Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis

Yiguan Wang[1*], Darren J. Obbard[1]

1. Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK
- Corresponding author: Yiguan.Wang@ed.ac.uk

Abstract

Mutation is the ultimate source of all genetic variation, and over the last ten years the easy availability of whole-genome sequencing has permitted direct estimation of the mutation rate for many non-model species from across the tree of life. In this meta-analysis we make a comprehensive search of the literature for mutation rate estimates in eukaryotes, identifying 138 mutation accumulation (MA) and parent-offspring (PO) sequencing studies covering 70 species. Based on these recently accumulated data, we revisit the mutation rate differences between different phylogenetic lineages and update the known relationships between mutation rate and generation time, population size, and genome size. We do this for single nucleotide mutations (SNM) in all species, and for short indels in the 29 species for which data are available. We do not find a significant difference between MA and PO in estimated mutation rates. However, we confirm that mammal and plant lineages have higher mutation rates than arthropods, and that unicellular eukaryotes have the lowest mutation rates. Generation time and genome size are each positively correlated with mutation rate, even when accounting for phylogenetic relationships. Meanwhile, indel mutation rates are also positively correlated with SNM rates and short deletions are generally more common than short insertions across all the species. However, we note that no estimates of either SNM or indel mutation rates are available for the majority of deeply-branching eukaryotic lineages—or even for most animal phyla. Even among charismatic megafauna, experimental mutation rate estimates remain unknown for reptiles and amphibia and scarce for birds and fish.

Keywords: *de novo* mutation rate, eukaryotes, indels, phylogeny

Abbreviations: MA, mutation accumulation experiment; PO, parent-offspring experiment; SNMs, single nucleotide mutations; indels, insertions and deletion mutations; PGLMM, phylogenetic generalised linear mixed model; Mb, Mega base pair(s)

Introduction

The *de novo* mutation rate, μ, is a key parameter in population and evolutionary genetics, appearing in almost every equation, either alone or combined with effective population size (N_e) in the compound parameter θ. However, it seems implicit that μ is of no interest to many evolutionary studies; researchers often use μ from one species when analysing another, simply for convenience—as if μ were exchangeable in a way that N_e is not (e.g. Wilding 2017). This is presumably because variation in μ among species is taken to be negligible compared to variation in parameters such as N_e. But, if the mutation rate varied among multicellular eukaryotes as much as the recombination rate varies—around 1000-fold (Stapley, et al. 2017)—there would be much less need to invoke variation in N_e in (e.g.) explaining variation in genetic diversity. One reason that we might expect mutation rate to vary little in comparison with recombination rate is if recombination is optimised differently by different selective regimes (Stapley, et al. 2017), but mutation is minimised to a lower limit imposed by drift (the drift barrier hypothesis; Lynch, et al. 2016). However, it might be that mutation rate is also optimised to well above the drift-barrier limit by selection, in favour of adaptability (Peck and Lauring 2018; Liu and Zhang 2021). If this were the case, then (as with recombination rates) we might find that mutation rates varied dramatically among even closely related taxa.

Early estimates of the mutation rate leveraged visible phenotypes, such as the occurrence of genetic diseases in humans (Trimble and Doughty 1974; Nute and Stamatoyannopoulos 1984). For example, using haemoglobin M disease Stamatoyannopoulos and Nute (1982) estimated the mutation rate in the human haemoglobin beta gene at 7.4×10^{-9} per bp per generation. However, these estimates are prone to bias, e.g. if the mutation is required to be autosomal dominant and non-lethal (Nute and Stamatoyannopoulos 1984), and such approaches are not easily applicable in organisms for which disease is hard to observe. More commonly, using the equivalence between neutral mutation and substitution rates (Kimura 1968), many studies have used time-calibrated sequence divergence to estimate mutation rates. For example, comparison between humans and chimpanzees give rise to estimates on the order of 10^{-8} per bp per generation (Kondrashov and Crow 1993; Drake, et al. 1998; Nachman and Crowell 2000). However, such indirect estimates are also limited (Kondrashov 2003). First, because generation time and calibration dates are prone to substantial uncertainty (e.g. Obbard, et al. 2012). Second, because it is hard to distinguish between sites that are unconstrained versus sites that are weakly constrained (Harmon, et al. 2021). Third, because mutation saturation in hotspots may lead to an underestimation, especially for sequences with a long divergence time

(Sigurðardóttir, et al. 2000). In addition, phylogenetic-calibration approaches reflect the long-term mutation rate, averaging across biologically important factors such as generation time, sex differences, life-stage differences, and inter-individual variation.

In contrast to phenotypic and phylogenetic approaches, mutation-accumulation (MA) and parent-offspring (PO) sequencing can provide direct estimates of the per-generation mutation rate. Mutation accumulation utilises the accumulation of spontaneous mutations in a single inbred genome over multiple generations, by minimising the effectiveness of natural selection (Keightley, et al. 2009). Given the low mutation rates in eukaryotes, the multiple generations in an MA experiment allow sufficient mutations to be easily detected at the end. Although the first use of MA to estimate mutation rates phenotypically dates back to the 1970s (Mukai and Cockerham 1977), the availability of whole genome sequencing and its declining price have recently made it more widely applicable. Mutation accumulation best suits easily maintained species with short generation times, for example nematodes (Denver, et al. 2004; Meier, et al. 2014; Konrad, et al. 2019; Saxena, et al. 2019), fruit flies (Haag-Liautard, et al. 2007; Keightley, et al. 2009; Schrider, et al. 2013; Huang, et al. 2016), green algae (Ness, et al. 2012; Lopez-Cortegano, et al. 2021), and water fleas (Ho, et al. 2020; Keith, et al. 2021). However, MA may underestimate mutation rates (Lynch and Walsh 1998) because natural selection will still come into play for highly deleterious or lethal mutations (Baer, et al. 2007).

By sequencing parents and offspring, and counting the mutations arising over a single generation, the parent-offspring approach avoids losing lethal mutations (unless they are dominant) and provides the most direct estimate of mutation rate. It is not limited to model species (Keightley, et al. 2014; Yang, et al. 2015; Wang, et al. 2022b), and is suitable for large animals outside of the laboratory (Harland, et al. 2017; Wang, et al. 2022a). Parent-offspring sequencing has been particularly widely used in humans, where MA is not applicable. However, the challenge in robustly identifying new mutations in natural heterozygous genomes is formidable, as the number of mutations arising within one generation is small when compared to the impact of sequencing and mapping errors. Minimising the rates of false positive and false negative mutation-calls is therefore critical for the PO approach (Yoder and Tiley 2021). For example, Bergeron, et al. (2022) compared the estimates from five different research groups for the same family of rhesus macaques and found that the highest estimates could be twice the lowest ones.

Relaxed-clock phylogenetic studies generally identify substantial among-lineage variation in the evolutionary rate, suggestive of variation in mutation rate among taxa (Ho, et al. 2015). However, such studies cannot separate the impact of generation time and mutation rate, and may be biased by variation in the action of (weak) selection (Harmon, et al. 2021). As more direct experimental

121 estimates have become available, it has become possible to draw comparisons within particular
122 clades such as primates (Chintalapati and Moorjani 2020), or vertebrates (Yoder and Tiley 2021;
123 Bergeron, et al. 2022). However, there has not been a wider analysis, nor one that includes both PO
124 and MA studies. Here we perform a comprehensive meta-analysis of published eukaryotic SNM and
125 indel mutation-rate estimates that utilise direct sequencing of multiple generations, either through
126 MA or PO comparison. We focus on the variation in mutation rate among species, and the role of the
127 host phylogeny and experimental approach in determining variation in rate estimates.

128 **Results and Discussion**

129 ***Studies included***

130 We searched for eukaryotic mutation-rate studies in Clarivate Web of Science, and then we
131 manually examined these studies and identified those containing mutation rate estimates (see
132 Methods). In total, we identified 138 studies using either MA or PO approaches for inclusion,
133 covering 70 species (see Availability of data and materials). Studies on human mutation rate
134 accounted for a third of the identified literature, followed by unicellular organisms (20%),
135 arthropods (17%), primates (8%), and plants (8%) (**Figure 1A)**. However, across all groups, the
136 majority of studies focused on model species; yeast in the unicellular group, *Drosophila* in
137 arthropods, mice in mammals, and *Arabidopsis* in plants. These estimates tended to derive from MA
138 analyses rather than PO studies (61 vs. 37 among non-human studies). Nevertheless, as sequencing
139 has become cheaper, there has been a rapid increase in the number of published experimental
140 estimates of the mutation rate, with an average of 10 studies each year in eukaryotic species since
141 2010 (**Figure 1B**).

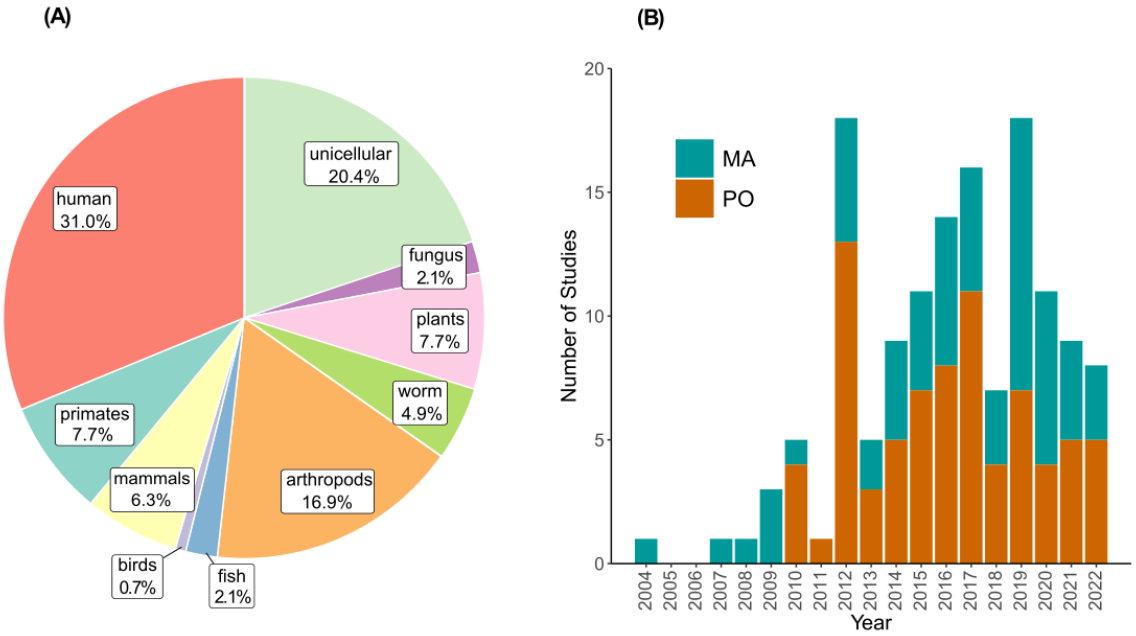


Figure 1: A summary of *de novo* mutation rate estimates included here. (A) The proportion of studies in different taxonomic groups. (B) The number of studies by year. MA: mutation accumulation studies, PO: parent-offspring studies.

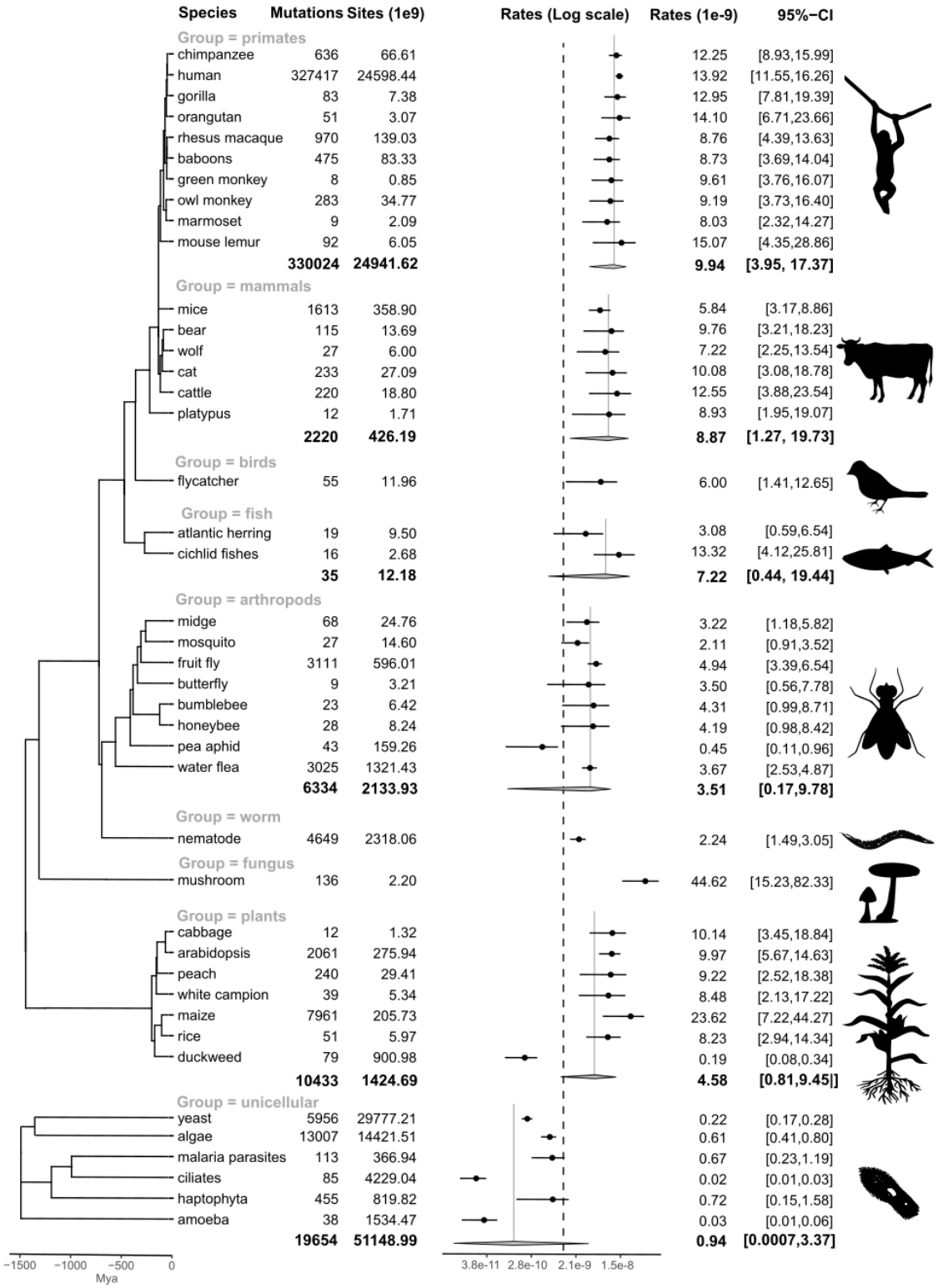
Single nucleotide mutation rates across eukaryotes

To summarise the single nucleotide mutation (SNM) rates in different lineages, we obtained a time-scaled phylogenetic tree of all represented species from http://www.timetree.org/ (Kumar, et al. 2022), and used this to fit the covariance among species in a Phylogenetic Generalised Linear Mixed Model (PGLMM) using the Bayesian mixed model package MCMCglmm (e.g. Hadfield and Nakagawa 2010). Treating species as a random effect, this provides a posterior estimate of mutation rate for each species (and each internal node) under a Brownian motion model of trait evolution (see Methods), given the observations and relationships to other species. In these data, the phylogenetic effect accounts for just 2.10% of the random-effects variance, with 95% HPD interval (0.93%, 3.44%).

Among the chosen species groupings, the unicellular species had the lowest mutation rate ranging from 0.02×10^{-9} per generation per base pair in ciliates to 0.72×10^{-9} in haptophyta. This estimate is more than an order of magnitude lower than rates in multicellular species (below; **Figure 2**), and may either be attributable to there being a single cell division per generation, or possibly to a lower limit set by the drift barrier in species with larger N_e (Sung, et al. 2012a). We also observed relatively

low SNM rates for a nematode at 2.24×10^{-9}, arthropods at 3.51×10^{-9}, fish at 7.22×10^{-9}, and a bird at 6.00×10^{-9} (random-effects estimates from our model, given the phylogeny). In particular, for arthropods (which were represented by multiple species) the lowest mutation rate was identified as 0.45×10^{-9} in pea aphids (per parthenogenic generation) and the highest one was only 4.94×10^{-9} in fruit flies. In contrast to these species, mammals and plants had relatively high mutation rates, including non-primate mammals at 8.87×10^{-9} and primates at 9.94×10^{-9}. The plants (excluding duckweed) and a multicellular fungus had the highest rates of all, ranging from 8.23 to 44.62×10^{-9} (**Figure 2**), and it is tempting to speculate that this might be attributable to the lack of a segregated germline (but see Wang, et al. (2019)).

To assess the impact of the experimental methods on rate estimates, we included the experimental approach (MA or PO) as a fixed effect in the PGLMM. Overall, we found that while the mean value from PO studies was higher than the value from MA studies, when accounting for variation among species and the covariance between them induced by the phylogeny, this effect was not significant (MCMC *p*-value = 0.646). We then examined comparisons between MA and PO in fruit flies and mice more closely, as these two species have been investigated using both methods. The rate estimated using MA in fruit flies was 5.28×10^{-9}, but 3.20×10^{-9} using PO (**supplementary Figure S1**), and in mice the MA estimate was 5.40×10^{-9} while the PO estimate was 4.03×10^{-9} (**supplementary Figure S2**), and these differences were likewise not significant. Note, however, that the mutation rate may be heterogeneous among different populations: for *D. melanogaster*, the estimated mutation rate from African populations was lower than that from European population and North American population (Chan, et al. 2012; Wang, et al. 2022c). To make a more robust analysis would require samples of multiple populations for each species, which are not yet available.



183

184 **Figure 2**. A phylogenetic meta-analysis of SNM rates for different species identified from the
185 literature. Mutation rates and 95% CIs were estimated using a PGLMM. For each group, we
186 estimated the rate for their most recent common ancestor (shown in bold under each group and

187 vertical solid grey lines). The vertical dashed line indicates the mutation rate of common ancestor of
188 the tree.

189 ***Factors associated with the SNM rates***

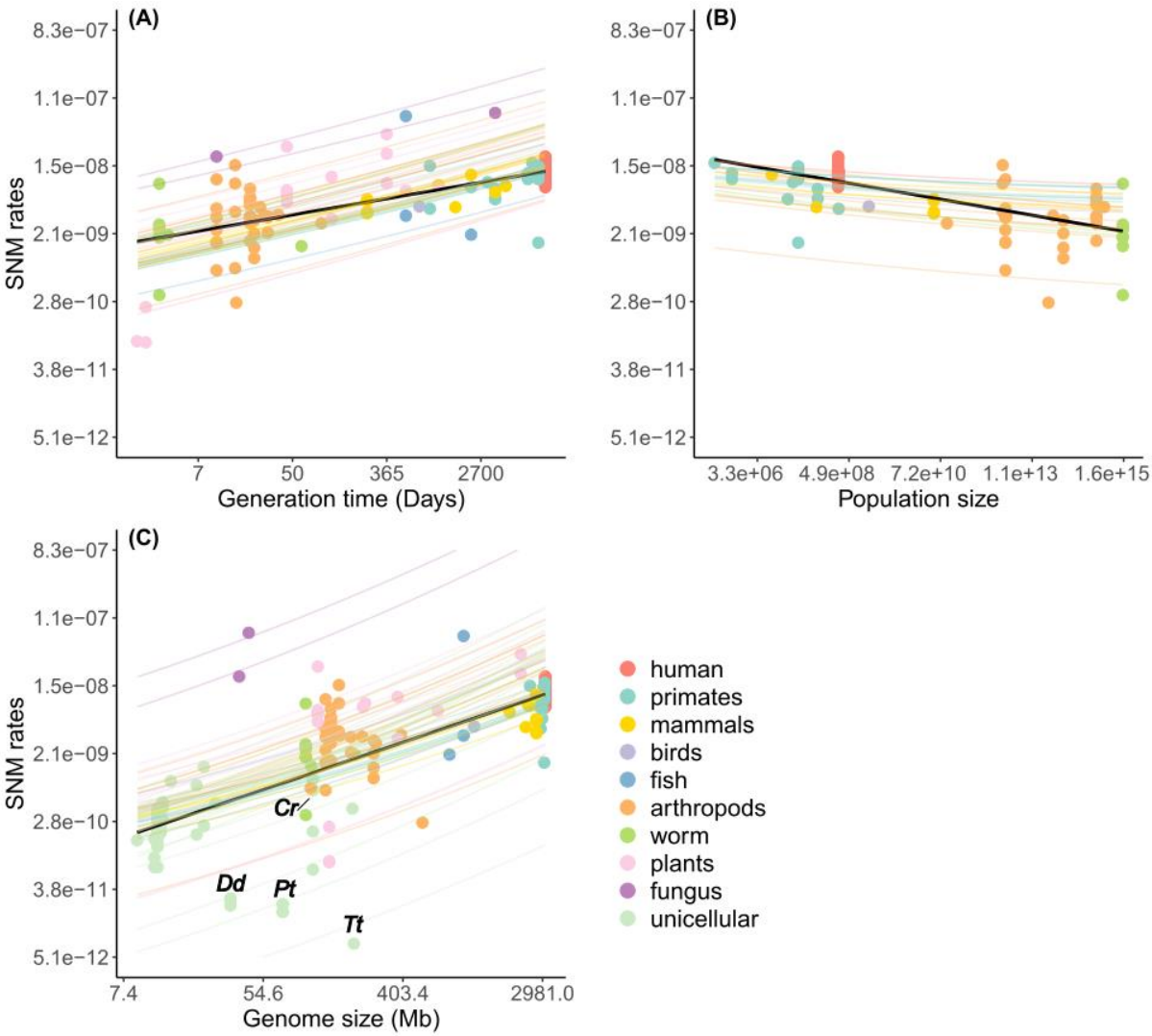
190 Many biological factors have been proposed to be associated with mutation rate, including
191 generation time (Wu and Li 1985; Li, et al. 1987; Bailey, et al. 1991; Keightley and Eyre-Walker 2000),
192 population size (Lanfear, et al. 2014; Krasovec, et al. 2020), and genome size (Drake 1991; Lynch
193 2010). To test for a dependence of mutation rate on each of these factors, we ran three further
194 models, including each of them separately as an additional fixed effect in a PGLMM.

195 We found a significant positive relationship with generation time (*p*MCMC = 0.0004 when
196 accounting for phylogeny; for simple non-phylogenetic linear model *p*-value < 2×10^{-16} **Figure 3A**),
197 such that species with a longer generation time had a higher mutation rate per generation. If
198 mutations resulted only from replication errors, and most species had a similar number of cell
199 divisions per generation, then the per-generation mutation rate would not scale with generation
200 time (Wu and Li 1985; Thomas and Hahn 2014). However, in reality the number of germline cell
201 divisions per generation varies within and between species. In humans, there are around 401 cell
202 divisions per 30-year generation in males and 31 in females (Drost and Lee 1995; Ohno 2019), but in
203 mice it is 62 per 9-month generation in males and 25 in females. In *Drosophila* the sexes are much
204 more similar, with an estimated at 35.5 cell divisions in a 18-day generation for males but 36.5 in an
205 25-day generation in females in *D. melanogaster* (Drost and Lee 1995). Within a species, a long
206 generation time is likely to permit more cell divisions and thus a higher mutation rate per
207 generation, and many studies in humans and primates have shown that mutations are heavily male-
208 biased and correlated with paternal ages (Kong, et al. 2012; Thomas, et al. 2018; Wang, et al. 2020;
209 Kaplanis, et al. 2022). Alternatively, mutations may not result predominantly from replication errors.
210 Recent studies suggest that not all germline mutations track cell divisions, and DNA damage may
211 contribute to the mutation rates (Wu, et al. 2020; Wu and Przeworski 2022). Independent of cell
212 divisions, this mutation rate would be correlated with calendar time, and thus species with a longer
213 generation time would have a higher rate. The general correlation we observe between generation
214 time and per-generation mutation rate could therefore be a consequence of either a greater number
215 of cell divisions or of accumulating damage over time, and it will be hard to separate these potential
216 causes (Gao, et al. 2016; Wu and Przeworski 2022).

217 A negative relationship is visible between per generation mutation rate and census population size
218 (**Figure 3B**)—suggesting that species with more individuals tend to have lower mutation rates—but
219 this was not significant in a phylogenetic mixed model (*p*MCMC = 0.315). It was, however, highly

220 significant under a simple linear model that does not account for phylogenetic non-independence (*p*-
221 value < 2.2×10^{-16}), demonstrating the importance of including species relationships in such models.
222 If census size is a predictor of effective population size (Buffalo 2021), then a significant negative
223 relationship could be consistent with a drift-barrier hypothesis; i.e. that mutation rates are
224 minimised by natural selection until the point at which the impact of selection for lower mutation
225 rate is counterbalanced by allele frequencies changes due to drift (Lynch 2010). And, if that were the
226 case, this could help to lessen the extent of "Lewontin's paradox", that diversity varies less than
227 expected among species, given the range of census population sizes. In that case, our observed (but
228 non-significant) negative correlation between mutation rate μ and N_e could mitigate the
229 unexpectedly small differences in nucleotide diversity across species, since $\theta_\pi = 4N_e\mu$ (Krasovec, et
230 al. 2020; Charlesworth and Jensen 2022).

231



232

233 **Figure 3**. The relationship between SNM rates and (A) generation time, (B) population size, and (C)
234 genome size. Both axes are shown in log scale. The solid black lines represent simple linear
235 regressions; the coloured lines represent different PGLMM regressions conditional on each of the
236 species random effects. Cr: *C. reinhardtii*; Pt: *P. tetraurelia*; Tt: *T. thermophila*; Dd: *D. discoideum*.

237 Finally, we found a positive relationship between genome size and per-generation mutation rate,
238 such that species with larger genomes also had higher mutation rates (*p*MCMC = 0.002; *p*-value for
239 simple linear model < 2×10^{-16} **Figure 3C**). Mutation rate has previously been found to correlate
240 negatively with genome size in prokaryotes (a phenomenon termed "Drake's rule" (Drake 1991;
241 Bourguignon, et al. 2020; Marais, et al. 2020)), but positively in eukaryotes (Lynch 2010). Two
242 'outlier' unicellular eukaryotes with large genomes but very low mutation rates have (*Paramecium*
243 *tetraurelia* and *Chlamydomonas reinhardtii*), have in the past been explained by appeal to their large
244 effective population size and the 'drift-barrier' hypothesis (Sung, et al. 2012a). With the increased
245 availability of mutation rates accrued over the past decade, we find that the mutation rate of *C.*
246 *reinhardtii* is not as low as previously thought, and this species now fits well to the positive trend in
247 eukaryotes (**Figure 3C**). However, the mutation rates of *Tetrahymena thermophila* (Long, et al.
248 2016), *P. tetraurelia*, and *Dictyostelium discoideum* (Kucukyildirim, et al. 2020) are indeed very low.
249 Under 'drift-barrier' hypothesis, these outliers could be caused by the large N_e in these species, and
250 *T. thermophila* should have the highest N_e. In fact, the N_e in *T. thermophila* is smaller than that in *P.*
251 *tetraurelia* (1.1×10^8 *vs.* 1.2×10^8) (Sung, et al. 2012b; Long, et al. 2016), suggesting that N_e seems
252 unable to explain these outliers (Kucukyildirim, et al. 2020).

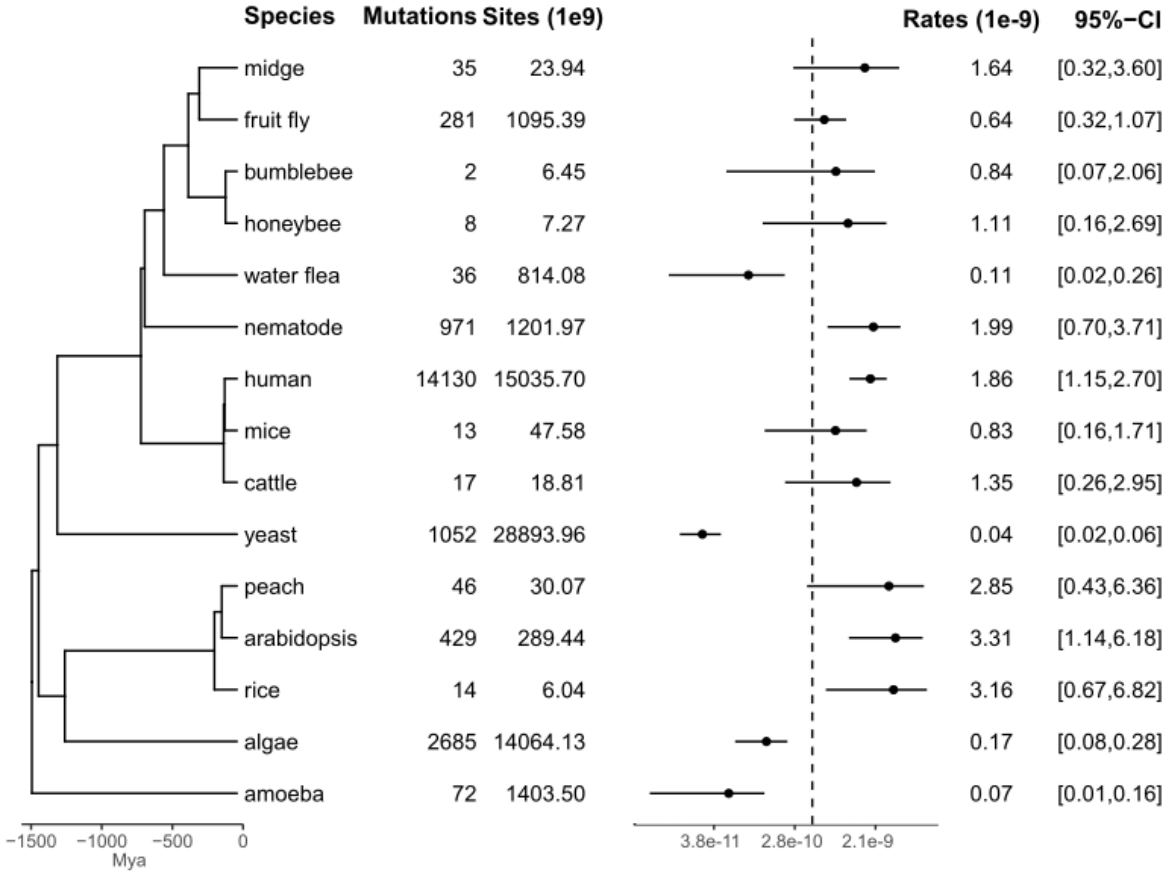
253 In each of the three models above, we fitted generation time, population size and genome size
254 individually as predictors of mutation rate. However, these factors tend to be highly correlated: large
255 organisms tend to have longer generation times, smaller population sizes, and often larger genomes
256 (Martin and Palumbi 1993; Buffalo 2021) and is therefore hard to distinguish between casual
257 relationships and confounding correlations. Indeed, including all three factors (generation time,
258 genome size, population size) as fixed terms in a single model, results in none of them being
259 significant, although generation time is marginal with *p*MCMC=0.08, versus *p*MCMC = 0.354 for
260 genome size and 0.602, for population size.

261

262 ***The mutation rates of short indels across eukaryotes***

263 Although the rate of indel mutation is estimated much less frequently than the SNM rate, our search
264 identified estimates for 29 species. We found that the indel mutation rate across eukaryotes

265 displayed a similar pattern to that of the SNM rate (**Figure 4**), with a Pearson's correlation coefficient
266 between the two of 0.58 (p-value = 8.32×10^{-8}, df = 70; ignoring any phylogenetic effect), with the
267 SNM rate being generally higher than the indel mutation rate, except in two lineages (nematodes
268 and amoebae) for which the indel mutation rate was higher (**Figure 5A**). The unicellular species had
269 the lowest indel mutation rate, from 0.04×10^{-9} in yeast to 0.17×10^{-9} in algae, followed by
270 arthropods with 0.65 (0.02, 1.57) $\times 10^{-9}$. Mammals had a relatively higher indel mutation rate with
271 1.23 (0.34, 2.36) $\times 10^{-9}$, with humans at the higher end of this range at 1.86 (1.15, 2.70) $\times 10^{-9}$. Of all
272 the studied species, plants had the highest indel mutation rate, which was as high as 2.85 (0.34,
273 6.21) $\times 10^{-9}$.



274

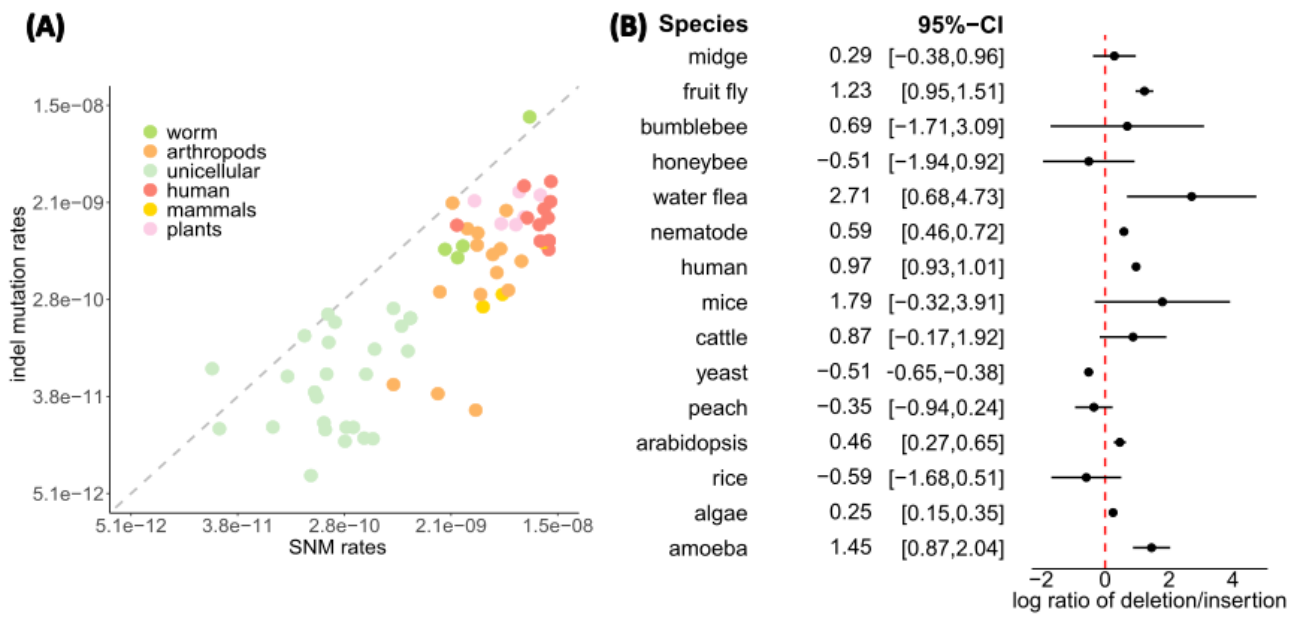
275 **Figure 4**. An analysis of short indel mutation rates for different species. The estimated mutation
276 rates are presented on a log scale for clarity.

277

278 Short deletions generally outnumbered short insertions across eukaryotes (**Figure 5B,**
279 **supplementary Figure S3**). The ratio of deletions over insertions was as high as 15.00 in the water

flea, 4.28 in amoeba, 3.43 in fruit flies, 2.64 in humans, 1.80 in nematode, 1.58 in Arabidopsis and 1.28 in algae. Honeybees and two plants (rice and peach) exhibited more insertions than deletions, but the difference wasn't statistically significant (*p*-value > 0.05, binomial test). Although yeast shows significantly more insertions overall, this estimate is dominated by two large studies and an excess of deletions was still reported in yeast across other studies (**supplementary Figure S3**).

A bias toward deletion events has been known for at least 40 years since de Jong and Rydén (1981) found a four-fold excess of deletions over insertions in protein sequences. Petrov (2002) argued that the deletion bias was likely due to a thermodynamic disparity between short deletions and insertions that makes deletion events easier. Studies in plants have suggested that unequal homologous recombination and illegitimate recombination may be another important cause of small deletions (Bennetzen, et al. 2005). Whatever the cause, the dominance of deletions over insertions seems convincing, and many early studies argued that these short deletions resulted in genome contraction (Petrov 2002; Gregory 2003; Thomas, et al. 2003). However, the role of deletions in shaping genome size remains controversial (Gregory 2003, 2004). A recent study in birds and mammals indicates that short deletions alone cannot explain the DNA content loss (Kapusta, et al. 2017), and we fail to find any evidence of a relationship between deletion rate and genome size here (**supplementary Figure S4;** but note that power is likely to be low). It seems probable that other structural mutations, such as large deletions, gene duplication and transposable elements, may have greater impacts on genome size. For example, in *Drosophila* Wang, et al. (2022c) found transposable element insertions to be more frequent than SNMs, and an order of magnitude more frequent than short indels. Together, these studies suggest that the impact of short deletions on genome size is likely to be small, despite their higher frequency relative to short insertions.



303 **Figure 5**. The correlation between indel mutation rates and SNM rates plotted on a log scale with a
304 1:1 ratio shown as a dashed line (A). The aggregated log ratios of deletion/insertion in eukaryotes,
305 (B).

306

307 **Concluding remarks**

308 All genetic variation arises from *de novo* mutations. To estimate the rate at which these arise,
309 studies have moved from investigating a few genes that result in phenotypic change, to whole
310 genome sequencing of accumulated lines or pedigrees. Ten years ago, Campbell and Eichler (2013)
311 called for more efforts to sequence genomes from non-human primate families to understand 'how
312 the rate has changed in different lineages'. And, in the present data-rich and tool-rich era we have
313 now seen the sequencing of not just multiple non-human primates, but many taxa representing
314 major clades across the tree of life. Nevertheless, the 'tree of mutations' is far from complete, at
315 either the large or small scale. Even within animals, the vast majority of phyla have never been
316 examined, and with the exception of primates, no groups of close relatives have been examined.

317 However, as progress is being made toward a tree of mutation rates, attention still needs to focus on
318 methodological heterogeneity. Different data filtering strategies may result in non-negligible
319 differences in mutation rate estimates (Bergeron, et al. 2022). Although many strategies have been
320 proposed (Bergeron, et al. 2022), no standardised pipeline has yet been agreed upon or widely
321 adopted. All studies must therefore carefully address the issues of false positives, by manual
322 curation and/or PCR follow-up, and false negatives, by simulation. Biases may also come from the
323 samples used in each study. Besides the biological factors that affect mutation rates (age, sex),
324 environmental factors can also exert influences on mutation rates, such as temperature (Belfield, et
325 al. 2021; Waldvogel and Pfenninger 2021), osmotic stress (Hasan, et al. 2022) or ultraviolet light (Xu,
326 et al. 2019). Samples collected from the wild rather than reared in laboratories may reflect the *real*
327 mutation rate for the studied species, but the variation that this introduces may also make it harder
328 to draw broad conclusions. Despite these challenges, the data to come are exciting. In particular, as
329 sequencing technology advances, we expect improved accuracy in long-read sequencing facilitating
330 the phasing of mutations and permitting the easy detection of complex mutations and larger-scale
331 structural changes.

332

333 **Methods**

334 We made an exhaustive literature search for publications on the experimental estimation of
335 mutation rate. We first searched Clarivate Web of Science up to the date of 21 September 2022 for
336 'Title/Keywords/Abstract' fields containing ("mutation rate" | "mutation rates" | "mutational rate" |
337 "mutational rates"). As much of the resulting literature was related to somatic mutations (e.g.,
338 cancers) or virus mutations, we excluded '("tumor" | "cancer" | "clinical" | "virus")' in the search
339 field of 'Title/Keywords/Abstract'. This search strategy resulted in 9462 studies with the earliest one
340 dated 1928. We also searched for references to mutation-rate estimates in other papers not
341 captured by our search. As bacteria are outside of our study, we manually removed studies related
342 to bacterial mutations. These filters resulted in 174 studies. Of these studies, we further excluded 20
343 studies that used phylogenetic approaches to infer mutation rates, 14 studies that used phenotypic
344 markers, and two other studies that used recombination landscape or the site frequency spectrum
345 to infer mutation rates statistically. This resulted in a total of 138 studies in our final dataset.

346 From each of the potential studies we extracted the year of the publication, the species, the
347 population, mutation type (SNM, short insertion, short deletion), the unit of mutation rate (per bp
348 per year, per bp per generation), the mean and 95% CI of mutation rate, the number of identified
349 mutation events, the number of callable sites, the methods used to infer mutation rate (PO, MA or
350 others), the sample size, the number of generations (if MA), the sequencing technology, the
351 sequencing depth and other pertinent facts comments. If a 95% confidence or credible interval was
352 not provided in the original studies, we calculated a 95% confidence interval using 'binconf()' from
353 the Hmisc R package (version 4.7-0), based on the number of mutation events and callable sites and
354 assuming a binomial distribution (Girard, et al. 2011). For studies in which the number of callable
355 sites was not reported, an estimate was made from the number of mutation events and the
356 reported mutation rate.

357 To compare the SNM rates across different species, we conducted a phylogenetic meta-analysis (i.e.
358 a combined analysis of published data) using the mutation data from MA or PO studies. For the
359 social amoeba (Saxer, et al. 2012), the estimated mutation rate was reported without the mutation
360 count or callable sites. Instead of discarding this study, we assumed one mutation event and
361 calculated the corresponding callable sites based on mutation rates. This will necessarily
362 underestimate the precision with which rates are estimated (minimising their weight in the analysis),
363 while keeping it in the meta-analysis with unchanged mean values.

364 We obtained the phylogenetic tree and relative divergence dates for the species included in this
365 study from http://www.timetree.org/ (Kumar, et al. 2022). We performed the analyses using a
366 phylogenetic mixed model with phylogeny as random effect, assuming a poisson distribution for the

occurrence of *de novo* mutations, off-set by the number of callable sites (i.e. enforcing a directly proportional relationship between the expected number of mutations and the number of sites observed). The analysis was conducted using Bayesian mixed-model R-package MCMCglmm (version 2.32) (Hadfield 2010) with a weak prior for the variance of random effects and a fully-informative prior at 1 for the fixed effect of 'callable sites', using the following code in R:

```
vv <- diag(2)
diag(vv) <- c(1e6, 1e-8)
prior <- list(B=list(mu=c(0,1), V=vv),
              G=list(G1=list(V=1, nu=1, alpha.mu=0, alpha.V=100)))
```

The syntax for the basic model in R was:

```
m1 <- MCMCglmm(Mutation ~ log(Callable),
          random= ~ Species,
          ginverse=list(Species=InverseTree),
          prior = prior, data=aa,
          nitt=5001000, thin=1000, burnin=10000,
          family = "poisson", pr=TRUE)
```

where 'Mutation' is the number of mutations and 'Callable' is the number of callable sites. 'Species' is the random effect and 'InverseTree' is the expected phylogenetic covariance among species based on relative branch lengths. To achieve stationarity, we ran 5010000 iterations with a thinning interval of 1000 and a burn-in of 10000 steps for the model. To investigate the effects of methods (MA or PO) used to estimate mutation rate, we additionally included 'Method' as a second fixed effect in the above model.

We also investigated the relationships between mutation rate and generation time, population size, genome size using phylogenetic mixed models that included these factors separately as fixed effects and phylogenetic covariance as a random effect. A log-transformation was made on these variables before fitting model and model predictions were made conditioning on each of the different species as random effect levels. Estimated generation time for each species was gathered from the literature (see Availability of data and materials). Estimates of census population size data were taken from

396 Buffalo (2021), who used body length and species range size to infer the population size. As the
397 estimation of population size is extremely challenging and subject to many uncertainties, we chose
398 to use only this source to avoid methodological heterogeneity. Genome sizes were taken to be
399 assembly size, and gathered from the NCBI database.

400

401 **Availability of data and materials**

402 The datasets and the code to analyse the data are available on GitHub:
403 https://github.com/Yiguan/mutation_literature

404 **Acknowledgements**

405 We thank the authors of the many studies we analyse for making their data available, and we thank
406 Jarrod Hadfield for statistical advice.

407 **Funding**

408 This work was funded by the UK Biotechnology and Biological Sciences Research Council through
409 grant BB/T007516/1. For the purpose of open access, the authors have applied a Creative Commons
410 Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

411 **Competing interests**

412 The authors declare that they have no competing interests.

413

414 **Supplementary figure legends**

415 **Supplementary Figure S1**: The SNM rates comparison in *D. melanogaster* across different studies
416 using MA or PO method to estimate mutation rates. 1a: line33; 1b: line39; a2: West African
417 population; 2b: European population.

418 **Supplementary Figure S2**: The SNM rates comparison in *M. musculus* across different studies using
419 MA or PO method to estimate mutation rates.

420 **Supplementary Figure S3**: The ratio of deletions over insertions across all the identified studies.

421 **Supplementary Figure S4:** The relationship between genome size and the ratio of deletion/insertion
422 without (A) or with (B) human and mammal groups. Log-transformed has been made on both
423 variables.

424

425

426 **References**

427 Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes:
428 causes and consequences. Nature Reviews Genetics 8:619-631.
429 Bailey WJ, Fitch D, Tagle DA, Czelusniak J, Slightom JL, Goodman M. 1991. Molecular evolution of the
430 psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown. Molecular Biology
431 and Evolution 8:155-184.
432 Belfield EJ, Brown C, Ding ZJ, Chapman L, Luo M, Hinde E, Van Es SW, Johnson S, Ning Y, Zheng SJ.
433 2021. Thermal stress accelerates Arabidopsis thaliana mutation rate. Genome Research
434 31:40-50.
435 Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering
436 plants. Annals of Botany 95:127-132.
437 Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL, Armstrong E, Riera M, Carlson
438 J, Chen HY, et al. 2022. The Mutationathon highlights the importance of reaching
439 standardization in estimates of pedigree-based germline mutation rates. Elife 11.
440 Bourguignon T, Kinjo Y, Villa-Martin P, Coleman NV, Tang Q, Arab DA, Wang Z, Tokuda G, Hongoh Y,
441 Ohkuma M. 2020. Increased mutation rate is linked to genome reduction in prokaryotes.
442 Current Biology 30:3848-3855. e3844.
443 Buffalo V. 2021. Quantifying the relationship between genetic diversity and population size suggests
444 natural selection cannot explain Lewontin's Paradox. Elife 10:e67509.
445 Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. Trends in
446 Genetics 29:575-584.
447 Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in
448 Drosophila melanogaster. Plos Genetics 8:e1003090.
449 Charlesworth B, Jensen JD. 2022. How Can We Resolve Lewontin's Paradox? Genome biology and
450 evolution 14:evac096.
451 Chintalapati M, Moorjani P. 2020. Evolution of the mutation rate across primates. Current Opinion in
452 Genetics & Development 62:58-64.
453 de Jong WW, Rydén L. 1981. Causes of more frequent deletions than insertions in mutations and
454 protein evolution. Nature 290:157-159.
455 Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of
456 insertions in the Caenorhabditis elegans nuclear genome. Nature 430:679-682.
457 Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proceedings of
458 the National Academy of Sciences 88:7160-7164.
459 Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics
460 148:1667-1686.
461 Drost JB, Lee WR. 1995. Biological basis of germline mutation: comparisons of spontaneous germline
462 mutation rates among drosophila, mouse, and human. Environmental and Molecular
463 Mutagenesis 25:48-64.
464 Gao Z, Wyman MJ, Sella G, Przeworski M. 2016. Interpreting the dependence of mutation rates on
465 age and time. Plos Biology 14:e1002355.

466 Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion
467 E, Diallo O, et al. 2011. Increased exonic de novo mutation rate in individuals with
468 schizophrenia. Nat Genet 43:860-863.
469 Gregory TR. 2004. Insertion–deletion biases and the evolution of genome size. Gene 324:15-34.
470 Gregory TR. 2003. Is small indel bias a determinant of genome size? Trends in Genetics 19:485-488.
471 Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD.
472 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in
473 Drosophila. Nature 445:82-85.
474 Hadfield J. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The
475 MCMCglmm R Package. Journal of Statistical Software 33:1-22.
476 Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology:
477 phylogenies, taxonomies and multi-trait models for continuous and categorical characters.
478 Journal of Evolutionary Biology 23:494-508.
479 Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges
480 M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell
481 divisions in cattle. bioRxiv:079863.
482 Harmon LJ, Pennell MW, Henao-Diaz LF, Rolland J, Sipley BN, Uyeda JC. 2021. Causes and
483 consequences of apparent timescaling across all estimated evolutionary rates. Annual
484 Review of Ecology, Evolution, and Systematics 52:587-609.
485 Hasan AR, Lachapelle J, El-Shawa SA, Potjewyd R, Ford SA, Ness RW. 2022. Salt stress alters the
486 spectrum of de novo mutation available to selection during experimental adaptation of
487 Chlamydomonas reinhardtii. Evolution.
488 Ho EKH, Macrae F, Latta LC, McIlroy P, Ebert D, Fields PD, Benner MJ, Schaack S. 2020. High and
489 Highly Variable Spontaneous Mutation Rates in Daphnia. Molecular Biology and Evolution
490 37:3258-3266.
491 Ho SY, Duchêne S, Duchêne D. 2015. Simulating and detecting autocorrelation of molecular
492 evolutionary rates among lineages. Molecular Ecology Resources 15:688-696.
493 Huang W, Lyman RF, Lyman RA, Carbone MA, Harbison ST, Magwire MM, Mackay TFC. 2016.
494 Spontaneous mutations and the origin and maintenance of quantitative genetic variation.
495 Elife 5.
496 Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, Prigmore E, Short P, Gallone G, McRae
497 J. 2022. Genetic and chemotherapeutic influences on germline hypermutation. Nature
498 605:503-508.
499 Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals.
500 Proceedings of the National Academy of Sciences 114:E1460-E1469.
501 Keightley PD, Eyre-Walker A. 2000. Deleterious mutations and the evolution of sex. Science 290:331-
502 333.
503 Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the Spontaneous Mutation Rate
504 per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. Genetics 196:313-+.
505 Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome
506 sequences of three Drosophila melanogaster spontaneous mutation accumulation lines.
507 Genome Research 19:1195-1201.
508 Keith N, Jackson CE, Glaholt SP, Young K, Lynch M, Shaw JR. 2021. Genome-Wide Analysis of
509 Cadmium-Induced, Germline Mutations in a Long-Term Daphnia pulex Mutation-
510 Accumulation Experiment. Environmental Health Perspectives 129.
511 Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624-626.
512 Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing
513 Mendelian diseases. Human Mutation 21:12-27.
514 Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation
515 rate. Human Mutation 2:229-234.

516 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A,
517 Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of
518 father's age to disease risk. Nature 488:471-475.
519 Konrad A, Brady MJ, Bergthorsson U, Katju V. 2019. Mutational Landscape of Spontaneous Base
520 Substitutions and Small Indels in Experimental Caenorhabditis elegans Populations of
521 Differing Size. Genetics 212:837-854.
522 Krasovec M, Rickaby RE, Filatov DA. 2020. Evolution of mutation rate in astronomically large
523 phytoplankton populations. Genome biology and evolution.
524 Kucukyildirim S, Behringer M, Sung W, Brock DA, Doak TG, Mergen H, Queller DC, Strassmann JE,
525 Lynch M. 2020. Low base-substitution mutation rate but high rate of slippage mutations in
526 the sequence repeat-rich genome of Dictyostelium discoideum. G3: Genes, Genomes,
527 Genetics 10:3445-3452.
528 Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022.
529 TimeTree 5: An expanded resource for species divergence times. Molecular Biology and
530 Evolution 39:msac174.
531 Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends in Ecology
532 & Evolution 29:33-41.
533 Li W-H, Tanimura M, Sharp PM. 1987. An evaluation of the molecular clock hypothesis using
534 mammalian DNA sequences. Journal of Molecular Evolution 25:330-342.
535 Liu H, Zhang J. 2021. The rate and molecular spectrum of mutation are selectively maintained in
536 yeast. Nature Communications 12:1-11.
537 Long H, Winter DJ, Chang AY-C, Sung W, Wu SH, Balboa M, Azevedo RB, Cartwright RA, Lynch M,
538 Zufall RA. 2016. Low base-substitution mutation rate in the germline genome of the ciliate
539 Tetrahymena thermophila. Genome biology and evolution 8:3629-3639.
540 Lopez-Cortegano E, Craig RJ, Chebib J, Samuels T, Morgan AD, Kraemer SA, Bondel KB, Ness RW,
541 Colegrave N, Keightley PD. 2021. De Novo Mutation Rate Variation and Its Determinants in
542 Chlamydomonas. Molecular Biology and Evolution 38:3709-3723.
543 Lynch M. 2010. Evolution of the mutation rate. Trends in Genetics 26:345-352.
544 Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift,
545 selection and the evolution of the mutation rate. Nature Reviews Genetics 17:704-714.
546 Lynch M, Walsh B. 1998. Genetics and analysis of quantitative traits: Sinauer Sunderland, MA.
547 Marais GA, Batut B, Daubin V. 2020. Genome evolution: mutation is the main driver of genome size
548 in prokaryotes. Current Biology 30:R1083-R1085.
549 Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock.
550 Proceedings of the National Academy of Sciences 90:4087-4091.
551 Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E,
552 Stratton MR, et al. 2014. C. elegans whole-genome sequencing reveals mutational signatures
553 related to carcinogens and DNA repair deficiency. Genome Research 24:1624-1636.
554 Mukai T, Cockerham CC. 1977. Spontaneous mutation rates at enzyme loci in Drosophila
555 melanogaster. Proceedings of the National Academy of Sciences 74:2514-2517.
556 Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics
557 156:297-304.
558 Ness RW, Morgan AD, Colegrave N, Keightley PD. 2012. Estimate of the Spontaneous Mutation Rate
559 in Chlamydomonas reinhardtii. Genetics 192:1447-+.
560 Nute PE, Stamatoyannopoulos G. 1984. Estimating Mutation-Rates Using Abnormal Human
561 Hemoglobins. Yearbook of Physical Anthropology 27:135-151.
562 Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating Divergence
563 Dates and Substitution Rates in the Drosophila Phylogeny. Molecular Biology and Evolution
564 29:3459-3473.
565 Ohno M. 2019. Spontaneous de novo germline mutations in humans and mice: rates, spectra, causes
566 and consequences. Genes & Genetic Systems 94:13-22.

567 Peck KM, Lauring AS. 2018. Complexities of viral mutation rates. Journal of Virology 92:e01031-
568 01017.
569 Petrov DA. 2002. Mutational equilibrium model of genome size evolution. Theoretical Population
570 Biology 61:531-544.
571 Saxena AS, Salomon MP, Matsuba C, Yeh SD, Baer CF. 2019. Evolution of the Mutational Process
572 under Relaxed Selection in Caenorhabditis elegans. Molecular Biology and Evolution 36:239-
573 251.
574 Saxer G, Havlak P, Fox SA, Quance MA, Gupta S, Fofanov Y, Strassmann JE, Queller DC. 2012. Whole
575 genome sequencing of mutation accumulation lines reveals a low mutation rate in the social
576 amoeba Dictyostelium discoideum.
577 Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and Genomic Consequences of Spontaneous
578 Mutational Events in Drosophila melanogaster. Genetics 194:937-+.
579 Sigurðardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. 2000. The mutation rate in the
580 human mtDNA control region. The American Journal of Human Genetics 66:1599-1609.
581 Stamatoyannopoulos G, Nute PE. 1982. De novo mutations producing unstable Hbs or Hbs M. II.
582 Direct estimates of minimum nucleotide mutation rates in man. Human Genetics 60:181-
583 188.
584 Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination
585 frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc
586 Lond B Biol Sci 372:20160455.
587 Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012a. Drift-barrier hypothesis and mutation-
588 rate evolution. Proceedings of the National Academy of Sciences 109:18488-18492.
589 Sung W, Tucker AE, Doak TG, Choi E, Thomas WK, Lynch M. 2012b. Extraordinary genome stability in
590 the ciliate Paramecium tetraurelia. Proceedings of the National Academy of Sciences
591 109:19339-19344.
592 Thomas GW, Hahn MW. 2014. The human mutation rate is increasing, even as it slows. Molecular
593 Biology and Evolution 31:253-257.
594 Thomas GW, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DS, Murali SC, Williams LE,
595 Doddapaneni H, Muzny DM. 2018. Reproductive longevity predicts mutation rates in
596 primates. Current Biology 28:3193-3197. e3195.
597 Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH,
598 Blanchette M, Siepel AC, Thomas PJ, McDowell JC, et al. 2003. Comparative analyses of
599 multi-species sequences from targeted genomic regions. Nature 424:788-793.
600 Trimble BK, Doughty JH. 1974. The amount of hereditary disease in human populations. Annals of
601 Human Genetics 38:199-223.
602 Waldvogel A-M, Pfenninger M. 2021. Temperature dependence of spontaneous mutation rates.
603 Genome Research 31:1582-1589.
604 Wang L, Ji Y, Hu Y, Hu H, Jia X, Jiang M, Zhang X, Zhao L, Zhang Y, Jia Y. 2019. The architecture of
605 intra-organism mutation rate variation in plants. Plos Biology 17:e3000191.
606 Wang RJ, Peña-Garcia Y, Bibby MG, Raveendran M, Harris RA, Jansen HT, Robbins CT, Rogers J, Kelley
607 JL, Hahn MW. 2022a. Examining the effects of hibernation on germline mutation rates in
608 grizzly bears. Genome biology and evolution.
609 Wang RJ, Thomas GW, Raveendran M, Harris RA, Doddapaneni H, Muzny DM, Capitanio JP,
610 Radivojac P, Rogers J, Hahn MW. 2020. Paternal age in rhesus macaques is positively
611 associated with germline mutation accumulation but not with measures of offspring
612 sociability. Genome Research 30:826-834.
613 Wang Y, Allen SL, Reddiex AJ, Chenoweth S. 2022b. Direct and indirect impacts of positive selection
614 on genomic variation in Drosophila serrata. bioRxiv.
615 Wang Y, McNeil P, Abdulazeez R, Pascual M, Johnston SE, Keightley PD, Obbard DJ. 2022c. Variation
616 in mutation, recombination, and transposition rates in Drosophila melanogaster and
617 Drosophila simulans. bioRxiv.

618 Wilding C. 2017. Genetic diversity of the African malaria vector Anopheles gambiae. Nature 552:96-
619 100.
620 Wu C-I, Li W-H. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man.
621 Proceedings of the National Academy of Sciences 82:1741-1745.
622 Wu F, Przeworski M. 2022. A paternal bias in germline mutation is widespread in amniotes and can
623 arise independently of cell division numbers. Elife 11:e80008-e80008.
624 Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P, Przeworski M. 2020. A comparison of humans
625 and baboons suggests germline mutation rates do not track cell divisions. Plos Biology
626 18:e3000838.
627 Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree KS, Gershenzon J, Widmer A, Huber M. 2019.
628 Low genetic variation is associated with low mutation rate in the giant duckweed. Nature
629 Communications 10:1-6.
630 Yang SH, Wang L, Huang J, Zhang XH, Yuan Y, Chen JQ, Hurst LD, Tian DC. 2015. Parent-progeny
631 sequencing indicates higher mutation rates in heterozygotes. Nature 523:463-U187.
632 Yoder AD, Tiley GP. 2021. The challenge and promise of estimating the de novo mutation rate from
633 whole-genome comparisons among closely related individuals. Molecular Ecology 30:6087-
634 6100.

635