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

Yiguan Wang, Darren J. Oppard

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 139 mutation accumulation (MA) and parent-offspring (PO) sequencing studies covering 71 species. Based on these recently accumulated data, we revisit differences in mutation rate between different phylogenetic lineages and update the known relationships between mutation rate and generation time, genome size, census population size, and neutral nucleotide diversity—while accounting for phylogenetic non-independence. We do not find a significant difference between MA and PO in estimated mutation rates, but we confirm that mammal and plant lineages have higher mutation rates than arthropods, and that unicellular eukaryotes have the lowest mutation rates. We find that mutation rates are higher in species with longer generation times and larger genome sizes, even when accounting for phylogenetic relationships. Moreover, although neutral nucleotide diversity is positively correlated with mutation rate, the gradient of the relationship is significantly less than one (on a logarithmic scale), consistent with higher mutation rates in populations with smaller effective size. For the 29 species for which data are available, we find that indel mutation rates are positively correlated with SNM rates, and that short deletions are generally more common than short insertions. Nevertheless, despite recent progress, 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); HPD, highest posterior density.
Introduction

The de novo mutation rate, \( \mu \), 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 \( \theta \). However, it seems implicit that \( \mu \) is of no interest to many evolutionary studies; researchers often use \( \mu \) from one species when analysing another, simply for convenience—as if \( \mu \) were exchangeable in a way that \( N_e \) is not (e.g. Wilding 2017). This is presumably because variation in \( \mu \) among species is taken to be negligible compared to variation in \( N_e \). 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 \( N_e \) when explaining variation in (e.g.) genetic diversity.

One reason that we might expect mutation rate to vary little compared to recombination rate, is if recombination is optimised differently by different selective regimes (Stapley, et al. 2017), whereas mutation is minimised to a lower limit imposed by drift (the drift barrier hypothesis; Lynch, et al. 2016). However, it might be that in reality mutation rate is optimised 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, even among 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 \( 7.4 \times 10^{-9} \) per bp per generation. However, such estimates are prone to bias, for example 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 gives 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 unconstrained and weakly constrained sites (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.


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 of 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. 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. Minimising the rates of false positive and false negative mutation-calls is therefore critical for the PO approach (Yoder and Tiley 2021).

Relaxed-clock phylogenetic studies generally identify substantial among-lineage variation in the evolutionary rate, suggestive of variation in mutation rates 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 estimates have become available, it has become possible to draw comparisons within particular clades such as primates (Chintalapati and Moorjani 2020), or vertebrates (Yoder and Tiley 2021; Bergeron, et al. 2022). However, there has not been a wider analysis, nor one that includes both PO and MA studies. Here we perform a comprehensive meta-analysis of published eukaryotic SNM and indel mutation-rate estimates that use direct sequencing of multiple generations, either through MA or PO comparison. We focus on the variation in mutation rate among species, and the role of the host phylogeny, experimental approach, generation time and population size in determining variation in estimates rates.

Results and Discussion

Studies included

We searched for eukaryotic mutation-rate studies in Clarivate Web of Science, and then we manually examined these studies and identified those containing mutation rate estimates (see Methods). In total, we identified 139 studies using either MA or PO approaches for inclusion, covering 71 species (see Availability of data and materials). Studies on human mutation rate accounted for a third of the identified literature, followed by unicellular organisms (20%), arthropods (17%), primates (8%), and plants (8%) (Figure 1A). However, across all groups, most studies focused on model species: yeast in the unicellular group, Drosophila in arthropods, mice in mammals, and Arabidopsis in plants. These estimates tended to derive from MA analyses rather than PO studies (61 vs. 38 among non-human studies). Nevertheless, as sequencing has become cheaper, there has been a rapid increase in the number of published experimental estimates of the mutation rate, with an average of 11 studies each year in eukaryotic species since 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/](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 94.3% of the random-effects variance, with 95% HPD interval (85.4%, 98.6%).

Among the chosen species groupings, the unicellular species had the lowest mutation rate ranging from $0.02 \times 10^{-9}$ per generation per base pair in ciliates to $0.74 \times 10^{-9}$ 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 $2.27 \times 10^{-9}$, arthropods at $3.56 \times 10^{-9}$, fish at $7.37 \times 10^{-9}$, and a bird at $6.15 \times 10^{-9}$ (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 $0.46 \times 10^{-9}$ in pea aphids (per parthenogenic generation) and the highest one was only $4.97 \times 10^{-9}$ in fruit flies. In contrast to these species, mammals and plants had relatively high mutation rates, including non-primate mammals at $8.76 \times 10^{-9}$ and primates at $9.60 \times 10^{-9}$. The plants (excluding duckweed) and a multicellular fungus had the highest rates of all, ranging from $8.32$ to $44.34 \times 10^{-9}$ (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 = 0.659). 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 $5.28 \times 10^{-9}$, but $3.20 \times 10^{-9}$ using PO (supplementary Figure S1), and in mice the MA estimate was $5.40 \times 10^{-9}$ while the PO estimate was $4.03 \times 10^{-9}$ (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.
**Figure 2.** A phylogenetic meta-analysis of SNM rates for different species identified from the literature. Mutation rates and 95% CIs were estimated using a PGLMM. For each group, we estimated the rate for their most recent common ancestor (shown in bold under each group and
vertical solid grey lines). The vertical dashed line indicates the mutation rate of common ancestor of the tree.

**Biological predictors of the SNM rate**

Many biological factors have been proposed to be associated with mutation rate, including generation time (Wu and Li 1985; Li, et al. 1987; Bailey, et al. 1991; Keightley and Eyre-Walker 2000), genome size (Drake 1991; Lynch 2010), and the rate of genetic drift (Lynch, et al. 2016) — perhaps through the impact of population size (Lanfear, et al. 2014; Krasovec, et al. 2020). To test for a dependence of mutation rate on each of these factors, we ran three further models, including each of them separately as an additional fixed effect in a PGLMM.

We found a significant positive relationship with generation time (Bayesian PGLMM pMCMC < 0.0002; Figure 3A), such that a longer generation time predicted a higher mutation rate per generation. If mutations resulted only from replication errors, and most species had a similar number of cell divisions per generation, then the per-generation mutation rate would not scale with generation time (Wu and Li 1985; Thomas and Hahn 2014). However, in reality, the number of germline cell divisions per generation varies within and between species. In humans, there are around 401 cell divisions per 30-year generation in males and 31 in females (Drost and Lee 1995; Ohno 2019), but in mice it is 62 per 9-month generation in males and 25 in females. In *Drosophila* the sexes are much more similar, with an estimated at 35.5 cell divisions in a 18-day generation for males but 36.5 in a 25-day generation in females in *D. melanogaster* (Drost and Lee 1995). Within a species, a long generation time is likely to permit more cell divisions and thus a higher mutation rate per generation, and many studies in humans and primates have shown that mutations are heavily male-biased and correlated with paternal ages (Kong, et al. 2012; Thomas, et al. 2018; Wang, et al. 2020; Kaplanis, et al. 2022).

Alternatively, mutations may not result predominantly from replication errors. Recent studies suggest that not all germline mutations track cell divisions, and DNA damage may contribute to the mutation rate (Wu, et al. 2020; Wu and Przeworski 2022). Independent of cell divisions, this mutation rate would be correlated with calendar time, and thus species with a longer generation time would have a higher rate. The relationship we observe between generation time and per-generation mutation rate could therefore be a consequence of either a greater number of cell divisions or of accumulating damage over time, and it will be hard to separate these potential causes (Gao, et al. 2016; Wu and Przeworski 2022).
We also found evidence that larger genome size predicts a higher per-generation mutation rate \((p_{\text{MCMC}} = 0.0020; \text{Figure 3B})\). Mutation rate has previously been found to correlate negatively with genome size in prokaryotes (a phenomenon termed “Drake’s rule” (Drake 1991; Bourguignon, et al. 2020; Marais, et al. 2020)), but positively in eukaryotes (Lynch 2010). Two ‘outlier’ unicellular eukaryotes with large genomes but very low mutation rates have \((\text{Paramecium tetraurelia} \text{ and Chlamydomonas reinhardtii})\), have in the past been explained by appeal to their large effective population size and the ‘drift-barrier’ hypothesis (Sung, et al. 2012a). With the increased availability of mutation rates accrued over the past decade, we find that the mutation rate of \(C.\ reinhardtii\) is not as low as previously thought, and this species now fits the positive trend in eukaryotes well \((\text{Figure 3B})\). However, the mutation rates of \(\text{Tetrahymena thermophila}\) (Long, et al. 2016), \(P.\ tetraurelia\), and \(\text{Dictyostelium discoideum}\) (Kucukyildirim, et al. 2020) are indeed very low. Under ‘drift barrier’ hypothesis, these outliers could be caused by the large \(N\) in these species, and \(T.\ thermophila\) would be expected to have the highest \(N\). In fact, the \(N_e\) in \(T.\ thermophila\) is smaller than that in \(P.\ tetraurelia\) \((1.1 \times 10^6 \text{ vs. } 1.2 \times 10^6)\) (Sung, et al. 2012b; Long, et al. 2016).

If census population size were a good predictor of the rate of genetic drift \((N_e)\) then a significant negative relationship between them could be consistent with the ‘drift barrier’ hypothesis; that is that mutation rates are minimised by natural selection until the point at which the impact of selection for lower mutation rate is counterbalanced by allele frequencies changes due to drift (Lynch 2010). A negative relationship was indeed superficially apparent in our data \((\text{Figure 3C})\), but this relationship was not significant when accounting for the non-independence among species in a phylogenetic mixed model \((p_{\text{MCMC}} = 0.3350)\). Nevertheless, given the weak link between \(N_e\) and census population size (Buffalo 2021), and the challenge of estimating census size (especially for smaller organisms), the lack of a relationship is perhaps unsurprising.

In each of the three models above, we fitted generation time, population size, and genome size individually as predictors of mutation rate in a regression model. However, these factors tend to be highly correlated: large organisms tend to have longer generation times, smaller population sizes, and often larger genomes (Martin and Palumbi 1993; Buffalo 2021) and is therefore hard to distinguish between casual relationships and confounding correlations. Indeed, including all three factors \((\text{generation time, genome size, population size})\) together as fixed terms in a single model resulted in none of them being significant, although the effect of generation time was marginal with \(p_{\text{MCMC}} = 0.0781\) (versus \(p_{\text{MCMC}} = 0.3855\) for genome size and 0.6844, for population size).

The drift-barrier hypothesis
The most direct test of the drift-barrier hypothesis would be a regression of SNM rate on $N_e$.

However, long-term $N_e$ can only be estimated from putatively neutral genetic diversity (i.e. $\pi_s$) by making use of the relationship $\theta_N = 4N_e\mu$, and unfortunately any uncertainty in $\mu$ could induce a spurious correlation in a regression of $\mu$ on $\pi/4\mu$. Instead, error in both $\pi_s$ and $\mu$ can be accounted for by fitting a bivariate linear mixed model, parameterised in terms of a regression of $\pi_s$ on $\mu$ (see Methods). Then, on a logarithmic scale, the gradient of this relationship would be expected to be 1 if $\mu$ and $N_e$ are independent of each other. Fitting only these terms, we found that $\pi_s$ increased with $\mu$, but that the gradient of this relationship was significantly less than 1 (gradient = 0.390 [0.078, 0.709]; $\rho_{MCMC} = 0.0002$; Figure 3D). This suggests that $\pi_s$ does not grow in direct proportion to $\mu$, such that populations with larger $N_e$ tend to have a lower mutation rate—as predicted by the drift barrier hypothesis (Lynch, et al. 2016).

Nevertheless, such a relationship could be driven by other factors. For example, if putatively silent sites are in fact weakly constrained, then the action of direct selection would lead to lower diversity in populations with larger $N_e$. Perhaps more seriously, if species with small $N_e$ tend to have a longer generation time, and longer generation time causes higher mutation rates (i.e. Figure 3A), then a higher $\mu$ in species with low $N_e$ could be driven by a mechanistic generation-time effect, rather than the drift barrier. We attempted to account for this latter hypothesis by fitting generation time as a predictor of both $\mu$ and $\pi_s$, thereby accounting for the direct impact of generation time on the relationship between them. In this final model, $\mu$ increased with generation time ($\rho_{MCMC} = 0.0004$) and $\pi$ did not ($\rho_{MCMC} = 0.9457$), but the gradient of the regression of $\pi_s$ on $\mu$ could no longer be distinguished from 1 (gradient = 0.688 [0.109, 1.230]; $\rho_{MCMC} = 0.1324$).
Figure 3. Phylogenetic linear regressions of SNM rate on (A) generation time, for 45 species, (B) genome size, for 63 species and (C) population size, for 25 species. The solid black lines represent simple linear regressions; the coloured lines represent different PGLMM regressions conditional on each of the species random effects. (D) a regression of pairwise nucleotide diversity ($\pi$) on SNM rate ($\mu$), fitted as a bivariate model in MCMCglmm for 41 species. Grey solid line indicates the model regression line; grey dashed line indicates slope = 1. All axes are plotted on a log scale; Cr: C. reinhardtii; Pt: P. tetraurelia; Tt: T. thermophila; Dd: D. discoideum.

The mutation rates of short indels across eukaryotes

Although the rate of indel mutation is estimated much less often than the SNM rate, our literature search identified estimates for 29 species. We found that the indel mutation rate across eukaryotes displayed a similar pattern to that of the SNM rate (Figure 4), with a correlation coefficient between
the two of 0.76 (95% HPD CI: 0.40, 0.99; \( \rho_{\text{MCMC}} = 0.0030 \)), with the SNM rate being generally higher than the indel mutation rate, except in two lineages (nematodes and amoebae) for which the indel mutation rate was higher (Figure 5A). The unicellular species had the lowest indel mutation rate, from \( 0.04 \times 10^{-9} \) in yeast to \( 0.17 \times 10^{-9} \) in algae, followed by arthropods with \( 0.65 (0.05, 1.61) \times 10^{-9} \).

Mammals had a relatively higher indel mutation rate with \( 1.23 (0.35, 2.33) \times 10^{-9} \), with humans at the higher end of this range at \( 1.87 (1.16, 2.74) \times 10^{-9} \). Of all the studied species, plants had the highest indel mutation rate, which was as high as \( 2.83 (0.40, 6.28) \times 10^{-9} \).

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

Short deletions generally outnumbered short insertions across eukaryotes (Figure 5B, 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.
Concluding remarks

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

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

Methods
We made an exhaustive literature search for publications on the experimental estimation of mutation rate. We first searched Clarivate Web of Science up to the date of 21 September 2022 for ‘Title/Keywords/Abstract’ fields containing (“mutation rate” | “mutation rates” | “mutational rate” | “mutational rates”). As much of the resulting literature was related to somatic mutations (e.g., cancers) or virus mutations, we excluded ("tumor" | "cancer" | "clinical" | "virus") in the search field of ‘Title/Keywords/Abstract’. This search strategy resulted in 9462 studies with the earliest one dated 1928. We also searched for references to mutation-rate estimates in other papers not captured by our search. As bacteria are outside of our study, we manually removed studies related to bacterial mutations. These filters resulted in 174 studies. Of these studies, we further excluded 20 studies that used phylogenetic approaches to infer mutation rates, 14 studies that used phenotypic markers, and two other studies that used recombination landscape or site frequency spectrum to infer mutation rates statistically. This resulted in a total of 139 studies in our final dataset.

From each of the potential studies we extracted the year of the publication, the species, the population, mutation type (SNM, short insertion, short deletion), the units of mutation rate (per bp per year, per bp per generation), the mean and 95% CI of mutation rate, the number of identified mutation events, the number of callable sites, the methods used to infer mutation rate (PO, MA or others), the sample size, the number of generations (if MA), the sequencing technology, the sequencing depth and other pertinent observations. If a 95% confidence or credible interval was not provided in the original study, we calculated a 95% confidence interval using ‘binconf()’ from the Hmisc R package (version 4.7-0), based on the number of mutation events and callable sites and assuming a binomial distribution (Girard, et al. 2011). If a study did not provide the mutation counts, we inferred an effective number of mutation events based on the reported mutation rate and its confidence interval. For studies in which the number of callable sites was not reported, an estimate was made from the number of mutation events and the reported mutation rate. If a study only reports mutation rate, we assumed one mutation event and calculated the corresponding callable sites based on mutation rates. This will necessarily underestimate the precision with which rates are estimated (minimising their weight in the analysis), while keeping it in the meta-analysis with unchanged mean values.

To compare the SNM rates across different species, we conducted a phylogenetic meta-analysis (i.e. a combined analysis of published data) using the mutation data from MA or PO studies. We obtained the phylogenetic tree and relative divergence dates for the species included in this study from http://www.timetree.org/ (Kumar, et al. 2022). We performed the analyses using a phylogenetic mixed model with a unit-scaled 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 the 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 fixed at 1 for the fixed effect of ‘callable sites’, using the following code in R:

\begin{verbatim}
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)
\end{verbatim}

where ‘Mutation’ was the number of mutations and ‘Callable’ was 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 5000000 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, and genome size using univariate phylogenetic mixed models that included these factors separately as fixed effects, and a phylogenetic species term as a random effect. A log-transformation was made on these variables before fitting the model, and model predictions were made conditioning on each of the different species as random effect levels. The estimated generation time for each species was gathered from the literature (see Availability of data and materials) and we obtained estimates for 45 species. Genome sizes were taken to be assembly size, and gathered from the NCBI database,
which resulted in 63 species with genome size available. Estimates of census population size data
were taken from Buffalo (2021), who used body length and species range size to infer the population
size. As the estimation of population size is extremely challenging and subject to many uncertainties,
we chose to use only this source to avoid methodological heterogeneity, which covered 25 species.

To test for a relationship between putatively neutral nucleotide diversity ($\pi_s$) and the SNM rate ($\mu$)
while accounting for uncertainty in both, we used MCMCglmm to fit a bivariate PGLMM with
mutation number and $\pi_s$ as responses. We collected nucleotide diversity data for 41 species either
from (Buffalo 2021) or literature (see Availability of data and materials). Because we were interested
in the gradient of regression of $\pi_s$ on $\mu$ (on the logarithmic scale), we modelled the covariance
between them using an antedependence structure that parameterises the covariance matrix in
terms of a regression (see supporting material to Thomson, et al. (2017)). As before, we modelled
the number of mutations as a Poisson distribution offset by the number of callable sites and
included a phylogenetic term to account for non-independence among species in either SNM rate or
$\pi_s$. The basic syntax for this model was:

```r
biv_model <- MCMCglmm(y ~ trait + at.level(trait, 'mu'):log_callable,
  random = ~ante1(trait):Species2_phylo,
  rcov = ~idh(trait):units,
  data = d12,
  ginverse = list(Species2_phylo=InverseTree),
  prior = prior,
  nitt=5001000, thin=1000, burnin=10000,
  family = NULL
)
```

In addition, to test for the potential role of generation time in driving this negative relationship, we
fitted a final model that additionally included log generation time as a fixed effect predicting both
SNM rate and $\pi_s$.

Availability of data and materials
The datasets and the code to analyse the data are available on GitHub:

https://github.com/Yiguan/mutation_literature

Acknowledgements

We thank the authors of the many studies we analyse for making their data available. We thank Jarrod Hadfield for suggesting the statistical approach and helping with MCMCglmm syntax and interpretation.

Funding

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

Competing interests

The authors declare that they have no competing interests.

Supplementary figure legends

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

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

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

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


Kucukyildirim S, Behringer M, Sung W, Brock DA, Doak TG, Mergen H, Queller DC, Strassmann JE, Lynch M. 2020. Low base-substitution mutation rate but high rate of slippage mutations in
the sequence repeat-rich genome of Dictyostelium discoideum. G3: Genes, Genomes, Genetics 10:3445-3452.


Thomas GW, Hahn MW. 2014. The human mutation rate is increasing, even as it slows. Molecular Biology and Evolution 31:253-257.


