# 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 Bhosle[1], David Jones[1], Raul E. Alcantara[1], Timothy M. Butler[1], Yvette Hooks[1], Kirsty Roberts[1], Elizabeth Anderson[1], Sharna Lunn[1], Edmund Flach[3], Simon Spiro[3], Inez Januszczak[3,4], Ethan Wrigglesworth[3], Hannah Jenkins[3], Tilly Dallas[3], Nic Masters[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) (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 contributing factor in ageing.

## 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] (**Supplementary Note**

96 **1**). If somatic mutations contribute to ageing, theory predicts that somatic mutation rates may
97 inversely correlate with lifespan across species[30,33]. This long-standing prediction has remained
98 largely untested due to the difficulty of measuring somatic mutation rates across species.
99
100 **Detection of somatic mutations across species**
101
102 The study of somatic mutations with standard whole-genome sequencing requires isolating clonal
103 groups of cells recently derived from a single cell[6–8]. To study somatic mutations across a diverse
104 set of mammals, we isolated 208 individual intestinal crypts from 56 individuals across 16 species
105 with a wide range of lifespans and body sizes: black-and-white colobus monkey, cat, cow, dog,
106 ferret, giraffe, harbour porpoise, horse, human, lion, mouse, naked mole-rat, rabbit, rat, ring-tailed
107 lemur, and tiger (**Supplementary Table 1**). We chose intestinal crypts for several reasons. First,
108 they are histologically identifiable units that line the epithelium of the colon and small intestine
109 and are amenable to laser microdissection. Second, human studies have confirmed that individual
110 crypts become clonally derived from a single stem cell and show a linear accumulation of mutations
111 with age, enabling the estimation of somatic mutation rates through genome sequencing of single
112 crypts[8]. Third, in most human crypts, the majority of somatic mutations are caused by endogenous
113 mutational processes common to other tissues, rather than by environmental mutagens[8,12].
114
115 A colon sample was collected from each individual, with the exception of a ferret from which only
116 a small intestine sample was available. This sample was included because results in humans have
117 shown the mutation rate of colorectal and small intestine epithelium to be similar[7,14] (**Extended**
118 **Data Figure 1**). We then used laser microdissection on histological sections to isolate individual
119 crypts for whole-genome sequencing with a low-input library preparation method[34] (**Fig. 1a,**
120 **Extended Data Figure 2, Supplementary Table 2**), with the exception of human crypts, for which
121 sequencing data were obtained from a previous study[8]. A bioinformatic pipeline was developed to
122 call somatic mutations robustly in all these species despite variable quality of their genome
123 assemblies (**Methods**). The distribution of variant allele fractions (VAFs) of the mutations detected
124 in each crypt confirmed that crypts are clonal units in all species, enabling the study of somatic
125 mutation rates and signatures (**Extended Data Figure 3**).
126

127 We found substantial variation in the number of somatic single-base substitutions (SBS) across
128 species and across individuals within each species (**Fig. 1b**). For five species with samples from
129 multiple individuals, linear regression confirmed a clear accumulation of somatic mutations with
130 age (**Fig. 1c, Extended Data Figure 4, Supplementary Table 3**). All linear regressions were also
131 consistent with a non-significant intercept. This resembles observations in humans[14] and suggests
132 that the time required for a single stem cell to drift to fixation within a crypt is a small fraction of
133 the lifespan of a species. This facilitates the estimation of somatic mutation rates across species by
134 dividing the number of mutations in a crypt by the age of the individual (**Supplementary Table**
135 **4**). The number of somatic insertions and deletions (indels) was consistently lower than that of SBS
136 in all crypts (**Fig. 1b**), in agreement with previous findings in humans[8].
137
138 **Similar mutational signatures across mammals**
139
140 Somatic mutations can be caused by different mutational processes, including multiple forms of
141 DNA damage and repair. Different processes cause characteristic frequencies of base substitution
142 types and indels at different sequence contexts, often referred to as mutational signatures, which
143 can be inferred from mutation data[35,36]. Across species, the mutational spectra showed clear
144 similarities, with a dominance of cytosine-to-thymine (C>T) substitutions at CpG sites, as observed
145 in human colon, but with considerable variation in the frequency of other substitution types (**Fig.**
146 **2a**). To quantify the contribution of different mutational processes to the observed spectra, we
147 applied mutational signature decomposition[8,35]. We used a Bayesian model to infer mutational
148 signatures *de novo*, while accounting for differences in genome sequence composition across
149 species, and using the COSMIC human signature SBS1 (C>T substitutions at CpG sites) as a fixed
150 prior to ensure its complete deconvolution[37] (**Methods**). This approach identified two signatures
151 beyond SBS1, labelled SBSB and SBSC, which resemble COSMIC human signatures SBS5 and
152 SBS18, respectively (cosine similarities 0.93 and 0.91) (**Fig. 2b**).
153
154 This analysis suggests that the same three signatures that dominate somatic mutagenesis in the
155 human colon are dominant in other mammals: SBS1, which is believed to result from the
156 spontaneous deamination of 5-methylcytosine[8,38]; SBSB/5, a common signature across human
157 tissues that may result from endogenous damage and repair[12,39]; and SBSC/18, dominated by C>A
158 substitutions and attributed to oxidative DNA damage[36]. Signature SBSC contains a minor

159 component of T>A substitutions (resembling COSMIC SBS34), which appear to be the result of
160 DNA polymerase slippage at the boundaries between adjacent adenine and thymine homopolymer
161 tracts but that could also reflect assembly errors at those sites[39]. While all species surveyed shared
162 the three mutational signatures, their contributions varied substantially across species (**Fig. 2c**).
163 SBSC was particularly prominent in mouse and ferret, and the ratio of SBS1 to SBSB (SBS5)
164 varied from approximately 1.2 in rat or rabbit to 6.4 in tiger. In several species with data from
165 multiple individuals, separate linear regressions for each signature confirmed that mutations from
166 all three signatures accumulate with age (**Fig. 2d, Extended Data Figure 5**).
167
168 Although signature deconvolution identified three signatures active across species, we noticed
169 some differences in the mutational profile of signature SBSB among species. To explore this
170 further, we inferred independent versions of SBSB from each species, while accounting for
171 differences in genome sequence composition (**Methods**). This revealed inter-species variability in
172 the mutational profile of this signature, particularly in the C>T component (**Extended Data Figure**
173 **6**). Species-specific versions of SBSB showed different similarities to the related human signatures
174 SBS5 and SBS40. For example, SBSB inferred from the human data showed a stronger similarity
175 with the reference SBS5 human signature (cosine similarities with SBS5 and SBS40: 0.93 and
176 0.84), whereas SBSB from rabbit more closely resembled the reference human SBS40 signature
177 (0.87 and 0.91). These observations are consistent with the hypothesis that SBS5 and SBS40 result
178 from a combination of correlated mutational processes, with some variation across human
179 tissues[12,39] and across species.
180
181 Analysis of the indel mutational spectra revealed a dominance of the human indel signatures ID1
182 and ID2, characterised by single-nucleotide insertions and deletions at A/T homopolymers, and
183 probably caused by strand slippage during DNA replication[36] (**Extended Data Figure 7**). The ratio
184 of insertions (ID1) to deletions (ID2) appears to vary across species, possibly reflecting a
185 differential propensity for slippage of the template and nascent DNA strands[36]. In addition, the
186 indel spectra suggest a potential contribution of signature ID9 (whose aetiology remains unknown)
187 to human, colobus, cow, giraffe and rabbit. Analysis of indels longer than 1 base pair (bp) also
188 suggested the presence of a signature of 4-bp insertions at tetrameric repeats, particularly prevalent
189 in mouse and tiger; a pattern of ≥5-bp insertions at repeats in colobus; and a pattern of ≥5-bp

190 deletions at repeats prominent in rabbit, which resembles ID8 (a signature possibly caused by
191 double-strand break repair via non-homologous end joining[36]) (**Extended Data Figure 7**).
192
193 **Other mutational processes and selection**
194
195 The apparent lack of additional mutational signatures is noteworthy. A previous study of 445
196 colorectal crypts from 42 human donors found many crypts to be affected by a novel signature that
197 was later attributed to colibactin, a genotoxin produced by *pks*+ strains of *Escherichia coli*[8,40,41].
198 Analysing the original human data and our non-human data with the same methodology, we found
199 evidence of colibactin mutagenesis in 21% of human crypts but only uncertain evidence of
200 colibactin in one non-human crypt (0.6%) (**Extended Data Figure 8, Methods**). This revealed a
201 significant depletion of colibactin mutagenesis in the non-human crypts studied (Fisher's exact
202 tests, $P=7\times10^{-14}$). The apparent difference in colibactin mutagenesis observed between species, or
203 between the cohorts studied, might result from a different prevalence of *pks*+ *E. coli* strains[42] or a
204 different expression of colibactin by *pks*+ *E.coli* across species[43]. Finally, we also searched for
205 evidence of APOBEC signatures (SBS2/13), which have been reported in a small number of human
206 crypts and are believed to be caused by APOBEC DNA-editing cytidine deaminases. We detected
207 APOBEC signatures in 2% (n = 9) of human crypts and only uncertain evidence in one non-human
208 crypt ($P$=0.30).
209
210 Beyond single-base substitutions and indels, crypts from the eight species with chromosome-level
211 genome assemblies were inspected for large-scale copy number changes (≥1 megabase, Mb)
212 (**Methods**). Studies in humans have found large-scale copy number changes to be relatively rare
213 in normal tissues, including colorectal epithelium[8]. Consistent with these results, we only identified
214 four large copy number changes across the 162 crypts included in this analysis: two megabase-
215 scale deletions in two crypts from the same cow, the loss of a chromosome X in a female mouse
216 crypt, and a 52-Mb segment with copy-neutral loss of heterozygosity in a human crypt (**Extended**
217 **Data Figure 9, Methods**). These results suggest that large-scale somatic copy number changes in
218 normal tissues are also rare in other mammalian species.
219
220 Previous analyses in humans have shown that most somatic mutations in colorectal crypts
221 accumulate neutrally, without clear evidence of negative selection against non-synonymous

222 mutations and with a low frequency of positively selected cancer-driver mutations[8]. To study
223 somatic selection in our data, we calculated the exome-wide ratio of non-synonymous to
224 synonymous substitution rates (dN/dS) in each of the 12 species with available genome annotation.
225 To do so and to detect genes under positive selection, while accounting for the effects of
226 trinucleotide sequence context and mutation rate variation across genes, we used the dNdScv
227 model[44] (**Methods**). Although the limited number of coding somatic mutations observed in most
228 species prevented an in-depth analysis of selection, exome-wide dN/dS ratios for somatic
229 substitutions were not significantly different from unity in any species, in line with previous
230 findings in humans[8] (**Extended Data Figure 10**). Gene-level analysis did not find genes under
231 significant positive selection in any species, although larger studies are likely to identify rare
232 cancer-driver mutations[8].
233
234 **Somatic mutation rates and life-history traits**
235
236 Whereas similar mutational processes operate across the species surveyed, the mutation rate per
237 genome per year varied widely. Across the 15 species with age information, we found that
238 substitution rates per genome ranged from 47 SBS/year in humans to 796 SBS/year in mice, and
239 indel rates from 2.5 to 158 indels/year, respectively (**Fig. 3a, Supplementary Table 4, Methods**).
240
241 To explore the relationship between somatic mutation rates, lifespan and other life-history traits,
242 we first estimated the lifespan of each species using survival curves. We used a large collection of
243 mortality data from animals in zoos to minimise the impact of extrinsic mortality (**Extended Data**
244 **Figure 11**). We defined lifespan as the age at which 80% of individuals reaching adulthood have
245 died, to reduce the impact of outliers and variable cohort sizes that affect maximum lifespan
246 estimates[45,46] (**Methods**). Remarkably, we found a tight anticorrelation between somatic mutation
247 rates per year and lifespan across species (**Fig. 3b**). A log-log allometric regression yielded a strong
248 linear anticorrelation between mutation rate per year and lifespan (fraction of variance explained,
249 FVE=0.85, $P$=1x10$^{-6}$), with a slope close to and non-significantly different from -1. This supports
250 a simple model in which somatic mutation rates per year are inversely proportional to the lifespan
251 of a species (rate ∝ 1/lifespan), such that the number of somatic mutations per cell at the end of the
252 lifespan (the end-of-lifespan burden, ELB) is similar in all species (**Fig. 3a-c**).
253

254 To further study the relationship between somatic mutation rates and life-history variables, we used
255 linear mixed-effects regression models. These models account for the hierarchical structure of the
256 data (with multiple crypts per individual and multiple individuals per species), as well as the
257 heteroscedasticity of somatic mutation rate estimates across species (**Methods**). Using these
258 models, we estimated that the inverse of lifespan explained 82% of the inter-species variance in
259 somatic substitution rates (*rate* = *k*/*lifespan*) ($P$=2.9×10$^{-9}$; **Fig. 3c**), with the slope of this regression
260 (*k*) representing the mean estimated ELB across species (3206.4 substitutions per genome per crypt,
261 95% confidence interval 2683.9–3728.9). Strikingly, despite uncertainty in the estimates of both
262 somatic mutation rates and lifespans, and despite the diverse life histories of the species surveyed,
263 including ~30-fold variation in lifespan and ~40,000-fold variation in body mass, the estimated
264 mutation load per cell at the end of lifespan varied by only ~3-fold across species (**Table 1**).
265 Analogous results were obtained repeating the analysis with protein-coding mutation rate
266 estimates, which may be a better proxy for the functional impact of somatic mutations (85% of
267 variance explained, ELB: 31 coding substitutions per crypt) (**Extended Data Figure 12**,
268 **Methods**).
269
270 We next explored the association between somatic mutation rates and adult body mass, which is
271 known to be a common confounder in correlations involving lifespan[47,48]. An anticorrelation
272 between somatic mutation rates and body mass may be expected if the modulation of cancer risk
273 across species of vastly different sizes has been a major factor in the evolution of somatic mutation
274 rates. We observed that log-transformed adult body mass was less strongly associated with the
275 somatic substitution rate (allometric regression FVE=0.21, **Fig. 3d**; linear mixed-effect regression
276 FVE=0.44, **Fig. 3e**). Given that lifespan is correlated with body mass, we performed two tests to
277 assess whether body mass explained any variation in somatic mutation rates not explained by
278 lifespan. First, including both lifespan and log-transformed adult body mass in the regression model
279 suggested that body mass does not explain a significant amount of variance in somatic mutation
280 rates across species after accounting for the effect of lifespan (likelihood ratio tests: $P$=0.16 for
281 body mass on a model with lifespan, $P$<10$^{-4}$ for lifespan on a model with body mass; **Fig. 3f,**
282 **Methods**). Second, partial correlation analyses using allometric regressions further confirmed that
283 the association between somatic mutation rates and lifespan is unlikely to be mediated by the effect
284 of body mass on both variables (lifespan residuals: $P$=3.2x10$^{-6}$, FVE=0.82, **Fig. 3b**; body mass
285 residuals: $P$=0.39, FVE=0.06, **Fig. 3d**; **Methods**).

286

287 The fact that the variation in somatic mutation rates across species appears to be dominated by
288 lifespan rather than body size is also apparent when looking at particularly informative species.
289 Giraffe and naked mole-rat, for instance, have similar somatic mutation rates (99 and 93
290 substitutions/year), in line with their similar lifespans (80th percentiles: 24 and 25 years), despite
291 a ~23,000-fold difference in adult body mass (**Fig. 3d,e**). Cows, giraffes and tigers weigh much
292 more than an average human, and yet have somatic mutation rates several fold higher, in line with
293 expectation from their lifespans but not their body mass. Altogether, the weak correlation between
294 body mass and somatic mutation rates after correction for lifespan suggests that the evolution of
295 larger body sizes may have relied on alternative or additional strategies to limit cancer risk, as has
296 been speculated (**Supplementary Note 2**)[23,49–51]. Of note, the low somatic mutation rate of naked
297 mole-rats, unusual for their body mass but in line with their long lifespan (**Fig. 3d,e**), might
298 contribute to the exceptionally low cancer incidence rates of this species[52,53].
299

300 We found similar results for other life-history variables that have been proposed to correlate with
301 lifespan, namely basal metabolic rate (BMR) and litter size[54,55] (**Fig. 3f**). With the caveat that
302 estimates for these variables vary in quality, they showed weaker correlations with the somatic
303 mutation rate as single predictors, and small non-significant increases in explanatory power when
304 considered together with lifespan (likelihood ratio tests, $P$=0.92 for litter size, P=0.083 for log-
305 BMR, $P$=0.79 for allometric BMR residuals; **Fig. 3f, Methods**). We note that the results above are
306 robust to the use of alternative measures of the somatic mutation rate, including the rate per exome
307 or mutations/Mb (**Extended Data Figure 12**, **Methods**); alternative estimates of lifespan,
308 including maximum lifespan (**Extended Data Figure 13**, **Methods**); alternative regression
309 models, including a Bayesian hierarchical model and a phylogenetic generalised least-squares
310 regression, which accounts for the effect of phylogenetic relationships (**Extended Data Figure**
311 **14a,b**, **Methods**); and bootstrapping analyses at the level of individuals or species (**Extended Data**
312 **Figure 14c**, **Methods**).
313

314 **Mutational processes and lifespan**
315

316 To explore whether a single biological process could drive the association between somatic
317 mutation rates and lifespan, we analysed each mutational signature separately. SBS1, SBSB/5 and

318 SBSC/18 are believed to result from different forms of DNA damage and are expected to be subject
319 to different DNA repair pathways[12,39,56]. They also appear to differ in their association with the rate
320 of cell division in humans, with SBS1 being more common in fast-proliferating tissues, such as
321 colon and embryonic or foetal tissues, and SBS5 dominating in post-mitotic cells in the absence of
322 cell division[7,12,14]. Overall, we found clear anticorrelations between mutation rates per year and
323 lifespan for the three SBS signatures and for indels, suggesting that a single biological process or
324 DNA repair pathway is unlikely to be responsible for this association (**Fig. 4**). The total mutation
325 burden also appears to show a closer fit with lifespan than individual mutational processes, as
326 measured by the range of end-of-lifespan-burden (ELB) for each process across species (**Fig. 4**).
327 This may be expected if the observed anticorrelation were the result of evolutionary pressure on
328 somatic mutation rates.
329
330 DNA damage and somatic mutations in the mitochondrial genome have also attracted considerable
331 interest in the ageing field[57,58]. Our whole-genome sequencing of individual crypts provided high
332 coverage of the mitochondrial genome, ranging from 2,188 to 29,691-fold. Normalised against the
333 nuclear coverage, these data suggest that colorectal crypts contained on the order of ~100–2,000
334 mitochondrial genomes per cell (**Extended Data Figure 15**). Using a mutation calling algorithm
335 sensitive to low-frequency variants, we found a total of 261 mitochondrial mutations across 199
336 crypts (**Methods**). The mutational spectra across species appeared broadly consistent with that
337 observed in humans, with a dominance of C>T and A>G substitutions believed to result from
338 mtDNA replication errors rather than DNA damage[59] (**Extended Data Figure 16**). While the low
339 number of mitochondrial mutations detected per species precludes a detailed analysis, the estimated
340 number of somatic mutations per copy of mtDNA also appears to show a significant anticorrelation
341 with lifespan. Across species, we obtained an average of 0.23 detectable mutations per copy of the
342 mitochondrial genome by the end of lifespan (**Fig. 4**, **Methods**), a considerable burden given the
343 coding-sequence density and functional relevance of the mitochondrial genome.
344
345 **Discussion**
346
347 Using whole-genome sequencing of 208 colorectal crypts from 56 individuals, we provide insights
348 into the somatic mutational landscape of 16 mammalian species. Despite their different diets and
349 life histories, we found remarkable similarities in their mutational spectra. Three main mutational

350 signatures explain the spectra across species, albeit with varying contributions and subtle variations
351 in the profile of signature SBSB. These results suggest that, at least in the colorectal epithelium, a
352 conserved set of mutational processes dominate somatic mutagenesis across mammals.
353
354 The most striking finding of this study is the inverse scaling of somatic mutation rates with lifespan,
355 a long-standing prediction of the somatic mutation theory of ageing[4,30,33]. Considering evolutionary
356 and mechanistic models of ageing together provides a framework for discussing the possible
357 implications of these results for ageing (**Supplementary Note 1**). Jointly, these models predict
358 ageing to be multifactorial, with multiple forms of molecular and cellular damage contributing to
359 organismal ageing due to evolutionary limits to selection acting on the rates of these processes.
360 The inverse scaling of somatic mutation rates and lifespan is consistent with somatic mutations
361 contributing to ageing and with somatic mutation rates being evolutionarily constrained, although
362 we discuss alternative explanations below. This interpretation is also supported by studies reporting
363 more efficient DNA repair in longer-lived species[60–62]. Somatic mutations could contribute to
364 ageing in different ways. Traditionally, they have been proposed to contribute to ageing through
365 deleterious effects on cellular fitness[4,63], but recent findings question this assumption
366 (**Supplementary Note 1**). Instead, the discovery of widespread clonal expansions in ageing human
367 tissues[13,64–66] raises the possibility that some somatic mutations contribute to ageing by increasing
368 the fitness of mutant cells at a cost to the organism[63,67,68]. Recent examples include the possible
369 links between clonal haematopoiesis and cardiovascular disease[69], between mutations in liver
370 disease and insulin resistance [70], and between driver mutations in cavernomas and brain
371 haemorrhages[63,67,68]. Detailed studies on the extent and impact of somatic mutations and clonal
372 expansions in age-related diseases may help clarify the precise role, if any, of somatic mutations
373 in ageing. Even if clear causal links between somatic mutations and ageing are established, ageing
374 is likely to be multifactorial. Other forms of molecular damage involved in ageing could be
375 expected to display similar anticorrelations with lifespan and, indeed, such anticorrelations have
376 been reported for telomere shortening and protein turnover[71,72].
377
378 Alternative non-causal explanations for the observed anticorrelation between somatic mutation
379 rates and lifespan need to be considered. One alternative explanation is that cell division rates scale
380 with lifespan and explain the observed somatic mutation rates. Available estimates of cell division
381 rates, though imperfect and limited to a few species, do not readily support this argument

382 (**Methods**). More importantly, studies in humans have shown that cell division rates are not a major
383 determinant of somatic mutation rates across human tissues[7,12]. Another alternative explanation for
384 the observed anticorrelation might be that selection acts to reduce germline mutation rates in
385 species with longer reproductive spans, in turn causing an anticorrelation of somatic mutation rates
386 and lifespan. While selective pressure on germline mutation rates could influence somatic mutation
387 rates, it is unlikely that germline mutation rates determine somatic mutation rates: somatic mutation
388 rates in humans are 10–20 times higher than germline mutation rates, show variability across cell
389 types, and are influenced by additional mutational processes[12,14]. Overall, the strong scaling of
390 somatic mutation rates with lifespan across mammals, despite variable contributions of different
391 mutational processes, suggests that somatic mutation rates themselves have been evolutionarily
392 constrained, possibly through selection on multiple DNA repair pathways. Alternative explanations
393 need to be able to explain the strength of the scaling despite these differences.
394
395 Altogether, this study provides an unprecedented description of somatic mutation across mammals,
396 identifying common and variable features and shedding light on long-standing hypotheses. Scaled
397 across the tree of life and across tissues, in species with vastly different physiologies, life histories,
398 genome compositions and mutagenic exposures, similar studies promise to transform our
399 understanding of somatic mutation and its impact on evolution, ageing, and disease.
400
401
402 **References**


403 1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
404 2. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
405 3. Szilard, L. ON THE NATURE OF THE AGING PROCESS. *Proc. Natl. Acad. Sci. U. S. A.* **45**, 30–
406 45 (1959).
407 4. Morley, A. A. The somatic mutation theory of ageing. *Mutat. Res.* **338**, 19–23 (1995).
408 5. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging.
409 *Cell* **182**, 12–23 (2020).
410 6. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–
411 278 (2012).

412 7. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*
413 **538**, 260–264 (2016).
414 8. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*
415 **574**, 532–537 (2019).
416 9. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**,
417 640–646 (2020).
418 10. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single
419 human neurons. *Science* **359**, 555–559 (2018).
420 11. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic
421 mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9014–
422 9019 (2019).
423 12. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* (2021)
424 doi:10.1038/s41586-021-03477-4.
425 13. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic
426 mutations in normal human skin. *Science* **348**, 880–886 (2015).
427 14. Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *Cold Spring Harbor*
428 *Laboratory* 2020.11.25.398172 (2020) doi:10.1101/2020.11.25.398172.
429 15. Garcia, A. M. *et al.* Age- and temperature-dependent somatic mutation accumulation in Drosophila
430 melanogaster. *PLoS Genet.* **6**, e1000950 (2010).
431 16. Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational
432 processes. *Nature* **513**, 422–425 (2014).
433 17. Milholland, B. *et al.* Differences between germline and somatic mutation rates in humans and mice.
434 *Nat. Commun.* **8**, 15183 (2017).
435 18. Schmid-Siegert, E. *et al.* Low number of fixed somatic mutations in a long-lived oak tree. *Nat Plants*
436 **3**, 926–929 (2017).
437 19. Orr, A. J. *et al.* A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant.

438 *Proc. Biol. Sci.* **287**, 20192364 (2020).
439 20. López, E. H. & Palumbi, S. R. Somatic Mutations and Genome Stability Maintenance in Clonal
440 Coral Colonies. *Mol. Biol. Evol.* **37**, 828–838 (2020).
441 21. Peto, R., Roe, F. J., Lee, P. N., Levy, L. & Clack, J. Cancer and ageing in mice and men. *Br. J.*
442 *Cancer* **32**, 411–426 (1975).
443 22. Caulin, A. F. & Maley, C. C. Peto's Paradox: evolution's prescription for cancer prevention. *Trends*
444 *Ecol. Evol.* **26**, 175–182 (2011).
445 23. Tollis, M., Boddy, A. M. & Maley, C. C. Peto's Paradox: how has evolution solved the problem of
446 cancer prevention? *BMC Biol.* **15**, 60 (2017).
447 24. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *Int. J. Epidemiol.* **45**,
448 621–637 (2016).
449 25. Caulin, A. F., Graham, T. A., Wang, L.-S. & Maley, C. C. Solutions to Peto's paradox revealed by
450 mathematical modelling and cross-species cancer gene analysis. *Philos. Trans. R. Soc. Lond. B Biol.*
451 *Sci.* **370**, (2015).
452 26. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging.
453 *Cell* **153**, 1194–1217 (2013).
454 27. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the
455 ageing process. *Nature* **592**, 695–703 (2021).
456 28. Medawar, P. B. UNSOLVED problem of biology. *Med. J. Aust.* **1**, 854–855 (1953).
457 29. Williams, G. C. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* **11**, 398–
458 411 (1957).
459 30. Kirkwood, T. B. & Holliday, R. The evolution of ageing and longevity. *Proc. R. Soc. Lond. B Biol.*
460 *Sci.* **205**, 531–546 (1979).
461 31. Partridge, L. & Barton, N. H. Optimality, mutation and the evolution of ageing. *Nature* **362**, 305–311
462 (1993).
463 32. Hughes, K. A. & Reynolds, R. M. Evolutionary and mechanistic theories of aging. *Annu. Rev.*

464 *Entomol.* **50**, 421–445 (2005).
465 33. Burnet, M. Intrinsic mutagenesis. (1974) doi:10.1007/978-94-011-6606-5.
466 34. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture
467 microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
468 35. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421
469 (2013).
470 36. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101
471 (2020).
472 37. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*
473 372896 (2020) doi:10.1101/372896.
474 38. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid.
475 *Biochemistry* **13**, 3405–3410 (1974).
476 39. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures
477 caused by replication errors and endogenous DNA damage. *Nat Cancer* **2**, 643–657 (2021).
478 40. Wilson, M. R. *et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**,
479 (2019).
480 41. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks+
481 E. coli. *Nature* **580**, 269–273 (2020).
482 42. Smati, M. *et al.* Quantitative analysis of commensal Escherichia coli populations reveals host-
483 specific enterotypes at the intra-species level. *Microbiologyopen* **4**, 604–615 (2015).
484 43. Oliero, M. *et al.* Oligosaccharides increase the genotoxic effect of colibactin produced by pks+
485 Escherichia coli strains. *BMC Cancer* **21**, 172 (2021).
486 44. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**,
487 1029–1041.e21 (2017).
488 45. Moorad, J. A., Promislow, D. E. L., Flesness, N. & Miller, R. A. A comparative assessment of
489 univariate longevity measures using zoological animal records. *Aging Cell* **11**, 940–948 (2012).

490 46. Tidière, M. *et al.* Comparative analyses of longevity and senescence reveal variable survival benefits
491 of living in zoos across mammals. *Sci. Rep.* **6**, 36361 (2016).
492 47. Speakman, J. R. Correlations between physiology and lifespan--two widely ignored problems with
493 comparative studies. *Aging Cell* **4**, 167–175 (2005).
494 48. de Magalhães, J. P., Costa, J. & Church, G. M. An analysis of the relationship between metabolism,
495 developmental schedules, and longevity using phylogenetic independent contrasts. *J. Gerontol. A*
496 *Biol. Sci. Med. Sci.* **62**, 149–160 (2007).
497 49. Risques, R. A. & Promislow, D. E. L. All's well that ends well: why large species have short
498 telomeres. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).
499 50. Tollis, M., Schneider-Utaka, A. K. & Maley, C. C. The Evolution of Human Cancer Gene
500 Duplications across Mammals. *Mol. Biol. Evol.* **37**, 2875–2886 (2020).
501 51. Vazquez, J. M. & Lynch, V. J. Pervasive duplication of tumor suppressors in Afrotherians during the
502 evolution of large bodies and reduced cancer risk. *Elife* **10**, (2021).
503 52. Buffenstein, R. Negligible senescence in the longest living rodent, the naked mole-rat: insights from
504 a successfully aging species. *J. Comp. Physiol. B* **178**, 439–445 (2008).
505 53. Smith, E. S. J., Schuhmacher, L.-N. & Husson, Z. The naked mole-rat as an animal model in
506 biomedical research: current perspectives. *Open Access Anim. Physiol.* 137 (2015)
507 doi:10.2147/oaap.s50376.
508 54. Millar, J. S. & Zammuto, R. M. Life Histories of Mammals: An Analysis of Life Tables. *Ecology* **64**,
509 631–635 (1983).
510 55. Speakman, J. R. Body size, energy metabolism and lifespan. *J. Exp. Biol.* **208**, 1717–1730 (2005).
511 56. Sanders, M. A. *et al.* Life without mismatch repair. *bioRxiv* 2021.04.14.437578 (2021)
512 doi:10.1101/2021.04.14.437578.
513 57. Lane, R. K., Hilsabeck, T. & Rea, S. L. The role of mitochondrial dysfunction in age-related
514 diseases. *Biochim. Biophys. Acta* **1847**, 1387–1400 (2015).
515 58. Kauppila, T. E. S., Kauppila, J. H. K. & Larsson, N.-G. Mammalian Mitochondria and Aging: An

516 Update. *Cell Metab.* **25**, 57–71 (2017).
517 59. Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in
518 human cancer. *Elife* **3**, (2014).
519 60. Hall, K. Y., Hart, R. W., Benirschke, A. K. & Walford, R. L. Correlation between ultraviolet-induced
520 DNA repair in primate lymphocytes and fibroblasts and species maximum achievable life span.
521 *Mech. Ageing Dev.* **24**, 163–173 (1984).
522 61. Tian, X. *et al.* SIRT6 Is Responsible for More Efficient DNA Double-Strand Break Repair in Long-
523 Lived Species. *Cell* **177**, 622–638.e22 (2019).
524 62. Zhang, L. *et al.* Maintenance of genome sequence integrity in long- and short-lived rodent species.
525 *Sci Adv* **7**, eabj3284 (2021).
526 63. Smith, J. M. Review Lectures on Senescence - I. The causes of ageing. *Proceedings of the Royal*
527 *Society of London. Series B. Biological Sciences* **157**, 115–127 (1962).
528 64. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J.*
529 *Med.* **371**, 2488–2498 (2014).
530 65. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**,
531 911–917 (2018).
532 66. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers.
533 *Nature* **565**, 312–317 (2019).
534 67. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, (2019).
535 68. Ren, A. A. *et al.* PIK3CA and CCM mutations fuel cavernomas through a cancer-like mechanism.
536 *Nature* **594**, 271–276 (2021).
537 69. Jaiswal, S. & Libby, P. Clonal haematopoiesis: connecting ageing and inflammation in
538 cardiovascular disease. *Nat. Rev. Cardiol.* **17**, 137–144 (2020).
539 70. Ng, S. W. K. *et al.* Convergent somatic mutations in metabolism genes in chronic liver disease.
540 *Nature* **598**, 473–478 (2021).
541 71. Swovick, K. *et al.* Interspecies Differences in Proteome Turnover Kinetics Are Correlated With Life

542 Spans and Energetic Demands. *Mol. Cell. Proteomics* **20**, 100041 (2021).
543 72. Whittemore, K., Vera, E., Martínez-Nevado, E., Sanpera, C. & Blasco, M. A. Telomere shortening
544 rate predicts species life span. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15122–15127 (2019).
545 73. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-*
546 *bio.GN]* (2013).
547 74. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic
548 Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
549 75. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from
550 Paired End Sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
551 76. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples
552 and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
553 77. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell
554 populations. *Nat. Commun.* **3**, 811 (2012).
555 78. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
556 *Bioinformatics* **26**, 841–842 (2010).
557 79. Tacutu, R. *et al.* Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.*
558 **46**, D1083–D1090 (2018).
559 80. Conde, D. A. *et al.* Data gaps and opportunities for comparative and conservation biology. *Proc.*
560 *Natl. Acad. Sci. U. S. A.* **116**, 9658–9664 (2019).
561 81. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-
562 wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
563 82. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees,
564 and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
565 83. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between
566 symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
567 84. Rijke, R. P., Plaisier, H. M. & Langendoen, N. J. Epithelial cell kinetics in the descending colon of

568 the rat. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.* **30**, 85–94 (1979).
569 85. Potten, C. S., Kellett, M., Rew, D. A. & Roberts, S. A. Proliferation in human gastrointestinal
570 epithelium using bromodeoxyuridine in vivo: data for different sites, proximity to a tumour, and
571 polyposis coli. *Gut* **33**, 524–529 (1992).
572 86. Bach, S. P., Renehan, A. G. & Potten, C. S. Stem cells: the intestinal stem cell as a paradigm.
573 *Carcinogenesis* **21**, 469–476 (2000).

574

575 **Acknowledgements**

576

577 We are grateful to the animals whose tissues were used in this study. We thank the staff of
578 Wellcome Sanger Institute Sample Logistics, Sequencing and Informatics facilities for their
579 contribution; CASM IT and administrative staff for their help processing the data. This research
580 was made possible by the worldwide information network of zoos and aquariums, which are
581 members of Species360 and is authorized by Species360 research data use and grant agreement
582 #60633. We thank Dalia Conde and Johanna Staerk for help accessing Species360 data.

583

584 **Funding**

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

588

589 **Author Contributions**

590 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
591 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.,
592 E.St.J.S., B.N., and E.P.M. performed and facilitated sample collection. A.C. performed the laser
593 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.,
594 and C.L processed the samples. A.C., A.B.-O., F.A., N.B., T.H.H.C., M.A.S., D.J., R.A. and K.R
595 processed the data. J.S. and D.C provided animal longevity records. A.C., A.B.-O. and N.B. led
596 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.
597 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.

## Figures

**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. 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.

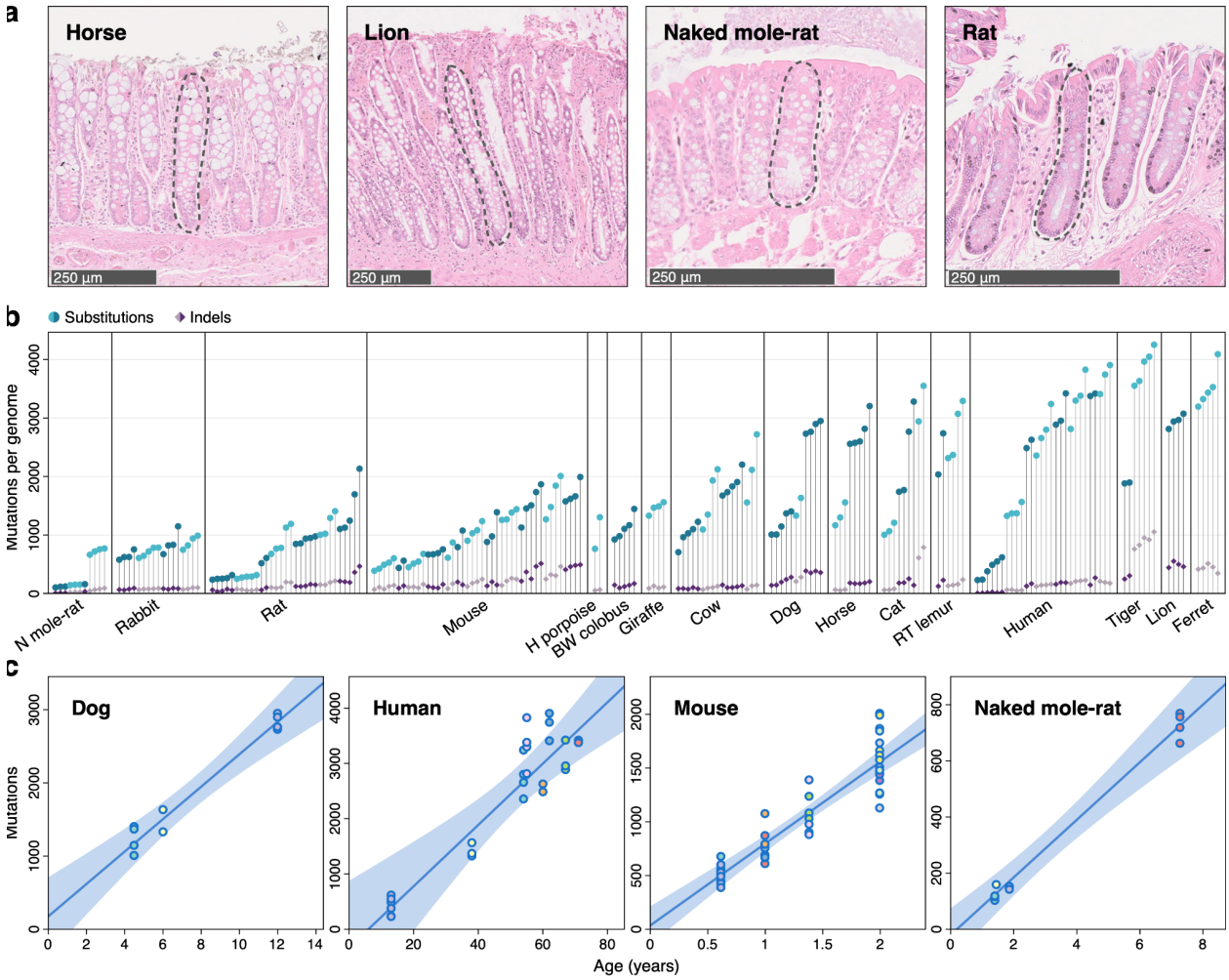


**Figure 2. Mutational processes in the mammalian colon. a**, Mutational spectra of somatic single-base substitutions in each species. The *x*-axis presents 96 mutation types on a trinucleotide context, coloured by base substitution type. **b**, Mutational signatures inferred from (SBSB, SBSC) or fitted 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.

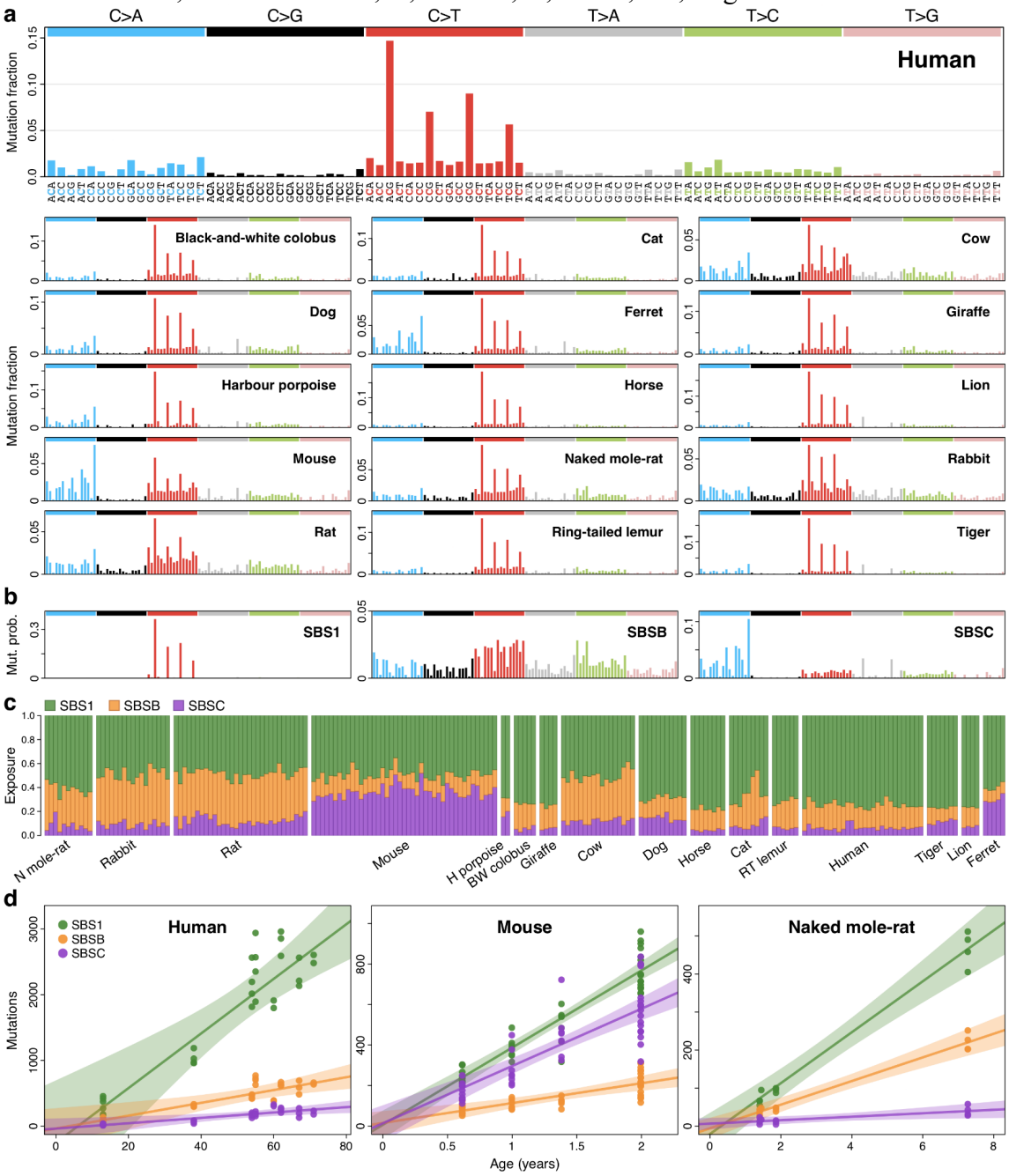
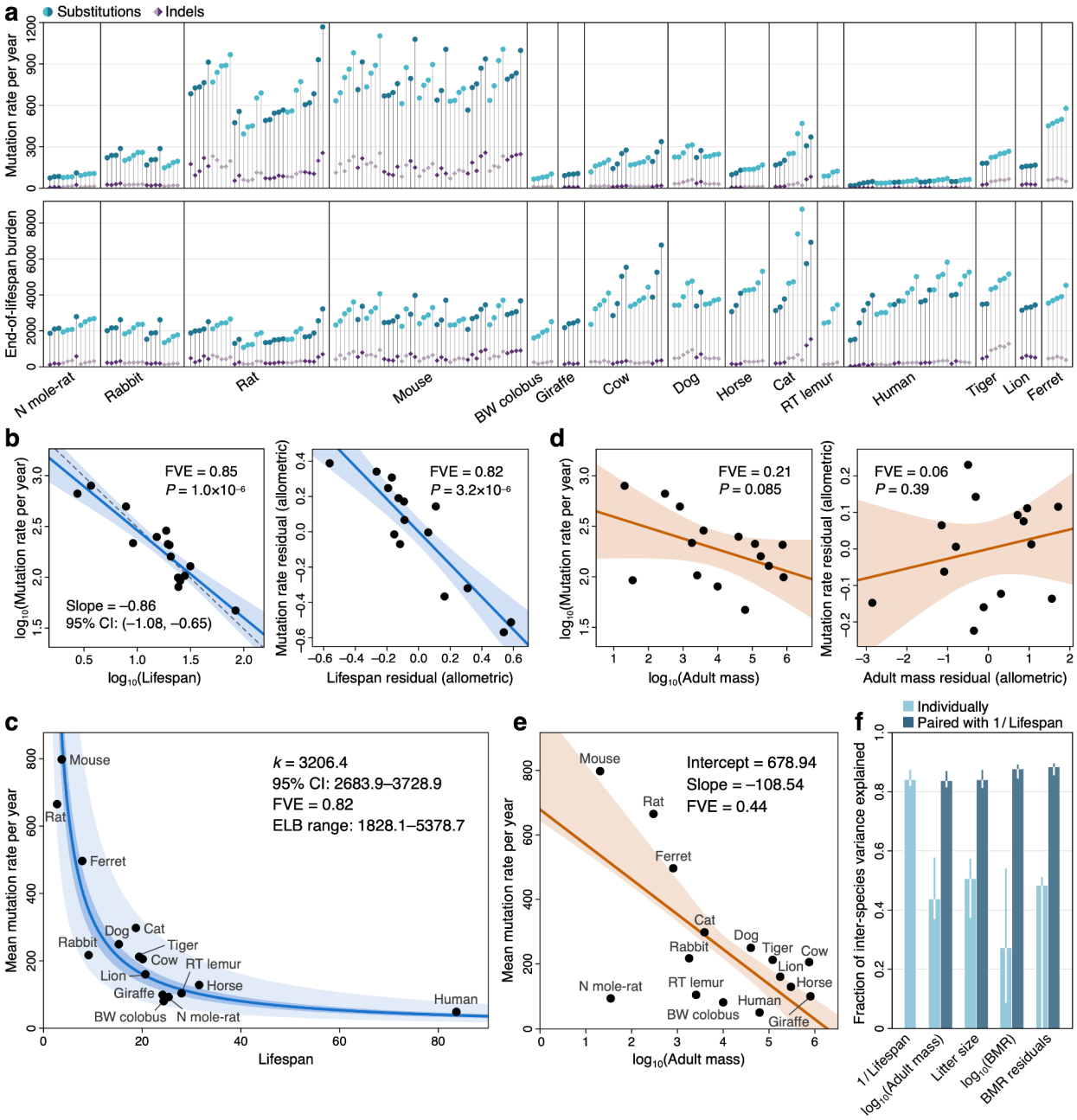


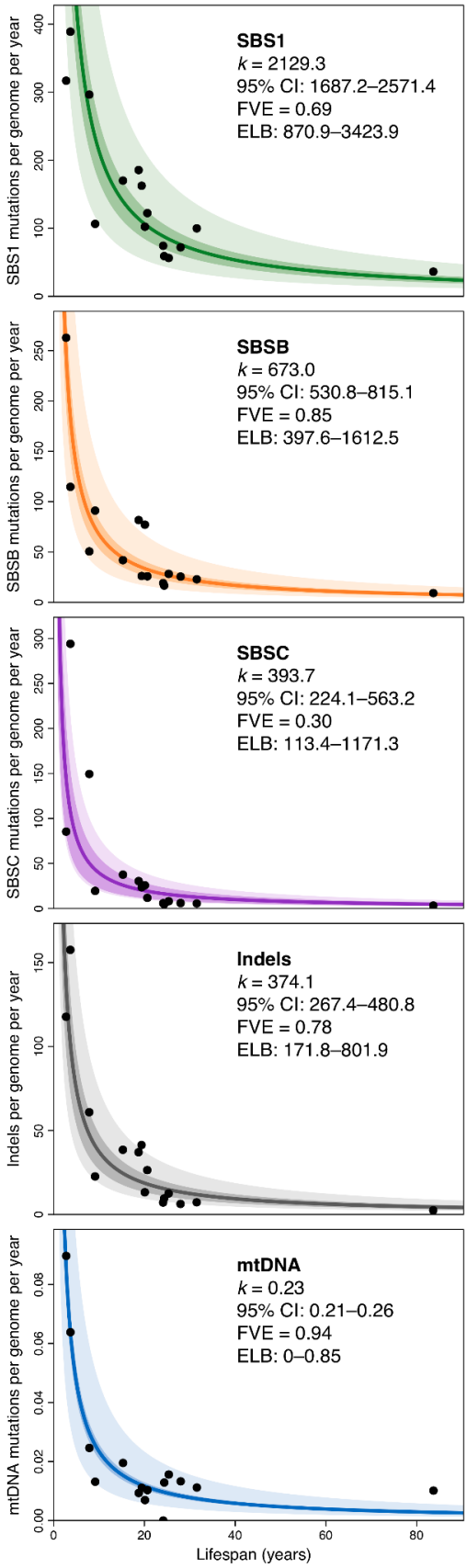

**Figure 3. Associations between somatic mutation rates and life-history variables. a**, Somatic mutation rate per year (top) and expected end-of-lifespan substitution burden in each crypt. Samples are grouped and sorted as in **Fig. 1b**; harbour porpoise samples were excluded due to the

unknown age of the individual. **b**, Allometric (log-log) regression of somatic mutation rate on lifespan (left), and regression of body-mass-adjusted residuals for somatic mutation rates and lifespan (partial correlation, **Methods**). Regressions were performed using a simple linear model on the mean mutation rate per species. Shaded areas represent 95% confidence intervals (CI) of the regression lines. Fraction of inter-species variance explained by the model (FVE) and *p*-value (*P*) are indicated (note that for simple linear regression FVE = $R^2$). Dashed line denotes a reference slope of –1. **c**, 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% CI of the regression line; lighter shaded area marks a two-fold deviation from the regression line. The point estimate and 95% CI of the slope (*k*), the FVE, and the range of end-of-lifespan burden (ELB) are shown. **d**, Allometric regression and linear regression of lifespan-adjusted residuals for somatic mutation rate and body mass. Elements are as described in **b**. **e**, Free-intercept LME regression of somatic mutation rate on log-transformed adult mass. The *y*-axis represents mean mutation rate per species, although mutation rates per crypt were used in the regression. Shaded area indicates the 95% bootstrap interval of the regression line. Point estimates of the regression intercept and slope coefficients, and model FVE, are indicated. **f**, Comparison of FVE values for free-intercept LME models using 1/lifespan and other life-history variables (alone or in combination with 1/lifespan) as explanatory variables. Error bars indicate 95% bootstrap intervals. BMR, basal metabolic rate; BW, black-and-white; N, naked; RT, ring-tailed.


a
Substitutions
Indels
Mutation rate per year
End-of-lifespan burden
N mole-rat
Rabbit
Rat
Mouse
BW colobus
Giraffe
Cow
Dog
Horse
Cat
RT lemur
Human
Tiger
Lion
Ferret
b
FVE = 0.85
P = 1.0×10⁻⁶
Slope = –0.86
95% CI: (–1.08, –0.65)
log10(Mutation rate per year)
log10(Lifespan)
FVE = 0.82
P = 3.2×10⁻⁶
Mutation rate residual (allometric)
Lifespan residual (allometric)
d
FVE = 0.21
P = 0.085
log10(Mutation rate per year)
log10(Adult mass)
FVE = 0.06
P = 0.39
Mutation rate residual (allometric)
Adult mass residual (allometric)
c
k = 3206.4
95% CI: 2683.9–3728.9
FVE = 0.82
ELB range: 1828.1–5378.7
Mean mutation rate per year
Lifespan
e
Intercept = 678.94
Slope = –108.54
FVE = 0.44
Mean mutation rate per year
log10(Adult mass)
f
Individually
Paired with 1/Lifespan
Fraction of inter-species variance explained
1/Lifespan
log10(Adult mass)
Litter size
log10(BMR)
BMR residuals

653

**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.



**Tables**

**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 Supplementary 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 |

# METHODS

## 1. Sample collection

We obtained colorectal epithelium and skin samples from a range of sources (**Supplementary 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 16 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

33 (**Supplementary Table 2**). Pre- and post-microdissection images of the tissue were recorded for
34 each crypt and muscle sample taken. Each microdissection was collected in a separate well of a
35 96-well plate.
36
37 Crypts were lysed using the Arcturus PicoPure Kit (Applied Biosystems) as previously
38 described[8,34]. Each crypt then underwent DNA library preparation, without a quantification step
39 to avoid loss of DNA, following the protocol described in Ellis et al.[34]. For some animals a
40 PAXgene fixed bulk skin biopsy was used as the germline control. For these skin samples, DNA
41 was extracted using the DNeasy Blood & Tissue Kit (Qiagen).
42
43 **3. Library preparation and sequencing**
44 Libraries from microdissected samples were prepared using enzymatic fragmentation, adapter
45 ligation and whole-genome sequencing following the method described in Ellis et al.[34]. Libraries
46 from skin samples were prepared using standard Illumina whole-genome library preparation.
47 Samples were multiplexed and sequenced using Illumina XTEN and Novaseq 6000 machines to
48 generate 150 base pair (bp) paired-end reads. Samples were sequenced to ~30× depth
49 (**Supplementary Table 2**).
50
51 **4. Sequence read alignment**
52 For each species sequences were aligned to a reference assembly (**Supplementary Table 2**)
53 using the BWA-MEM algorithm[73] as implemented in BWA version 0.7.17-r1188, with options '-
54 T 30 -Y -p -t 8'. The aligned reads were sorted using the bamsort tool from the biobambam2
55 package, version 2.0.86 (gitlab.com/german.tischler/biobambam2), with options 'fixmates=1
56 level=1 calmdnm=1 calmdnmrecompindetonly=1 calmdnmreference=<reference_fasta>
57 outputthreads=7 sortthreads=7'. Duplicate reads were marked using the bammarkduplicates2 tool
58 from biobambam2, with option 'level=0'.
59
60 **5. Variant calling**
61 Identification of somatic single-base substitutions (SBS) and short insertions and deletions
62 (indels) was divided into two steps: variant calling, and variant filtering to remove spurious calls
63 (see 'Variant filtering' below). For human colorectal crypts, we obtained previously sequenced
64 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[74], 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 (**Supplementary 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[75], 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[74,75]. 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

97 statistics. Variants without at least four high-quality fragments (alignment score ≥ 40 and base
98 Phred quality score ≥ 30) were excluded. Variants were excluded if reads supporting the variant
99 had a secondary alignment score that was greater than the primary alignment score. This filter
100 was not applied to indel calls.
101 **Hairpin filter.** To remove variants introduced by erroneous processing of cruciform DNA during
102 the enzymatic digestion we applied a custom filter to remove variants in inverted repeats[34]. This
103 filter was not applied to indel calls.
104 **Chromosome and contig filter.** For species with chromosome-level assemblies, we discarded
105 variants located in non-chromosomal contigs, including the mitochondrial genome (calling of
106 mitochondrial variants is described in the section 'Mitochondrial variant calling and filtering').
107 For males, variants on the Y chromosome were excluded for species where the Y chromosome
108 was annotated in the assembly.
109 **N-tract and contig-end filter.** To reduce artefactual calls due to read misalignment, we
110 discarded variants located within 1 kilobase (kb) of a tract of 50 or more consecutive N bases in
111 the reference assembly, as well as variants within 1 kb of the start or end of a contig (this implies
112 discarding all variants in contigs shorter than 2 kb).
113 **Sequencing coverage filter.** A sample-specific read depth filter was designed to exclude sites
114 with coverage above the 99th coverage percentile in the sample or its matched normal control, as
115 well as sites with coverage < 10× in the sample or its matched normal control.
116 **Allelic strand bias filter.** We discarded variants without any supporting reads on either the
117 forward or reverse strand.
118 **Indel proximity filter.** We discarded variants for which the total number of reads supporting the
119 presence of an indel within 10 bp of the variant was more than 3 times larger than the number of
120 reads supporting the presence of the variant. This filter was not applied to indel calls.
121 **Spatial clustering filter.** Visual assessment of variant calls and mutational spectra showed
122 spatially clustered variants to be highly enriched for artefacts. Therefore, we discarded groups of
123 two or more variants located within 1 kb of each other.
124 **Beta-binomial filter.** For each crypt, an artefact filter based on the beta-binomial distribution
125 was applied, which exploits read count information in other crypts from the same individual.
126 More specifically, for each sample, we fitted a beta-binomial distribution to the variant allele
127 counts and sequencing depths of somatic variants across samples from the same individual. The
128 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 ($\rho$) 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[76]. 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.

**Maximum indel VAF filter.** For each sample, we discarded indels presenting VAF > 0.9, as such indels were found to be highly enriched in spurious calls in some species. This filter was not applied to SBS calls.

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

161 evidence of polyclonality, or for which the possibility of polyclonality could not be excluded. An
162 expectation–maximization (EM) algorithm was employed to determine the optimal number of
163 variant allele fraction (VAF) clusters within each crypt sample, as well as each cluster's location
164 and relative contribution to the overall VAF distribution. The algorithm considered a range of
165 numbers of clusters (1–5), with the optimal number being that which minimised the Bayesian
166 Information Criterion (BIC). As the minimum number of supporting reads to call a variant was 4,
167 the binomial probability distribution was truncated to incorporate this minimum requirement for
168 the number of successes, and subsequently re-normalised. The EM algorithm returned the
169 inferred optimal number of clusters, the mean VAF (location) and mixing proportion
170 (contribution) of each clone, and an assignment of each input variant to the most likely cluster.
171 After applying this model to the somatic substitutions identified in each sample, sample filtering
172 was performed on the basis of the following three criteria.
173 **Low mutation burden.** We discarded samples presenting fewer than 50 somatic variants, which
174 was indicative of low DNA quality or sequencing issues.
175 **High mutation burden.** We discarded samples with a number of somatic variants greater than
176 three times the median burden of samples from the same individual (excluding samples with less
177 than 50 variants). This served to exclude a small minority of samples presenting evident
178 sequencing quality problems (such as low sequencing coverage), but which did not fulfill the
179 low-VAF criterion for exclusion (see below).
180 **Low VAF.** We discarded samples in which less than 70% of the somatic variants were assigned
181 to clusters with VAF $\geq 0.3$. However, this rule was not applied to those cases in which all the
182 samples from the same individual had primary clusters with mean VAF $< 0.3$; this was done to
183 prevent the removal of samples from individuals presenting high fractions of non-epithelial cells,
184 but whose crypts were nonetheless dominated by a single clone.
185 These criteria led to the exclusion of 41 out of 249 samples. On the basis of visual assessment of
186 sequencing coverage and VAF distributions, we decided to preserve three samples
187 (ND0003c_lo0004, ND0003c_lo0011, TIGRD0001b_lo0010) which we considered to be clonal,
188 but which would have been discarded based on the criteria above.
189
190 **8. Mitochondrial variant calling and filtering**
191 For six species whose reference genome assemblies did not include the mitochondrial sequence,
192 mitochondrial reference sequences were obtained from the GenBank database (**Supplementary**

193 **Table 5**). For each species, alignment to the reference genome was performed using BWA
194 version 0.7.17-r1188, as described above (see 'Sequence read alignment'). Pileup files were
195 generated for mtDNA genomes using the 'bam2R' function in the deepSNV (v1.32.0) R
196 package[76,77]. The mapping quality cut-off was set to 0, taking advantage of the fact that the
197 mitochondrial genome coverage for most samples was >100-fold higher than the nuclear genome
198 coverage, and hence most reads with poor mapping scores should be of mitochondrial origin.
199 Mitochondrial variants were called using the Shearwater algorithm[76] (deepSNV package
200 v1.32.0). Multiple rounds of filtering were applied to identify and remove false positives. The
201 first set of filters removed germline polymorphisms, applied a maximum false discovery rate
202 (FDR) threshold of $q > 0.01$, required that mismatches should be supported by at least one read
203 on both the forward and reverse strands, and merged consecutive indel calls. Further filtering
204 steps were as follows.
205 **Minimum VAF filter.** Only variants with VAF > 0.01 were considered for analysis, based on the
206 quality of the mutational spectra.
207 **Sequencing coverage filter.** Due to species-specific mtDNA regions of poor mappability, we
208 discarded sites with read coverage < 500×.
209 **D-loop filter.** Analysis of the distribution of mutations along the mitochondrial genome revealed
210 clusters of mutations within the hypervariable region of mtDNA known as the D-loop. To obtain
211 estimates of the mutation burden in mtDNA unaffected by hypermutation of the D-loop,
212 mutations in the D-loop region (coordinates MT:1–576 and MT:16,024–16,569 in human) were
213 excluded from this analysis
214 **High mutation burden.** We discarded samples having a number of somatic mtDNA variants
215 greater than four times the mean mtDNA burden across all samples. This served to exclude a
216 small minority of samples that were suspected of enrichment in false positive calls. Visual
217 inspection of these samples in a genome browser confirmed the presence of high numbers of
218 variants found on sequence reads with identical start positions and/or multiple base mismatches,
219 suggestive of library preparation or sequencing artefacts.
220
221 We examined the mutational spectra of somatic mtDNA substitutions on a trinucleotide sequence
222 context (**Extended Data Figure 16**). The specificity of the filtered variant calls was supported by
223 the observation that the mutational spectra across species were broadly consistent with those

224 previously observed in studies of human tissues[59], with a dominance of C>T and T>C
225 transversions and a strong replication strand bias.
226
227 **9. Mitochondrial copy number analysis**
228 Sequence reads from each sample were separately mapped to the species-specific mtDNA
229 reference sequence in order to estimate average mtDNA sequencing coverage. Excluding nuclear
230 reference sequences from the alignment enabled obtaining even coverage across the
231 mitochondrial genome by preventing mismapping of sequence reads to inherited nuclear
232 insertions of mitochondrial DNA (known as NuMTs). Next, coverage information from
233 individual mtDNA and whole-genome alignment (BAM) files was obtained using the genomecov
234 tool in the bedtools suite (v2.17.0)[78]. Mitochondrial copy number was calculated according to the
235 formula
236 $$depth_{mtDNA} \times ploidy \,/\, depth_{gDNA},$$
237 where $depth_{mtDNA}$ and $depth_{gDNA}$ are the mean coverage values for mtDNA and the nuclear
238 genome, respectively, and $ploidy = 2$ (assuming normal somatic cells to be diploid). For
239 simplicity, the sex chromosomes were excluded from the calculation of the mean nuclear genome
240 coverage.

241 **10. Analysable genome size calculation**
242 To estimate the somatic mutation rate, it was first necessary to establish the size of the analysable
243 nuclear genome (i.e. the portion of the genome where variant calling could be performed reliably)
244 for each sample (**Supplementary Table 4**). For both single-base substitutions and indels, the
245 analysable genome of a sample was defined as the complement of the union of the following
246 genomic regions: regions reported as 'not analysed' by the CaVEMan variant caller; regions
247 failing the 'chromosome and contig' filter; regions failing the 'N-tract and contig-end' filter; and
248 regions failing the 'sequencing coverage' filter (see 'Variant filtering'). For the analysis of
249 mitochondrial variants, the analysable genome of a sample was defined as the portion of mtDNA
250 satisfying the 'sequencing coverage' filter (see 'Mitochondrial variant calling and filtering'), after
251 subtracting the hypervariable region (D-loop).
252
253 **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 (**Supplementary 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[79]. This resulted in a mean of 2,681 animal lifespan records per species for the species in the study (minimum 309, maximum 8403; **Supplementary 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 (**Supplementary 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 13. 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

286 mean of the three countries as our estimate of the human 80% lifespan (83.7 years)
287 (**Supplementary Table 6**).
288
289 To test the impact of different estimates of lifespan on our results, we also obtained maximum
290 longevity estimates for each species from a range of databases[80] and a survey of the literature
291 (**Supplementary Table 6**). Other life-history metrics were obtained from the AnAge database[79]
292 (**Supplementary Table 6**).
293
294 **12. Mutational signature analysis**
295 Mutational signatures of single-base substitutions on a trinucleotide sequence context were
296 inferred from sets of somatic mutation counts using the sigfit (v2.0) R package[37]. Initially,
297 signature extraction was performed *de novo* for a range of numbers of signatures ($N = 2, ..., 10$),
298 using counts of mutations grouped per sample, per individual and per species. To account for
299 differences in sequence composition across samples, and especially across species, mutational
300 opportunities per sample, per individual and per species were calculated from the reference
301 trinucleotide frequencies across the analysable genome of each sample (see 'Analysable genome
302 size calculation'), and supplied to the 'extract_signatures' function in sigfit. The
303 'convert_signatures' function in sigfit was subsequently used to transform the extracted
304 signatures to a human-relative representation (**Fig. 2b**), by scaling the mutation probability values
305 using the corresponding human genome trinucleotide frequencies. The best-supported number of
306 signatures, on the basis of overall goodness-of-fit[37] and consistency with known COSMIC
307 signatures (cancer.sanger.ac.uk/signatures), was found to be $N = 3$. The cleanest deconvolution of
308 the three signatures was achieved when using the mutation counts grouped by species, rather than
309 by sample or individual. The three extracted signatures (labelled SBSA, SBSB, SBSC) were
310 found to be highly similar to COSMIC signatures SBS1 (cosine similarity 0.96), SBS5 (0.89),
311 and SBS18 (0.91), respectively. These signatures were independently validated using the
312 MutationalPatterns (v1.12.0) R package[81], which produced comparable signatures (respective
313 cosine similarities 0.999, 0.98 and 0.89).
314
315 This *de novo* signature extraction approach, however, failed to deconvolute signatures SBSA and
316 SBSB entirely from each other, resulting in a general overestimation of the exposure to SBSA
317 (**Extended Data Figure 18**). To obtain more accurate estimates of signature exposure, the

318 deconvolution was repeated using an alternative approach that combines signature fitting and
319 extraction in a single inference process[37]. More specifically, the 'fit_extract_signatures' function
320 in sigfit was used to fit COSMIC signature SBS1 (retrieved from the COSMIC v3.0 signature
321 catalogue; cancer.sanger.ac.uk/signatures) to the mutation counts grouped by species (with
322 species-specific mutational opportunities), while simultaneously extracting two additional
323 signatures *de novo* (SBSB and SBSC). Before this operation, COSMIC SBS1 was transformed
324 from its human-relative representation to a genome-independent representation using the
325 'convert_signatures' function in sigfit. By completely deconvoluting SBS1 and SBSB, this
326 approach yielded a version of SBSB that was more similar to COSMIC SBS5 (cosine similarity
327 0.93); the similarity of SBSC to COSMIC SBS18 was the same under both approaches (0.91).
328
329 Finally, the inferred signatures were re-fitted to the mutational spectra of mutations in each
330 sample (using the 'fit_signatures' function in sigfit with sample-specific mutational
331 opportunities) to estimate the exposure of each sample to each signature. The fitting of the three
332 signatures yielded spectrum reconstruction similarity values (measured as the cosine similarity
333 between the observed mutational spectrum and a spectrum reconstructed from the inferred
334 signatures and exposures) with median 0.98 and interquartile range 0.96–0.99. Although the
335 purely *de novo* extraction approach and the 'fitting and extraction' approach yielded comparable
336 versions of signatures SBSB and SBSC, the fixing of COSMIC SBS1 in the latter approach
337 resulted in lower SBS1 exposures and higher SBSB exposures in the majority of samples, due to
338 the cleaner deconvolution of these two signatures (**Fig. 2, Extended Data Figure 18**).
339
340 To examine potential variation in the spectrum of signature SBS5 across species, the following
341 procedure was conducted for each species: individual-specific mutation counts and mutational
342 opportunities were calculated for each individual in the species, and the 'fit_extract_signatures'
343 function was used to fit COSMIC signatures SBS1, SBS18 and SBS34 (transformed to a
344 genome-independent representation using the 'convert_signatures' function) to the mutational
345 spectra of each individual, while simultaneously inferring one additional signature
346 (corresponding to signature SBS5 as manifested in that species; **Extended Data Figure 6**).
347
348 To assess the presence in non-human colorectal crypts of mutational signatures caused by
349 APOBEC or colibactin, which have been previously observed in human crypts[8], we used an

350 expectation–maximisation algorithm for signature fitting, in combination with likelihood ratio
351 tests (LRTs). More specifically, for each non-human sample, we tested for exposure to colibactin
352 (signature SBS88, COSMIC v3.2) by comparing the log-likelihoods of (*i*) a model fitting
353 COSMIC signatures SBS1, SBS5, SBS18, SBS34 and SBS88, and (*ii*) a reduced model fitting
354 only the first four signatures. Benjamini–Hochberg multiple-testing correction was applied to the
355 *p*-values resulting from the LRTs, and colibactin exposure was considered significant in a sample
356 if the corresponding corrected *q*-value was less than 0.05. We followed the same approach to
357 assess exposure to APOBEC (SBS2 and SBS13), using two separate sets of LRTs for models
358 including either SBS2 or SBS13, in addition to SBS1, SBS5, SBS18 and SBS34. APOBEC
359 exposure was considered significant in a sample if its *q*-values for the models including SBS2
360 and SBS13 were both less than 0.05. This analysis identified 1/180 crypts with significant
361 exposure to each of colibactin and APOBEC (although the evidence for the presence of the
362 relevant signatures in these two crypts was not conclusive). To test for depletion of colibactin or
363 APOBEC exposure in non-human crypts relative to human crypts, we first applied the LRT-
364 based method described above to a published set of 445 human colorectal crypts[8], identifying 92
365 colibactin-positive and 9 APOBEC-positive crypts. We then compared the numbers of colibactin-
366 and APOBEC-positive crypts in the human and non-human sets using two separate Fisher's exact
367 tests ('fisher.test' function in R). This revealed the difference in colibactin exposure to be highly
368 significant ($P$=7×10$^{-14}$), unlike the difference in APOBEC exposure ($P$=0.30).
369
370 Mutational spectra of somatic indels identified in each species were generated using the
371 'indel.spectrum' function in the Indelwald tool for R (24/09/2021 version;
372 github.com/MaximilianStammnitz/Indelwald).
373
374 **13. Selection analysis**
375 Evidence of selection was assessed using the ratio of nonsynonymous to synonymous substitution
376 rates (dN/dS) in the somatic mutations called in each species. The dNdScv (v0.0.1.0) R package[44]
377 was used to estimate dN/dS ratios for missense and truncating substitutions in each species
378 separately. Reference CDS databases for the dNdScv package were built for those species with
379 available genome annotation in Ensembl (www.ensembl.org; **Supplementary Table 2**), using
380 the 'buildref' function in dNdScv. For each species, the 'dndscv' function was applied to the list
381 of somatic substitutions called in samples of that species, after de-duplicating any substitutions

382 that were shared between samples from the same individual in order to avoid counting shared
383 somatic mutations multiple times. In addition, the analysis was restricted to genes that were fully
384 contained in the analysable genomes of all samples from the species (a condition satisfied by the
385 vast majority of protein-coding genes). Genome-wide and gene-specific dN/dS ratios were
386 obtained for missense and truncating substitutions in each species; no genes with statistically
387 significant dN/dS ≠ 1 were observed.
388
389 **14. Copy number analysis**
390 For species with chromosome-level assemblies (cat, cow, dog, horse, human, mouse, rabbit, rat),
391 total and allele-specific copy number (CN) were assessed in each sample adapting a likelihood
392 model previously applied to the detection of subclonal CN changes in normal human skin[13]. This
393 method exploits two sources of evidence: relative sequencing coverage and B-allele fraction
394 (BAF; the fraction of reads covering a heterozygous SNP that support one of the alleles). Human
395 samples PD36813x15 and PD36813x16 were excluded from this analysis due to the poor quality
396 of their SNP data.
397
398 For each sample, sequencing coverage was measured in non-overlapping 100-kilobase (kb) bins
399 along the species' reference genome, using the coverageBed tool in the bedtools suite (v2.17.0)[78].
400 For each bin, the coverage per base pair was calculated by dividing the number of reads mapping
401 to the bin by the bin length, and multiplying the result by the read length (150 bp). A normalised
402 sample–normal coverage ratio was then calculated for each bin by dividing the bin coverage in
403 the sample by the corresponding coverage in its matched normal control (see 'Sample
404 processing'). Heterozygous SNPs were isolated for each sample by selecting germline SNPs with
405 a BAF between 0.4 and 0.6 in the matched normal sample, and a coverage of at least 15 reads in
406 both the target sample and its matched normal sample. After assigning each SNP to its
407 corresponding 100-kb genome bin, the bins in each sample were divided into two sets: (*i*) bins
408 with coverage ≥ 10 in both the target sample and its matched normal, and at least one
409 heterozygous SNP; and (*ii*) bins with coverage ≥ 10 in both the target sample and its matched
410 normal, and no heterozygous SNPs. For the first set, estimates of total and allele-specific CN
411 were inferred by maximising the joint likelihood of a beta-binomial model for BAF and a
412 negative binomial model for relative coverage, as previously described[13]. The most likely
413 combination of allele CN values was obtained for each bin by conducting an exhaustive search of

414 CN values between 0 and 4, and selecting the combination maximising the joint likelihood
415 (calculated on the basis of expected BAF and relative coverage values). A penalty matrix was
416 used to penalise more complex solutions over simpler ones, as previously described[13]. For the
417 second set of bins (bins without SNPs), only estimates of total CN were inferred, by maximising
418 the likelihood of a negative binomial model for relative coverage. The most substantial
419 differences between these methods and the one previously published are: (*i*) SNPs were obtained
420 from the variant calling output, instead of from a public database; (*ii*) relative coverage was
421 calculated per 100-kb bin, rather than per SNP; (*iii*) SNPs were not phased within each gene, but
422 within each bin; (*iv*) no reference bias was assumed (i.e. the underlying BAF of heterozygous
423 SNPs was assumed to be 0.5); (*v*) the minimum sample purity was raised to 0.85; (*vi*) putative
424 CN changes were not subjected to significance testing, but selected according to their likelihood,
425 and subsequently filtered by means of a segmentation algorithm (see below).
426
427 Estimates of total and allele-specific CN per bin were merged into CN segments, which were
428 defined as contiguous segments composed of five or more bins with identical CN states.
429 Segmentation was performed separately for total and allele-specific CN estimates in each sample.
430 After this process, any pair of adjacent segments with the same CN assignment, and separated by
431 a distance shorter than 5 bins, was merged into a single segment. Finally, within each species,
432 segments presenting CN values other than 2 (or 1/1 for allele-specific CN), and being either
433 shorter than 10 bins (1 megabase), or shared among two or more samples, were discarded,
434 resulting in the removal of nearly all spurious CN changes.
435
436 **15. Mutation rate estimation**
437 For each sample, the somatic mutation density (mutations per bp) was calculated by dividing the
438 somatic mutation burden (total number of mutations called) by the analysable genome size for the
439 sample (see 'Analysable genome size calculation'). The adjusted somatic mutation burden
440 (number of mutations per whole genome) was then calculated by multiplying the mutation
441 density by the total genome size of the species (see below). The somatic mutation rate per year
442 (mutations per genome per year) was obtained by dividing this adjusted mutation burden by the
443 age of the individual, expressed in years (**Supplementary Table 2**). The expected end-of-
444 lifespan burden (ELB) for each sample was calculated by multiplying the somatic mutation rate
445 by the estimated lifespan of the species (see 'Life history data').

446

447 The total genome size of a species was estimated as the total size of its reference genome
448 assembly. Across species, the mean genome size was 2.67 gigabases (Gb), ranging between 2.41
449 Gb and 3.15 Gb and with a standard deviation of 221 megabases (**Supplementary Table 4**). This
450 suggests that inter-species variation in genome size should not have a substantial influence on the
451 somatic mutation rate estimates. For an assessment of alternative measures of mutation rate, see
452 'Association of mutation rate and end-of-lifespan burden with lifespan'.
453

454 **16. Association of mutation rate with life-history traits**

455 The association of the somatic mutation rate with different life-history traits was assessed using
456 linear mixed-effects (LME) models. In particular, associations with the following traits were
457 examined: lifespan (in years), adult mass (or adult weight, in grams), basal metabolic rate (BMR,
458 in watts), and litter size (see 'Life history data'). Associations for lifespan, adult mass and BMR
459 were assessed using the following transformed variables: 1/Lifespan, $\log_{10}$(Adult mass), and
460 $\log_{10}$(BMR). To account for the potentially confounding effect of the correlation between
461 metabolic rate and body mass, the residuals of a fitted allometric regression model of BMR on
462 adult mass (equivalent to a simple linear regression of $\log_{10}$(BMR) on $\log_{10}$(Adult mass)) were
463 employed as a mass-adjusted measure of metabolic rate, referred to as 'BMR residuals'.
464

465 For each variable, an LME model was implemented for the regression of somatic mutation rates
466 per sample on the variable of interest, using the 'lme' function in the nlme R package (v3.1-137;
467 cran.r-project.org/package=nlme). To account for non-independence of the samples, both at the
468 individual level and at the species level, the model included fixed effects (intercept and slope
469 parameters) for the variable of interest, and random effects (slope parameters) at the individual
470 and species levels. In addition, to account for the heteroscedasticity of mutation rate estimates
471 across species, the usual assumption of constant response variance was replaced with explicit
472 species-specific variances, to be estimated within the model.
473

474 To determine the fraction of inter-species variance in mutation rate explained by each life-history
475 variable individually, the LME model described above was used to produce predictions of the
476 mean mutation rate per species; only fixed effects were employed when obtaining these
477 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:

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

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

510 allometric regressions among these three variables) were used to examine the associations
511 between somatic mutation rate and either body mass or lifespan, after accounting for the effect of
512 the third variable (partial correlation analysis). For instance, to account for the potential influence
513 of body mass on the relationship between somatic mutation rate and lifespan, the residuals of an
514 allometric regression between mutation rate and adult mass, and the residuals of an allometric
515 regression between lifespan and adult mass, were analysed using simple linear regression. This
516 analysis supported a strong association between somatic mutation rate and lifespan
517 (independently of the effect of mass; FVE=0.82, $P$=3.2×10$^{-6}$; **Fig. 3c**), and a non-significant
518 association between somatic mutation rate and body mass (independently of the effect of
519 lifespan). Therefore, the relationship between somatic mutation rate and lifespan does not appear
520 to be mediated by the effect of body mass on both variables. Importantly, this result remains after
521 excluding naked mole-rat: after removing this species, partial correlation analysis still reveals a
522 strong association between somatic mutation rate and lifespan (FVE=0.77, $P$=4.1×10$^{-5}$), and a
523 non-significant association between somatic mutation rate and body mass ($P$=0.84). This
524 demonstrates that the observed relationships are not dependent on the presence of naked mole-rat
525 in the study.
526
527 To assess the robustness of the LME regression analyses described above, we performed
528 bootstrap analysis on each LME model, at the level of both individuals and species. More
529 specifically, for each level we used each of the LME models to perform regression on 10,000
530 bootstrap replicates, produced by resampling either species or individuals with replacement. We
531 then assessed the distributions of FVE across bootstrap replicates (**Extended Data Figure 14c**).
532 Additionally, we performed a similar bootstrap analysis using a collection of maximum longevity
533 estimates obtained from the literature (see section 'Life history data'). We applied the zero-
534 intercept LME model described above (for regressing mutation rate on inverse lifespan) on a set
535 of 5,000 bootstrap replicates, each of which used a set of species lifespan estimates randomly
536 sampled from the collection of literature-derived estimates (**Extended Data Figure 13**).
537
538 The results obtained with the LME models were additionally validated using an independent
539 hierarchical Bayesian model, in which the mean somatic mutation burden of each individual was
540 modelled as following a normal distribution with mean defined as a linear predictor containing a
541 species-specific slope parameter and a multiplicative offset (corresponding to the individual's

542 age; inclusion of this offset minimises the heteroscedasticity of the mutation rate across species,
543 which results from dividing mutation burdens by age). Species-specific slope parameters were in
544 turn modelled as normally distributed around a global slope parameter, equivalent to the fixed-
545 effect slope estimated by the LME model. This hierarchical model produced very similar results
546 to those of the LME model for all life-history variables (**Extended Data Figure 14a**).
547
548 We note that samples CATD0002b_lo0003 and MD6267ab_lo0003 were excluded from all
549 regression analyses, due to the fact that each shared the majority of its somatic variants with
550 another sample from the same individual (indicating, in each case, that both samples were closely
551 related), hence violating the assumption of independence among samples. The inclusion of these
552 two samples, however, had no effect on the outcome of the analyses.
553
554 **17. Association of mutation rate and end-of-lifespan burden with lifespan**
555 The relationship between somatic mutation rate and species lifespan was further explored by
556 adapting the LME model described in the previous section to perform constrained (zero-
557 intercept) regression of the adjusted mutation rate per year on the inverse of lifespan, 1/Lifespan
558 (see 'Life history data', 'Mutation rate estimation' and 'Association of mutation rate with life-
559 history traits'). The use of zero-intercept regression was motivated by the prediction that, if
560 somatic mutation is a determinant of maximum lifespan, then it would be expected for all species
561 to end their lifespans with a similar somatic mutation burden. Indeed, this was confirmed via
562 simple linear regression of the species mean end-of-lifespan mutation burden against species
563 lifespan (slope $P$=0.39). Thus, if $m$ is the mutation rate per year, and $L$ is the species' lifespan, the
564 expected relationship is of the form
565 $$m\,L \approx k,$$
566 where $k$ is a constant representing the typical end-of-lifespan mutation burden across species.
567 According to this relationship, the mutation rate per year is linearly related to the inverse of
568 lifespan,
569 $$m \approx k\,(1/L).$$
570 Therefore, the cross-species average end-of-lifespan burden ($k$), can be estimated as the slope
571 parameter of a zero-intercept linear regression model with the mutation rate per year ($m$) as the
572 dependent variable, and the inverse of lifespan ($1/L$) as the explanatory variable. To this purpose,
573 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[82], 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 14b**), suggesting that phylogenetic non-independence of the samples does not have a substantial effect on the association analyses.

606 **18. Cell division analysis**

607 To investigate the extent to which differences in cell division rate could explain differences in
608 mutation rate and burden across species, we obtained estimates of intestinal crypt cell division
609 rates from mouse[83], rat[84] and human[85,86] (**Supplementary Table 7**). Using these cell division
610 rates, our lifespan estimates and the observed SBS mutation rates, we calculated the number of
611 cell divisions at the end of lifespan and the corresponding number of mutations per cell division
612 expected under a simple model assuming that all mutations occur during cell division
613 (**Supplementary Table 7**).
614
615 We investigated whether differences in the number of cell divisions among species could explain
616 the observed differences in mutation burden. Although colorectal cell division rate estimates are
617 lacking for most species, existing estimates from mouse, rat and human indicate that the total
618 number of stem cell divisions per crypt in a lifetime varies greatly across species, for example
619 with ~6–31 fold more divisions per intestinal stem cell in a human than a rat over their respective
620 lifetimes, depending on the cell division rate estimates used (**Supplementary Table 7**). Mouse
621 intestinal stem cells are estimated to divide once every 24 hours[83], while estimates of the human
622 intestinal stem cell division rate vary from once every 48 hours[85] to once every 264 hours[86]. Thus
623 mouse intestinal stem cells divide 2–11 times faster than human intestinal stem cells. By the end
624 of lifespan, an intestinal stem cell is predicted to have divided ~1351 times in a mouse, ~486
625 times in a rat and 2774–15,257 times in a human (depending on the cell division rate estimate
626 used). Applying our somatic mutation burden and lifespan data, this implies that the somatic
627 mutation rate per cell division in a mouse is ~1.5–8.4 fold higher than in a human. However, the
628 observed fold difference in somatic mutation rate between these two species is 16.9 (**Table 1**).
629 Therefore, differences in cell division rate appear unable to fully account for the observed
630 differences in mutation rate across species. Nevertheless, we note that accurate cell division rate
631 estimates for basal intestinal stem cells are lacking for the majority of species.
632
633
634 **19. Data reporting**

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

## 20. Ethics statement

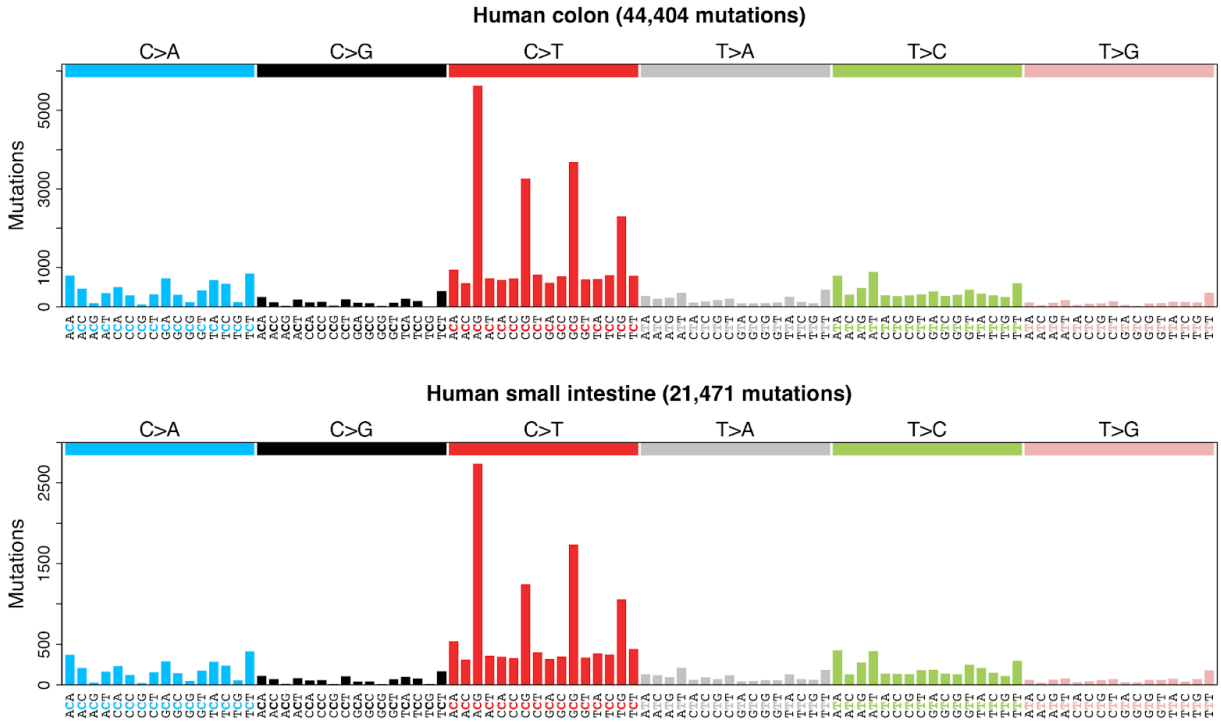
All animal samples were obtained with the approval of the local ethical review committee (AWERB) at the Wellcome Sanger Institute and those at the holding institutions.

## 21. Data and code availability

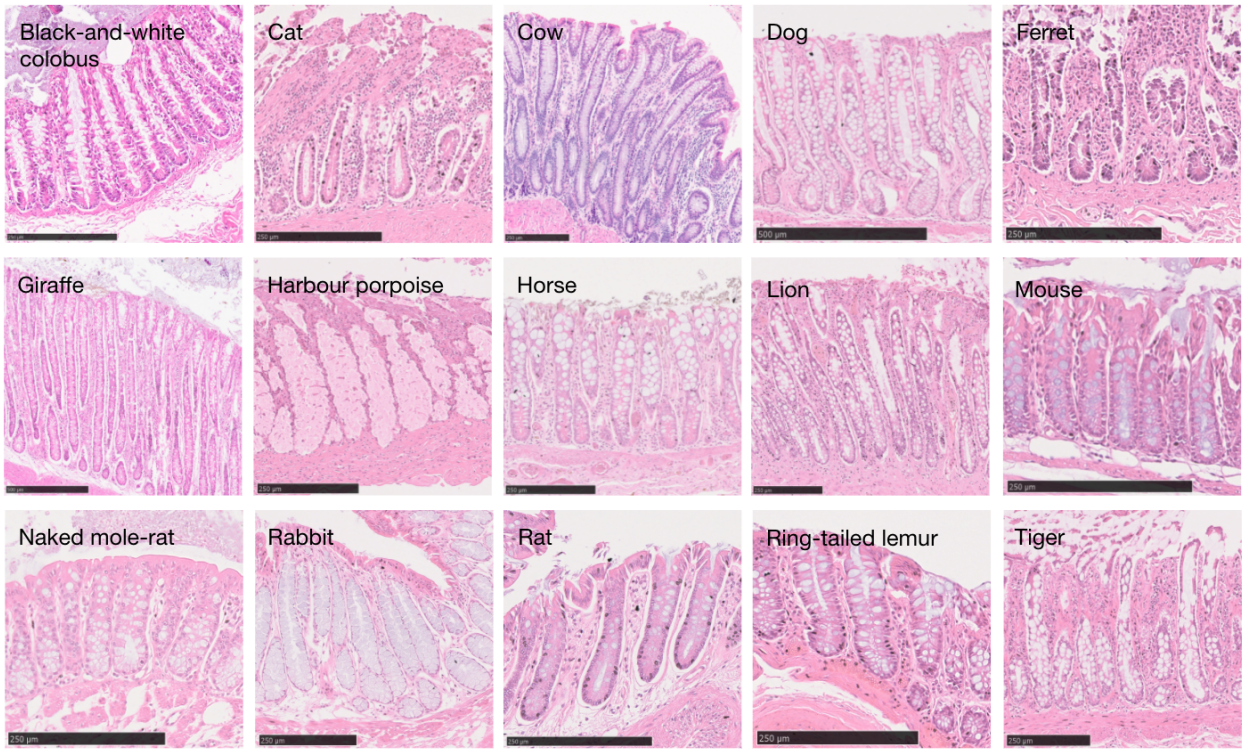
DNA sequence data have been deposited in the European Genome-Phenome Archive (ega-archive.org) under overarching accession EGAD00001008032. The computer code and data files used in the analyses are available on Zenodo (doi.org/10.5281/zenodo.5554801; doi.org/10.5281/zenodo.5554777) and GitHub (github.com/baezortega/CrossSpecies2021).

## Extended Data Figures

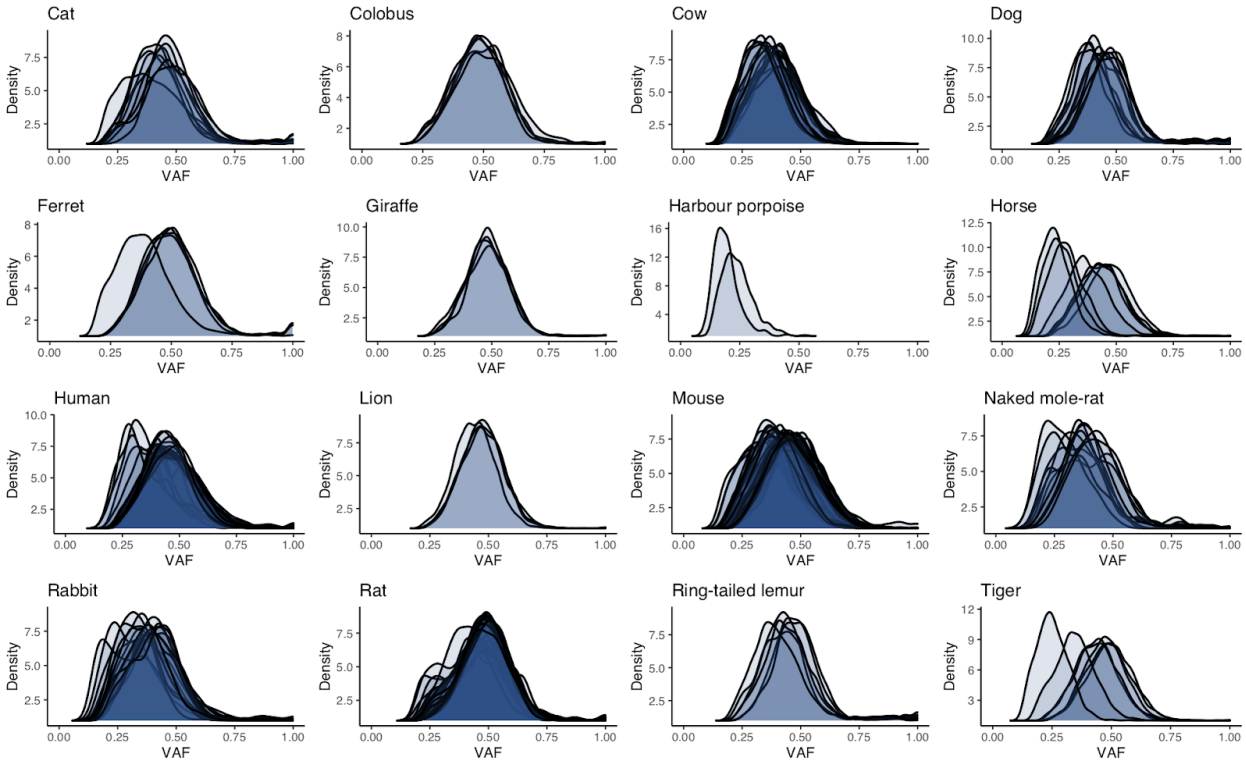
**Extended Data Figure 1. Somatic mutational spectra of the human colon and small intestine.** Trinucleotide-context mutational spectra of somatic single-base substitutions from human adult stem cells in colon (top) and small intestine, using mutation calls obtained from Blokzijl *et al.*[7].



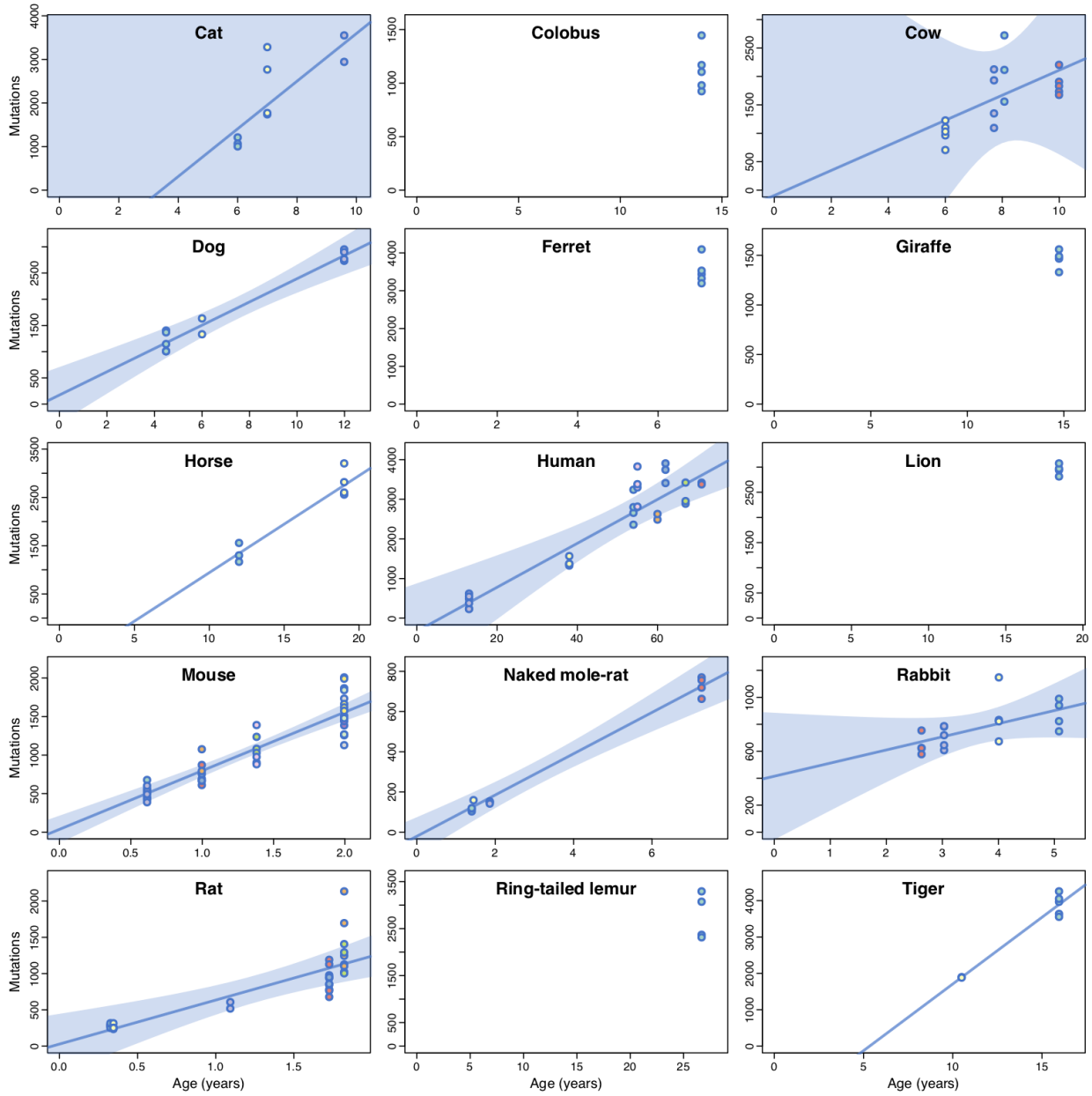
**Extended Data Figure 2. Histology images of intestinal crypts across species.** Histological images of the colorectal or intestinal (ferret) epithelium for each non-human species. Scale bars are provided at the bottom of each image.



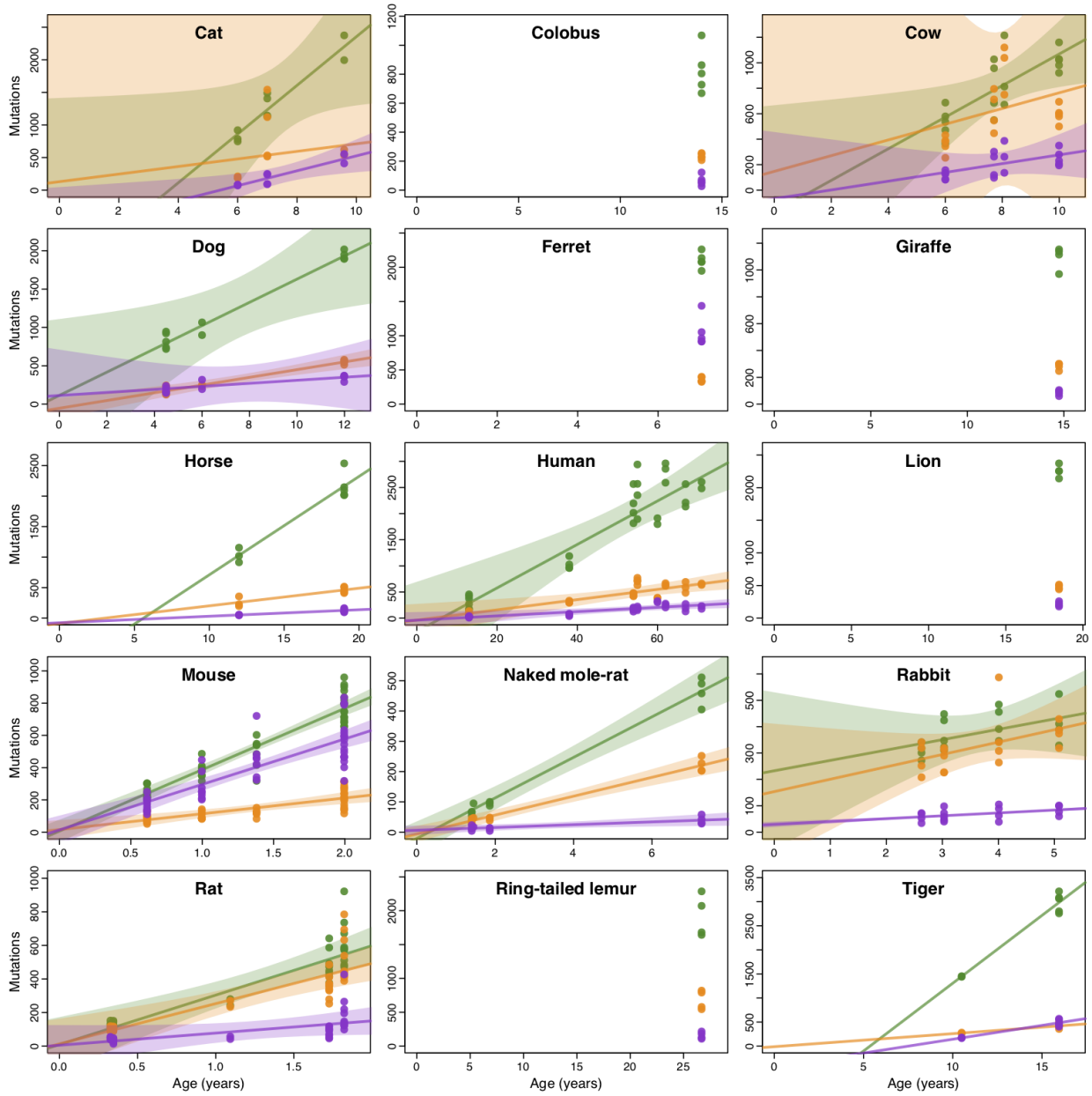
**Extended Data Figure 3. Somatic VAF distributions per species.** Distributions of variant allele fraction (VAF) for somatic substitutions in each crypt for each species. Each distribution refers to the variants in a single sequenced crypt.



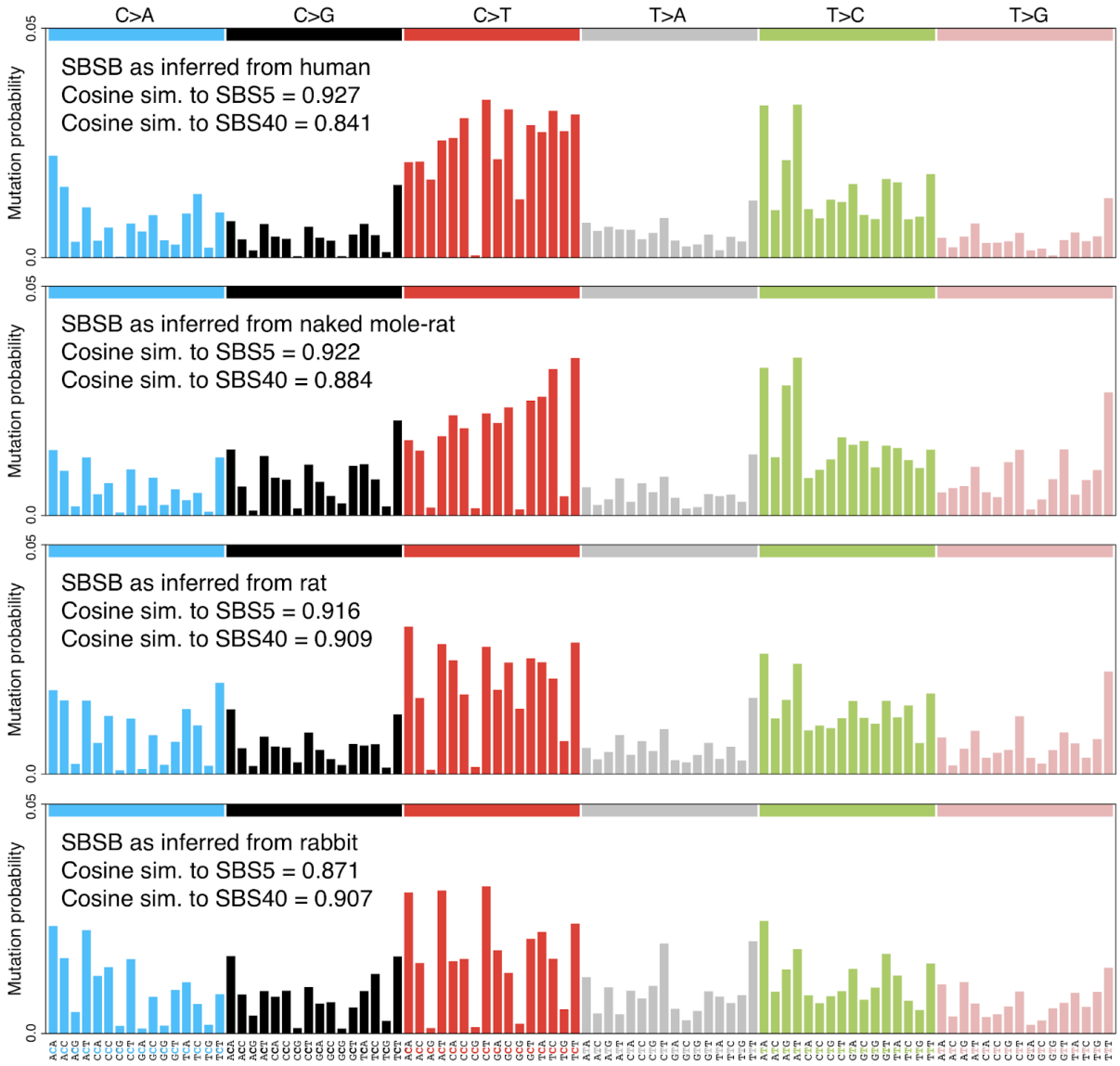
**Extended Data Figure 4. Somatic mutation accumulation across species.** Each panel presents somatic substitution burdens per genome for a given species. Each dot represents a crypt sample, with samples from the same individual sharing the same colour. For species with two or more individuals, the estimated 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.



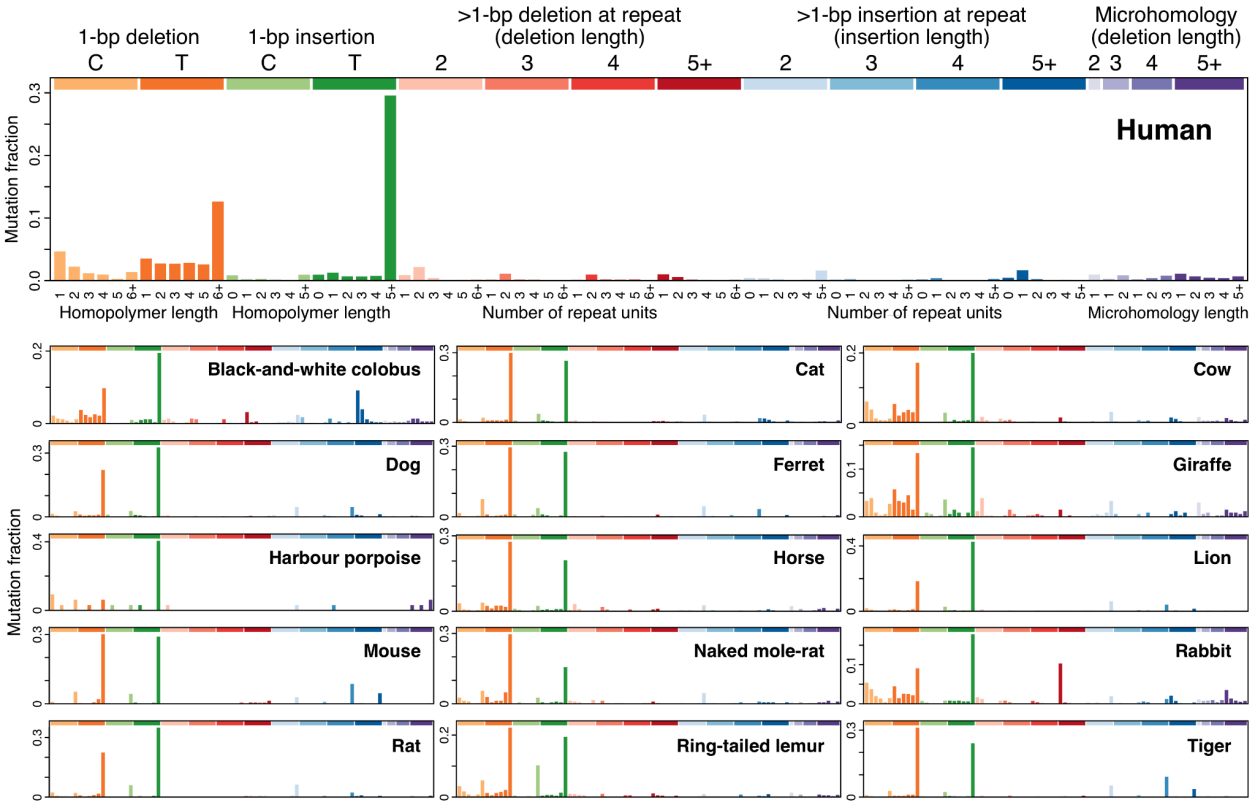
**Extended Data Figure 5. Signature-specific mutation accumulation across species.** Each panel presents somatic substitution burdens per genome for mutational signatures SBS1 (green), SBSB (yellow) and SBSC (purple) in a given species. For species with two or more individuals, the estimated regression 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.



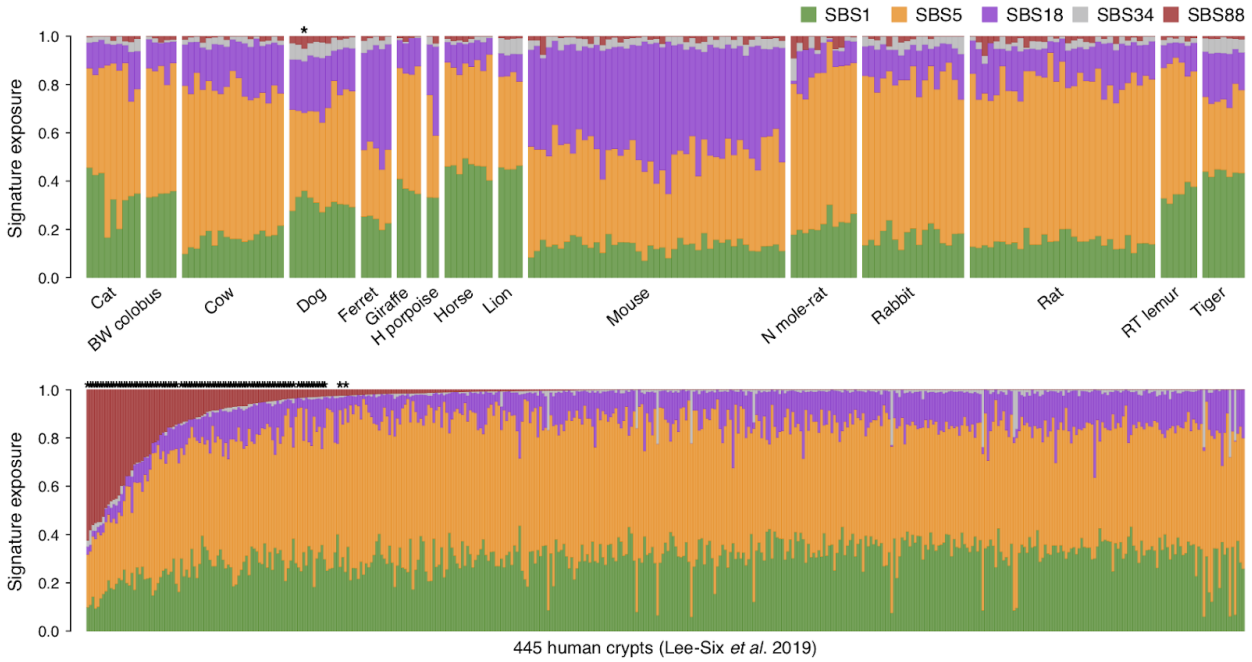
**Extended Data Figure 6. Profiles of signature SBSB as inferred from different species.** Trinucleotide-context mutational spectra of signature SBSB, as inferred independently from somatic mutations in crypts from four representative species (top to bottom): human, naked mole-rat, rat, and rabbit (Methods). 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 in each species.** Horizontal axis presents 83 insertion/deletion types, coloured by type and length[36].

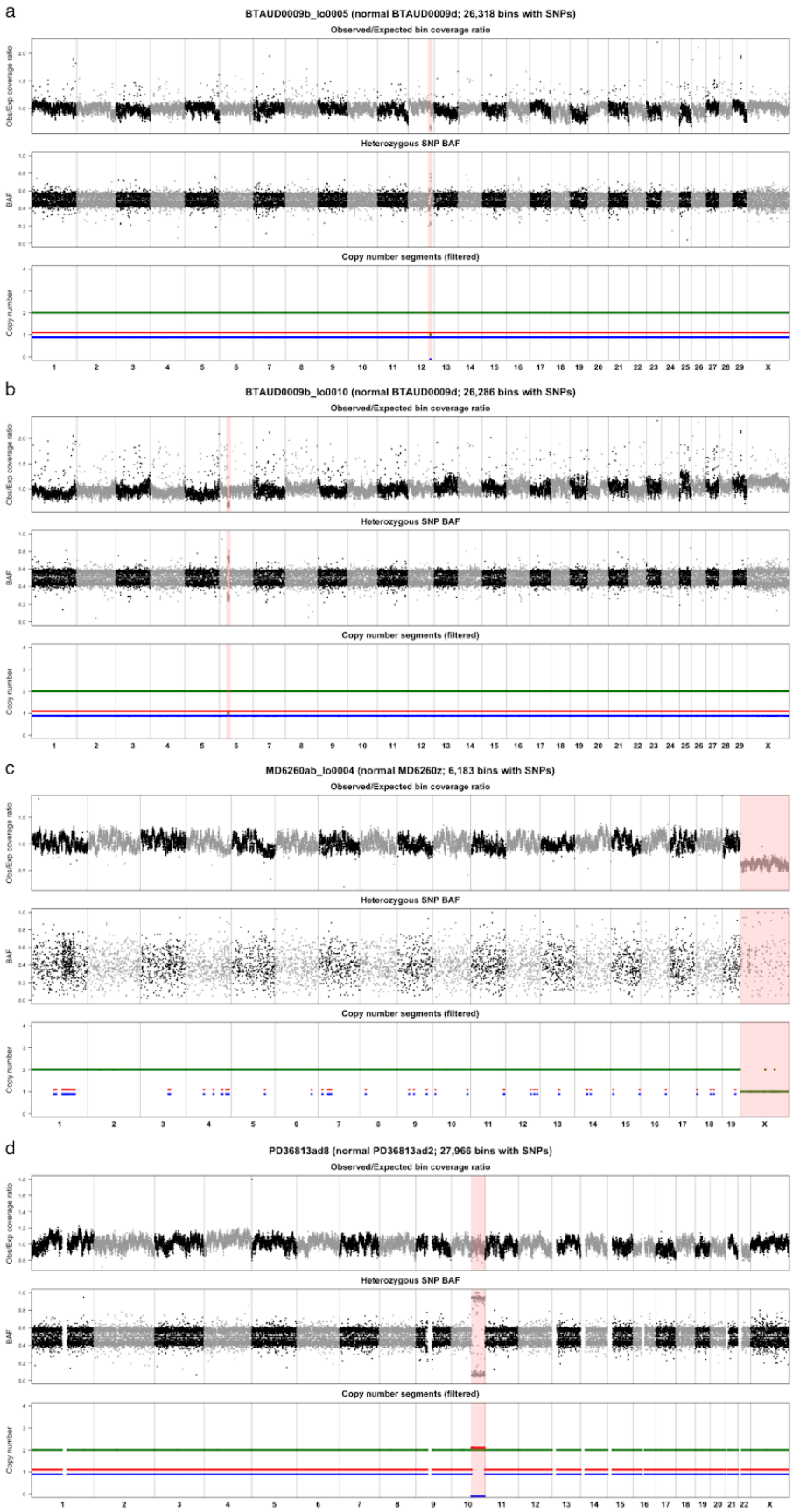


716 **Extended Data Figure 8. Colibactin exposure in non-human and human colorectal crypts.**
717 Exposures to mutational signatures SBS1, SBS5, SBS18, SBS34 and SBS88, as inferred via
718 expectation–maximisation, for 180 non-human crypts (top) and 445 human crypts sequenced in a
719 previous study[8]. Asterisks indicate samples with statistically significant colibactin (SBS88)
720 exposure, based on a likelihood ratio test (Methods). BW, black-and-white; H, harbour; N, naked;
721 RT, ring-tailed.
722



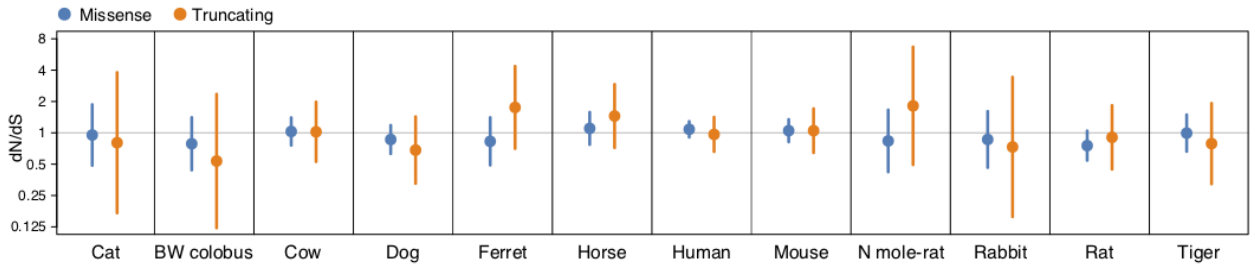

723
724
725

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


a
BTAUD0009b_lo0005 (normal BTAUD0009d; 26,318 bins with SNPs)
Observed/Expected bin coverage ratio
Obs/Exp coverage ratio
Heterozygous SNP BAF
BAF
Copy number segments (filtered)
Copy number
b
BTAUD0009b_lo0010 (normal BTAUD0009d; 26,286 bins with SNPs)
Observed/Expected bin coverage ratio
Obs/Exp coverage ratio
Heterozygous SNP BAF
BAF
Copy number segments (filtered)
Copy number
c
MD6260ab_lo0004 (normal MD6260z; 6,183 bins with SNPs)
Observed/Expected bin coverage ratio
Obs/Exp coverage ratio
Heterozygous SNP BAF
BAF
Copy number segments (filtered)
Copy number
d
PD36813ad8 (normal PD36813ad2; 27,966 bins with SNPs)
Observed/Expected bin coverage ratio
Obs/Exp coverage ratio
Heterozygous SNP BAF
BAF
Copy number segments (filtered)
Copy number

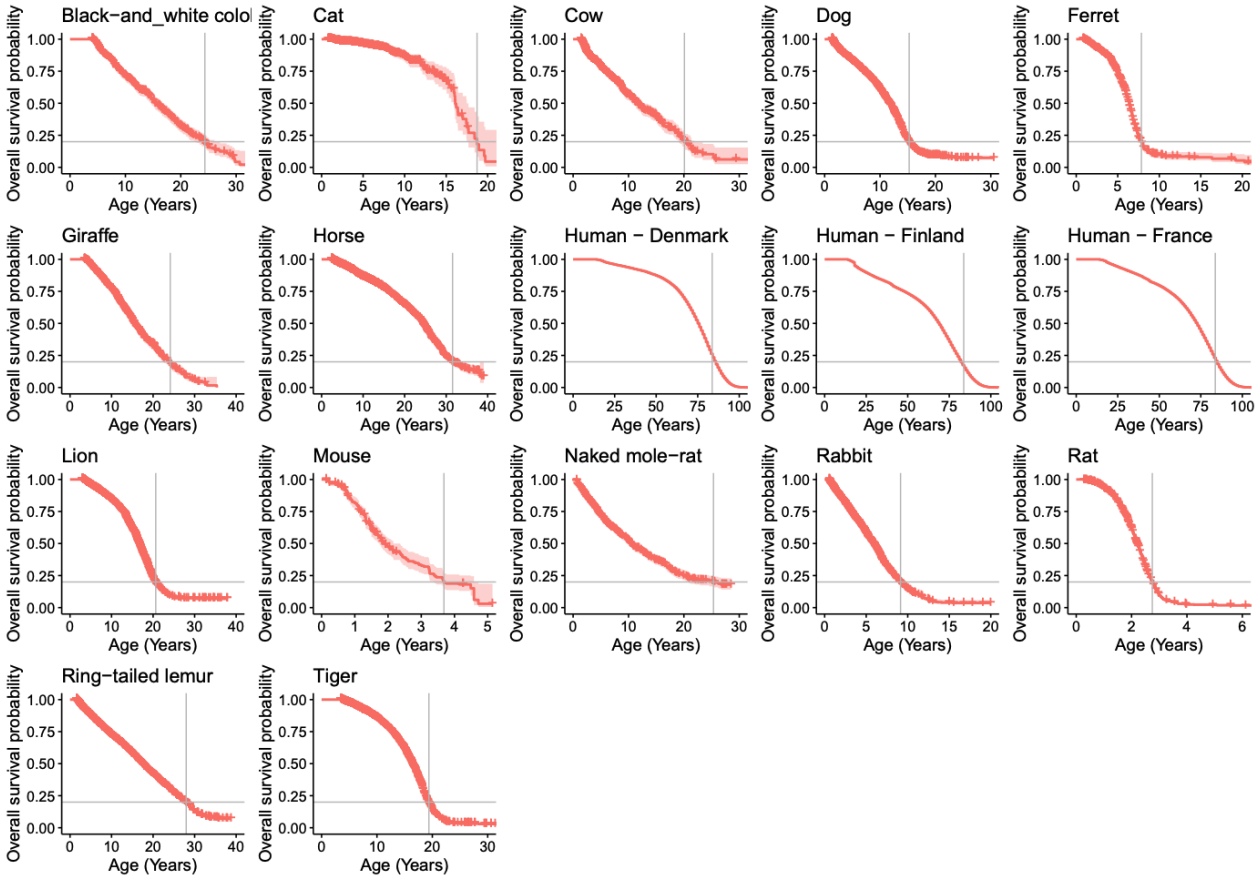
735

736 **Extended Data Figure 10. Somatic dN/dS.** Estimates of dN/dS for somatic missense and
737 truncating mutation in each of the species with available genome annotation. Dots and error bars
738 represent maximum likelihood estimates and 95% confidence intervals, respectively. Note the
739 logarithmic scale of the vertical axis.
740

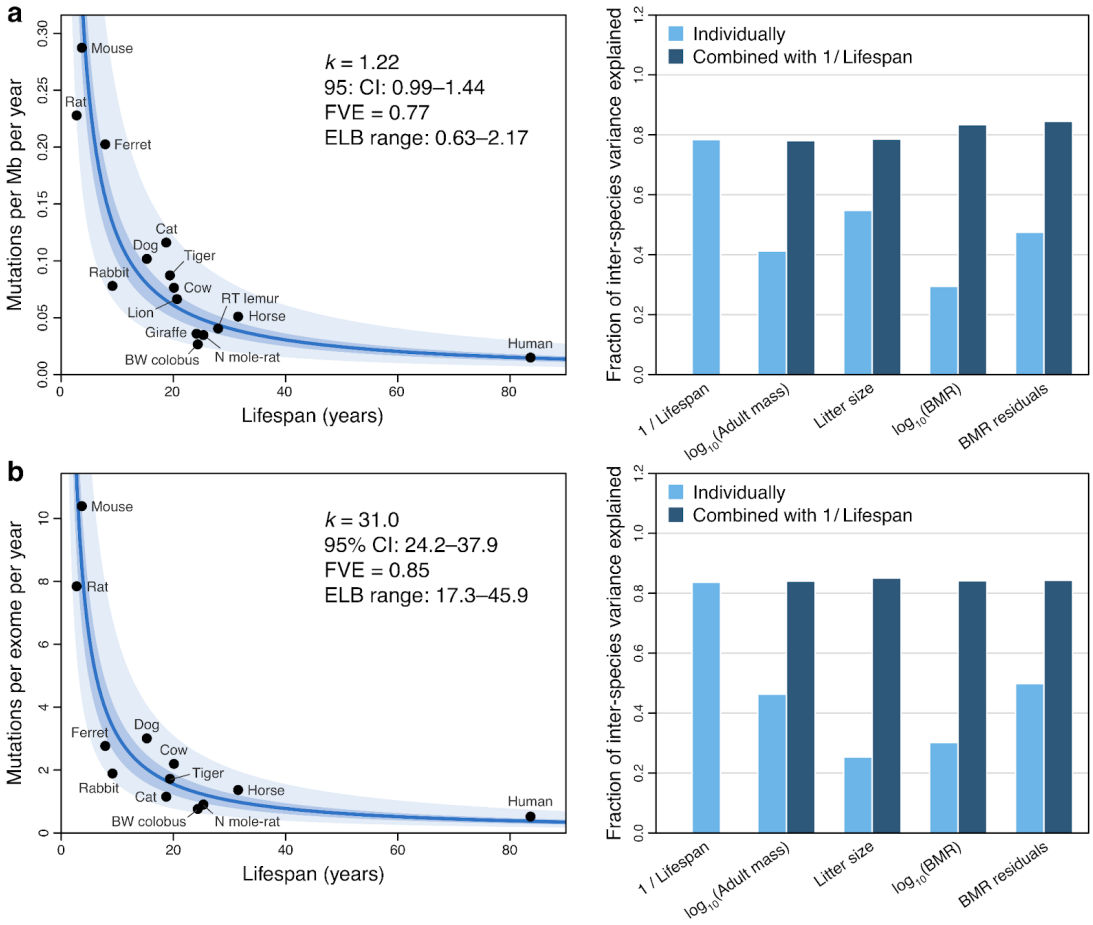



741
742
743
744

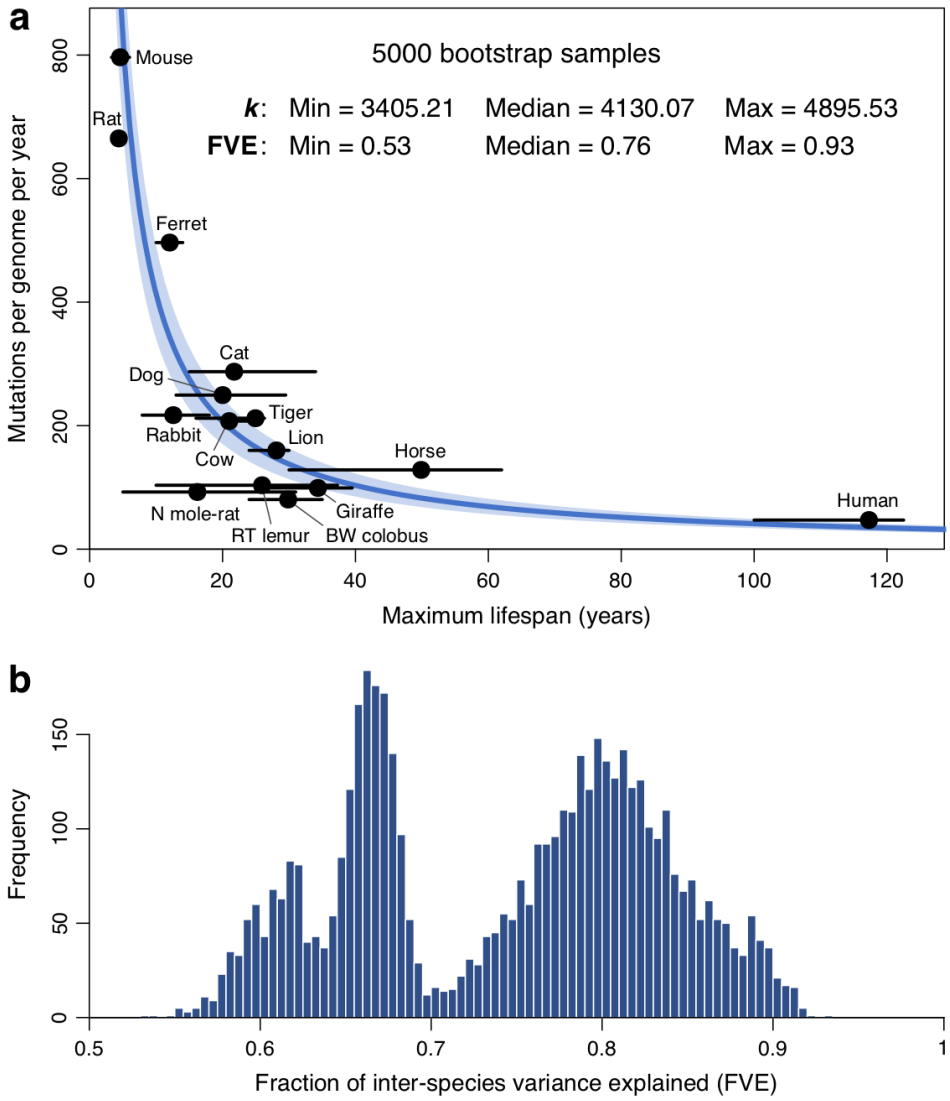
**Extended Data Figure 11. Kaplan-Meier curves of longevity in captivity.** Kaplan Meier survival curves for each species using captive lifespan data from Species360 for non-human species and census record data for humans (**Methods**). The red shaded areas represent 95% confidence intervals of the curve. A horizontal 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. 3d,e**, 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.

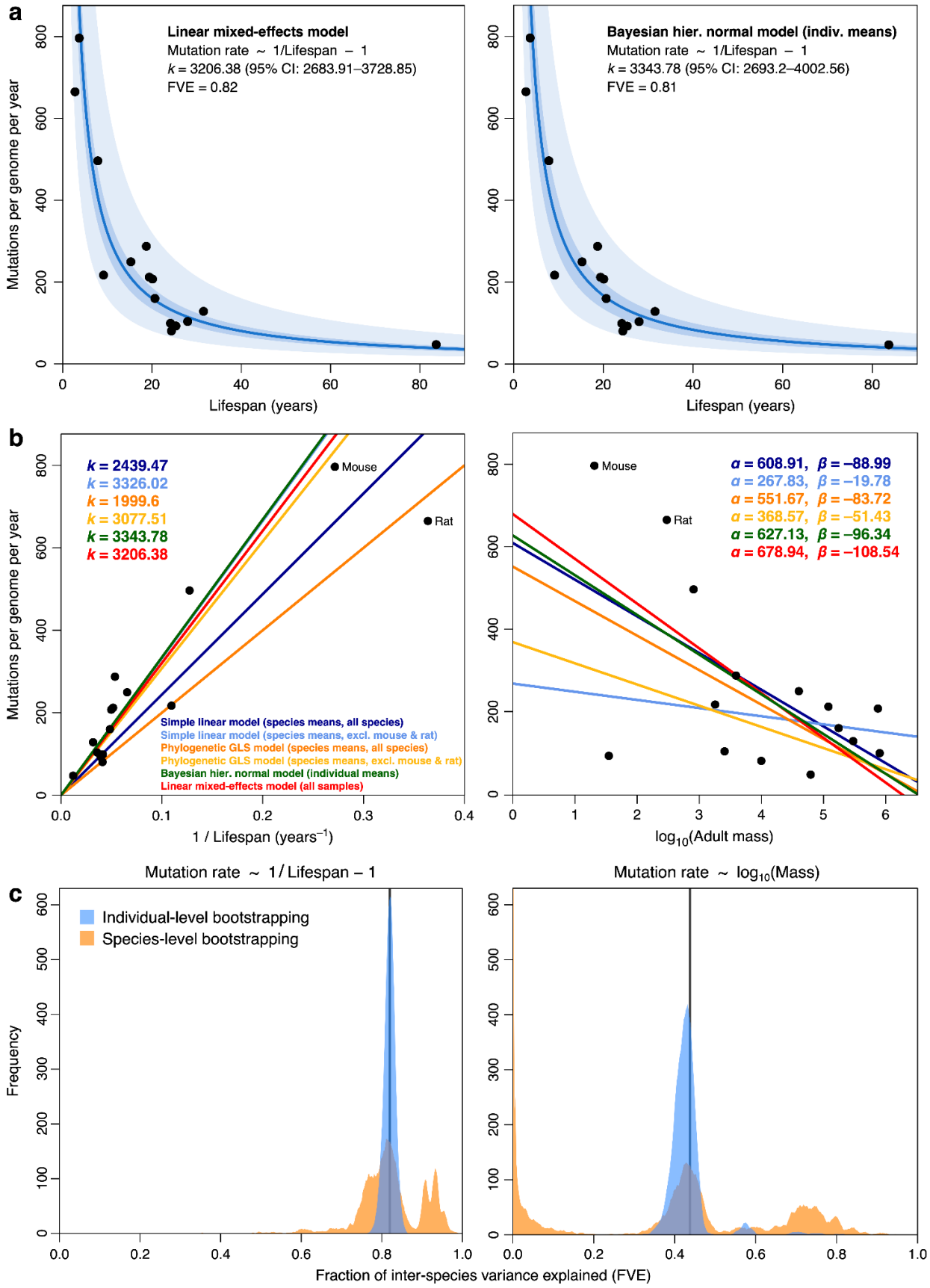


769 **Extended Data Figure 13. Bootstrapped regression of somatic mutation rates on published**
770 **lifespan estimates. a**, Bootstrapped regression of somatic substitution rates on the inverse of
771 lifespan (1/Lifespan), using a zero-intercept linear mixed-effects model (**Methods**). For each of
772 5000 bootstrap samples, lifespan values per species were randomly selected from a set of
773 published maximum longevity estimates (**Supplementary Table 6**). The blue line indicates the
774 median regression slope (*k*) across bootstrap samples, and the blue shaded area depicts the range
775 of estimates of *k* across bootstrap samples. Black dots and error bars indicate the mean and range,
776 respectively, of published longevity estimates for each species. The median and range of both *k*
777 and the fraction of inter-species variance explained (FVE) are provided. **b**, Histogram of FVE
778 values across the 5000 bootstrap samples.
779



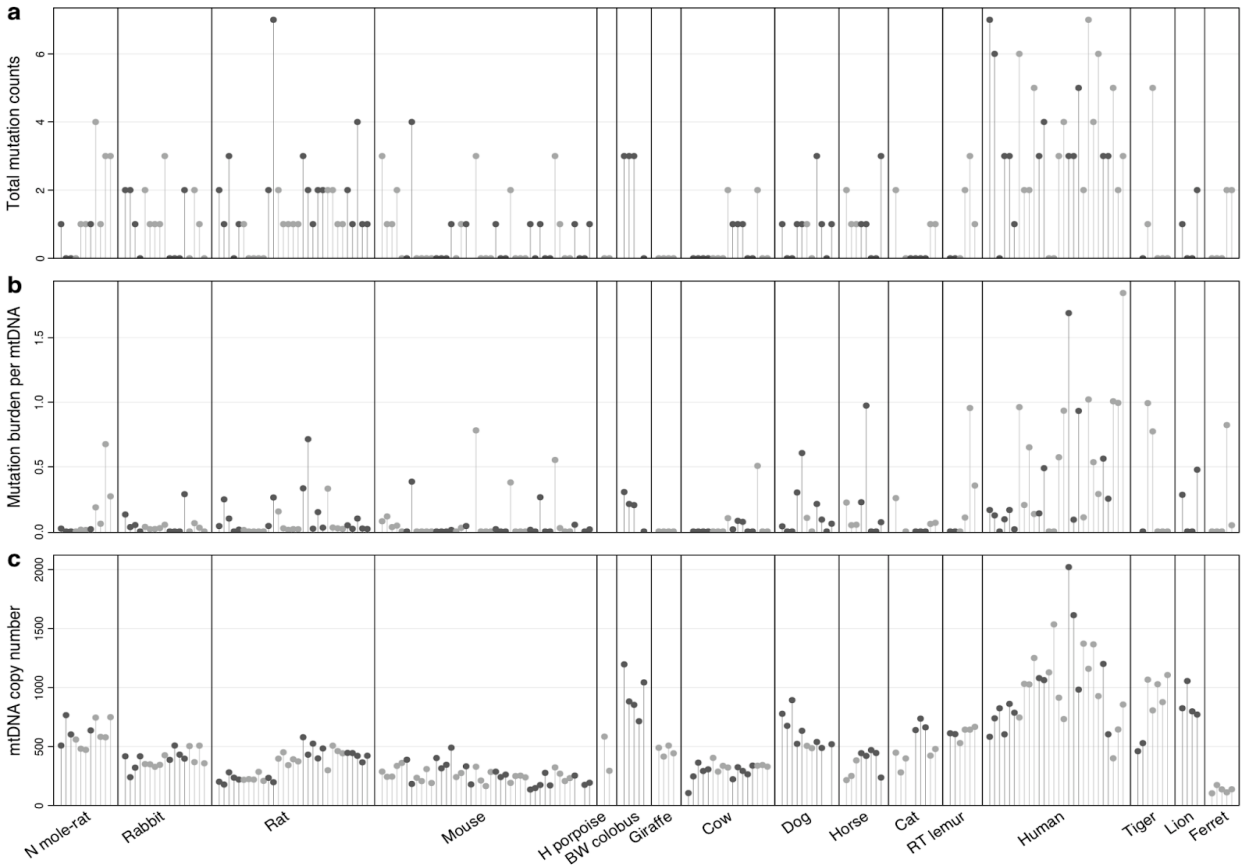

780

**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 (LME) 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 LME models (red). Point estimates of the regression coefficients for each model are provided. **c**, Distributions of regression FVE under individual- and species-level bootstrapping. For the LME models regressing somatic mutation rates on inverse lifespan (zero intercept; left) and log-mass (free intercept), the curves present distributions of FVE from 10,000 bootstrap replicates, obtained through random resampling of either individuals (blue) or species (orange) (**Methods**). Vertical lines indicate the FVE values obtained using the entire data set.

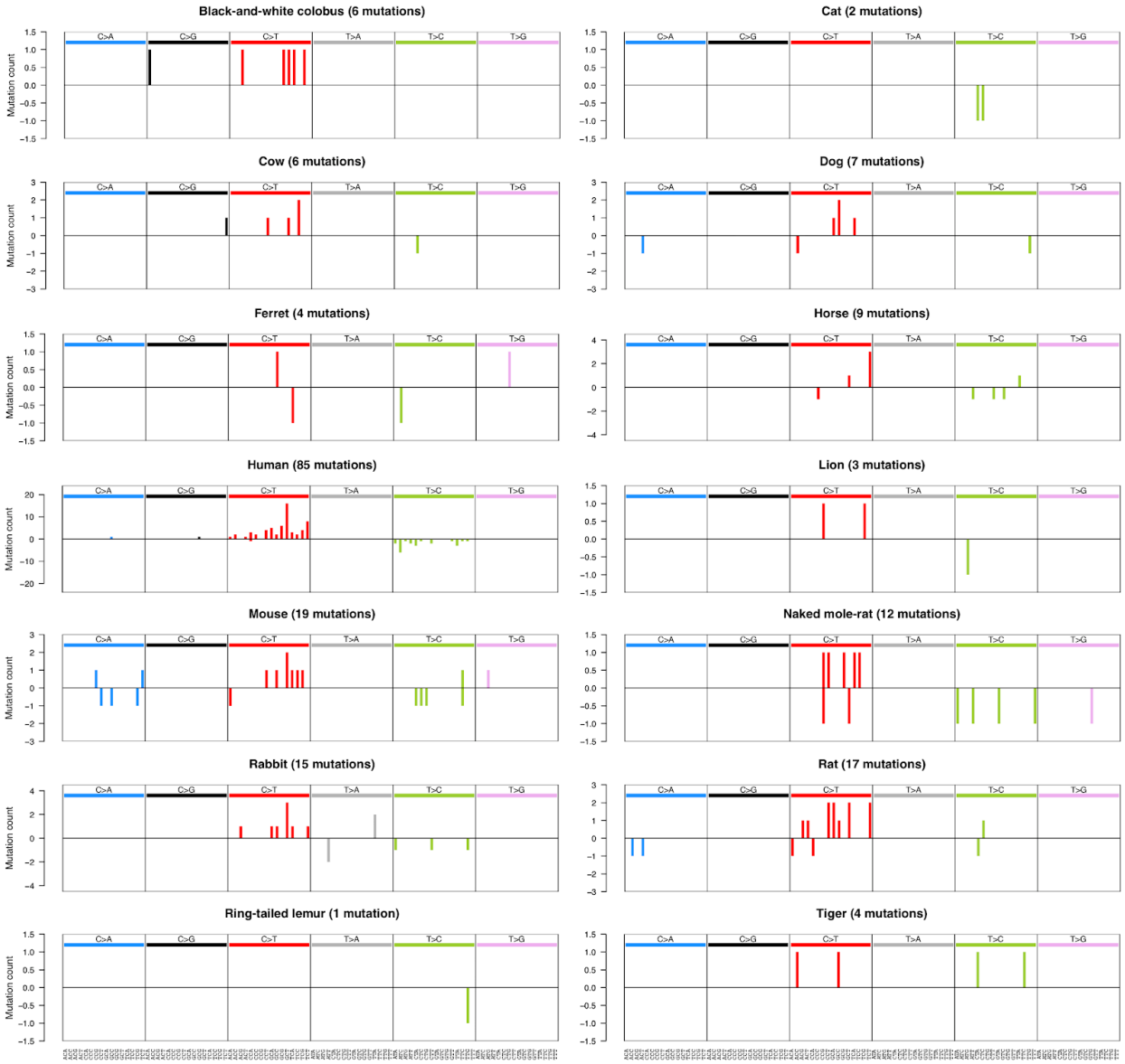

a
Linear mixed-effects model
Mutation rate ~ 1/Lifespan – 1
k = 3206.38 (95% CI: 2683.91–3728.85)
FVE = 0.82
Bayesian hier. normal model (indiv. means)
Mutation rate ~ 1/Lifespan – 1
k = 3343.78 (95% CI: 2693.2–4002.56)
FVE = 0.81
Mutations per genome per year
Lifespan (years)
b
k = 2439.47
k = 3326.02
k = 1999.6
k = 3077.51
k = 3343.78
k = 3206.38
Mouse
Rat
Simple linear model (species means, all species)
Simple linear model (species means, excl. mouse & rat)
Phylogenetic GLS model (species means, all species)
Phylogenetic GLS model (species means, excl. mouse & rat)
Bayesian hier. normal model (individual means)
Linear mixed-effects model (all samples)
1 / Lifespan (years$^{-1}$)
α = 608.91, β = –88.99
α = 267.83, β = –19.78
α = 551.67, β = –83.72
α = 368.57, β = –51.43
α = 627.13, β = –96.34
α = 678.94, β = –108.54
log$_{10}$(Adult mass)
Mutation rate ~ 1 / Lifespan – 1
Mutation rate ~ log$_{10}$(Mass)
c
Individual-level bootstrapping
Species-level bootstrapping
Frequency
Fraction of inter-species variance explained (FVE)

799

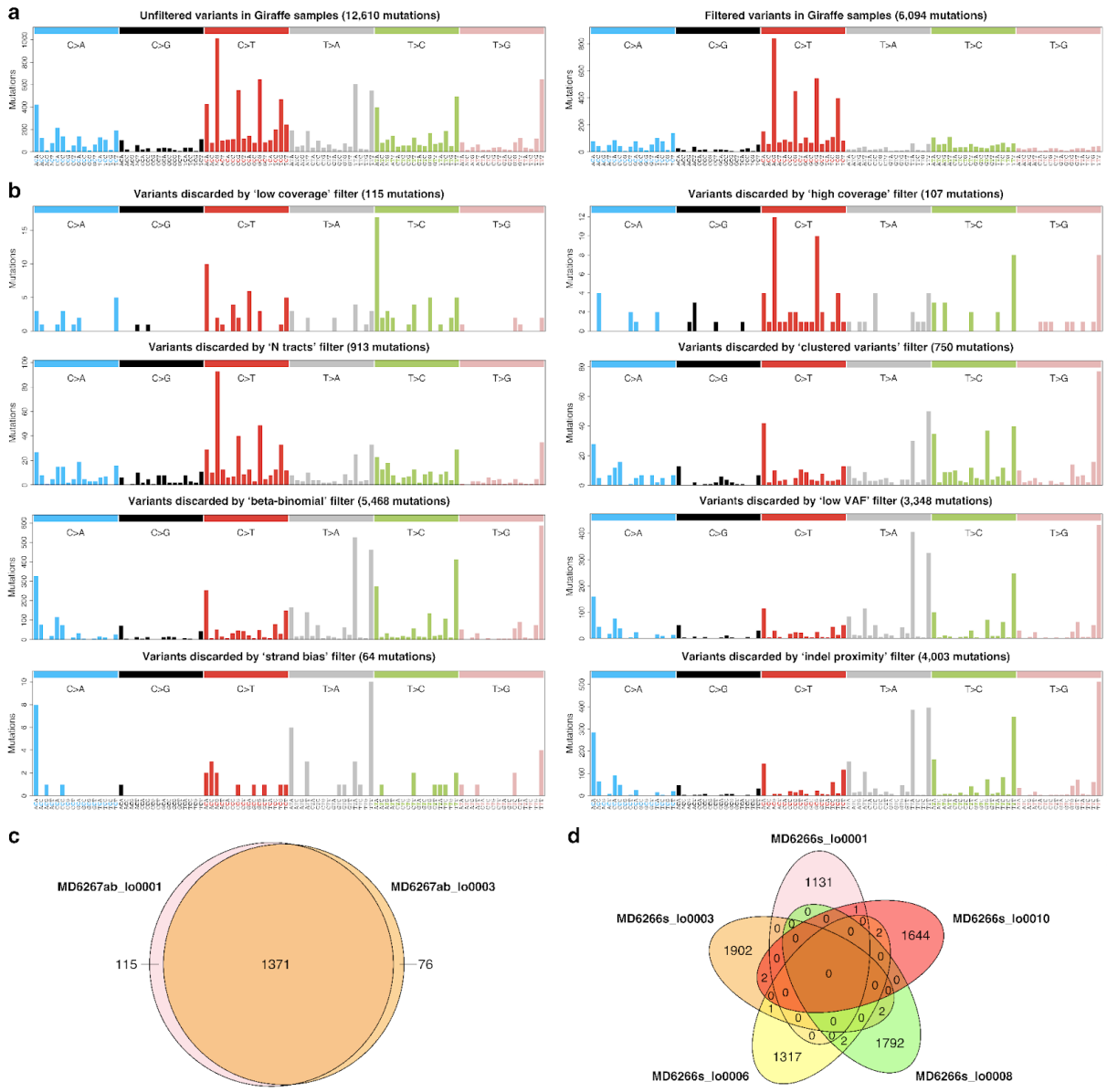
**Extended Data Figure 15. mtDNA mutation burden and copy number. a**, Total somatic mtDNA mutations (substitutions and indels) called in each sample. **b**, Somatic mutation burden per mitochondrial genome copy per sample. **c**, Estimated mtDNA copy number per sample. Samples are arranged as in **Fig.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



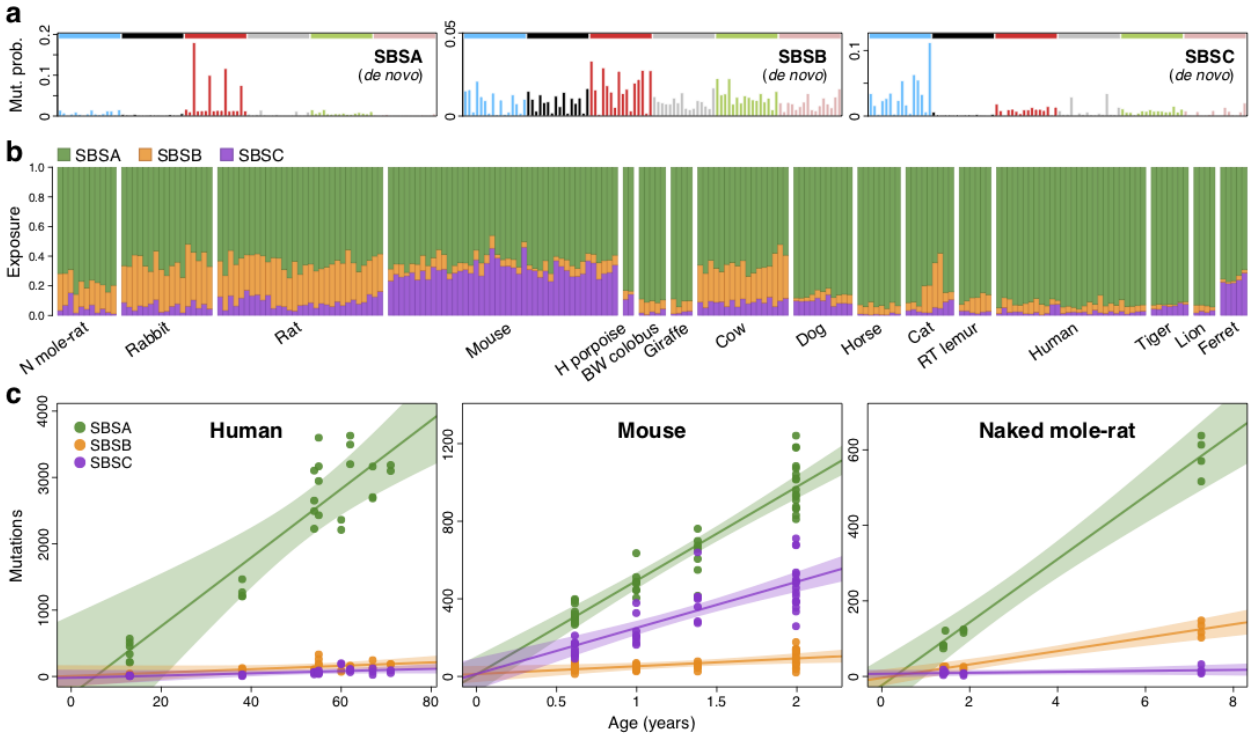
**Extended Data Figure 16. Mutational spectra of mtDNA substitutions in each species.** Horizontal axis presents 96 mutation types on a trinucleotide context, coloured by base substitution type. Mutations on the upper and lower halves of the spectrum represent substitutions in which the pyrimidine base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



843 **Supplementary Tables**

844

845 **Supplementary Table 1. Species information.** For each of the (sub)species in the study, the
846 table provides: common name, scientific name, number of individuals in the study, number of
847 colorectal crypts sequenced, range of individual ages, and source institution.

848

849 **Supplementary Table 2. Sample information.** For each colorectal crypt sample in the study,
850 the table provides: sample ID, individual ID, species name, matched normal sample ID, matched
851 normal sample type, and median sequencing depth.

852

853 **Supplementary Table 3. Mutation rate and burden regression coefficients per species.** For
854 each species in the study (except harbour porpoise), the table provides the mean observed values
855 of the rate of somatic substitutions per genome per year, and point estimates and 95% confidence
856 intervals for simple linear regression of mean substitution burdens per individual on individual
857 ages. The estimated regression slopes correspond to the estimated mutation rate per year for each
858 species. Estimates are provided for constrained-intercept linear models applied to all species, and
859 for free-intercept linear models applied to the eight species with at least three individuals.

860

861 **Supplementary Table 4. Somatic mutation burdens, rates and signature exposures.** For each
862 colorectal crypt sample in the study, the table provides: sample ID, individual ID, species name,
863 individual age, total genome size, coding genome size, analysable genome size, analysable
864 mtDNA size, mutational signature exposures (SBS1, SBSB, SBSC); somatic mutation burdens
865 per genome for single-base substitutions, indels, signature-specific substitutions (SBS1, SBSB,
866 SBSC), and mtDNA mutations; and somatic mutation rates per genome per year for single-base
867 substitutions, indels, signature-specific substitutions (SBS1, SBSB, SBSC), and mtDNA
868 mutations.

869

870 **Supplementary Table 5. Reference genome information.** For each species in the study, the
871 table provides: reference genome version used, reference mtDNA sequence used, Ensembl
872 genome annotation version used (where applicable), reference genome file source, reference
873 mtDNA file source, reference genome file URL, reference mtDNA file URL.

874

875 **Supplementary Table 6. Life history data.** For each species in the study (except harbour
876 porpoise), the table provides: adult mass (g), basal metabolic rate (W), litter/clutch size,
877 maximum longevity (years); and maximum likelihood estimate and 95% confidence limits for the
878 estimated 80% lifespan, together with the corresponding sample size. The source of each estimate
879 is given in brackets.
880
881 **Supplementary Table 7. Number of cell divisions per lifespan.** For mouse, rat and human, the
882 table provides: estimated rate of colorectal cell division (hours), estimated lifespan (years),
883 estimated number of cell division at the end of lifespan, estimated number of mutations per cell
884 division (obtained using our mutation rate estimates), and the reference for the cell division rate.
885 For humans, two estimates of cell division rate are included.
886
887