Somatic mutation rates scale with lifespan across mammals

Authors

Alex Cagan[1+]*, Adrian Baez-Ortega[1]*, Natalia Brzozowska[1], Federico Abascal[1], Tim H. H. Coorens[1], Mathijs A. Sanders[1,2], Andrew R. J. Lawson[1], Luke M. R. Harvey[1], Shriram G. Bhosle[1], David Jones[1], Raul E. Alcantara[1], Timothy M. Butler[1], Yvette Hooks[1], Kirsty Roberts[1], Elizabeth Anderson[1], Edmund Flach[3], Simon Spiro[3], Inez Januszczak[3,4], Ethan Wrigglesworth[3], Matthew W. Perkins[5], Robert Deaville[5], Megan Druce[6,7], Ruzhica Bogeska[6,7], Michael D. Milsom[6,7], Björn Neumann[8,9], Frank Gorman[10], Fernando Constantino-Casas[10], Laura Peachey[10,11], Diana Bochynska[10,12], Ewan St. John Smith[13], Moritz Gerstung[14], Peter J. Campbell[1], Elizabeth P. Murchison[10], Michael R. Stratton[1], Iñigo Martincorena[1+]

* These authors contributed equally

[+] Correspondence to: ac36@sanger.ac.uk (A.C.) and im3@sanger.ac.uk (I.M.)

Author information

(1) Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Hinxton, CB10 1SA, UK

(2) Department of Hematology, Erasmus MC Cancer Institute, 3015 CN Rotterdam, Netherlands

(3) Wildlife Health Services, Zoological Society of London, Regent's Park, London, NW1 4RY, UK

(4) The Natural History Museum, Cromwell Road, London, SW7 5BD, UK

(5) Institute of Zoology, Zoological Society of London, Regent's Park, London, NW1 4RY, UK

(6) Division of Experimental Hematology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

(7) Heidelberg Institute for Stem Cell Technology and Experimental Medicine GmbH (HI-STEM), 69120 Heidelberg, Germany

(8) Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, CB2 0AW, UK

(9) Department of Clinical Neurosciences, University of Cambridge, Cambridge, CB2 0QQ, UK
(10) Department of Veterinary Medicine, University of Cambridge, Cambridge, CB3 0ES, UK
(11) Bristol Veterinary School, Faculty of Health Sciences, University of Bristol, Langford, BS40 5DU, UK
(12) Pathology Department, Faculty of Veterinary Medicine, Universitatea de Stiinte Agricole si Medicina Veterinara, Cluj-Napoca 400372, Romania
(13) Department of Pharmacology, University of Cambridge, Cambridge CB2 1PD, UK
(14) European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, CB10 1SD, UK

Abstract

The rates and patterns of somatic mutation in normal tissues are largely unknown outside of humans. Comparative analyses can shed light on the diversity of mutagenesis across species and on long-standing hypotheses regarding the evolution of somatic mutation rates and their role in cancer and ageing. Here, we used whole-genome sequencing of 208 intestinal crypts from 56 individuals to study the landscape of somatic mutation across 16 mammalian species. We found somatic mutagenesis to be dominated by seemingly endogenous mutational processes in all species, including 5-methylcytosine deamination and oxidative damage. With some differences, mutational signatures in other species resembled those described in humans, although the relative contribution of each signature varied across species. Remarkably, the somatic mutation rate per year varied greatly across species and exhibited a strong inverse relationship with species lifespan, with no other life-history trait studied displaying a comparable association. Despite widely different life histories among the species surveyed, including ~30-fold variation in lifespan and ~40,000-fold variation in body mass, the somatic mutation burden at the end of lifespan varied only by a factor of ~3. These data unveil common mutational processes across mammals and suggest that somatic mutation rates are evolutionarily constrained and may be a determinant of lifespan.

Introduction

Somatic mutations accumulate in healthy cells throughout life. They underpin cancer development[1,2] and, for decades, have been speculated to contribute to ageing[3–5]. Directly studying somatic mutations in normal tissues has been challenging due to the difficulty of detecting mutations present in single cells or small clones in a tissue. Only recent technological developments, such as *in vitro* expansion of single cells into colonies[6,7], microdissection of histological units[8,9] or single-cell or single-molecule sequencing[10–12], are beginning to enable the study of somatic mutation in normal tissues.

Over the last few years, studies in humans have started to provide a detailed understanding of somatic mutation rates and the contribution of endogenous and exogenous mutational processes across normal tissues[6–8,13,14]. These studies are also revealing how, as we age, some human tissues are colonised by mutant cells carrying cancer-driving mutations, and how this clonal composition changes with age and disease. With the exception of some initial studies, far less is known about somatic mutation in other species[15–20]. Yet, comparative analyses of somatic mutagenesis would shed light on the diversity of mutagenic processes across species, and on long-standing questions regarding the evolution of somatic mutation rates and their role in cancer and ageing.

A decades-long hypothesis on the evolution of somatic mutation rates pertains to the relationship between body mass and cancer risk. Some models predict that the risk of cancer should increase proportionally to the number of cells at risk of transformation. However, there appears to be no correlation between body mass and cancer risk across species[21–23]. This observation, known as Peto's paradox, suggests that the evolution of larger body sizes likely requires the evolution of stronger cancer suppression mechanisms[23,24]. Whether evolutionary reduction of cancer risk across species is partly achieved by a reduction of somatic mutation rates remains unknown[25].

A second long-standing hypothesis on the evolution of somatic mutation rates relates to the proposed role of somatic mutations in ageing. Multiple forms of molecular damage, including somatic mutations, telomere attrition, epigenetic drift or loss of proteostasis, have been proposed

to contribute to ageing, but their causal roles and relative contributions remain debated[26,27]. Evolutionary theory predicts that species will evolve protection or repair mechanisms against life-threatening damage to minimise death from intrinsic causes, but that selection is too weak to delay ageing far beyond the typical life expectancy of an organism in the wild[28–32]. If somatic mutations contribute to ageing, theory predicts that somatic mutation rates may inversely correlate with lifespan across species[30,33]. This long-standing prediction has remained largely untested due to the difficulty of measuring somatic mutation rates across species.

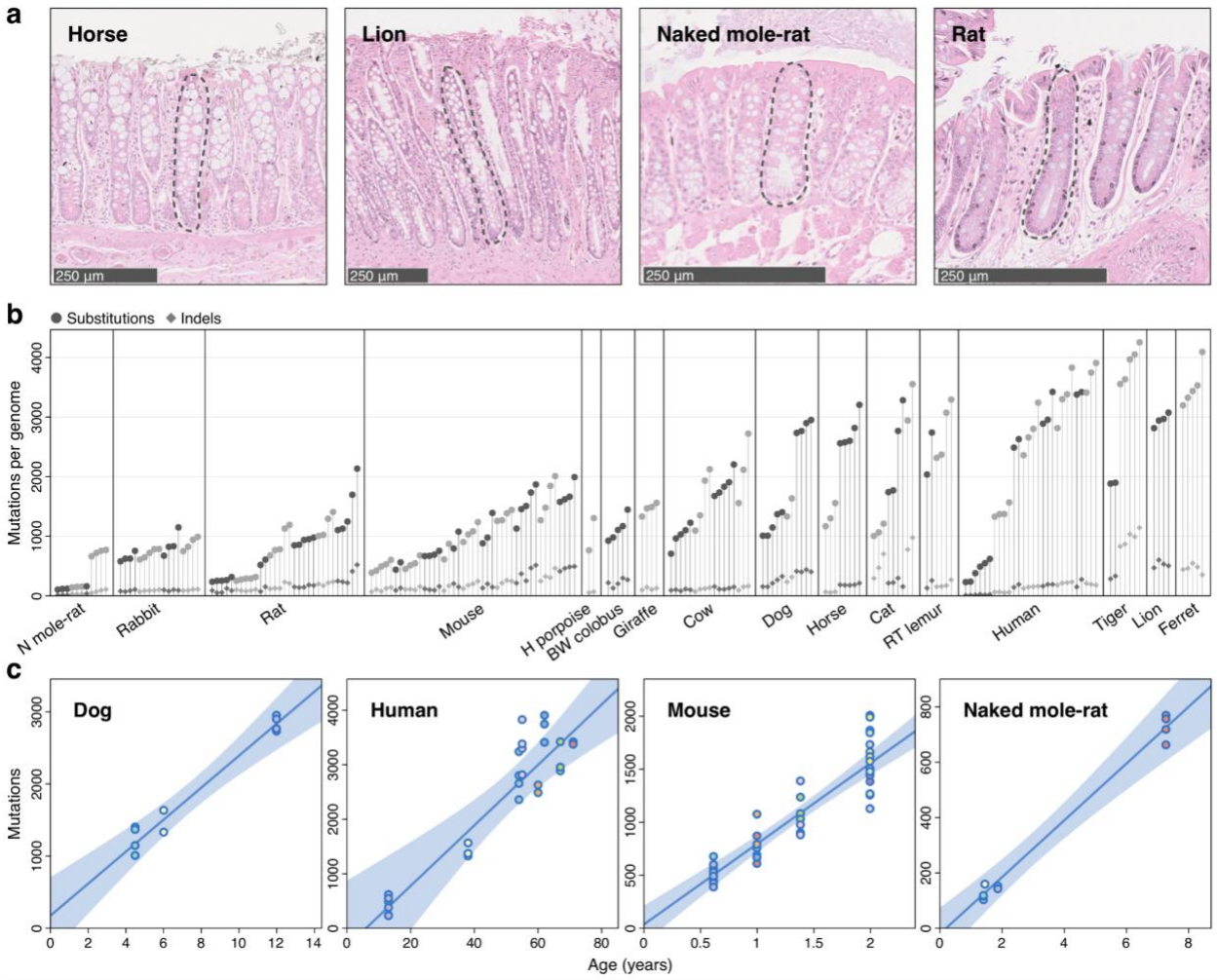
Detection of somatic mutations across species

The study of somatic mutations with standard whole-genome sequencing requires isolating clonal groups of cells recently derived from a single cell[6–8]. To study somatic mutations across a diverse set of mammals, we isolated 208 individual intestinal crypts from 56 individuals across 16 species with a wide range of lifespans and body sizes: black-and-white colobus monkey, cat, cow, dog, ferret, giraffe, harbour porpoise, horse, human, lion, mouse, naked mole-rat, rabbit, rat, ring-tailed lemur, and tiger (**Extended Data Table 1**). We chose intestinal crypts for several reasons. First, they are histologically identifiable units that line the epithelium of the colon and small intestine and are amenable to laser microdissection. Second, human studies have confirmed that individual crypts become clonally derived from a single stem cell and show a linear accumulation of mutations with age, enabling the estimation of somatic mutation rates through genome sequencing of single crypts[8]. Third, in most human crypts, the majority of somatic mutations are caused by endogenous mutational processes common to other tissues, rather than by environmental mutagens[8,12].

A colon sample was collected from each individual, with the exception of a ferret from which only a small intestine sample was available. This sample was included because results in humans have shown the mutation rate of colorectal and small intestine epithelium to be similar[7,14] (**Extended Data Figure 1**). We then used laser microdissection on histological sections to isolate individual crypts for whole-genome sequencing with a low-input library preparation method[34] (**Fig. 1a, Extended Data Figure 2, Extended Data Table 2**), with the exception of human crypts, for which sequencing data were obtained from a previous study[8]. A bioinformatic pipeline was developed to

call somatic mutations robustly in all these species despite variable quality of their genome assemblies (**Methods**). The distribution of variant allele fractions (VAFs) of the mutations detected in each crypt confirmed that crypts are clonal units in all species, enabling the study of somatic mutation rates and signatures (**Extended Data Figure 3**).

Figure 1. Somatic mutation burden in mammalian colorectal crypts. a, Histology images from horse, lion, naked mole-rat and rat colon samples, with one colorectal crypt marked in each. **b,** Burden of somatic single-base substitutions and indels per genome in each colorectal crypt sample (corrected for the size of the analysable genome). Samples are grouped by individual, with samples from the same individual coloured in the same shade of grey. Species, and individuals within each species, are sorted by mean mutation burden **c**, Linear regression of somatic substitution burden (corrected for analysable genome size) on individual age for dog, human, mouse and naked mole-rat samples. Samples from the same individual are shown in the same colour. Regression was performed using mean mutation burden per individual. Shaded areas indicate 95% confidence intervals of the regression line. BW: black-and-white, H: harbour, N: naked, RT: ring-tailed.

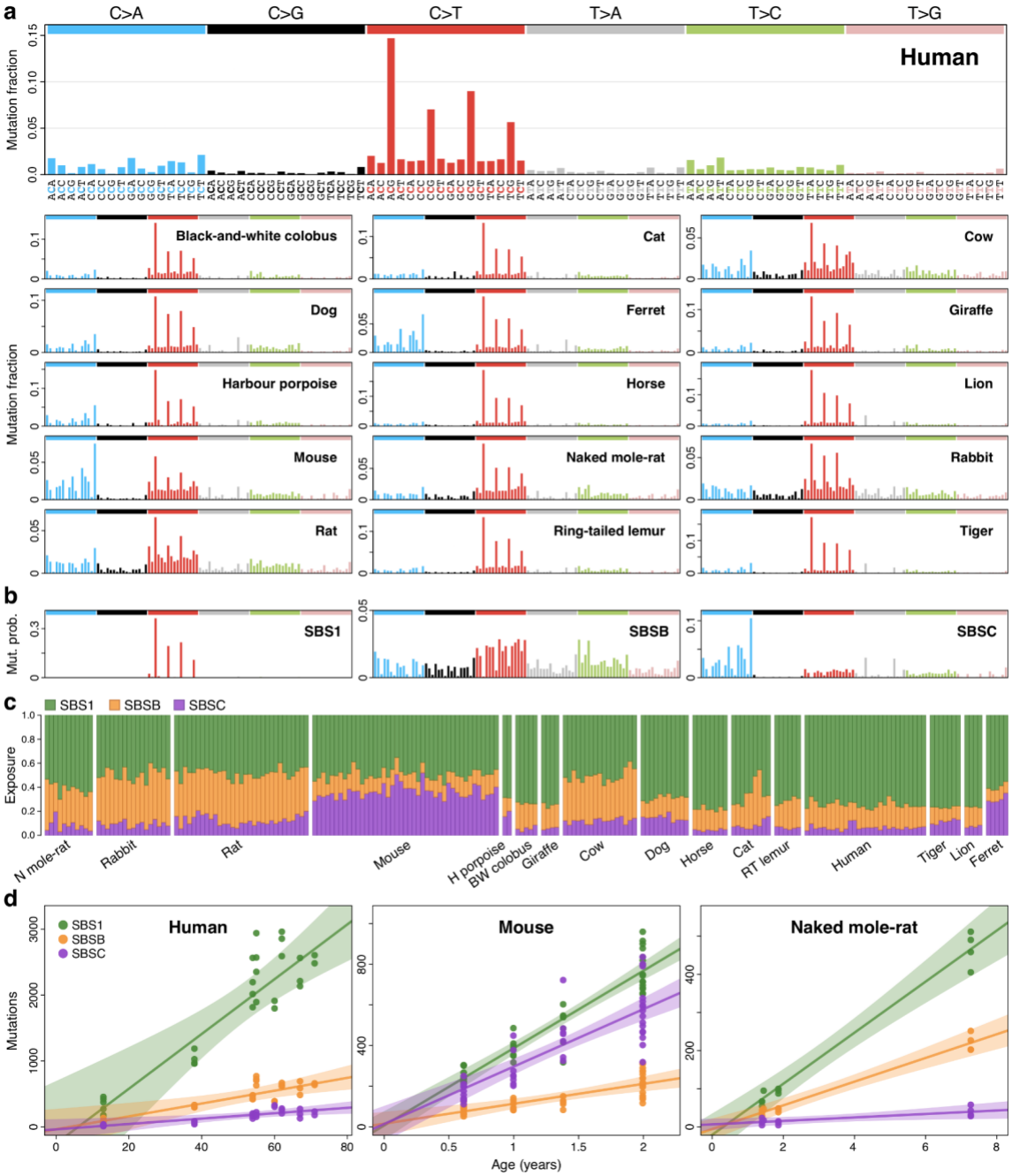


We found substantial variation in the number of somatic single-base substitutions (SBS) across species and across individuals within each species (**Fig. 1b**). For five species with samples from multiple individuals, linear regression confirmed a clear accumulation of somatic mutations with age (**Fig. 1c, Extended Data Figure 4, Extended Data Table 3**). All linear regressions were also consistent with a non-significant intercept. This resembles observations in humans[14] and suggests that the time required for a single stem cell to drift to fixation within a crypt is a small fraction of the lifespan of a species. This facilitates the estimation of somatic mutation rates across species by dividing the number of mutations in a crypt by the age of the individual (**Extended Data Table 4**). The number of somatic insertions and deletions (indels) was consistently lower than that of SBS in all crypts (**Fig. 1b**), in agreement with previous findings in humans[8].

Similar mutational signatures across mammals

To explore the mechanisms behind somatic mutagenesis in colorectal crypts across species, we first classified substitutions and indels according to their sequence context[35,36]. The mutational spectra showed clear similarities across species, with a dominance of cytosine-to-thymine (C>T) substitutions at CpG sites, as observed in human colon, but with considerable variation in the frequency of other substitution types (**Fig. 2a**). To quantify the contribution of different mutational processes to the observed spectra, we applied mutational signature decomposition[8,35]. We used a Bayesian model to infer mutational signatures *de novo*, while accounting for differences in genome sequence composition across species, and using the COSMIC human signature SBS1 (C>T substitutions at CpG sites) as a fixed prior to ensure its complete deconvolution[37] (**Methods**). This approach identified two signatures beyond SBS1, labelled SBSB and SBSC, which resemble COSMIC human signatures SBS5 and SBS18, respectively (cosine similarities 0.93 and 0.91) (**Fig. 2b**).

Figure 2. Mutational processes in the mammalian colon. a, Mutational spectra of somatic single-base substitutions in each species. Horizontal axis presents 96 mutation types on a trinucleotide context, coloured by base substitution type. **b**, Mutational signatures inferred from (SBSB, SBSC) or fitted to (SBS1) the species mutational spectra shown in **a**, and normalised to the human genome trinucleotide frequencies. **c**, Estimated contribution of each signature to each sample. Samples are arranged horizontally as in **Fig. 1b**. **d**, Regression of signature-specific mutation burdens on individual age for human, mouse and naked mole-rat samples. Regression was performed using mean mutation burden per individual. Shaded areas indicate 95% confidence intervals of the regression line. BW, black-and-white; H, harbour; N, naked; RT, ring-tailed.



This analysis suggests that the same three signatures that dominate somatic mutagenesis in the human colon are dominant in other mammals: SBS1, which is believed to result from the spontaneous deamination of 5-methylcytosine[8,38]; SBSB/5, a common signature across human tissues that may result from endogenous damage and repair[12,39]; and SBSC/18, dominated by C>A substitutions and attributed to oxidative DNA damage[36]. Signature SBSC contains a minor component of T>A substitutions (resembling COSMIC SBS34), which appear to be the result of DNA polymerase slippage at the boundaries between adjacent adenine and thymine homopolymer tracts but that could also reflect assembly errors at those sites[39]. While all species surveyed shared the three mutational signatures, their contributions varied substantially across species (**Fig. 2c**). SBSC was particularly prominent in mouse and ferret, and the ratio of SBS1 to SBSB (SBS5) varied from approximately 1.2 in rat or rabbit to 6.4 in tiger. In several species with data from multiple individuals, separate linear regressions for each signature confirmed that mutations from all three signatures accumulate with age (**Fig. 2d, Extended Data Figure 5**).

Although signature deconvolution identified three signatures active across species, we noticed some differences in the mutational profile of signature SBSB among species. To explore this further, we inferred independent versions of SBSB from each species, while accounting for differences in genome sequence composition (**Methods**). This revealed inter-species variability in the mutational profile of this signature, particularly in the C>T component (**Extended Data Figure 6**). Species-specific versions of SBSB showed different similarities to the related human signatures SBS5 and SBS40. For example, SBSB inferred from the human data showed a stronger similarity with the reference SBS5 human signature (cosine similarities with SBS5 and SBS40: 0.93 and 0.84), whereas SBSB from rabbit more closely resembled the reference human SBS40 signature (0.87 and 0.91). These observations are consistent with the hypothesis that SBS5 and SBS40 result from a combination of correlated mutational processes, with some variation across human tissues[12,39] and across species.

Analysis of the indel mutational spectra revealed a dominance of the human indel signatures ID1 and ID2, which are characterised by single-nucleotide insertions and deletions at A/T homopolymers, probably caused by strand slippage during DNA replication[36] (**Extended Data Figure 7**). The ratio of insertions (ID1) to deletions (ID2) appears to vary across species, possibly

reflecting a differential propensity for slippage of the template and nascent DNA strands[36]. In addition, the human indel spectrum suggests a potential contribution of signature ID9, whose aetiology remains unknown. Analysis of indels longer than 1 base pair (bp) also suggested the presence of a signature of 4-bp insertions at tetrameric repeats, particularly prevalent in the mouse (**Extended Data Figure 7**).

Other mutational processes and selection

The apparent lack of additional mutational signatures is noteworthy. A previous study of 445 colorectal crypts from 42 human donors found many crypts to be affected by a novel signature that was later attributed to colibactin, a genotoxin produced by *pks*+ strains of *Escherichia coli*[8,40,41]. Analysing the original human data and our non-human data with the same methodology, we found evidence of colibactin mutagenesis in 21% of human crypts but only uncertain evidence of colibactin in one non-human crypt (0.6%) (**Extended Data Figure 8, Methods**). This revealed a significant depletion of colibactin mutagenesis in the non-human crypts studied (Fisher's exact tests, $P=7\times10^{-14}$). The apparent difference in colibactin mutagenesis observed between species, or between the cohorts studied, might result from a different prevalence of *pks*+ *E. coli* strains[42] or a different expression of colibactin by *pks*+ *E.coli* across species[43]. Finally, we also searched for evidence of APOBEC signatures (SBS2/13), which have been reported in a small number of human crypts and are believed to be caused by APOBEC DNA-editing cytidine deaminases. We detected APOBEC signatures in 2% (n = 9) of human crypts and only uncertain evidence in one non-human crypt ($P=0.30$).

Beyond single-base substitutions and indels, crypts from the eight species with chromosome-level genome assemblies were inspected for large-scale copy number changes (≥1 megabase, Mb) (**Methods**). Studies in humans have found large-scale copy number changes to be relatively rare in normal tissues, including colorectal epithelium[8]. Consistent with these results, we only identified four large copy number changes across the 162 crypts included in this analysis: two megabase-scale deletions in two crypts from the same cow, the loss of a chromosome X in a female mouse crypt, and a 52-Mb segment with copy-neutral loss of heterozygosity in a human crypt

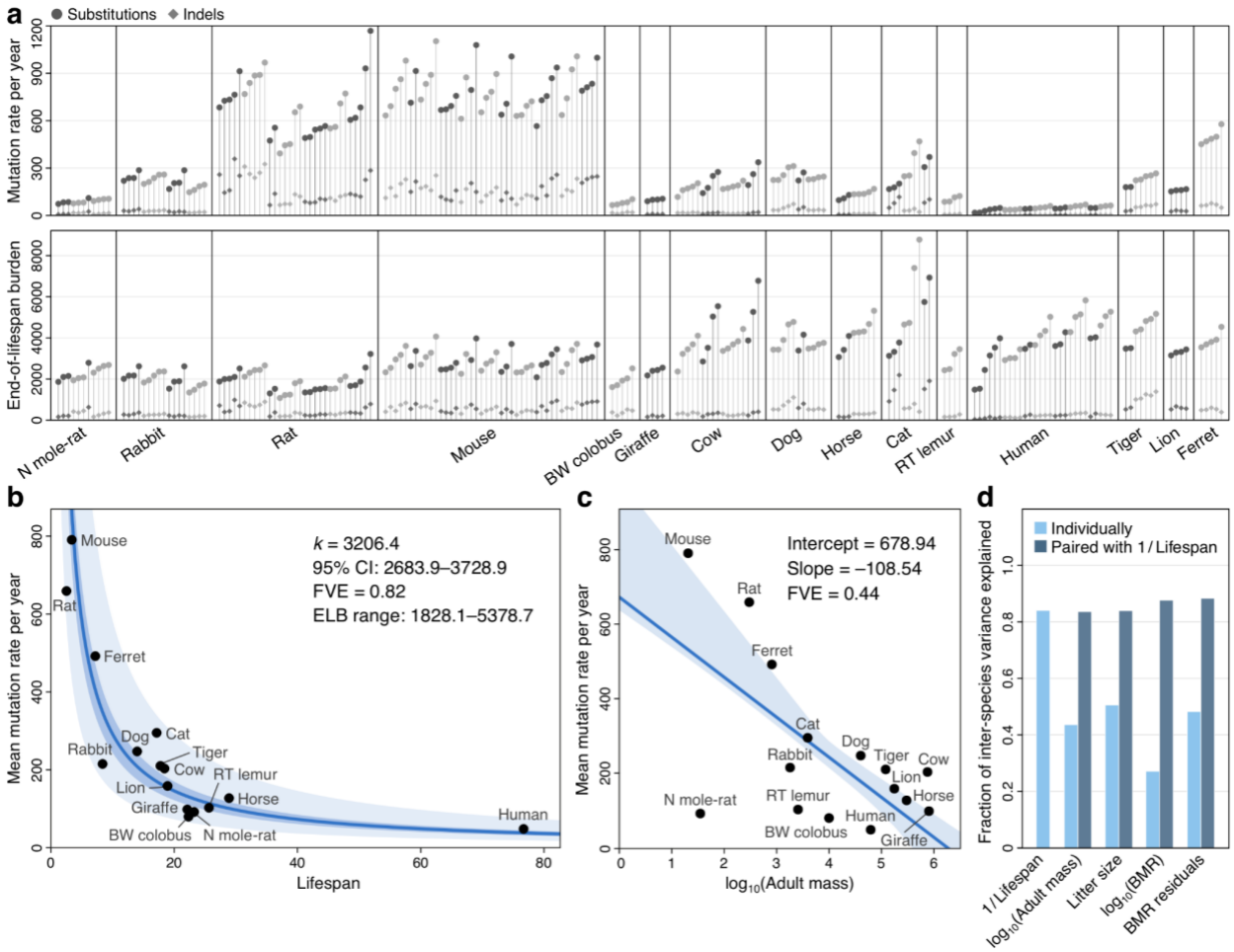
(**Extended Data Figure 9, Methods**). These results suggest that large-scale somatic copy number changes in normal tissues are also rare in other mammalian species.

Previous analyses in humans have shown that most somatic mutations in colorectal crypts accumulate neutrally, without clear evidence of negative selection against non-synonymous mutations and with a low frequency of positively selected cancer-driver mutations[8]. To study somatic selection in our data, we calculated the exome-wide ratio of non-synonymous to synonymous substitution rates (dN/dS) in each of the 12 species with available genome annotation. To do so and to detect genes under positive selection, while accounting for the effects of trinucleotide sequence context and mutation rate variation across genes, we used the dNdScv model[44] (**Methods**). Although the limited number of coding somatic mutations observed in most species prevented an in-depth analysis of selection, exome-wide dN/dS ratios for somatic substitutions were not significantly different from unity in any species, in line with previous findings in humans[8] (**Extended Data Figure 10**). Gene-level analysis did not find genes under significant positive selection in any species, although larger studies are likely to identify rare cancer-driving mutations[8].

Somatic mutation rates and life-history traits

Whereas similar mutational processes operate across the species surveyed, the mutation rate per genome per year varied widely. Across the 15 species with age information, we found that substitution rates per genome ranged from 47 SBS/year in humans to 796 SBS/year in mice, and indel rates from 2.6 to 159 indels/year, respectively (**Fig. 3a, Extended Data Table 4, Methods**).

Figure 3. Associations between somatic mutation rate and life-history variables. a, Somatic mutation rate per year (top) and expected end-of-lifespan mutation burden (bottom) in each crypt sample (derived from single-base substitutions). Samples are grouped and sorted as in **Fig. 1b**; harbour porpoise samples were excluded because the individual was of unknown age. **b**, Zero-intercept linear mixed-effects (LME) regression of somatic mutation rate on inverse lifespan (1/lifespan), presented on the scale of untransformed lifespan (x-axis). For simplicity, the y-axis represents mean mutation rate per species, although mutation rates per crypt were used in the regression. Darker shaded area indicates 95% confidence interval (CI) of the regression line; lighter shaded area marks a two-fold deviation from the regression line. Point estimate and 95% CI of the regression slope coefficient (*k*), fraction of inter-species variance explained by the model (FVE), and range of end-of-lifespan burden (ELB) are provided. **c**, Free-intercept LME regression of somatic mutation rate on log-transformed adult mass. y-axis represents mean mutation rate per species, although mutation rates per crypt were used in the regression. Shaded area indicates 95% bootstrap interval of the regression line. Point estimates of the regression intercept and slope coefficients, and model FVE, are provided. **d**, Comparison of FVE values achieved by free-intercept LME models using inverse lifespan and other life-history variables (alone or in combination with inverse lifespan) as explanatory variables. BMR, basal metabolic rate; BW, black-and-white; N, naked; RT, ring-tailed.



To explore the relationship between somatic mutation rates, lifespan and other life-history traits, we first estimated the lifespan of each species using mortality curves. We used a large collection of mortality data from animals in zoos to minimise the impact of extrinsic mortality (**Extended Data Figure 11**). We defined lifespan as the age at which 80% of individuals reaching adulthood have died, to reduce the impact of outliers and variable cohort sizes that affect maximum lifespan estimates[45,46] (**Methods**). Remarkably, we found a tight anticorrelation between somatic mutation rates per year and lifespan across species (**Fig. 3b**). The shape of this relationship followed a simple model in which somatic mutation rates per year are inversely proportional to the lifespan of a species ($rate \propto 1/lifespan$), such that the number of somatic mutations per cell at the end of the lifespan (the end-of-lifespan burden, ELB) is similar in all species (**Fig. 3a,b**).

To study the relationship between somatic mutation rates and life-history variables formally, we used linear mixed-effects regression models. These models account for the hierarchical structure of the data (with multiple crypts per individual and multiple individuals per species), as well as the heteroscedasticity of somatic mutation rate estimates across species (**Methods**). Using these models, we estimated that the inverse of lifespan explained 82% of the inter-species variance in somatic substitution rates ($rate = k/lifespan$) (**Fig. 3b**), with the slope of this regression (k) representing the mean estimated ELB across species (3206.4 substitutions per genome per crypt, 95% confidence interval 2683.9–3728.9). Strikingly, despite uncertainty in the estimates of both somatic mutation rates and lifespans, and despite the diverse life histories of the species surveyed, including ~30-fold variation in lifespan and ~40,000-fold variation in body mass, the estimated mutation load per cell at the end of lifespan varied by only ~3-fold across species (**Table 1**). Analogous results were obtained repeating the analysis with protein-coding mutation rate estimates, which may be a better proxy for the functional impact of somatic mutations (85% of variance explained, ELB: 31 coding substitutions per crypt) (**Extended Data Figure 12**, **Methods**).

Table 1. Variation in adult mass, lifespan, mutation rate and end-of-lifespan burden across the 16 mammalian species surveyed. Species-level estimates are provided in Extended Data Tables 3 and 6**.**

Variable	Minimum	Maximum	Fold variation
Adult mass (g)	20.50	800,000.00	39,024.39
Lifespan (yr)	2.75	83.67	30.44
Mutation rate per year (SBS/genome)	47.12	796.42	16.90
End-of-lifespan burden (SBS/genome)	1828.08	5378.73	2.94

We next explored the association between somatic mutation rates and body mass, which is known to be a common confounder in correlations involving lifespan[47,48]. An anticorrelation between somatic mutation rates and body mass may be expected if the modulation of cancer risk across species of vastly different sizes has been a major factor in the evolution of somatic mutation rates. We observed that log-transformed adult body mass is significantly associated with the somatic substitution rate (**Fig. 3c**). However, the fraction of inter-species variance explained (FVE) by body mass was 0.44, considerably lower than that explained by the inverse of lifespan (0.82). Including both variables in the regression model suggested that body mass does not explain a significant amount of variance in somatic mutation rates across species after accounting for the effect of lifespan (likelihood ratio tests, P=0.16 for body mass on a model with lifespan, $P<10^{-4}$ for lifespan on a model with body mass; **Fig. 3d, Methods**). Partial correlation analysis using an allometric model further confirmed that the association between somatic mutation rates and lifespan is unlikely to be mediated by the effect of body mass on both variables (**Methods**).

The fact that the variation in somatic mutation rates across species appears to be dominated by lifespan rather than body size is also apparent when looking at particularly informative species. Giraffe and naked mole-rat, for instance, have similar somatic mutation rates (99 and 93 substitutions/year), in line with their similar lifespans (80th percentiles: 24 and 25 years), despite a ~23,000-fold difference in adult body mass (**Fig. 3b,c**). Cows, giraffes and tigers weigh much more than an average human, and yet have somatic mutation rates several fold higher, in line with expectation from their lifespans but not their body mass. Altogether, the weak correlation between body mass and somatic mutation rates after correction for lifespan suggests that the evolution of larger body sizes may have relied on alternative or additional strategies to limit cancer risk, as has been speculated[23,49–51]. Of note, the low somatic mutation rate of naked mole-rats, unusual for their body mass but in line with their long lifespan (**Fig. 3b,c**), might contribute to the exceptionally low cancer incidence rates of this species[52,53].

We found similar results for other life-history variables that have been proposed to correlate with lifespan, namely basal metabolic rate (BMR) and litter size[54,55] (**Fig. 3d**). With the caveat that estimates for these variables vary in quality, they showed weaker correlations with the somatic mutation rate as single predictors, and small or non-significant increases in explanatory power when considered together with lifespan (likelihood ratio tests, P=0.92 for litter size, P=0.083 for log-BMR, P=0.79 for allometric BMR residuals; **Fig. 3d, Methods**). This result is perhaps unsurprising, as metabolic rate is more strongly dependent on body mass than on lifespan[48]. We note that the results above are robust to the use of alternative estimates of lifespan, including maximum lifespan (**Extended Data Figure 13**, **Methods**), alternative measures of somatic mutation rate, including the rate per exome or mutations/Megabase (**Extended Data Figure 12**, **Methods**), and alternative regression models, including a Bayesian hierarchical model and phylogenetic generalised least-squares regression (**Extended Data Figure 14**, **Methods**).

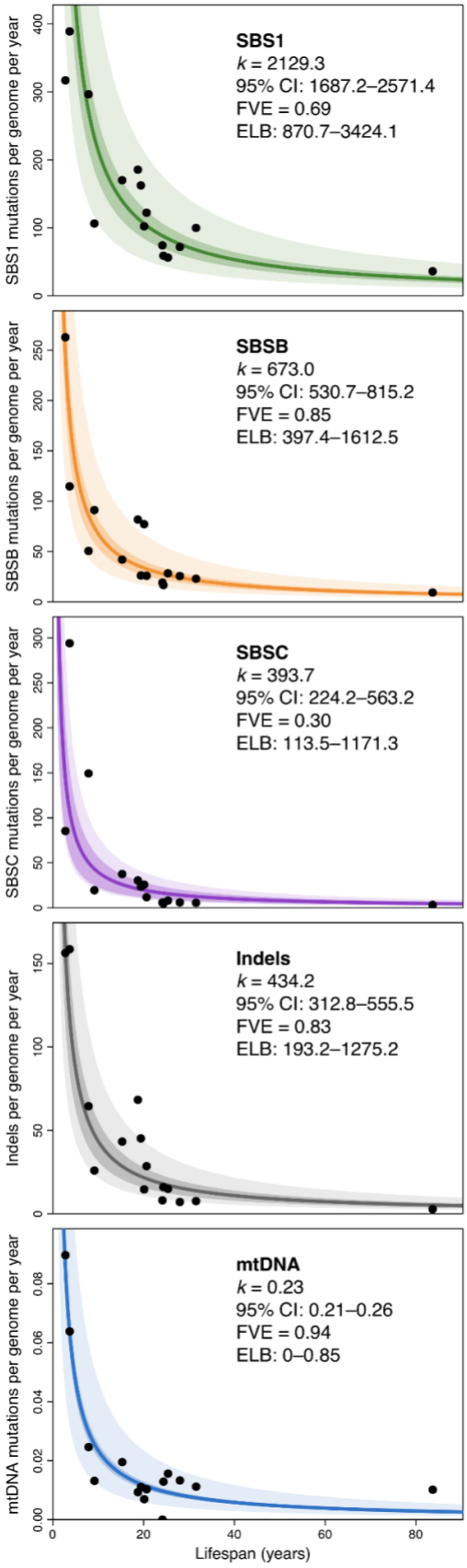
Mutational processes and lifespan

To explore whether a single biological process could drive the association between somatic mutation rates and lifespan, we analysed each mutational signature separately. SBS1, SBSB/5 and SBSC/18 are believed to result from different forms of DNA damage and are expected to be subject

to different DNA repair pathways[12,39,56]. They also appear to differ in their association with the rate of cell division in humans, with SBS1 being more common in fast-proliferating tissues, such as colon and embryonic or foetal tissues, and SBS5 dominating in post-mitotic cells in the absence of cell division[7,12,14]. Overall, we found clear anticorrelations between mutation rates per year and lifespan for the three SBS signatures and for indels, suggesting that a single biological process or DNA repair pathway is unlikely to be responsible for this association (**Fig. 4**). The total mutation burden also appears to show a closer fit with lifespan than individual mutational processes, as measured by the range of end-of-lifespan-burden (ELB) for each process across species (**Fig. 4**). This may be expected if the observed anticorrelation were the result of evolutionary pressure on somatic mutation rates.

DNA damage and somatic mutations in the mitochondrial genome have also attracted considerable interest in the ageing field[57,58]. Our whole-genome sequencing of individual crypts provided high coverage of the mitochondrial genome, ranging from 2,188 to 29,691-fold. Normalised against the nuclear coverage, these data suggest that colorectal crypts contained on the order of ~100–2,000 mitochondrial genomes per cell (**Extended Data Figure 15**). Using a mutation calling algorithm sensitive to low-frequency variants, we found a total of 261 mitochondrial mutations across 199 crypts (**Methods**). The mutational spectra across species appeared broadly consistent with that observed in humans, with a dominance of C>T and A>G substitutions believed to result from mtDNA replication errors rather than DNA damage[59](**Extended Data Figure 16**). While the low number of mitochondrial mutations detected per species precludes a detailed analysis, the estimated number of somatic mutations per copy of mtDNA also appears to show a significant anticorrelation with lifespan. Across species, we obtained an average of 0.23 detectable mutations per copy of the mitochondrial genome by the end of lifespan (**Fig. 4**, **Methods**).

Figure 4. Association between mutation rate subtypes and species lifespan. Zero-intercept linear mixed-effects regression of somatic rates of signature-specific substitutions, indels and mtDNA mutations (top to bottom) on inverse lifespan (1/lifespan), presented on the scale of untransformed lifespan (x-axis). For simplicity, y-axes represent mean mutation rate per species, although mutation rates per crypt were used in the regressions. Darker shaded areas indicate 95% confidence intervals (CI) of the regression lines; lighter shaded areas mark a two-fold deviation from the regression lines. Point estimates of the regression slope coefficient (*k*), fraction of inter-species variance explained by each model (FVE), and ranges of end-of-lifespan burden (ELB) are shown.



Discussion

Using whole-genome sequencing of 208 colorectal crypts from 56 individuals, we provide insights into the somatic mutational landscape of 16 mammalian species. Despite their different diets and life histories, we found remarkable similarities in their mutational spectra. Three mutational signatures explain the spectra across species, albeit with varying contributions and subtle variations in the profiles of the SBSB signature. These results suggest that, at least in the colorectal epithelium, a conserved set of mutational processes dominate somatic mutagenesis across mammals.

The most striking finding of this study is the inverse scaling of somatic mutation rates with lifespan. This has been a long-standing prediction of the somatic mutation theory of ageing[4,30,33]. The observation is consistent with a causative role of somatic mutations in mammalian ageing, although alternative explanations are discussed below. Selection pressure to limit the incidence of cancer, the best-understood pathogenic consequence of somatic mutations, may have contributed to the evolution of somatic mutation rates across species. However, the weak correlation between body mass and somatic mutation rates after correction for lifespan suggests that other adaptations play a role in modulating cancer risk across species[23,49] and that somatic mutations may contribute to ageing in other ways. Traditionally, somatic mutations have been proposed to contribute to ageing by deleterious effects on cellular fitness[4,60], however, the recent discovery of widespread clonal expansions in ageing human tissues[13,61–63] raises the possibility that some somatic mutations contribute to ageing by increasing the fitness of mutant cells at a cost to the organism[60,64,65]. Driver mutations leading to clonal expansions could affect tissue function by causing cell type imbalances[66], cellular senescence[67] or progressive loss of coordination and functional capacity in a tissue. Even if somatic mutations contribute causally to ageing, organismal ageing is a multifactorial process[26,60], and other forms of molecular damage involved in ageing could be expected to display similar anticorrelations with lifespan. Indeed, such anticorrelations have been reported for telomere shortening and protein turnover[68,69].

Alternative explanations for the observed anticorrelation need to be considered. It also remains to be shown to what extent this result extends to other tissues and other species. One alternative

explanation is that cell division rates scale with lifespan and explain the observed somatic mutation rates. Available estimates of cell division rates are imperfect and limited to a few species, but they do not readily support this argument (**Methods**). More importantly, studies in humans have shown that cell division rates are not a major determinant of somatic mutation rates across human tissues[7,12]. An alternative explanation for the observed anticorrelation could be that selection acts to reduce germline mutation rates in species with longer reproductive spans, in turn causing an anticorrelation of somatic mutation rates and lifespan. While this could influence somatic mutation rates, it is unlikely that somatic mutagenesis is tightly determined by germline mutation rates: somatic mutation rates in humans are 10-20 times higher than germline mutation rates and are influenced by additional mutational processes[12,14]. Overall, the strong scaling of somatic mutation rates with lifespan across mammals, despite differences in the contribution of mutational signatures between the germline and the soma and among species, suggests that somatic mutation rates themselves have been evolutionarily constrained, through selection on multiple DNA repair pathways. Alternative explanations would need to be able to explain the strength of the scaling despite the variable contribution of mutational processes across species.

Altogether, this study provides an unprecedented description of somatic mutation across mammals, identifying common and variable features and shedding light on long-standing hypotheses. Scaled across the tree of life and across tissues, in species with vastly different physiologies, life histories, genome compositions and mutagenic exposures, similar studies promise to transform our understanding of somatic mutation and its impact on evolution, ageing, and disease.

References

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
2. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
3. Szilard, L. On the nature of the aging process. *Proc. Natl. Acad. Sci. U. S. A.* **45**, 30–45 (1959).
4. Morley, A. A. The somatic mutation theory of ageing. *Mutat. Res.* **338**, 19–23 (1995).
5. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12–23 (2020).
6. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
7. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
8. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
9. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
10. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
11. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9014–9019 (2019).
12. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* (2021) doi:10.1038/s41586-021-03477-4.
13. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
14. Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *bioRxiv* 2020.11.25.398172 (2020) doi:10.1101/2020.11.25.398172.
15. Garcia, A. M. *et al.* Age- and temperature-dependent somatic mutation accumulation in Drosophila melanogaster. *PLoS Genet.* **6**, e1000950 (2010).
16. Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and

mutational processes. *Nature* **513**, 422–425 (2014).

17. Milholland, B. *et al.* Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
18. Schmid-Siegert, E. *et al.* Low number of fixed somatic mutations in a long-lived oak tree. *Nat Plants* **3**, 926–929 (2017).
19. Orr, A. J. *et al.* A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *Proc. Biol. Sci.* **287**, 20192364 (2020).
20. López, E. H. & Palumbi, S. R. Somatic Mutations and Genome Stability Maintenance in Clonal Coral Colonies. *Mol. Biol. Evol.* **37**, 828–838 (2020).
21. Peto, R., Roe, F. J., Lee, P. N., Levy, L. & Clack, J. Cancer and ageing in mice and men. *Br. J. Cancer* **32**, 411–426 (1975).
22. Caulin, A. F. & Maley, C. C. Peto's Paradox: evolution's prescription for cancer prevention. *Trends Ecol. Evol.* **26**, 175–182 (2011).
23. Tollis, M., Boddy, A. M. & Maley, C. C. Peto's Paradox: how has evolution solved the problem of cancer prevention? *BMC Biol.* **15**, 60 (2017).
24. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *Int. J. Epidemiol.* **45**, 621–637 (2016).
25. 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**, (2015).
26. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
27. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**, 695–703 (2021).
28. Medawar, P. B. UNSOLVED problem of biology. *Med. J. Aust.* **1**, 854–855 (1953).
29. Williams, G. C. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* **11**, 398–411 (1957).
30. Kirkwood, T. B. & Holliday, R. The evolution of ageing and longevity. *Proc. R. Soc. Lond. B Biol. Sci.* **205**, 531–546 (1979).
31. Partridge, L. & Barton, N. H. Optimality, mutation and the evolution of ageing. *Nature* **362**, 305–311 (1993).

32. Hughes, K. A. & Reynolds, R. M. Evolutionary and mechanistic theories of aging. *Annu. Rev. Entomol.* **50**, 421–445 (2005).
33. Burnet, M. Intrinsic mutagenesis. (1974) doi:10.1007/978-94-011-6606-5.
34. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* (2020) doi:10.1038/s41596-020-00437-6.
35. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
36. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
37. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* 372896 (2020) doi:10.1101/372896.
38. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).
39. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat Cancer* **2**, 643–657 (2021).
40. Wilson, M. R. *et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, (2019).
41. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* **580**, 269–273 (2020).
42. Smati, M. *et al.* Quantitative analysis of commensal Escherichia coli populations reveals host-specific enterotypes at the intra-species level. *Microbiologyopen* **4**, 604–615 (2015).
43. Oliero, M. *et al.* Oligosaccharides increase the genotoxic effect of colibactin produced by pks+ Escherichia coli strains. *BMC Cancer* **21**, 172 (2021).
44. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
45. Moorad, J. A., Promislow, D. E. L., Flesness, N. & Miller, R. A. A comparative assessment of univariate longevity measures using zoological animal records. *Aging Cell* **11**, 940–948 (2012).
46. Tidière, M. *et al.* Comparative analyses of longevity and senescence reveal variable survival

benefits of living in zoos across mammals. *Sci. Rep.* **6**, 36361 (2016).

47. Speakman, J. R., Selman, C., McLaren, J. S. & Harper, E. J. Living fast, dying when? The link between aging and energetics. *J. Nutr.* **132**, 1583S–97S (2002).
48. de Magalhães, J. P., Costa, J. & Church, G. M. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J. Gerontol. A Biol. Sci. Med. Sci.* **62**, 149–160 (2007).
49. Risques, R. A. & Promislow, D. E. L. All's well that ends well: why large species have short telomeres. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).
50. Tollis, M., Schneider-Utaka, A. K. & Maley, C. C. The Evolution of Human Cancer Gene Duplications across Mammals. *Mol. Biol. Evol.* **37**, 2875–2886 (2020).
51. Vazquez, J. M. & Lynch, V. J. Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. *Elife* **10**, (2021).
52. Buffenstein, R. Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *J. Comp. Physiol. B* **178**, 439–445 (2008).
53. Smith, E. S. J., Schuhmacher, L.-N. & Husson, Z. The naked mole-rat as an animal model in biomedical research: current perspectives. *Open Access Anim. Physiol.* 137 (2015) doi:10.2147/oaap.s50376.
54. Millar, J. S. & Zammuto, R. M. Life Histories of Mammals: An Analysis of Life Tables. *Ecology* **64**, 631–635 (1983).
55. Speakman, J. R. Body size, energy metabolism and lifespan. *J. Exp. Biol.* **208**, 1717–1730 (2005).
56. Sanders, M. A. *et al.* Life without mismatch repair. *bioRxiv* 2021.04.14.437578 (2021) doi:10.1101/2021.04.14.437578.
57. Lane, R. K., Hilsabeck, T. & Rea, S. L. The role of mitochondrial dysfunction in age-related diseases. *Biochim. Biophys. Acta* **1847**, 1387–1400 (2015).
58. Kauppila, T. E. S., Kauppila, J. H. K. & Larsson, N.-G. Mammalian Mitochondria and Aging: An Update. *Cell Metab.* **25**, 57–71 (2017).
59. Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3**, (2014).
60. Smith, J. M. Review Lectures on Senescence - I. The causes of ageing. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **157**, 115–127 (1962).

61. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
62. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
63. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
64. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, (2019).
65. Ren, A. A. *et al.* PIK3CA and CCM mutations fuel cavernomas through a cancer-like mechanism. *Nature* **594**, 271–276 (2021).
66. Bodmer, W. F. & Crouch, D. J. M. Somatic selection of poorly differentiating variant stem cell clones could be a key to human ageing. *J. Theor. Biol.* **489**, 110153 (2020).
67. Gorgoulis, V. *et al.* Cellular Senescence: Defining a Path Forward. *Cell* **179**, 813–827 (2019).
68. Swovick, K. *et al.* Interspecies Differences in Proteome Turnover Kinetics Are Correlated With Life Spans and Energetic Demands. *Mol. Cell. Proteomics* **20**, 100041 (2021).
69. Whittemore, K., Vera, E., Martínez-Nevado, E., Sanpera, C. & Blasco, M. A. Telomere shortening rate predicts species life span. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15122–15127 (2019).

Acknowledgements

We are grateful to the animals that contributed to this study. We thank the staff of Wellcome Sanger Institute Sample Logistics, Sequencing and Informatics facilities for their contribution; CASM IT and administrative staff for their help processing the data. We thank Nic Masters, Hannah Jenkins, Tilly Dallas and Sharna Lunn for their help obtaining samples. This research was made possible by the worldwide information network of zoos and aquariums, which are members of Species360 and is authorized by Species360 research data use and grant agreement #60633. We thank Dalia Conde and Johanna Staerk for help accessing Species360 data.

Funding

This research was funded by the Wellcome Trust (Grant number 206194). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author Contributions

A.C., E.P.M., M.R.S. and I.M. conceived the project. I.M., E.P.M., and M.R.S. supervised the project. E.F., S.S., I.J., E.W., N.M., R.D., M.W.P., M.D., R.B., M.D.M., F.G., F.C.-C., L.P., D.B., E.St.J.S., B.N., and E.P.M. performed and facilitated sample collection. A.C. performed the laser capture microdissection. A.C., L.M.R.H., A.R.J.L., Y.H., K.R., E.A., S.L., M.M, P.S., J.F., L.O., and C.L processed the samples. A.C., A.B.-O., F.A., N.B., T.H.H.C., M.A.S., D.J., S.G.B, R.A. and K.R processed the data. J.S. and D.C provided animal longevity records. A.C., A.B.-O. and N.B. led the analysis with help from F.A., T.H.H.C., M.A.S., A.R.J.L., T.M.B., T.D., H.J., E.P.M. and I.M. The manuscript was written by A.C., A.B.-O., N.B. and I.M. with input from all the authors.

Competing Interests

The authors declare no competing interests.

METHODS

1. Sample collection

We obtained colorectal epithelium and skin samples from a range of sources (**Extended Data Table 1**). For comparability across species a ~1 cm biopsy of the colorectal epithelium was taken from the terminal colon during necropsy. All necropsies occurred as soon as possible post-mortem to minimise tissue and DNA degradation. Tissue samples taken later than 24 hours post-mortem typically showed extensive degradation of the colorectal epithelium, making identification of colorectal crypts challenging. These samples were also associated with poor DNA yields and so were not included in his study. Sampled tissue was fixed in PAXgene FIX (PreAnalytiX, Hombrechtikon, Switzerland), a commercially available fixative, during the necropsy. After 24 hours in the fixative at room temperature samples were transferred into the PAXgene STABILIZER and stored at -20°C until further processing.

2. Sample processing

Samples were processed using a workflow designed for detection of somatic mutations in solid tissues by laser-capture microdissection (LCM) using low-input DNA sequencing. For a more detailed description see the paraffin workflow described in Ellis et al[34]. Briefly, PAXgene-fixed tissue samples of the colorectal epithelium were paraffin-embedded using a Sakura Tissue-Tek VIP tissue processor. Sections of 10 µm were cut using a microtome, mounted on PEN-membrane slides, and stained with Gill's haematoxylin and eosin by sequential immersion in the following: xylene (two minutes, twice), ethanol (100%, 1 minute, twice), deionised water (1 minute, once), Gill's haematoxylin (10 seconds, once), tap water (20 seconds, twice), eosin (5 seconds, once), tap water (20 seconds, once), ethanol (70%, 20 seconds, twice) and xylene or Neo-Clear, a xylene substitute (20 seconds, twice).

High-resolution scans were obtained from representative sections of each species. Example images are shown in **Fig. 1a** and **Extended Data Figure 2**. Individual colorectal crypts were isolated from sections on polyethylene naphthalate (PEN) membrane slides by laser-capture microdissection with a Leica LMD7 microscope. Haematoxylin and eosin histology images were reviewed by a veterinary pathologist. For some samples we also cut a section of muscle

tissue from below the colorectal epithelium of the section to use as a germline control for variant calling (**Extended Data Table 2**). Pre- and post-microdissection images of the tissue were recorded for each crypt and muscle sample taken. Each microdissection was collected in a separate well of a 96-well plate.

Crypts were lysed using the Arcturus PicoPure Kit (Applied Biosystems) as previously described[8,34]. Each crypt then underwent DNA library preparation, without a quantification step to avoid loss of DNA, following the protocol described in Ellis et al.[34]. For some animals a PAXgene fixed bulk skin biopsy was used as the germline control. For these skin samples, DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen).

3. Library preparation and sequencing

Libraries from microdissected samples were prepared using enzymatic fragmentation, adapter ligation and whole-genome sequencing following the method described in Ellis et al.[34]. Libraries from skin samples were prepared using standard Illumina whole-genome library preparation. Samples were multiplexed and sequenced using Illumina XTEN and Novaseq 6000 machines to generate 150 base pair (bp) paired-end reads. Samples were sequenced to ~30× depth (**Extended Data Table 2**).

4. Sequence read alignment

For each species sequences were aligned to a reference assembly (**Extended Data Table 2**) using the BWA-MEM algorithm[70] as implemented in BWA version 0.7.17-r1188, with options '-T 30 -Y -p -t 8'. The aligned reads were sorted using the bamsort tool from the biobambam2 package, version 2.0.86 (gitlab.com/german.tischler/biobambam2), with options 'fixmates=1 level=1 calmdnm=1 calmdnmrecompindetonly=1 calmdnmreference=<reference_fasta> outputthreads=7 sortthreads=7'. Duplicate reads were marked using the bammarkduplicates2 tool from biobambam2, with option 'level=0'.

5. Variant calling

Identification of somatic single-base substitutions (SBS) and short insertions and deletions (indels) was divided into two steps: variant calling, and variant filtering to remove spurious calls

(see 'Variant filtering' below). For human colorectal crypts, we obtained previously sequenced and mapped reads from a study where colorectal crypts were isolated by laser-capture microdissection[8], and processed them using the sample variant calling and filtering process that was applied to the non-human samples.

Substitutions were identified using the Cancer Variants through Expectation Maximization (CaVEMan) algorithm[71], version 1.13.15. CaVEMan uses a naive Bayesian classifier to perform a comparative analysis of the sequence data from a target and control sample from the same individual to derive a probabilistic estimate for putative somatic substitutions at each site. The copy number options were set to 'major copy number = 5' and 'minor copy number = 2', as in our experience this maximises the sensitivity to detect substitutions in normal tissues. CaVEMan identifies and excludes germline variants shared in the target (colorectal crypt) and matched normal (skin or muscle tissue) samples, and produces a list of putative somatic mutations present only in the target sample. CaVEMan was run separately for each colorectal crypt, using either bulk skin or muscle microdissected from the sample colorectal biopsy as the matched normal control (**Extended Data Table 2**). For two human donors where an alternative tissue was not available, a colonic crypt not included as a target sample was used as the matched normal control.

Indels were identified using the Pindel algorithm[72], version 3.3.0, using a second sample from the same individual as a matched control. The indel calls produced by Pindel were subsequently re-genotyped using the vafCorrect tool (github.com/cancerit/vafCorrect), which performs a local sequence assembly to address alignment errors for indels located at the end of sequence reads, and produces corrected counts of sequence reads supporting the indel and corrected estimates of variant allele fraction (VAF).

6. Variant filtering

A number of post-processing filters were applied to the variant calls to remove false positives (**Extended Data Figure 17a,b**).

Quality flag filter. CaVEMan and Pindel annotate variant calls using a series of quality flags, with the 'PASS' flag denoting no quality issues affecting the call[71,72]. Variant calls presenting any flag other than 'PASS' were discarded.

Alignment quality filter. Variants were excluded if more than half of the supporting reads were clipped. The library preparation methods create short insert size libraries that can result in reads overlapping. To avoid the risk of double-counting mutant reads we used fragment-based statistics. Variants without at least four high-quality fragments (alignment score ≥ 40 and base Phred quality score ≥ 30) were excluded. Variants were excluded if reads supporting the variant had a secondary alignment score that was greater than the primary alignment score. This filter was not applied to indel calls.

Hairpin filter. To remove variants introduced by erroneous processing of cruciform DNA during the enzymatic digestion we applied a custom filter to remove variants in inverted repeats[34]. This filter was not applied to indel calls.

Chromosome and contig filter. For species with chromosome-level assemblies, we discarded variants located in non-chromosomal contigs, including the mitochondrial genome (calling of mitochondrial variants is described in the section 'Mitochondrial variant calling and filtering'). For males, variants on the Y chromosome were excluded for species where the Y chromosome was annotated in the assembly.

N-tract and contig-end filter. To reduce artefactual calls due to read misalignment, we discarded variants located within 1 kilobase (kb) of a tract of 50 or more consecutive N bases in the reference assembly, as well as variants within 1 kb of the start or end of a contig (this implies discarding all variants in contigs shorter than 2 kb).

Sequencing coverage filter. A sample-specific read depth filter was designed to exclude sites with coverage above the 99th coverage percentile in the sample or its matched normal control, as well as sites with coverage < 10× in the sample or its matched normal control.

Allelic strand bias filter. We discarded variants without any supporting reads on either the forward or reverse strand.

Indel proximity filter. We discarded variants for which the total number of reads supporting the presence of an indel within 10 bp of the variant was more than 3 times larger than the number of reads supporting the presence of the variant. This filter was not applied to indel calls.

Spatial clustering filter. Visual assessment of variant calls and mutational spectra showed spatially clustered variants to be highly enriched for artefacts. Therefore, we discarded groups of two or more variants located within 1 kb of each other.

Beta-binomial filter. For each crypt, an artefact filter based on the beta-binomial distribution was applied, which exploits read count information in other crypts from the same individual. More specifically, for each sample, we fitted a beta-binomial distribution to the variant allele counts and sequencing depths of somatic variants across samples from the same individual. The beta-binomial distribution was used to determine whether read support for a mutation varies across samples from an individual, as expected for genuine somatic mutations but not for artefacts. Artefacts tend to be randomly distributed across samples and can be modelled as drawn from a binomial or a lowly overdispersed beta-binomial distribution. True somatic variants will be present at a high VAF in some samples, but absent in others, and are hence best captured by a highly overdispersed beta-binomial. For each variant site, the maximum likelihood estimate of the overdispersion factor (ρ) was calculated using a grid-based method, with values ranging between 10^{-6} and $10^{-0.05}$. Variants with $\rho > 0.3$ were considered to be artefactual and discarded. The code for this filter is based on the Shearwater variant caller[73]. We found this to be one of the most effective filters against spurious calls (**Extended Data Figure 17b**).

Minimum VAF filter. For each sample, we discarded variants whose variant allele fraction (VAF) was less than half the median VAF of variants passing the beta-binomial filter (see above) in that sample.

To validate our variant calling strategy, we used LCM to microdissect two sections from the same mouse colorectal crypt. We expected to detect a high fraction of shared somatic variants in these two sections, since their cells should be derived from the same ancestral epithelial stem cell. Both sections were submitted for independent library preparation, genome sequencing, variant calling and filtering using our pipeline. The majority of SBS variant calls (2742 of 2933, 93.5%) were shared between both sections (**Extended Data Figure 17c**). In contrast, when comparing five separate crypts from a mouse, a maximum of two variants were shared between two crypts, and no variants were shared by three or more crypts (**Extended Data Figure 17d**).

7. Sample filtering

Our method for estimation of mutation rates assumes monoclonality of colorectal crypt samples. This assumption can be violated due to several causes, including contamination from other colorectal crypts during microdissection or library preparation, contamination with non-epithelial cells located in or near the crypt, insufficient time for a stem cell to drift to clonality within the crypt, or the possibility that in some species, unlike in humans[8], polyclonal crypts are the norm. Therefore, a truncated binomial mixture model was applied in order to remove crypts that showed evidence of polyclonality, or for which the possibility of polyclonality could not be excluded. An expectation–maximization (EM) algorithm was employed to determine the optimal number of variant allele fraction (VAF) clusters within each crypt sample, as well as each cluster's location and relative contribution to the overall VAF distribution. The algorithm considered a range of numbers of clusters (1–5), with the optimal number being that which minimised the Bayesian Information Criterion (BIC). As the minimum number of supporting reads to call a variant was 4, the binomial probability distribution was truncated to incorporate this minimum requirement for the number of successes, and subsequently re-normalised. The EM algorithm returned the inferred optimal number of clusters, the mean VAF (location) and mixing proportion (contribution) of each clone, and an assignment of each input variant to the most likely cluster. After applying this model to the somatic substitutions identified in each sample, sample filtering was performed on the basis of the following three criteria.

Low mutation burden. We discarded samples presenting fewer than 50 somatic variants, which was indicative of low DNA quality or sequencing issues.

High mutation burden. We discarded samples with a number of somatic variants greater than three times the median burden of samples from the same individual (excluding samples with less than 50 variants). This served to exclude a small minority of samples presenting evident sequencing quality problems (such as low sequencing coverage), but which did not fulfill the low-VAF criterion for exclusion (see below).

Low VAF. We discarded samples in which less than 70% of the somatic variants were assigned to clusters with VAF ≥ 0.3. However, this rule was not applied to those cases in which all the samples from the same individual had primary clusters with mean VAF < 0.3; this was done to prevent the removal of samples from individuals presenting high fractions of non-epithelial cells, but whose crypts were nonetheless dominated by a single clone.

These criteria led to the exclusion of 41 out of 249 samples. On the basis of visual assessment of sequencing coverage and VAF distributions, we decided to preserve three samples (ND0003c_lo0004, ND0003c_lo0011, TIGRD0001b_lo0010) which we considered to be clonal, but which would have been discarded based on the criteria above.

8. Mitochondrial variant calling and filtering

For six species whose reference genome assemblies did not include the mitochondrial sequence, mitochondrial reference sequences were obtained from the GenBank database (**Extended Data Table 5**). For each species, alignment to the reference genome was performed using BWA version 0.7.17-r1188, as described above (see 'Sequence read alignment'). Pileup files were generated for mtDNA genomes using the 'bam2R' function in the deepSNV (v1.32.0) R package[73,74]. The mapping quality cut-off was set to 0, taking advantage of the fact that the mitochondrial genome coverage for most samples was >100-fold higher than the nuclear genome coverage, and hence most reads with poor mapping scores should be of mitochondrial origin. Mitochondrial variants were called using the Shearwater algorithm[73] (deepSNV package v1.32.0). Multiple rounds of filtering were applied to identify and remove false positives. The first set of filters removed germline polymorphisms, applied a maximum false discovery rate (FDR) threshold of $q > 0.01$, required that mismatches should be supported by at least one read on both the forward and reverse strands, and merged consecutive indel calls. Further filtering steps were as follows.

Minimum VAF filter. Only variants with VAF > 0.01 were considered for analysis, based on the quality of the mutational spectra.

Sequencing coverage filter. Due to species-specific mtDNA regions of poor mappability, we discarded sites with read coverage < 500×.

D-loop filter. Analysis of the distribution of mutations along the mitochondrial genome revealed clusters of mutations within the hypervariable region of mtDNA known as the D-loop. To obtain estimates of the mutation burden in mtDNA unaffected by hypermutation of the D-loop, mutations in the D-loop region (coordinates MT:1–576 and MT:16,024–16,569 in human) were excluded from this analysis

High mutation burden. We discarded samples having a number of somatic mtDNA variants greater than four times the mean mtDNA burden across all samples. This served to exclude a

small minority of samples that were suspected of enrichment in false positive calls. Visual inspection of these samples in a genome browser confirmed the presence of high numbers of variants found on sequence reads with identical start positions and/or multiple base mismatches, suggestive of library preparation or sequencing artefacts.

We examined the mutational spectra of somatic mtDNA substitutions on a trinucleotide sequence context (**Extended Data Figure 16**). The specificity of the filtered variant calls was supported by the observation that the mutational spectra across species were broadly consistent with those previously observed in studies of human tissues[59], with a dominance of C>T and T>C transversions and a strong replication strand bias.

9. Mitochondrial copy number analysis

Sequence reads from each sample were separately mapped to the species-specific mtDNA reference sequence in order to estimate average mtDNA sequencing coverage. Excluding nuclear reference sequences from the alignment enabled obtaining even coverage across the mitochondrial genome by preventing mismapping of sequence reads to inherited nuclear insertions of mitochondrial DNA (known as NuMTs). Next, coverage information from individual mtDNA and whole-genome alignment (BAM) files was obtained using the genomecov tool in the bedtools suite (v2.17.0)[75]. Mitochondrial copy number was calculated according to the formula

$$depth_{mtDNA} \times ploidy \,/\, depth_{gDNA},$$

where $depth_{mtDNA}$ and $depth_{gDNA}$ are the mean coverage values for mtDNA and the nuclear genome, respectively, and $ploidy = 2$ (assuming normal somatic cells to be diploid). For simplicity, the sex chromosomes were excluded from the calculation of the mean nuclear genome coverage.

10. Analysable genome size calculation

To estimate the somatic mutation rate, it was first necessary to establish the size of the analysable nuclear genome (i.e. the portion of the genome where variant calling could be performed reliably) for each sample (**Extended Data Table 4**). For both single-base substitutions and indels, the analysable genome of a sample was defined as the complement of

the union of the following genomic regions: regions reported as 'not analysed' by the CaVEMan variant caller; regions failing the 'chromosome and contig' filter; regions failing the 'N-tract and contig-end' filter; and regions failing the 'sequencing coverage' filter (see 'Variant filtering'). For the analysis of mitochondrial variants, the analysable genome of a sample was defined as the portion of mtDNA satisfying the 'sequencing coverage' filter (see 'Mitochondrial variant calling and filtering'), after subtracting the hypervariable region (D-loop).

11. Life history data

Obtaining accurate lifespan estimates is challenging; while point estimates of maximum lifespan are available for many species, their veracity is often difficult to assess and estimates can vary widely for the same species (**Extended Data Table 6**). There can be many causes for this variation, including errors in recording and real variation in longevity between populations (i.e. captive *versus* wild). As we were interested in whether the somatic mutation burden has an association with lifespan in the absence of extrinsic mortality, we sought to obtain estimates of longevity from individuals under human care, to minimise the impact of external factors such as predation or infection.

Mortality records for 14 species were obtained from the Species360 database, authorized by Species360 research data use agreement #60633 [Species360 Zoological Information Management System (ZIMS) (2020), zims.Species360.org]. This database contains lifespan data of zoo animals from international zoo records. Using records from 1980 to the present, we excluded animals for which the date of birth or death was unknown or uncertain. To avoid infant mortality influencing the longevity estimates for each species, we removed animals that died before the age of female sexual maturity, as defined by the AnAge database[76]. This resulted in a mean of 2,681 animal lifespan records per species for the species in the study (minimum 309, maximum 8403; **Extended Data Table 6**). For the domestic dog, we combined records for domestic dogs (*Canis lupus familiaris*) and wolves (*Canis lupus*) because of the paucity of records for domestic dogs in Species360. Although the data are curated, they are still vulnerable to the presence of inaccurate records, which can bias the lifespan estimates. To reduce the impact of these outliers, for each species lifespan was estimated as the age at which 80% of the adults from that species had died (**Extended Data Table 6**)[46].

Human longevity estimates were obtained using census birth and death record data from Denmark, (1900–2020), Finland (1900–2019) and France (1900–2018), retrieved from the Human Mortality Database [University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany); www.mortality.org, www.humanmortality.de]. We selected these countries because they had census records going back at least 100 years. To remove the impact of infant mortality, we excluded individuals who died before the age of 16. For each country, we selected the cohort born in 1900 and calculated the age at which 80% of the individuals had died (Denmark, 87 years; Finland, 83 years; France, 81 years). We then used the mean of the three countries as our estimate of the human 80% lifespan (83.7 years) (**Extended Data Table 6**).

To test the impact of different estimates of lifespan on our results, we also obtained maximum longevity estimates for each species from a range of databases[77] and a survey of the literature (**Extended Data Table 6**). Other life-history metrics were obtained from the AnAge database[76] (**Extended Data Table 6**).

12. Mutational signature analysis

Mutational signatures of single-base substitutions on a trinucleotide sequence context were inferred from sets of somatic mutation counts using the sigfit (v2.0) R package[37]. Initially, signature extraction was performed *de novo* for a range of numbers of signatures (N = 2, ..., 10), using counts of mutations grouped per sample, per individual and per species. To account for differences in sequence composition across samples, and especially across species, mutational opportunities per sample, per individual and per species were calculated from the reference trinucleotide frequencies across the analysable genome of each sample (see 'Analysable genome size calculation'), and supplied to the 'extract_signatures' function in sigfit. The 'convert_signatures' function in sigfit was subsequently used to transform the extracted signatures to a human-relative representation (**Fig. 2b**), by scaling the mutation probability values using the corresponding human genome trinucleotide frequencies. The best-supported number of signatures, on the basis of overall goodness-of-fit[37] and consistency with known COSMIC signatures (cancer.sanger.ac.uk/signatures), was found to be N = 3. The cleanest

deconvolution of the three signatures was achieved when using the mutation counts grouped by species, rather than by sample or individual. The three extracted signatures (labelled SBSA, SBSB, SBSC) were found to be highly similar to COSMIC signatures SBS1 (cosine similarity 0.96), SBS5 (0.89), and SBS18 (0.91), respectively. These signatures were independently validated using the MutationalPatterns (v1.12.0) R package[78], which produced comparable signatures (respective cosine similarities 0.999, 0.98 and 0.89).

This *de novo* signature extraction approach, however, failed to deconvolute signatures SBSA and SBSB entirely from each other, resulting in a general overestimation of the exposure to SBSA (**Extended Data Figure 18**). To obtain more accurate estimates of signature exposure, the deconvolution was repeated using an alternative approach that combines signature fitting and extraction in a single inference process[37]. More specifically, the 'fit_extract_signatures' function in sigfit was used to fit COSMIC signature SBS1 (retrieved from the COSMIC v3.0 signature catalogue; cancer.sanger.ac.uk/signatures) to the mutation counts grouped by species (with species-specific mutational opportunities), while simultaneously extracting two additional signatures *de novo* (SBSB and SBSC). Before this operation, COSMIC SBS1 was transformed from its human-relative representation to a genome-independent representation using the 'convert_signatures' function in sigfit. By completely deconvoluting SBS1 and SBSB, this approach yielded a version of SBSB that was more similar to COSMIC SBS5 (cosine similarity 0.93); the similarity of SBSC to COSMIC SBS18 was the same under both approaches (0.91).

Finally, the inferred signatures were re-fitted to the mutational spectra of mutations in each sample (using the 'fit_signatures' function in sigfit with sample-specific mutational opportunities) to estimate the exposure of each sample to each signature. The fitting of the three signatures yielded spectrum reconstruction similarity values (measured as the cosine similarity between the observed mutational spectrum and a spectrum reconstructed from the inferred signatures and exposures) with median 0.98 and interquartile range 0.96–0.99. Although the purely *de novo* extraction approach and the 'fitting and extraction' approach yielded comparable versions of signatures SBSB and SBSC, the fixing of COSMIC SBS1 in the latter approach resulted in lower SBS1 exposures and higher SBSB exposures in the majority of samples, due to the cleaner deconvolution of these two signatures (**Fig. 2, Extended Data Figure 18**).

To examine potential variation in the spectrum of signature SBS5 across species, the following procedure was conducted for each species: individual-specific mutation counts and mutational opportunities were calculated for each individual in the species, and the 'fit_extract_signatures' function was used to fit COSMIC signatures SBS1, SBS18 and SBS34 (transformed to a genome-independent representation using the 'convert_signatures' function) to the mutational spectra of each individual, while simultaneously inferring one additional signature (corresponding to signature SBS5 as manifested in that species; **Extended Data Figure 6**).

To assess the presence in non-human colorectal crypts of mutational signatures caused by APOBEC or colibactin, which have been previously observed in human crypts[8], we used an expectation–maximisation algorithm for signature fitting, in combination with likelihood ratio tests (LRTs). More specifically, for each non-human sample, we tested for exposure to colibactin (signature SBS88, COSMIC v3.2) by comparing the log-likelihoods of (*i*) a model fitting COSMIC signatures SBS1, SBS5, SBS18, SBS34 and SBS88, and (*ii*) a reduced model fitting only the first four signatures. Benjamini–Hochberg multiple-testing correction was applied to the *p*-values resulting from the LRTs, and colibactin exposure was considered significant in a sample if the corresponding corrected *q*-value was less than 0.05. We followed the same approach to assess exposure to APOBEC (SBS2 and SBS13), using two separate sets of LRTs for models including either SBS2 or SBS13, in addition to SBS1, SBS5, SBS18 and SBS34. APOBEC exposure was considered significant in a sample if its *q*-values for the models including SBS2 and SBS13 were both less than 0.05. This analysis identified 1/180 crypts with significant exposure to each of colibactin and APOBEC (although the evidence for the presence of the relevant signatures in these two crypts was not conclusive). To test for depletion of colibactin or APOBEC exposure in non-human crypts relative to human crypts, we first applied the LRT-based method described above to a published set of 445 human colorectal crypts[8], identifying 92 colibactin-positive and 9 APOBEC-positive crypts. We then compared the numbers of colibactin- and APOBEC-positive crypts in the human and non-human sets using two separate Fisher's exact tests ('fisher.test' function in R). This revealed the difference in colibactin exposure to be highly significant ($P=7\times10^{-14}$), unlike the difference in APOBEC exposure ($P=0.30$).

Mutational spectra of somatic indels identified in human, mouse, rat and dog samples were generated using the SigProfilerMatrixGenerator (v1.1) software[79], and the 17 indel signatures available in COSMIC (v3.0) were fitted to these mutational spectra using the 'fit_signatures' function in sigfit.

13. Selection analysis

Evidence of selection was assessed using the ratio of nonsynonymous to synonymous substitution rates (dN/dS) in the somatic mutations called in each species. The dNdScv (v0.0.1.0) R package[44] was used to estimate dN/dS ratios for missense and truncating substitutions in each species separately. Reference CDS databases for the dNdScv package were built for those species with available genome annotation in Ensembl (www.ensembl.org; **Extended Data Table 2**), using the 'buildref' function in dNdScv. For each species, the 'dndscv' function was applied to the list of somatic substitutions called in samples of that species, after de-duplicating any substitutions that were shared between samples from the same individual in order to avoid counting shared somatic mutations multiple times. In addition, the analysis was restricted to genes that were fully contained in the analysable genomes of all samples from the species (a condition satisfied by the vast majority of protein-coding genes). Genome-wide and gene-specific dN/dS ratios were obtained for missense and truncating substitutions in each species; no genes with statistically significant dN/dS $\neq 1$ were observed.

14. Copy number analysis

For species with chromosome-level assemblies (cat, cow, dog, horse, human, mouse, rabbit, rat), total and allele-specific copy number (CN) were assessed in each sample adapting a likelihood model previously applied to the detection of subclonal CN changes in normal human skin[13]. This method exploits two sources of evidence: relative sequencing coverage and B-allele fraction (BAF; the fraction of reads covering a heterozygous SNP that support one of the alleles). Human samples PD36813x15 and PD36813x16 were excluded from this analysis due to the poor quality of their SNP data.

For each sample, sequencing coverage was measured in non-overlapping 100-kilobase (kb) bins along the species' reference genome, using the coverageBed tool in the bedtools suite (v2.17.0)[75]. For each bin, the coverage per base pair was calculated by dividing the number of reads mapping to the bin by the bin length, and multiplying the result by the read length (150 bp). A normalised sample–normal coverage ratio was then calculated for each bin by dividing the bin coverage in the sample by the corresponding coverage in its matched normal control (see 'Sample processing'). Heterozygous SNPs were isolated for each sample by selecting germline SNPs with a BAF between 0.4 and 0.6 in the matched normal sample, and a coverage of at least 15 reads in both the target sample and its matched normal sample. After assigning each SNP to its corresponding 100-kb genome bin, the bins in each sample were divided into two sets: (*i*) bins with coverage ≥ 10 in both the target sample and its matched normal, and at least one heterozygous SNP; and (*ii*) bins with coverage ≥ 10 in both the target sample and its matched normal, and no heterozygous SNPs. For the first set, estimates of total and allele-specific CN were inferred by maximising the joint likelihood of a beta-binomial model for BAF and a negative binomial model for relative coverage, as previously described[13]. The most likely combination of allele CN values was obtained for each bin by conducting an exhaustive search of CN values between 0 and 4, and selecting the combination maximising the joint likelihood (calculated on the basis of expected BAF and relative coverage values). A penalty matrix was used to penalise more complex solutions over simpler ones, as previously described[13]. For the second set of bins (bins without SNPs), only estimates of total CN were inferred, by maximising the likelihood of a negative binomial model for relative coverage. The most substantial differences between these methods and the one previously published are: (*i*) SNPs were obtained from the variant calling output, instead of from a public database; (*ii*) relative coverage was calculated per 100-kb bin, rather than per SNP; (*iii*) SNPs were not phased within each gene, but within each bin; (*iv*) no reference bias was assumed (i.e. the underlying BAF of heterozygous SNPs was assumed to be 0.5); (*v*) the minimum sample purity was raised to 0.85; (*vi*) putative CN changes were not subjected to significance testing, but selected according to their likelihood, and subsequently filtered by means of a segmentation algorithm (see below).

Estimates of total and allele-specific CN per bin were merged into CN segments, which were defined as contiguous segments composed of five or more bins with identical CN states.

Segmentation was performed separately for total and allele-specific CN estimates in each sample. After this process, any pair of adjacent segments with the same CN assignment, and separated by a distance shorter than 5 bins, was merged into a single segment. Finally, within each species, segments presenting CN values other than 2 (or 1/1 for allele-specific CN), and being either shorter than 10 bins (1 megabase), or shared among two or more samples, were discarded, resulting in the removal of nearly all spurious CN changes.

15. Mutation rate estimation

For each sample, the somatic mutation density (mutations per bp) was calculated by dividing the somatic mutation burden (total number of mutations called) by the analysable genome size for the sample (see 'Analysable genome size calculation'). The adjusted somatic mutation burden (number of mutations per whole genome) was then calculated by multiplying the mutation density by the total genome size of the species (see below). The somatic mutation rate per year (mutations per genome per year) was obtained by dividing this adjusted mutation burden by the age of the individual, expressed in years (**Extended Data Table 2**). The expected end-of-lifespan burden (ELB) for each sample was calculated by multiplying the somatic mutation rate by the estimated lifespan of the species (see 'Life history data').

The total genome size of a species was estimated as the total size of its reference genome assembly. Across species, the mean genome size was 2.67 gigabases (Gb), ranging between 2.41 Gb and 3.15 Gb and with a standard deviation of 221 megabases (**Extended Data Table 4**). This suggests that inter-species variation in genome size should not have a substantial influence on the somatic mutation rate estimates. For an assessment of alternative measures of mutation rate, see 'Association of mutation rate and end-of-lifespan burden with lifespan'.

16. Association of mutation rate with life-history traits

The association of the somatic mutation rate with different life-history traits was assessed using linear mixed-effects (LME) models. In particular, associations with the following traits were examined: lifespan (in years), adult mass (or adult weight, in grams), basal metabolic rate (BMR, in watts), and litter size (see 'Life history data'). Associations for lifespan, adult mass and BMR were assessed using the following transformed variables: 1/Lifespan, $\log_{10}$(Adult mass), and

$\log_{10}$(BMR). To account for the potentially confounding effect of the correlation between metabolic rate and body mass, the residuals of a fitted allometric regression model of BMR on adult mass (equivalent to a simple linear regression of $\log_{10}$(BMR) on $\log_{10}$(Adult mass)) were employed as a mass-adjusted measure of metabolic rate, referred to as 'BMR residuals'.

For each variable, an LME model was implemented for the regression of somatic mutation rates per sample on the variable of interest, using the 'lme' function in the nlme R package (v3.1-137; cran.r-project.org/package=nlme). To account for non-independence of the samples, both at the individual level and at the species level, the model included fixed effects (intercept and slope parameters) for the variable of interest, and random effects (slope parameters) at the individual and species levels. In addition, to account for the heteroscedasticity of mutation rate estimates across species, the usual assumption of constant response variance was replaced with explicit species-specific variances, to be estimated within the model.

To determine the fraction of inter-species variance in mutation rate explained by each life-history variable individually, the LME model described above was used to produce predictions of the mean mutation rate per species; only fixed effects were employed when obtaining these predictions, random effects being ignored. The variance of these predictions was then compared to the variance in observed mean mutation rates; the latter were calculated for each species as the mean of the observed mean rates per individual, to avoid individuals with larger numbers of samples exerting a stronger influence on the species mean. The fraction of inter-species variance explained by the model was calculated using the standard formula for the coefficient of determination,

$$R^2 = \text{ESS} / (\text{ESS} + \text{RSS}),$$

where ESS is the explained sum of squares, and RSS is the residual sum of squares:

$$\text{ESS} = \sum_i \quad (\hat{y}_i - \bar{y})^2, \quad \text{RSS} = \sum_i \quad (y_i - \hat{y}_i)^2.$$

In this formulation, y_i and $\hat{y}_i$ denote the observed and predicted mutation rates for species *i*, respectively, and $\bar{y}$ is the overall mean rate. This definition of R^2 coincides with the fraction of variance explained (FVE), defined as 1 minus the fraction of variance unexplained (FVU):

$$\text{FVE} = 1 - \text{FVU} = 1 - [\text{RSS} / (\text{ESS} + \text{RSS})] = \text{ESS} / (\text{ESS} + \text{RSS}) = R^2.$$

As the predicted and observed values correspond to mean mutation rates per species, rather than mutation rates per sample, FVE provides a measure of the fraction of inter-species variance explained by the fixed effects of the LME model. Among the variables considered, 1/Lifespan was found to have the greatest explanatory power (FVE = 0.84, using a free-intercept model).

To compare the explanatory power of variables other than 1/Lifespan when considered either individually or in combination with 1/Lifespan, the method described above was also applied to two-variable combinations of 1/Lifespan and each of the remaining variables, using an LME model with fixed effects for both variables and random effects for 1/Lifespan only. The R^2 formula above was used to measure the fraction of inter-species variance explained by each model. In addition, to test whether the inclusion of a second explanatory variable was justified by the increase in model fit, likelihood ratio tests between each two-variable LME model and a reduced LME model including only 1/Lifespan were performed using the 'anova' function in the nlme R package.

To further assess the potential effects of body mass and lifespan on each other's association with the somatic mutation rate, allometric regression models (equivalent to simple linear models under logarithmic transformation of both variables) were fitted to the mean somatic mutation rate per species, using either adult mass or lifespan as the explanatory variable. In addition, the 'allometric residuals' of mutation rate, adult mass and lifespan (i.e. the residuals of pairwise allometric regressions among these three variables) were used to examine the associations between somatic mutation rate and either body mass or lifespan, after accounting for the effect of the third variable (partial correlation analysis). For instance, to account for the potential influence of body mass on the relationship between somatic mutation rate and lifespan, the residuals of an allometric regression between mutation rate and adult mass, and the residuals of an allometric regression between lifespan and adult mass, were analysed using simple linear regression. This analysis supported a strong association between somatic mutation rate and lifespan (independently of the effect of mass; FVE=0.82, P=3.2×10^{-6}; **Extended Data Figure 19**), and a non-significant association between somatic mutation rate and body mass (independently of the effect of lifespan). Therefore, the relationship between somatic mutation rate and lifespan does not appear to be mediated by the effect of body mass on both variables.

The results obtained with the LME models described above were additionally validated using an independent hierarchical Bayesian model, in which the mean somatic mutation burden of each individual was modelled as following a normal distribution with mean defined as a linear predictor containing a species-specific slope parameter and a multiplicative offset (corresponding to the individual's age; inclusion of this offset minimises the heteroscedasticity of the mutation rate across species, which results from dividing mutation burdens by age). Species-specific slope parameters were in turn modelled as normally distributed around a global slope parameter, equivalent to the fixed-effect slope estimated by the LME model. This hierarchical model produced very similar results to those of the LME model for all life-history variables (**Extended Data Figure 14**).

We note that samples CATD0002b_lo0003 and MD6267ab_lo0003 were excluded from all regression analyses, due to the fact that each shared the majority of its somatic variants with another sample from the same individual (indicating, in each case, that both samples were closely related), hence violating the assumption of independence among samples. The inclusion of these two samples, however, had no effect on the outcome of the analyses.

17. Association of mutation rate and end-of-lifespan burden with lifespan

The relationship between somatic mutation rate and species lifespan was further explored by adapting the LME model described in the previous section to perform constrained (zero-intercept) regression of the adjusted mutation rate per year on the inverse of lifespan, 1/Lifespan (see 'Life history data', 'Mutation rate estimation' and 'Association of mutation rate with life-history traits'). The use of zero-intercept regression was motivated by the prediction that, if somatic mutation is a determinant of maximum lifespan, then it would be expected for all species to end their lifespans with a similar somatic mutation burden. Thus, if m is the mutation rate per year, and L is the species' lifespan, the expected relationship is of the form

$$m\,L \approx k,$$

where k is a constant representing the typical end-of-lifespan mutation burden across species. According to this relationship, the mutation rate per year is linearly related to the inverse of lifespan,

$$m \approx k\,(1/L).$$

Therefore, the cross-species average end-of-lifespan burden (k), can be estimated as the slope parameter of a zero-intercept linear regression model with the mutation rate per year (m) as the dependent variable, and the inverse of lifespan ($1/L$) as the explanatory variable. To this purpose, the LME model described in the previous section was altered by removing the fixed-effect intercept parameter, thus considering only fixed- and random-effect slope parameters for 1/Lifespan.

The zero-intercept LME model estimated a value of k = 3210.52 (95% confidence interval 2686.89–3734.15). The fraction of inter-species variance explained by the zero-intercept model (FVE) was 0.82, while the LME model described in the previous section (which estimated k = 2869.98, and an intercept of 14.76) achieved FVE = 0.84 (see 'Association of mutation rate with life-history traits'). To test whether the increase in model fit justifies the inclusion of an intercept, both models were compared using a likelihood ratio test (as implemented by the 'anova' function in the nlme R package [v3.1-137]). This yielded P=0.23, indicating that the free-intercept model does not achieve a significantly better fit than the zero-intercept model. Similarly, the zero-intercept model yielded lower values for both the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). Notably, equivalent analyses using somatic mutation rates per megabase and per protein-coding exome (instead of per whole genome) yielded comparable results (**Extended Data Figure 12**).

To investigate the possibility of phylogenetic relationships between the species sampled confounding the analysis, a phylogenetic generalised linear model was used to regress the mean mutation rate of each species on the inverse of its lifespan ($1/L$), while accounting for the phylogenetic relationships among species. A phylogenetic tree of the 15 species examined was obtained from the TimeTree resource[80], and the phylogenetic linear model was fitted using the 'pgls' function in the caper R package (v1.0.1; cran.r-project.org/package=caper). The estimates produced by zero-intercept regression of mean mutation rates per species on 1/Lifespan were compared between this phylogenetic generalised linear model and a simple linear model ('lm' function in R). The use of this simple model, as well as the use of mean mutation rates per species (rather than mutation rates per sample), was necessary due to the impossibility of

replicating the heteroscedastic mixed-effects structure of the LME model employed for the main association analyses (see 'Association of mutation rate with life-history traits') within the phylogenetic linear model. Both the phylogenetic linear model and the simple linear model produced similar estimates (**Extended Data Figure 14**), suggesting that phylogenetic non-independence of the samples does not have a substantial effect on the association analyses.

18. Cell division analysis

To investigate the extent to which differences in cell division rate could explain differences in mutation rate and burden across species, we obtained estimates of intestinal crypt cell division rates from mouse[81], rat[82] and human[83,84] (**Extended Data Table 7**). Using these cell division rates, our lifespan estimates and the observed SBS mutation rates, we calculated the number of cell divisions at the end of lifespan and the corresponding number of mutations per cell division expected under a simple model assuming that all mutations occur during cell division (**Extended Data Table 7**).

We investigated whether differences in the number of cell divisions among species could explain the observed differences in mutation burden. Although colorectal cell division rate estimates are lacking for most species, existing estimates from mouse, rat and human indicate that the total number of stem cell divisions per crypt in a lifetime varies greatly across species, for example with ~6–31 fold more divisions per intestinal stem cell in a human than a rat over their respective lifetimes, depending on the cell division rate estimates used (**Extended Data Table 7**). Mouse intestinal stem cells are estimated to divide once every 24 hours[81], while estimates of the human intestinal stem cell division rate vary from once every 48 hours[83] to once every 264 hours[84]. Thus mouse intestinal stem cells divide 2–11 times faster than human intestinal stem cells. By the end of lifespan, an intestinal stem cell is predicted to have divided ~1351 times in a mouse, ~486 times in a rat and 2774–15,257 times in a human (depending on the cell division rate estimate used). Applying our somatic mutation burden and lifespan data, this implies that the somatic mutation rate per cell division in a mouse is ~1.5–8.4 fold higher than in a human. However, the observed fold difference in somatic mutation rate between these two species is 16.9 (**Table 1**). Therefore, differences in cell division rate appear unable to fully account for the observed

differences in mutation rate across species. Nevertheless, we note that accurate cell division rate estimates for basal intestinal stem cells are lacking for the majority of species.

19. Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomised and the investigators were not blinded to allocation during experiments and outcome assessment.

20. Ethics statement

All animal samples were obtained with the approval of the local ethical review committee (AWERB) at the Wellcome Sanger Institute and those at the holding institutions.

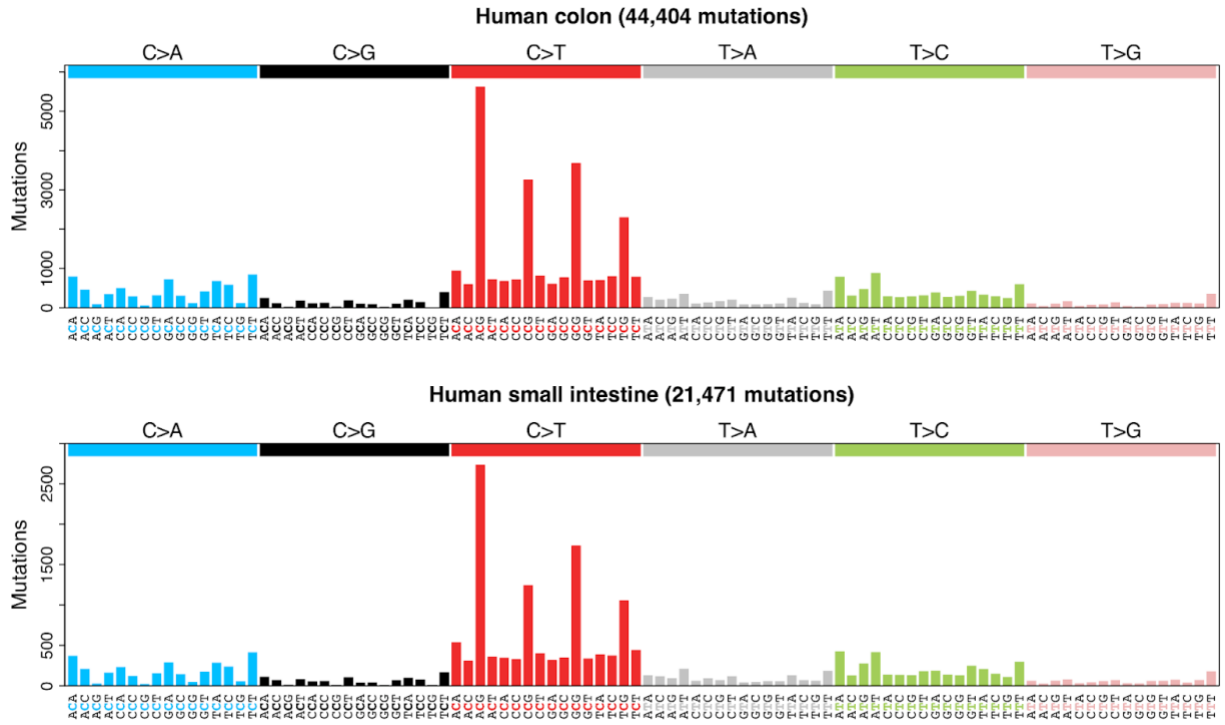
Methods references

70. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
71. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
72. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
73. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
74. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
76. Tacutu, R. *et al.* Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.* **46**, D1083–D1090 (2018).
77. Conde, D. A. *et al.* Data gaps and opportunities for comparative and conservation biology. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9658–9664 (2019).
78. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
79. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).

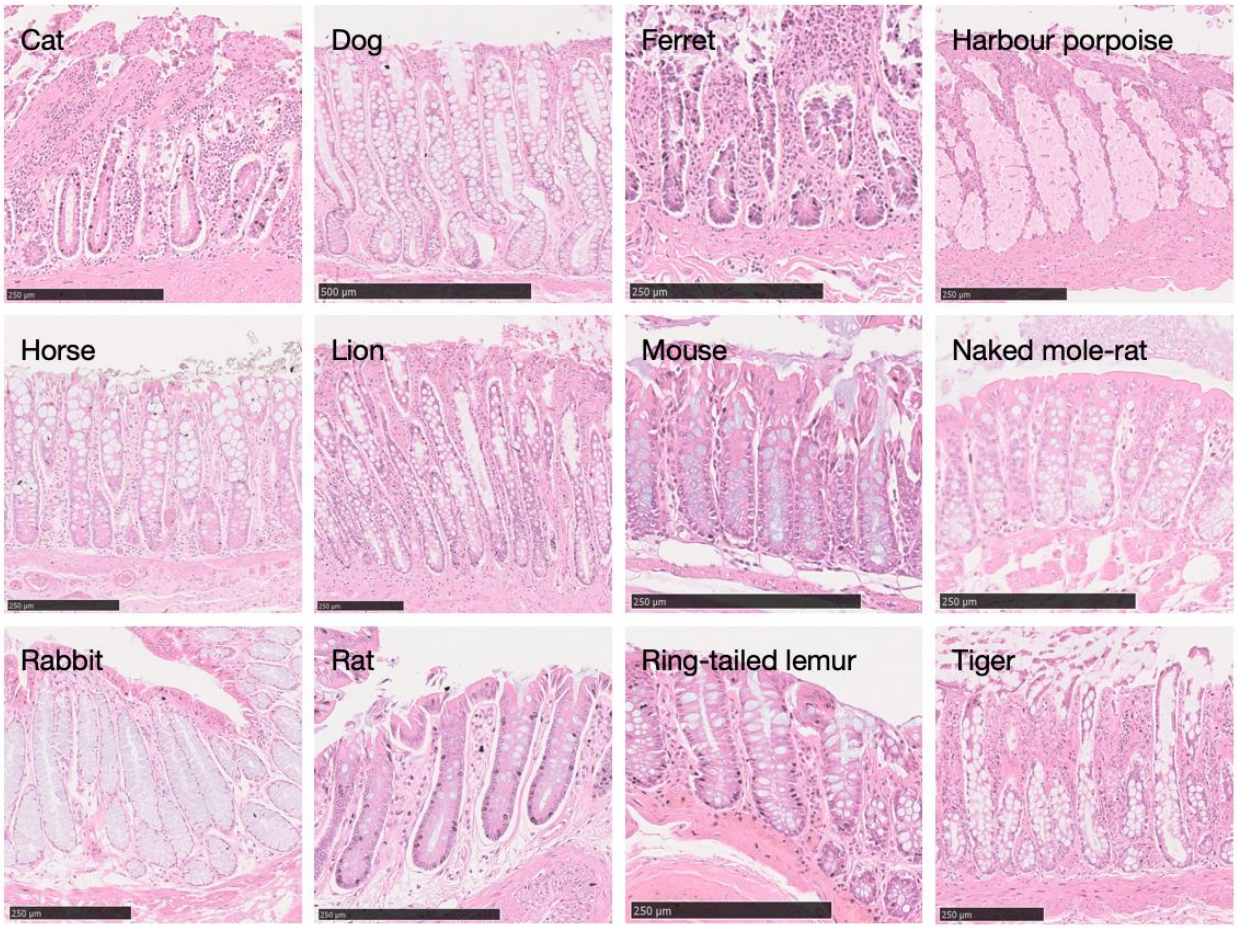
80. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
81. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
82. Rijke, R. P., Plaisier, H. M. & Langendoen, N. J. Epithelial cell kinetics in the descending colon of the rat. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.* **30**, 85–94 (1979).
83. Potten, C. S., Kellett, M., Rew, D. A. & Roberts, S. A. Proliferation in human gastrointestinal epithelium using bromodeoxyuridine in vivo: data for different sites, proximity to a tumour, and polyposis coli. *Gut* **33**, 524–529 (1992).
84. Bach, S. P., Renehan, A. G. & Potten, C. S. Stem cells: the intestinal stem cell as a paradigm. *Carcinogenesis* **21**, 469–476 (2000).

Extended Data Figures

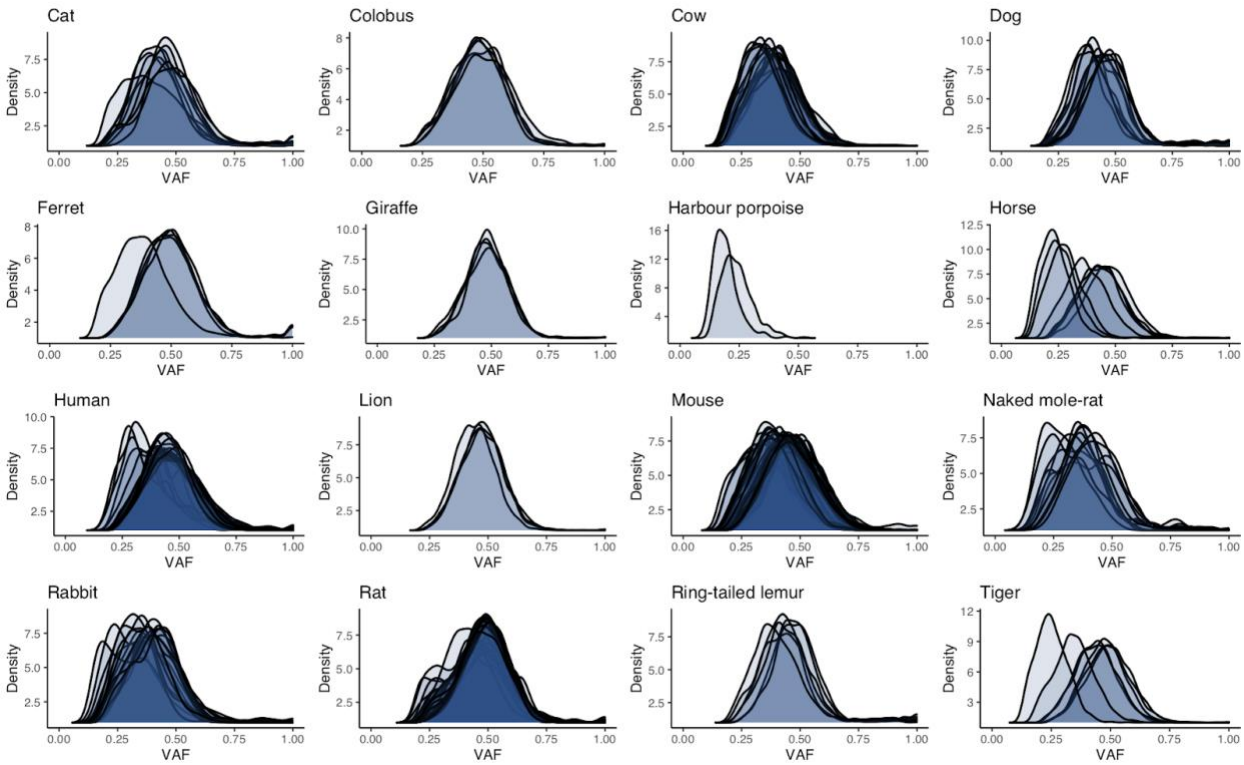
Extended Data Figure 1. Somatic mutational spectra of the human colon and small intestine. Trinucleotide-context mutational spectra of somatic single-base substitutions from human adult stem cells in colon (top) and small intestine, using mutation calls obtained from Blokzijl *et al.*[7].



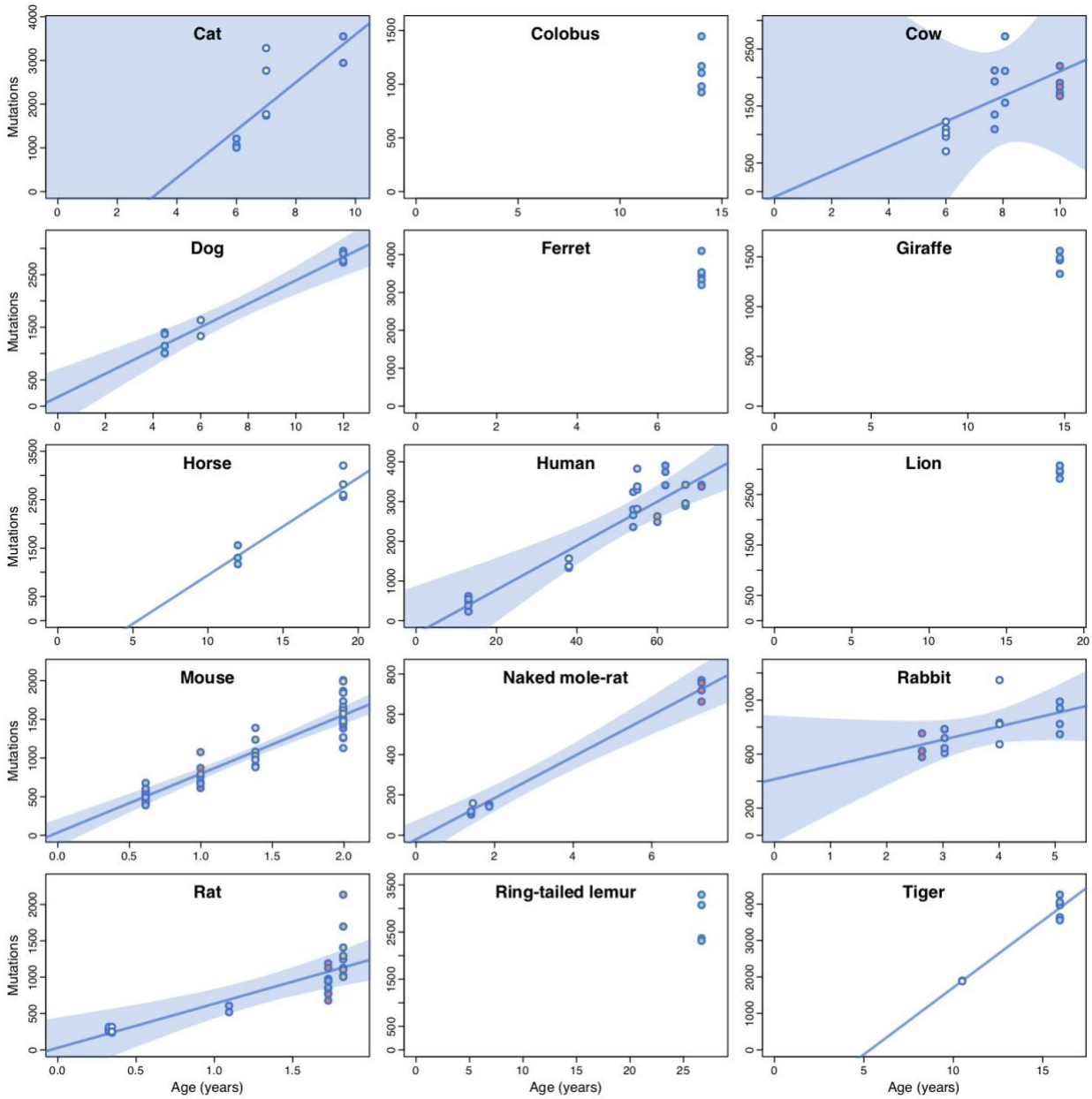
Extended Data Figure 2. Histology images of intestinal crypts across species. Histological images of the colorectal or intestinal (ferret) epithelium for each species. Scale bars are provided at the bottom of each image.



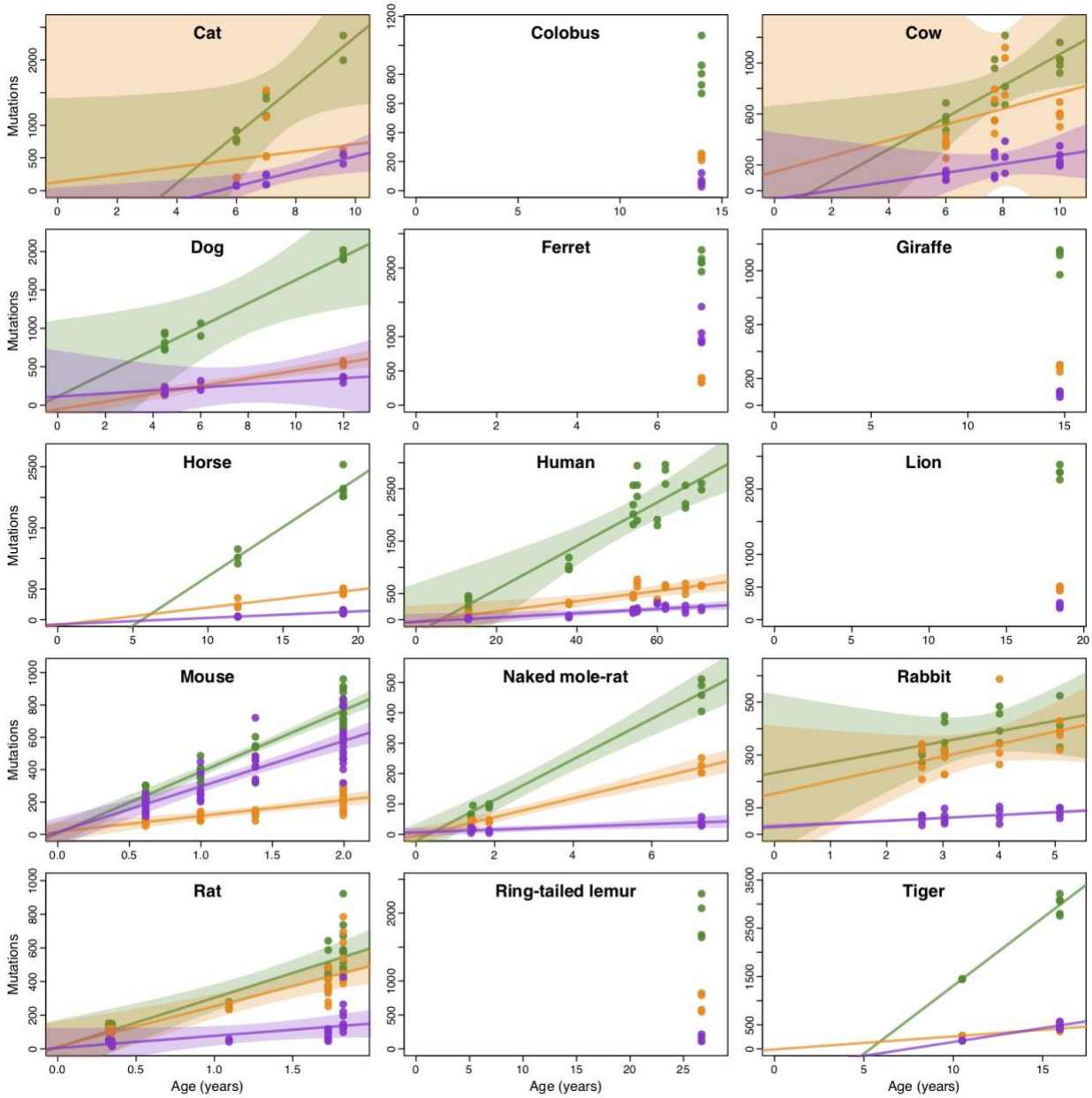
Extended Data Figure 3. Somatic VAF distributions per species. Distributions of variant allele fraction (VAF) for somatic substitutions in each crypt for each species. Each distribution refers to the variants in a single sequenced crypt.



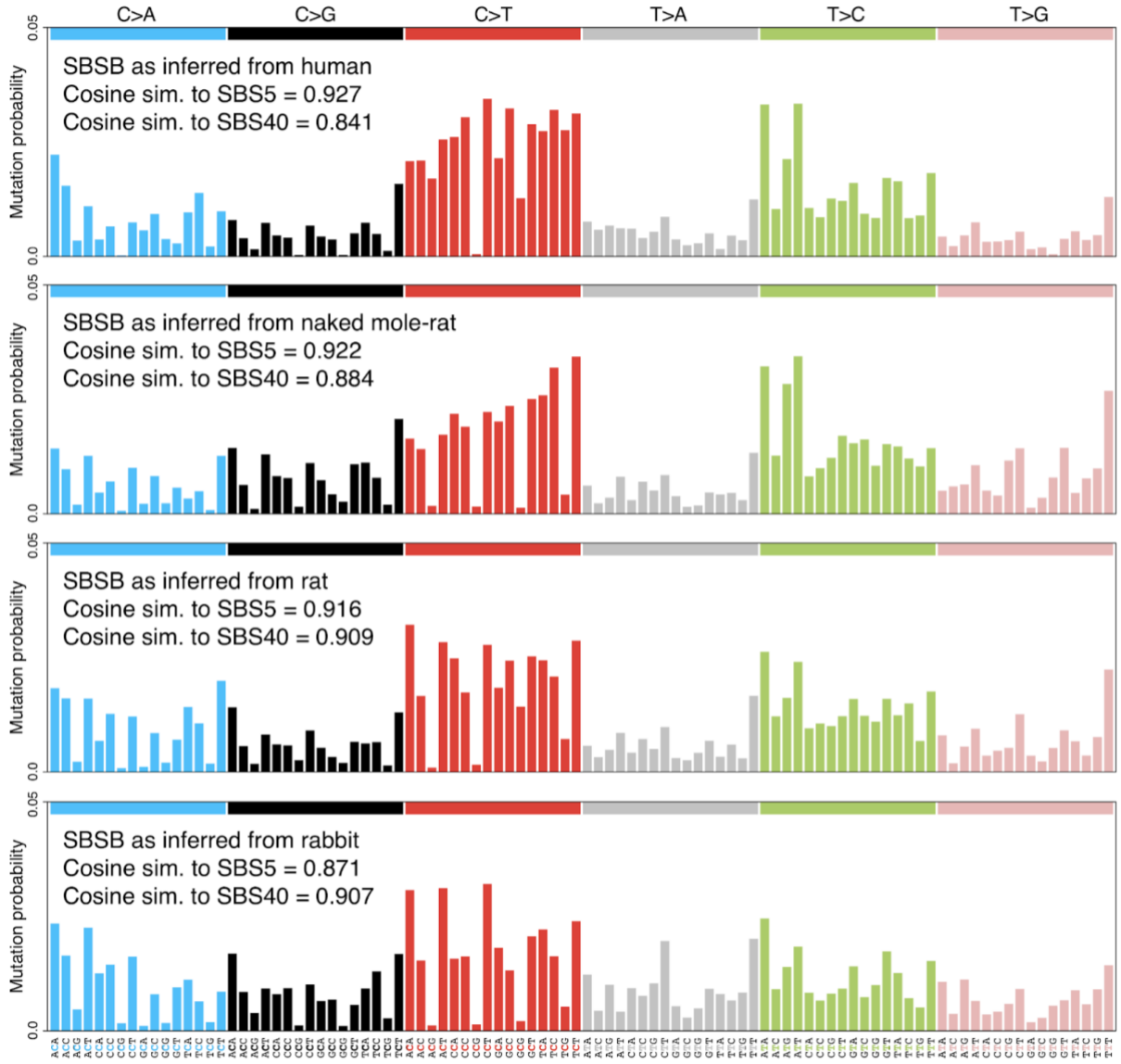
Extended Data Figure 4. Somatic mutation accumulation across species. Each panel presents somatic substitution burdens per genome for a given species. Each dot represents a crypt sample, with samples from the same individual sharing the same colour. For species with two or more individuals, the estimated regression line from a simple linear regression model on individual mean burdens is shown. For species with three or more individuals, blue shaded regions indicate the 95% confidence intervals of the regression line. Harbour porpoise samples were excluded because the individual was of unknown age.



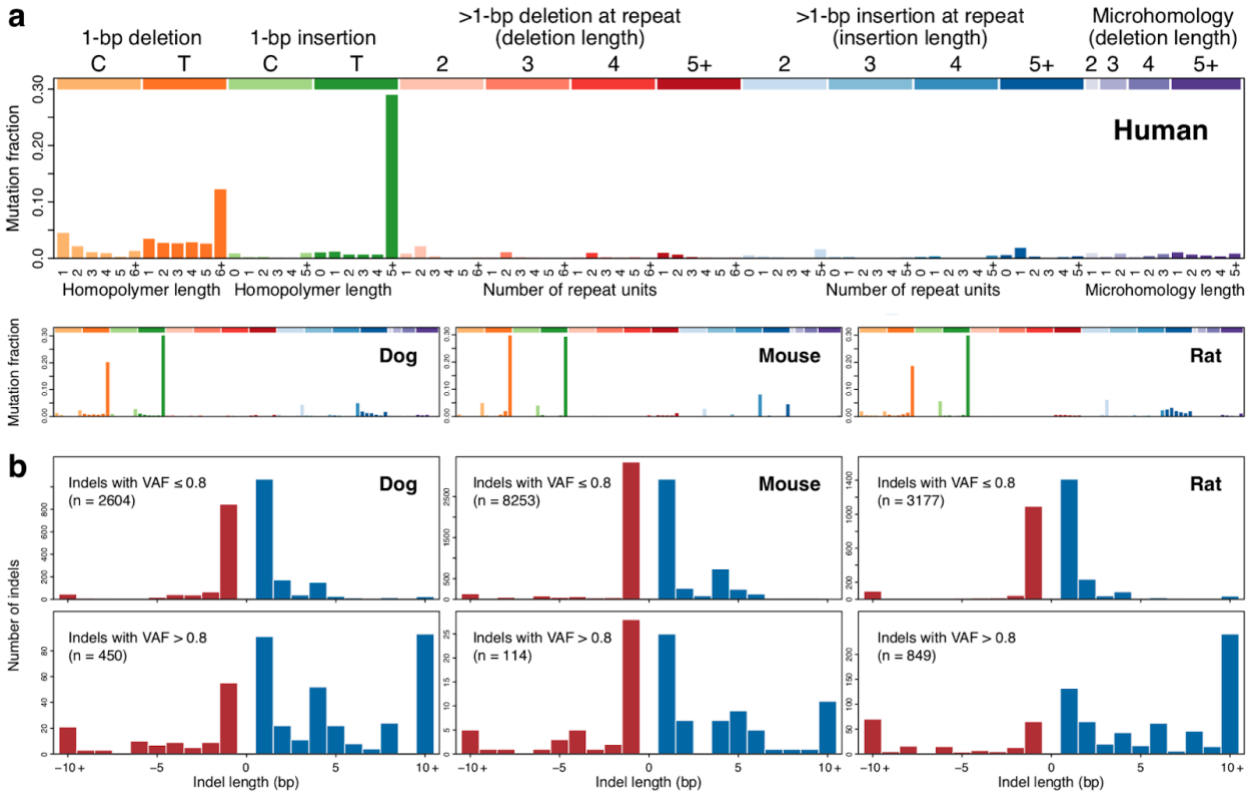
Extended Data Figure 5. Signature-specific mutation accumulation across species. Each panel presents somatic substitution burdens per genome for mutational signatures SBS1 (green), SBSB (yellow) and SBSC (purple) in a given species. For species with two or more individuals, the estimated regression lines from a simple linear regression model on individual mean burdens per signature are shown. For species with three or more individuals, shaded regions indicate the 95% confidence intervals of the regression lines. Harbour porpoise samples were excluded because the individual was of unknown age.



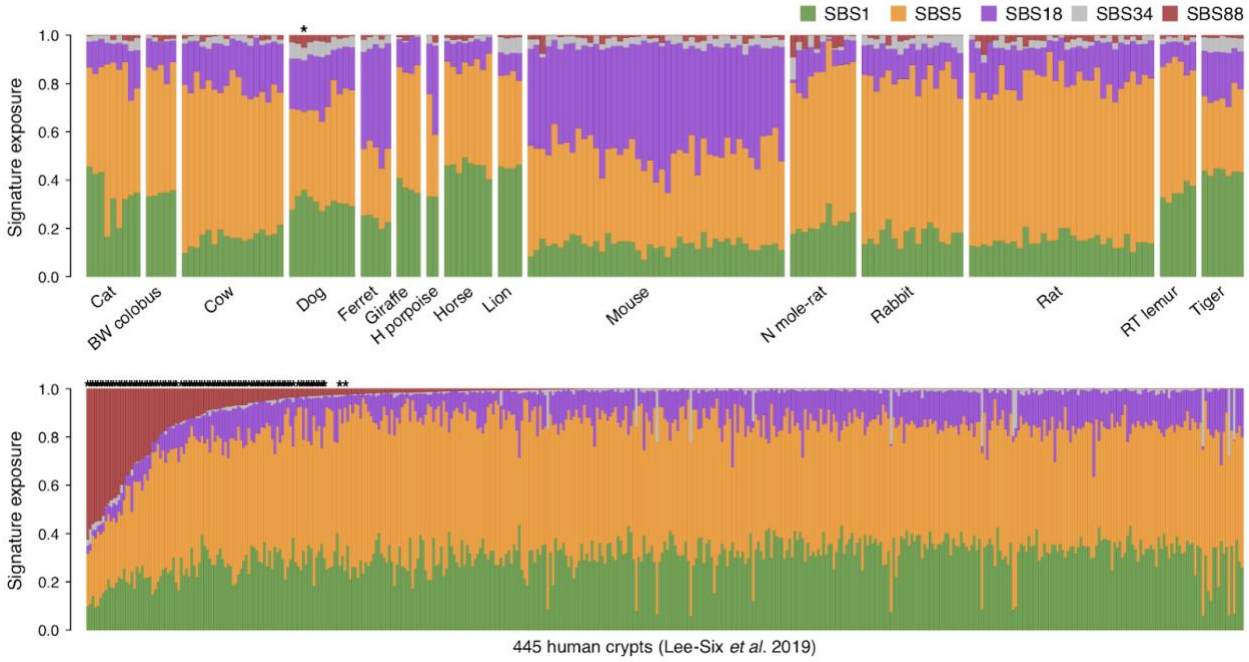
Extended Data Figure 6. Profiles of signature SBSB as inferred from different species. Trinucleotide-context mutational spectra of versions of signature SBSB extracted independently from somatic mutations in (top to bottom) human, naked mole-rat, rat and rabbit colorectal crypts (Methods). Signatures are shown in a human-genome-relative representation. Cosine similarities between each signature and the human COSMIC SBS5 and SBS40 are provided.



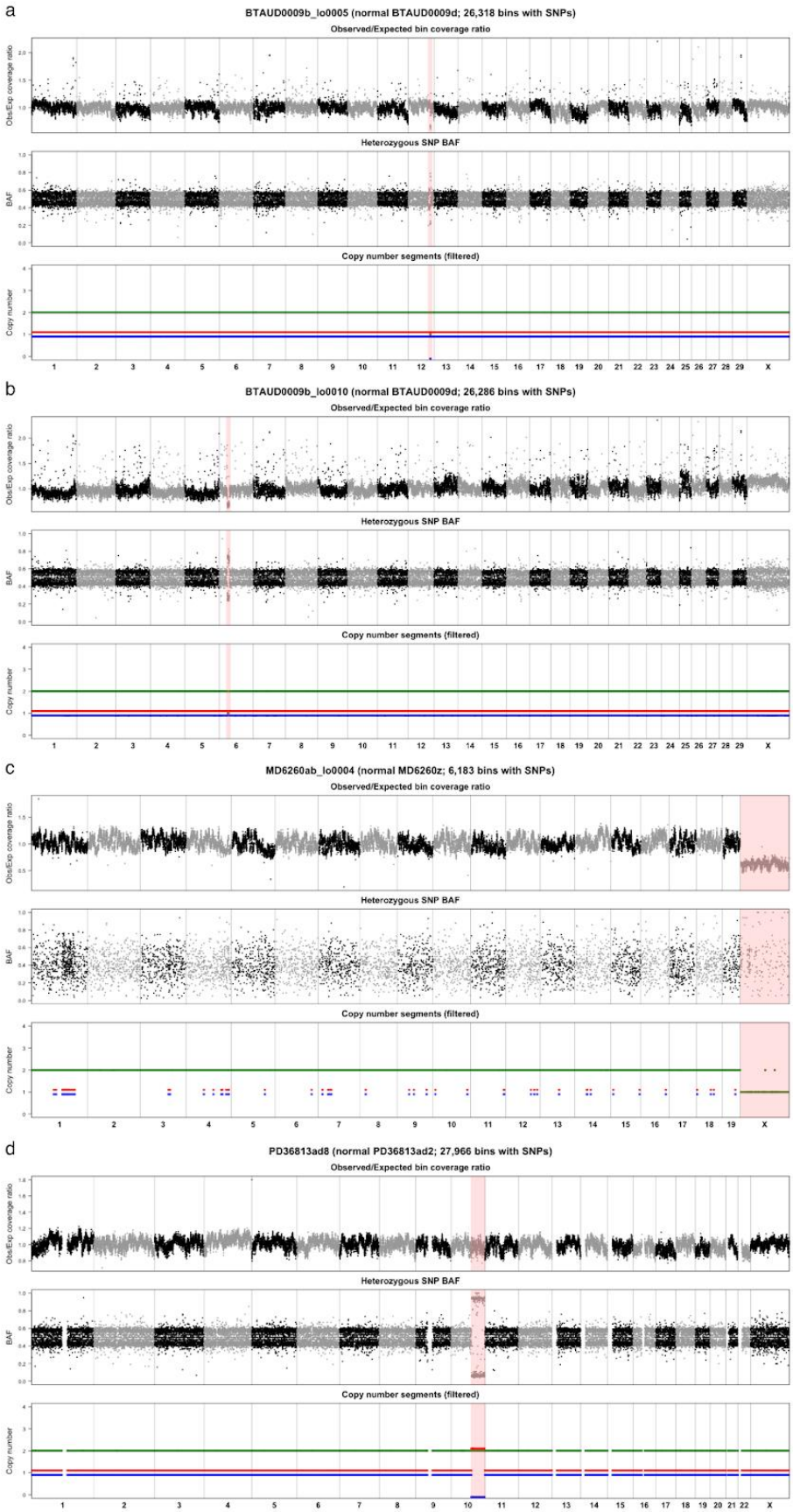
Extended Data Figure 7. Mutational spectra of somatic indels. a, Spectra of somatic indels in human, dog, mouse and rat crypts. Horizontal axis presents 83 insertion/deletion types, coloured by type and length[36]. **b**, Histograms of indel length, showing the number of insertions (blue) and deletions (red) with length ranging from 1 to 10+ bp in dog, mouse and rat. Top and bottom panels in each column present indels with VAF ≤ 0.8 and VAF > 0.8, respectively; the latter are more likely to be artefacts.



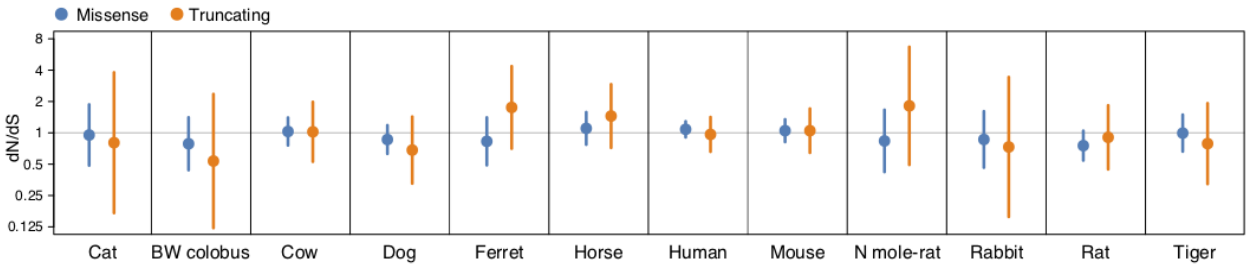
Extended Data Figure 8. Colibactin exposure in non-human and human colorectal crypts. Exposures to mutational signatures SBS1, SBS5, SBS18, SBS34 and SBS88, as inferred via expectation–maximisation, for 180 non-human crypts (top) and 445 human crypts sequenced in a previous study[8]. Asterisks indicate samples with statistically significant colibactin (SBS88) exposure, based on a likelihood ratio test (Methods). BW, black-and-white; H, harbour; N, naked; RT, ring-tailed.



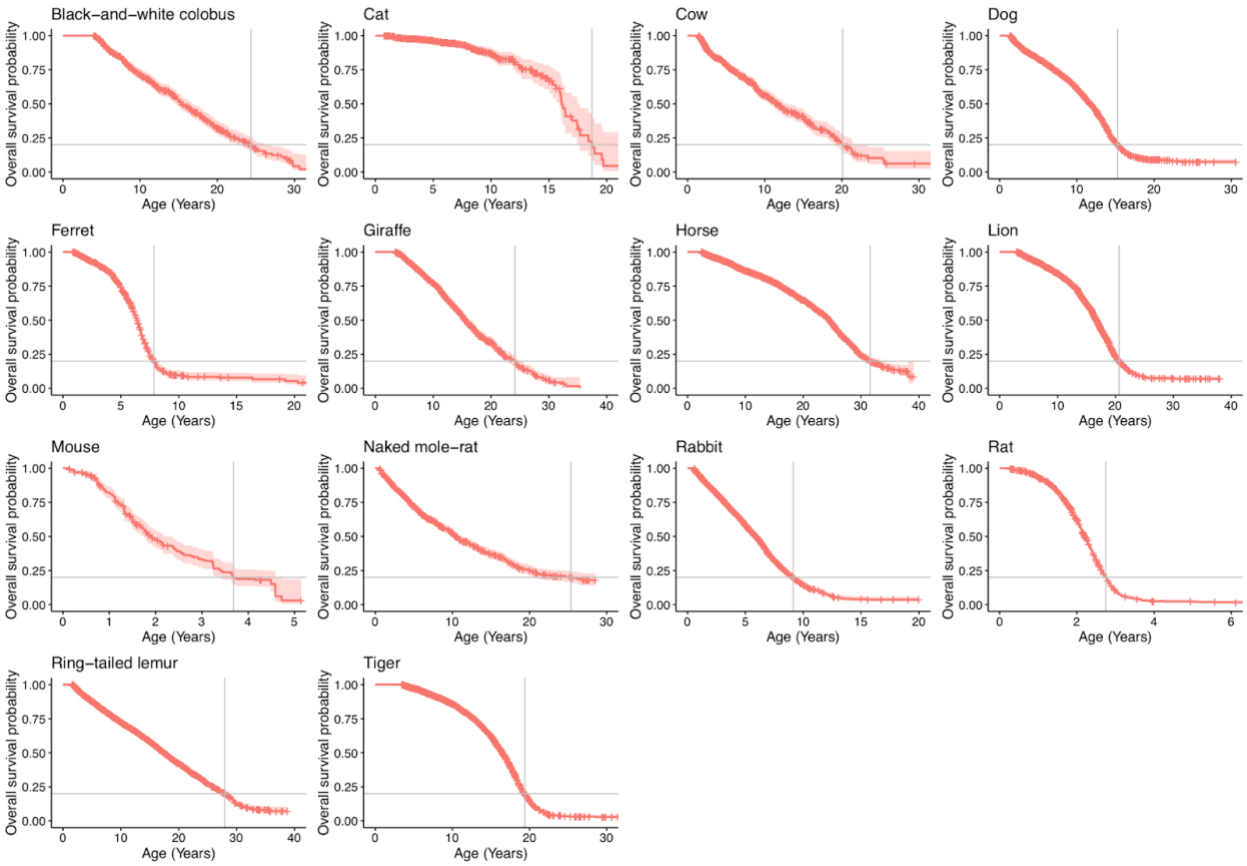
Extended Data Figure 9. Identified copy number changes. Somatic copy number changes in cow (**a**, **b**), mouse (**c**) and human (**d**) colorectal crypts. For each case, chromosomes are presented along the horizontal axis, and each point represents a 100-kb genomic bin. The top panel presents the ratio between observed and expected sequencing coverage per bin; the middle panel shows the median B-allele fraction (BAF) of heterozygous germline SNPs per bin; and the bottom panel presents the inferred segments of total copy number (green) and allele-specific copy number (red/blue). Regions of copy number change are highlighted in pink. The sparsity of BAF and allele-specific copy number values in the mouse crypt (**c**) are due to the fact that mouse samples generally had very low numbers of germline SNPs.



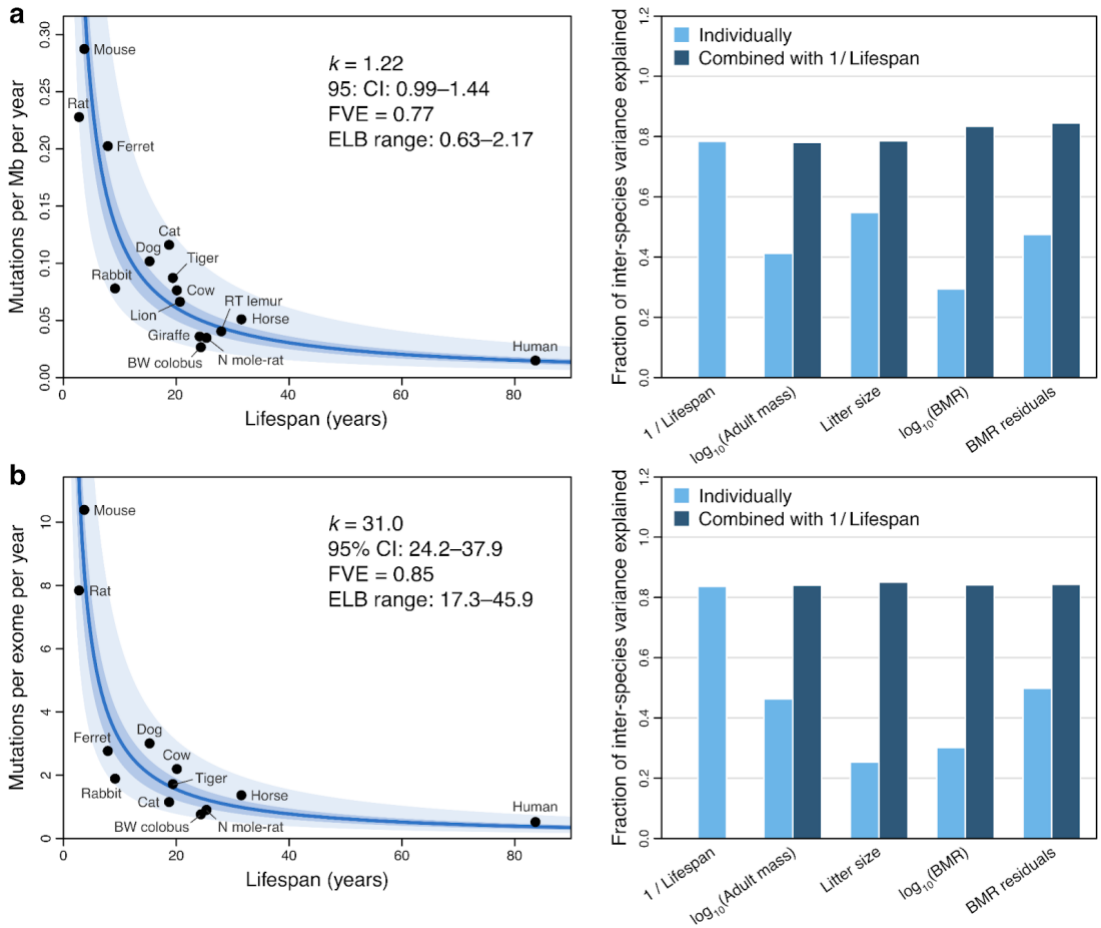
Extended Data Figure 10. Somatic dN/dS. Estimates of dN/dS for somatic missense and truncating mutation in each of the species with available genome annotation. Dots and error bars represent maximum likelihood estimates and 95% confidence intervals, respectively. Note the logarithmic scale of the vertical axis.



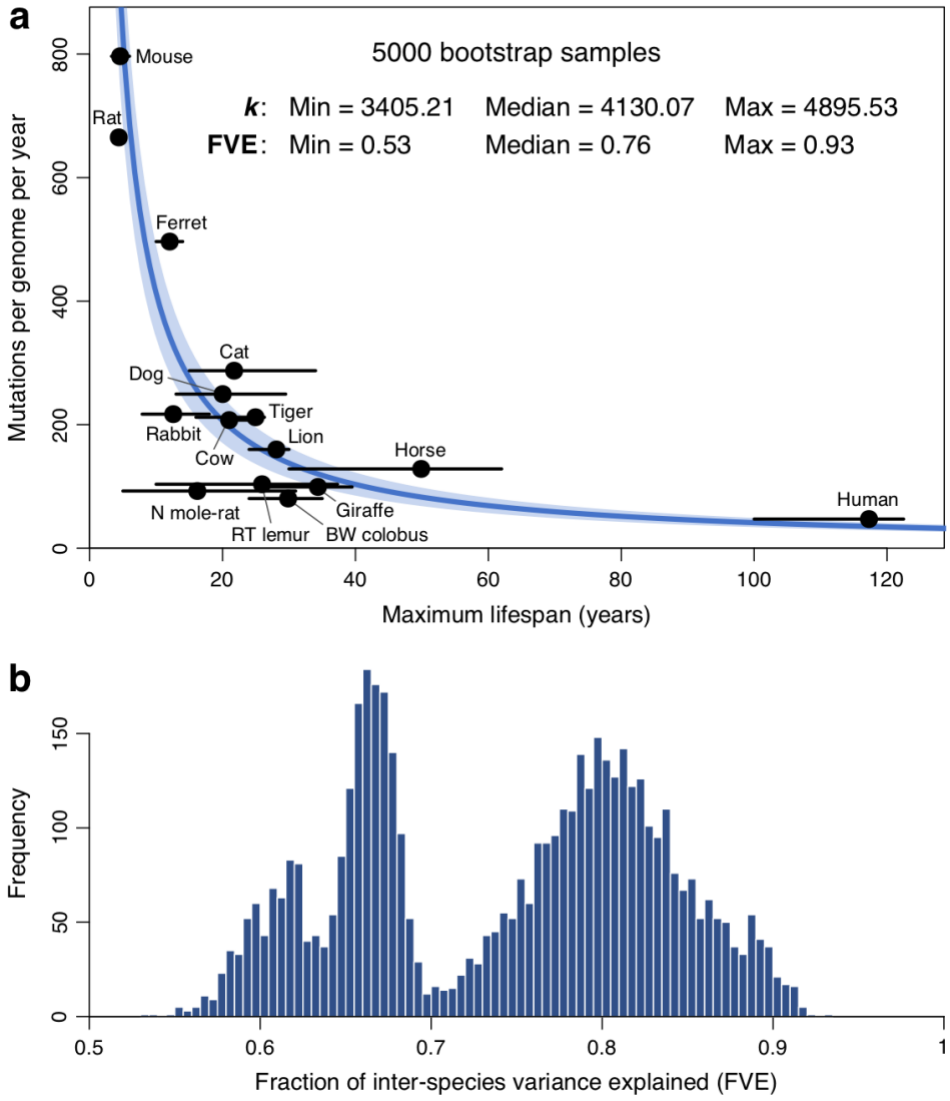
Extended Data Figure 11. Kaplan-Meier curves of longevity in captivity. Kaplan Meier survival curves for each species using captive lifespan data from Species360 (**Methods**). The red shaded areas represent 95% confidence intervals of the curve. A vertical grey bar indicates the 80th percentile, which was employed as a robust estimate of species lifespan.



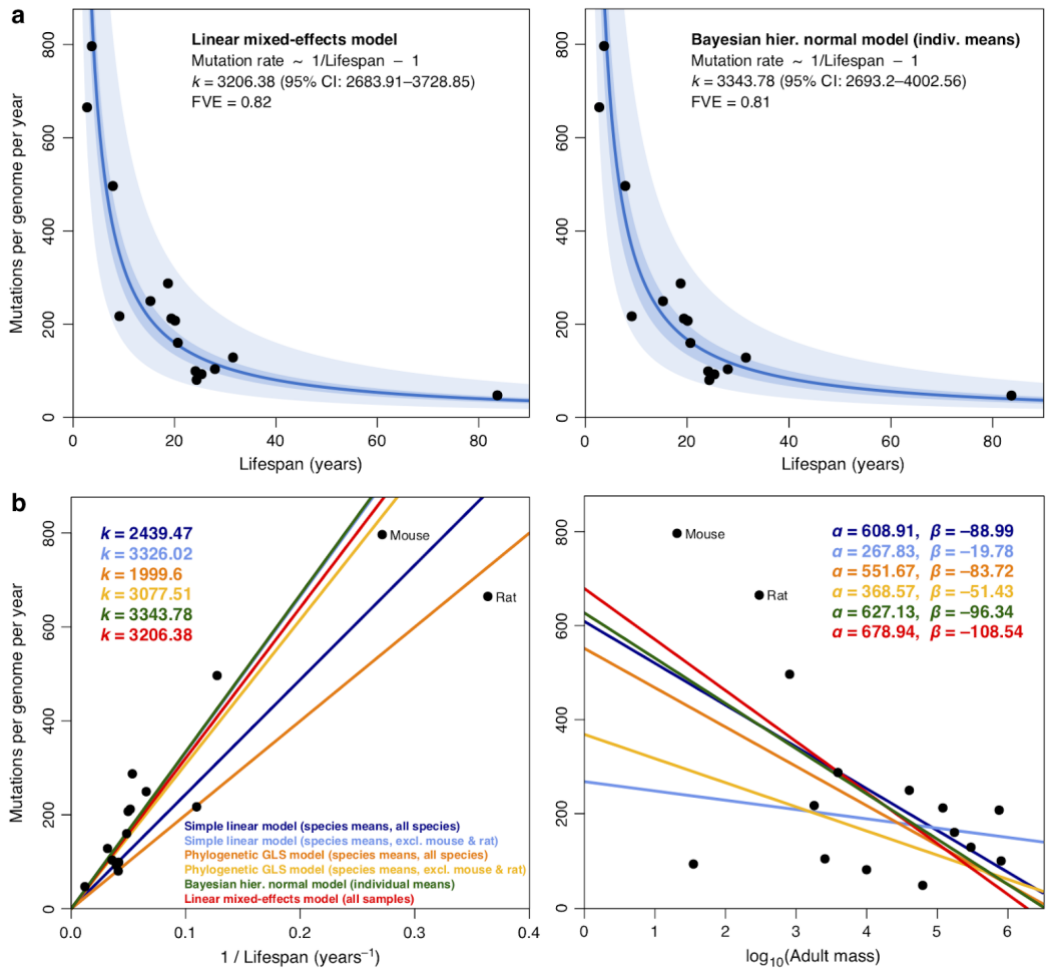
Extended Data Figure 12. Associations between life-history variables and alternative measures of somatic mutation rate. The figure presents the same analyses as **Fig. 3b,d**, but employing somatic mutation rates per megabase (**a**), or per protein-coding exome (**b**), rather than per genome (**Methods**). Leftmost panels show zero-intercept linear mixed-effects (LME) regressions of somatic mutation rates on inverse lifespan (1/Lifespan), presented on the scale of untransformed lifespan (horizontal axes). Vertical axes present mean mutation rate per species, although mutation rates per crypt were used in the regressions. Darker shaded areas indicate 95% confidence interval (CI) of the regression lines; lighter shaded areas mark a two-fold deviation from the regression line. Point estimate and 95% CI of the regression slope coefficient (*k*), fraction of inter-species variance explained by the model (FVE), and range of end-of-lifespan burden (ELB) are provided. Rightmost panels show comparisons of FVE values achieved by free-intercept LME models using inverse lifespan and other life-history variables (alone or in combination with inverse lifespan) as explanatory variables. BMR, basal metabolic rate; BW, black-and-white; Mb, megabase; N, naked; RT, ring-tailed.



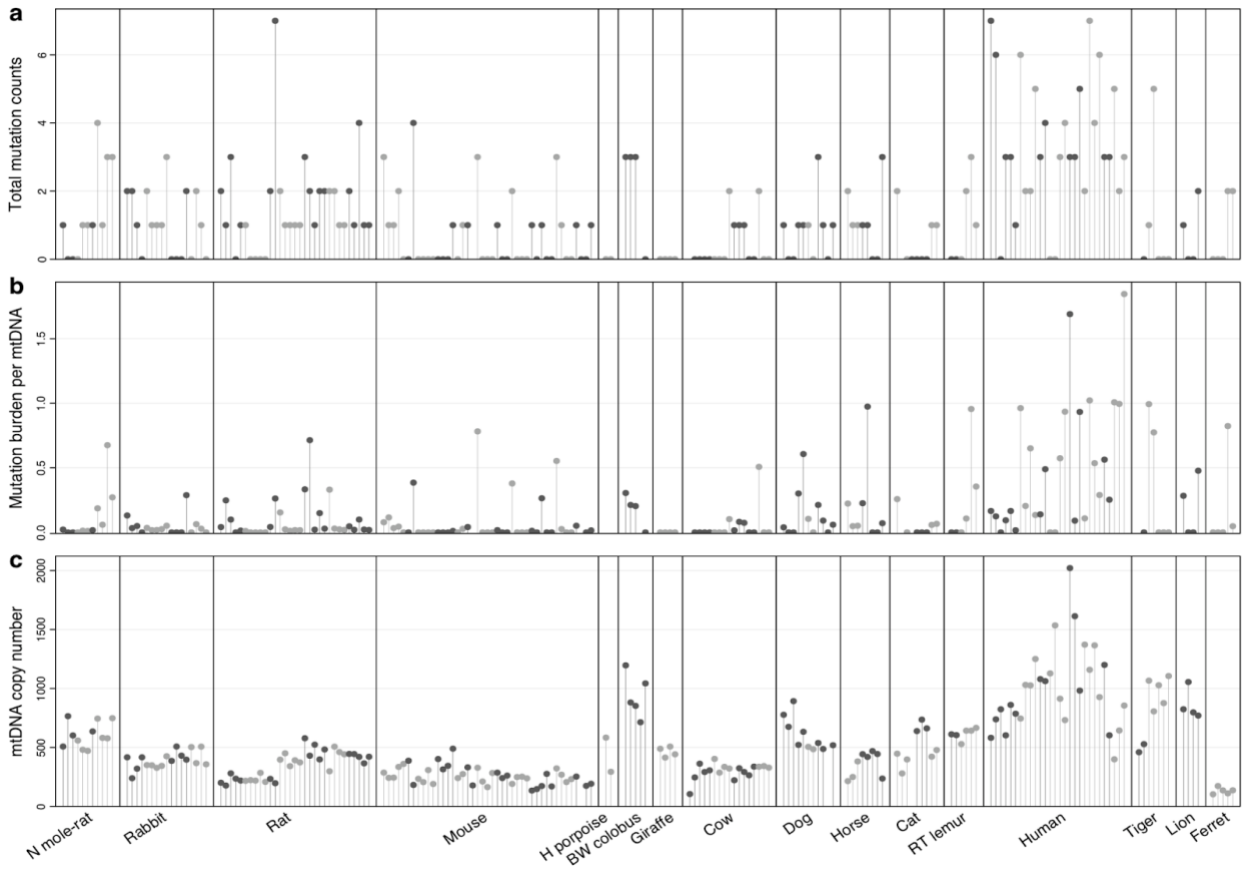
Extended Data Figure 13. Bootstrapped regression of somatic mutation rates on published lifespan estimates. a, Bootstrapped regression of somatic substitution rates on the inverse of lifespan (1/Lifespan), using a zero-intercept linear mixed-effects model (**Methods**). For each of 5000 bootstrap samples, lifespan values per species were randomly selected from a set of published maximum longevity estimates (**Extended Data Table 6**). The blue line indicates the median regression slope (*k*) across bootstrap samples, and the blue shaded area depicts the range of estimates of *k* across bootstrap samples. Black dots and error bars indicate the mean and range, respectively, of published longevity estimates for each species. The median and range of both *k* and the fraction of inter-species variance explained (FVE) are provided. **b**, Histogram of FVE values across the 5000 bootstrap samples.



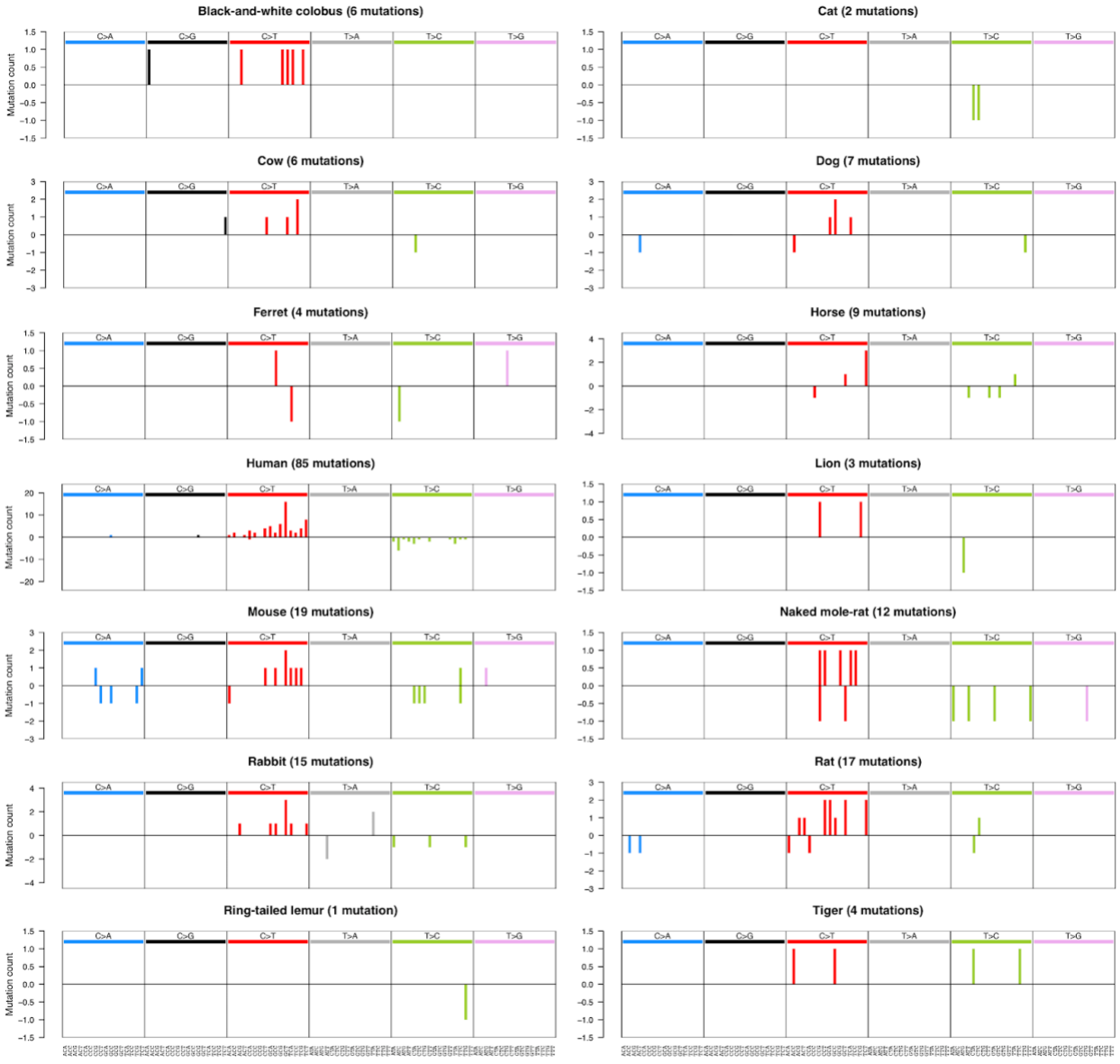
Extended Data Figure 14. Comparison of regression models for somatic mutation rates. a, Constrained-intercept regression of somatic substitution rates on the inverse of lifespan (1/Lifespan), using a linear mixed-effects model applied to the rates per crypt (left) and a Bayesian hierarchical normal regression model applied to the mean rates per individual. For simplicity, black dots present mean mutation rates per species. Darker shaded area indicates 95% confidence/credible interval (CI) of the regression line; lighter shaded area marks a two-fold deviation from the regression line. Point estimates and 95% CI of the regression slopes (*k*) and fraction of inter-species variance explained (FVE) are provided. **b**, Comparison of regression lines for the of somatic substitution rates on 1/Lifespan (left; zero intercept) and log-transformed adult mass (right; free intercept), using simple linear models (dark and light blue), phylogenetic generalised least-squares models (orange and yellow), Bayesian hierarchical normal models (green) and linear mixed-effects models (red). Point estimates of the regression coefficients for each model are provided.



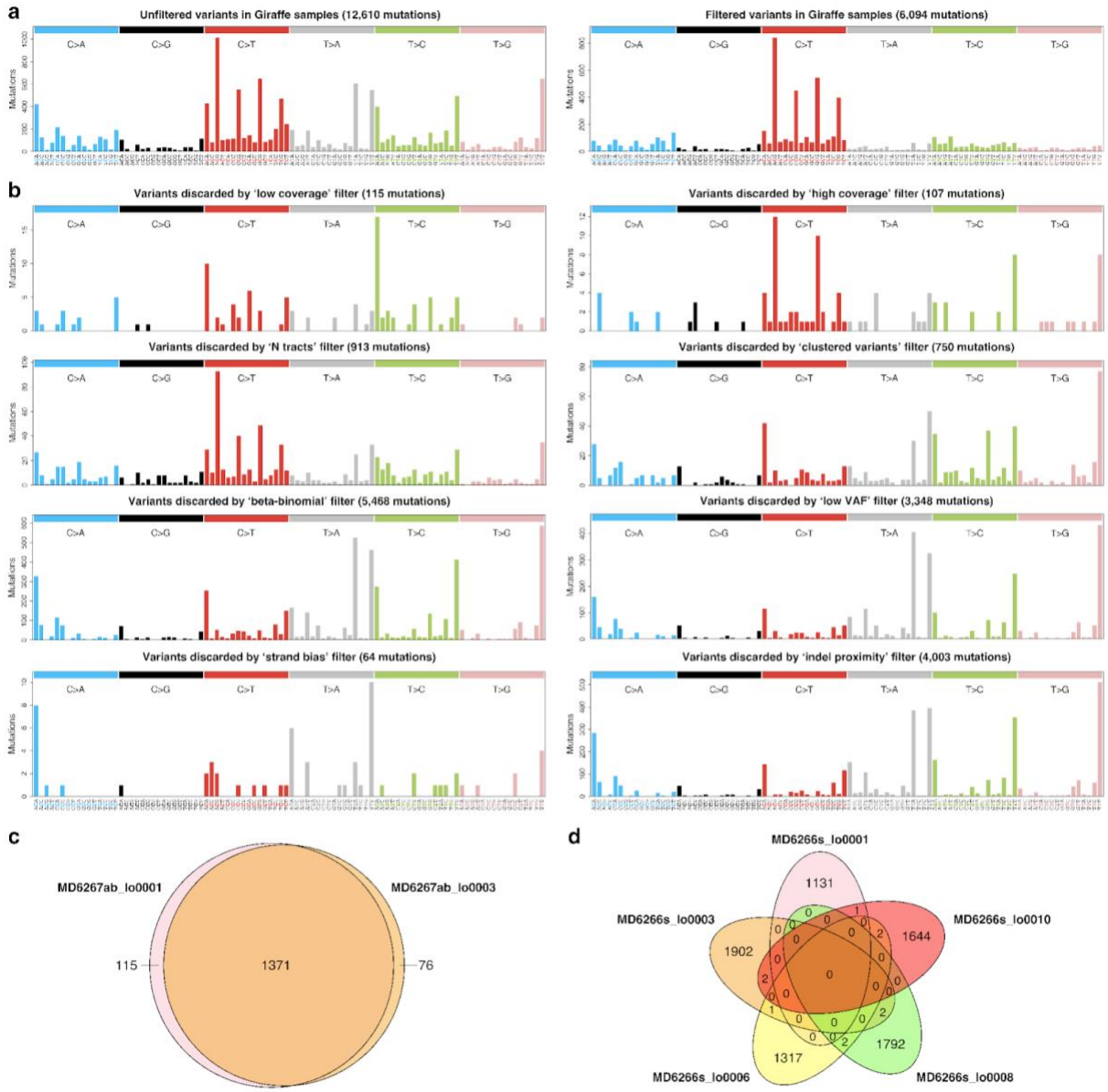
Extended Data Figure 15. mtDNA mutation burden and copy number. a, Total somatic mtDNA mutations (substitutions and indels) called in each sample. **b**, Somatic mutation burden per mitochondrial genome copy per sample. **c**, Estimated mtDNA copy number per sample. Samples are arranged as in **Fig. 1b**, with samples from the same individual coloured in the same shade of grey.



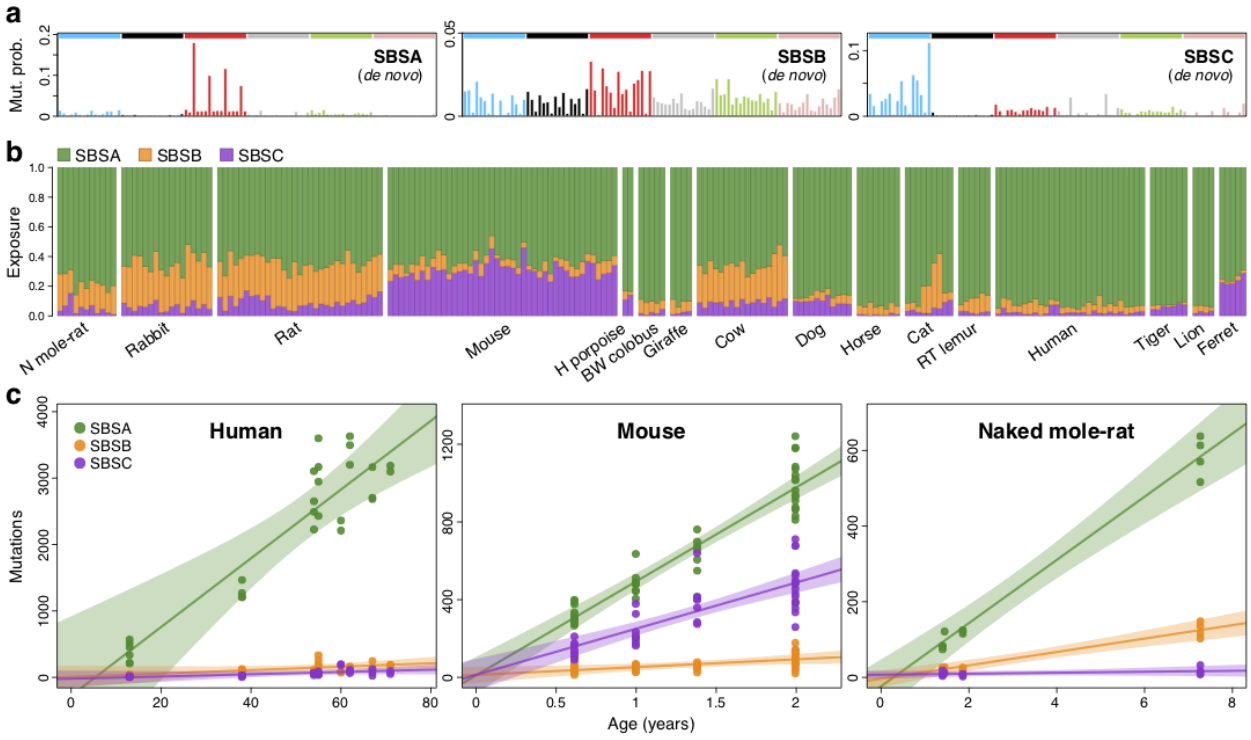
Extended Data Figure 16. Mutational spectra of mtDNA substitutions in each species. Horizontal axis presents 96 mutation types on a trinucleotide context, coloured by base substitution type. Mutations on the upper and lower halves of the spectrum represent substitutions in which the pyrimidine base is located in the heavy and light strands of mtDNA, respectively.



Extended Data Figure 17. Assessment of variant calling and filtering. a, Spectra of SBS calls before (left) and after application of the final eight variant filters, across all giraffe samples. Note that the set of 'unfiltered' variants (left) has gone through the three early filters named 'quality flag filter', 'alignment quality filter' and 'hairpin filter' (Methods). **b**, Spectra of calls flagged as artefactual by each of the final eight variant filters, across all giraffe samples. Sets of variants flagged by different filters are not mutually exclusive. **c**, Venn diagram showing the number of variant calls shared between two LCM sections from the same mouse colorectal crypt. **d**, Venn diagram showing the numbers of variant calls shared between five different colorectal crypts from the same mouse.



Extended Data Figure 18. Mutational signatures and exposures inferred *de novo*. a, Mutational signatures inferred *de novo* from the species mutational spectra shown in **Fig. 2a**. Signatures are shown in a human-genome-relative representation. SBSA is the *de novo* equivalent of COSMIC signature SBS1 (**Fig. 2b**). **b**, Exposure of each sample to each of the mutational signatures shown in **a**. Samples are arranged horizontally as in **Fig. 1b**. **c**, Regression of signature-specific mutation burdens on individual age for human, mouse and naked mole-rat samples. Regression was performed using mean mutation burden per individual. Shaded areas indicate 95% confidence intervals of the regression line. BW, black-and-white; H, harbour; N, naked; RT, ring-tailed.



Extended Data Figure 19. Allometric regressions of somatic mutation rate, body mass and lifespan. a, Allometric (log-log) regressions of somatic mutation rate on body mass (left) and lifespan. Regressions were performed using a simple linear model on the mean mutation rate per species. **b**, Simple linear regressions of the allometric-regression residuals for somatic mutation rate and body mass (from allometric regressions of each variable on lifespan; left), and for somatic mutation rate and lifespan (from allometric regressions of each variable on body mass). Shaded areas represent 95% confidence intervals of the regression lines. Fraction of inter-species variance explained (FVE) and *p*-value (*P*) are provided on the top-right corner for each model.