephants. *Gajah* **35**, 15–20 (2011).
7. Thouless, C. R. *et al.* African elephant status report 2016: an update from the African Elephant Database. *Occas. Pap. IUCN Surviv. Comm.* **60**, (2016).
8. Roca, A. L. *et al.* Elephant natural history: a genomic perspective. *Annu. Rev. Anim. Biosci.* **3**, 139–167 (2015).
9. Supple, M. A. & Shapiro, B. Conservation of biodiversity in the genomics era. *Genome Biol.* **19**, 131 (2018).
10. Lynch, V. J. *et al.* Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the arctic. *Cell Rep.* 1–13 (2015) doi:10.1016/j.celrep.2015.06.027.
11. Fry, E. *et al.* Functional architecture of deleterious genetic variants in the genome of a wrangel island mammoth. *Genome Biol. Evol.* **12**, 48–58 (2020).
12. Rogers, R. L. & Slatkin, M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet.* **13**, e1006601 (2017).
13. Ferris, E., Abegglen, L. M., Schiffman, J. D. & Gregg, C. Accelerated evolution in distinctive species reveals candidate elements for clinically relevant traits, including mutation and cancer resistance. *Cell Rep.* **22**, 2742–2755 (2018).
14. Jachmann, H., Berry, P. S. M. & Imae, H. Tusklessness in African elephants: a future trend. *Afr. J. Ecol.* **33**, 230–235 (1995).
15. Allendorf, F. W. & Hard, J. J. Human-induced evolution caused by unnatural selection through harvest of wild animals. *Proc. Natl. Acad. Sci.* **106**, 9987–9994 (2009).
16. Raubenheimer, E. J. & Miniggio, H. D. Ivory harvesting pressure on the genome of the African elephant: a phenotypic shift to tusklessness. *Head Neck Pathol.* **10**, 332–335 (2016).
17. Hayward, G. S. Conservation: clarifying the risk from herpesvirus to captive Asian elephants. *Vet. Rec.* **170**, 202–203 (2012).
18. Abegglen, L. M. *et al.* Ppotential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans. *JAMA* **314**, 1850–1860 (2015).
19. Sulak, M. *et al.* TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *eLife* **5**, 1850 (2016).

20. Vazquez, J. M., Sulak, M., Chigurupati, S. & Lynch, V. J. A zombie LIF gene in elephants is upregulated by tp53 to induce apoptosis in response to DNA damage. *Cell Rep.* **24**, 1765–1776 (2018).
21. Booker, B. M. *et al.* Bat accelerated regions identify a bat forelimb specific enhancer in the HoxD locus. *PLOS Genet.* **12**, e1005738 (2016).
22. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
23. Reddy, P. C. *et al.* Comparative sequence analyses of genome and transcriptome reveal novel transcripts and variants in the Asian elephant *Elephas maximus*. *J. Biosci.* **40**, 891–907 (2015).
24. Zimmermann, A., Bernuit, D., Gerlinger, C., Schaefers, M. & Geppert, K. Prevalence, symptoms and management of uterine fibroids: an international internet-based survey of 21,746 women. *BMC Womens Health* **12**, 6 (2012).
25. Montali, R. J. *et al.* Ultrasonography and pathology of genital tract leiomymoas in captive asian elephants: implications for reproductive soundness. *Erkrangungen Zootiere* **38**, (1997).
26. Caulin, A. F., Graham, T. A., Wang, L.-S. & Maley, C. C. Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140222 (2015).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2009).
28. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
29. Meyer, M. *et al.* Palaeogenomes of Eurasian straight-tusked elephants challenge the current view of elephant evolution. *eLife* **6**, 1175 (2017).
30. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
31. Pavlidis, P., Živkovic, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).
32. Haug, H. Comparative studies of the brains of men, elephants and toothed whales. *Verh. Anat. Ges.* **64**, 191–195 (1970).
33. Park, S.-Y., Kim, S.-Y., Jung, M.-Y., Bae, D.-J. & Kim, I.-S. Epidermal growth factor-like domain repeat of stabilin-2 recognizes phosphatidylserine during cell corpse clearance. *Mol. Cell. Biol.* **28**, 5288–5298 (2008).
34. Balkwill, F. Tumour necrosis factor and cancer. *Nat. Rev. Cancer* **9**, 361–371 (2009).
35. Caulin, A. F. & Maley, C. C. Peto's Paradox: evolution's prescription for cancer prevention. *Trends Ecol. Evol.* **26**, 175–182 (2011).
36. Srivorakul, S. *et al.* Possible roles of monocytes/macrophages in response to elephant endotheliotropic herpesvirus (EEHV) infections in Asian elephants (Elephas maximus). *PLOS ONE* **14**, e0222158 (2019).
37. Greenwald, R. *et al.* Highly accurate antibody assays for early and rapid detection of tuberculosis in African and Asian elephants. *Clin. Vaccine Immunol.* **16**, 605–612 (2009).
38. Jorch, S. K. & Kubes, P. An emerging role for neutrophil extracellular traps in noninfectious disease. *Nat. Med.* **23**, 279–287 (2017).
39. Goggs, R., Jeffery, U., LeVine, D. N. & Li, R. H. L. Neutrophil-extracellular traps, cell-free DNA, and immunothrombosis in companion animals: a review. *Vet. Pathol.* **57**, 6–23 (2020).
40. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).
41. Chen, Y.-A., Lin, C.-C., Wang, C.-D., Wu, H.-B. & Hwang, P.-I. An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics* **8**, 416 (2007).

42. MacCallum, I. *et al.* ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.* **10**, R103 (2009).
43. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513–1518 (2011).
44. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, btv351-3212 (2015).
45. Smit, A. F. A., Hubley, R. M. & Green, P. RepeatModeler Open-1.0 2008-2015. *http://www.repeatmasker.org*.
46. Smit, A. F. A., Hubley, R. M. & Green, P. RepeatMasker Open-4.0 2013-2015. *http://www.repeatmasker.org*.
47. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
48. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204-12 (2015).
49. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
50. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
51. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* **24**, 637–644 (2008).
52. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
53. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**, 1236–1240 (2014).
54. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
55. Boddy, A. M. *et al.* Lifetime cancer prevalence and life history traits in mammals. *Evol. Med. Public Health.* eoaa015. (2020)
56. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
57. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
58. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
59. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**, W7-14 (2015).
60. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).
61. Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontol. Electron.* **18**, 1–106 (2015).
62. Harris, R. S. Improved pairwise alignment of genomic DNA. (The Pennsylvania State University, 2007).
63. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100**, 11484–11489 (2003).
64. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
65. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).

66. Marchetto, M. C. *et al.* Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *eLife* **8**, e37527 (2019).
67. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
68. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).
69. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis -regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
70. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
71. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
72. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).
73. Palkopoulou, E. *et al.* Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol. CB* **25**, 1395–1400 (2015).
74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).
75. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
76. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics.* **27**, 2156–2158 (2011).
77. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
78. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
79. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
80. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* **18**, 337–338 (2002).
81. Dowle, M., Srinivasan, A. data.table: Extension of 'data.frame'. R pacakge version 1.12.8 (2019).
82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **26**, 841–842 (2010).
83. Jensen, J. D., Foll, M. & Bernatchez, L. The past, present and future of genomic scans for selection. *Mol. Ecol.* **25**, 1–4 (2016).
84. Rohland, N. *et al.* Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biol.* **8**, e1000564 (2010).
85. Comstock, K. E. *et al.* Patterns of molecular genetic variation among African elephant populations. *Mol. Ecol.* **11**, 2489–2498 (2002).
86. Fleischer, R. C., Perry, E. A., Muralidharan, K., Stevens, E. E. & Wemmer, C. M. Phylogeography of the Asian elephant (*Elephas maximus*) based on mitochondrial DNA. *Evolution* **55**, 1882–1892 (2001).
87. Vidya, T. N. C., Sukumar, R. & Melnick, D. J. Range-wide mtDNA phylogeography yields insights into the origins of Asian elephants. *Proc. R. Soc. B Biol. Sci.* **276**, 893–902 (2009).
88. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).

89. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* **6**, e21800 (2011).
90. Huang, D. W., Sherman, B. T. & RA Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
91. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
92. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
93. Weng, M.-P. & Liao, B.-Y. modPhEA: model organism Phenotype Enrichment Analysis of eukaryotic gene sets. *Bioinformatics.* **33**, 3505–3507 (2017).