

RESEARCH ARTICLE

# Interpreting the Dependence of Mutation Rates on Age and Time

Ziyue Gao<sup>1\*¶</sup>, Minyoung J. Wyman<sup>2</sup>, Guy Sella<sup>2</sup>, Molly Przeworski<sup>2,3\*</sup>

**1** Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, Illinois, United States of America, **2** Department of Biological Sciences, Columbia University, New York, New York, United States of America, **3** Department of Systems Biology, Columbia University, New York, New York, United States of America

¶ Current address: Howard Hughes Medical Institute, Stanford University, Stanford, California, United States of America

\* [ziyuegao@stanford.edu](mailto:ziyuegao@stanford.edu) (ZG); [mp3284@columbia.edu](mailto:mp3284@columbia.edu) (MP)



click for updates

## Abstract

Mutations can originate from the chance misincorporation of nucleotides during DNA replication or from DNA lesions that arise between replication cycles and are not repaired correctly. We introduce a model that relates the source of mutations to their accumulation with cell divisions, providing a framework for understanding how mutation rates depend on sex, age, and cell division rate. We show that the accrual of mutations should track cell divisions not only when mutations are replicative in origin but also when they are non-replicative and repaired efficiently. One implication is that observations from diverse fields that to date have been interpreted as pointing to a replicative origin of most mutations could instead reflect the accumulation of mutations arising from endogenous reactions or exogenous mutagens. We further find that only mutations that arise from inefficiently repaired lesions will accrue according to absolute time; thus, unless life history traits co-vary, the phylogenetic “molecular clock” should not be expected to run steadily across species.

## OPEN ACCESS

**Citation:** Gao Z, Wyman MJ, Sella G, Przeworski M (2016) Interpreting the Dependence of Mutation Rates on Age and Time. *PLoS Biol* 14(1): e1002355. doi:10.1371/journal.pbio.1002355

**Academic Editor:** Nick H. Barton, Institute of Science and Technology Austria (IST Austria), AUSTRIA

**Received:** August 31, 2015

**Accepted:** December 11, 2015

**Published:** January 13, 2016

**Copyright:** © 2016 Gao et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information file.

**Funding:** ZG was supported in part by the William Rainey Harper Fellowship from the University of Chicago. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Author Summary

We relate how mutations arise to how they accumulate in different sexes, with age and with cell division. This model provides a single framework within which to interpret emerging results from evolutionary biology, human genetics, and cancer genetics. We show that the accrual of mutations should track cell divisions not only when mutations originate during DNA replication but also when they arise through non-replicative mechanisms and are repaired efficiently. This realization means that previous observations of correlations between mutation and cell division rates actually provide little support to the commonly held belief that most germline and somatic mutations arise from replication errors. We further find that only mutations that arise from inefficiently repaired lesions will accrue according to absolute time; thus, without covariation in life history traits, the phylogenetic “molecular clock” should not be expected to run at constant rates across species.

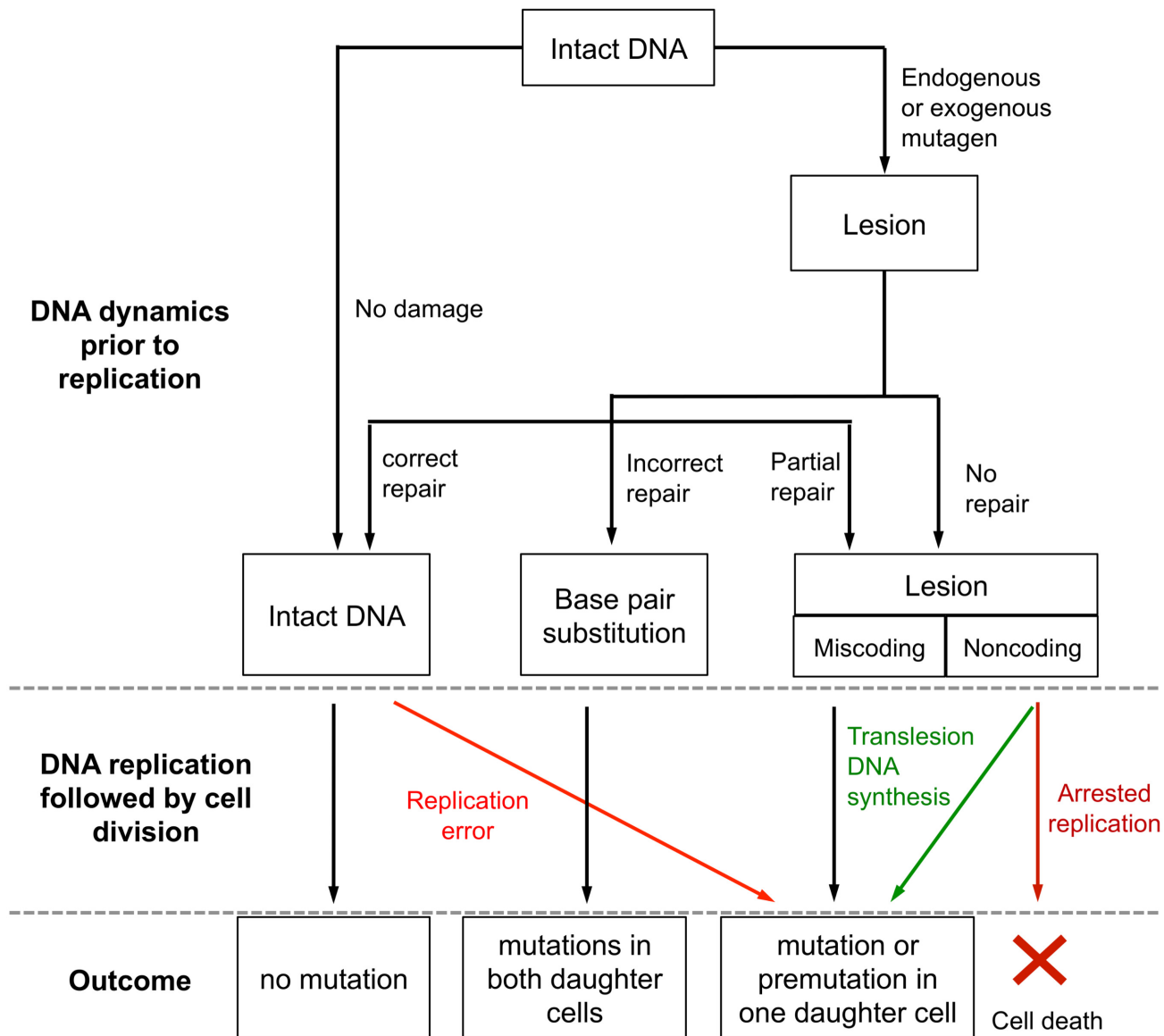
## Introduction

Because mutations are the ultimate source of all genetic variation, deleterious and advantageous, mutagenesis has been of central interest even before the discovery of DNA as the genetic material (e.g., [1]), and developing a model of mutational heterogeneity along the genome is a major focus of current disease mapping studies [2,3]. From many decades of research into mechanisms of DNA replication, damage, and repair, we know that mutations can arise from errors during replication, such as the incorporation of a non-complementary nucleotide opposite an intact template nucleotide during DNA synthesis [4], or from DNA damage caused by exogenous mutagens or endogenous reactions at any time during normal growth of a cell (Fig 1). If uncorrected by the next round of DNA replication, these lesions will lead to arrested replication and cell death, or to mutations in the descendent cells (either because of incorrect template information or due to lesion bypass by error-prone DNA polymerase) [5].

While the fraction of mutations that is non-replicative in origin remains unknown, the common assumption is that mutations are predominantly replicative [6–9]. The basis for this assumption is a set of observations from disparate fields suggesting that, at least in mammals, mutations seem to track cell divisions. First, in phylogenetic studies, it has been observed repeatedly that species with longer generation times tend to have lower substitution rates, which under neutrality reflects lower mutation rates per unit time (“the generation-time effect”) (e.g., [7,10]). Second, based on comparisons of X, Y chromosomes and autosomes, it has been inferred that substantially more mutations arise in the male than in the female germline (e.g., [6,8,11]). In human genetics, pedigree resequencing studies have confirmed a male bias in mutation of approximately 3:1 at a paternal age of 30, and revealed a linear increase in the number of mutations in the child with the father’s age (e.g., [12,13]). These observations are all qualitatively consistent with mutations arising from the process of copying DNA: all else being equal, organisms with shorter generation times should undergo more germ cell divisions per unit time; in mammals, oocytogenesis is completed by birth whereas spermatogenesis is ongoing since puberty throughout the male lifespan, resulting in more germ cell divisions in males than females (Fig 2A) [14,15].

An informative exception to the “generation time effect” seen in phylogenetic studies is transitions at CpG sites, which represent approximately a fifth of de novo germline mutations [12], and show relatively constant substitution rates across species [16–18]. Their more “clock-like” behavior may reflect their distinct molecular origin [16], as CpG transitions are believed to be due primarily to the spontaneous deamination of the 5-methylcytosine (5mC) [19]. This case demonstrates the potential importance of non-replicative sources in germline mutations and raises the possibility that, despite the usual assumption (e.g., [20,21]), not all non-CpG mutations arise from mistakes in replication.

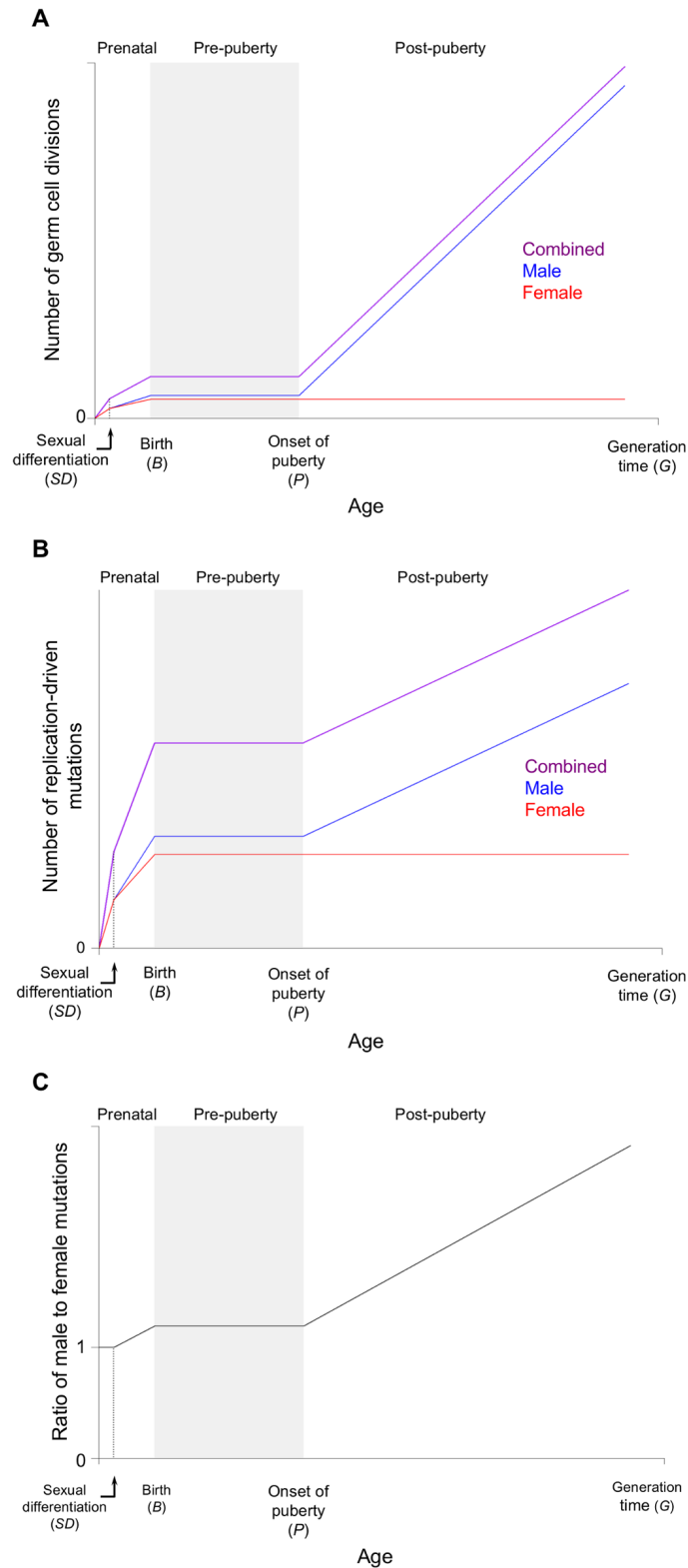
A third argument for the preponderance of replication errors has been made recently in cancer genetics, on the basis of two observations: (i) that somatic mutations tend to accrue more rapidly in tissues with higher renewal rates [22] and (ii) that, across tissues, the lifetime risk of cancer is associated with the total number of stem cell divisions [9]. Together, these findings were interpreted as indicating that in humans, random errors that occur during DNA replication are the source of most somatic mutations, and hence the main determinant of the odds of developing driver mutations that lead to cancer [9]. However, sequencing of tumor samples also revealed characteristic mutation patterns (“mutational signatures”) that reflect known DNA damage processes by endogenous or exogenous sources [23]. Moreover, environmental mutagens are known to influence the incidence of a subset of cancers, implying a role of mutations of non-replicative origins (e.g., [24,25]). These apparently conflicting observations again point to the importance of understanding how mutations arise in somatic tissues as well as in the germline.



**Fig 1. An overview of the mutagenesis process, which involves DNA damage, repair, and replication (adapted from [5]).** Explanation of terms: **Lesion**: chemically altered base; **noncoding lesion**: lesion that cannot pair properly with any regular DNA bases; **miscoding lesion**: lesion that pairs with regular DNA bases that differ from the original one; **correct repair**: repair that completely reverses the lesion to the original state; **incorrect repair**: repair that recognizes the mismatch caused by lesion but alters the undamaged base by mistake; **partial repair**: incomplete repair that leads to abasic sites or other base alterations; **replication error**: misincorporation of nucleotide in the newly synthesized strand despite intact template; **translesion DNA synthesis**: damage tolerance mechanism that allows the DNA replication to bypass lesions and is often mutagenic; **point mutation**: base pair substitution; **premutation**: a base pair at which a lesion is present on one strand and the base on the other strand is substituted, as a result of DNA synthesis from incorrect template information.

doi:10.1371/journal.pbio.1002355.g001

Because, to date, arguments for the replicative origin of mutations have been qualitative and often based on implicit assumptions, we decided to model how the source of mutations relates to their rate of accumulation over cell divisions. For replication-driven mutations, we describe how mutations are expected to accumulate with age, and hence how the generation time relates to the yearly neutral mutation rate. This simple derivation allows us to show that, all else being equal, increases in the generation time will lead to decrease in the mutation rate only under very specific conditions on other parameters. For non-replicative mutations, we relate the



**Fig 2. The accumulation of replication-driven mutations with sex and age.** (A) An illustration of the increase in the number of germ cell divisions with age in humans. For legibility, the plot is not exactly to scale and the final four cell divisions in males needed to complete spermatogenesis are not shown. The origin is the time of fertilization, and SD, B, P, and G are the times of sexual differentiation, birth, onset of puberty, and

reproduction (i.e., generation time), respectively. (B) The increase in the number of mutations due to replication errors with sex and age. (C) The ratio of mutations that occurred in the male versus the female germline (the “male bias”) as a function of increasing parental age.

doi:10.1371/journal.pbio.1002355.g002

mutation rate to rates of DNA damage, repair, and cell division. We show that only when the repair of DNA lesions is highly inefficient will mutations accrue according to absolute time. Otherwise, the accrual of mutations is expected to depend not only on absolute time but also on the rate of cell divisions—a feature previously thought to be specific to replication-driven mutations. By providing explicit expectations for how mutations should accumulate with sex, age, and cell division, these models provide a framework within which to interpret observations from evolutionary biology, human genetics, and cancer genetics.

## Results

### The Accumulation of Mutations Due to Replication Errors

The mutation rate per generation, i.e., the total number of germline mutations between two consecutive generations, is the sum of mutations inherited from both parents, which arose in the lineages of germ cells that gave rise to the child. If mutations are introduced by replication errors, their accumulation will track rounds of DNA replication. In each developmental stage, the number of replication-driven mutations can then be expressed as the product of the number of cell divisions and the mutation rate per cell division. Although a constant mutation rate per cell division is often assumed, explicitly or implicitly [6,26], this need not hold, especially when the cell lineage goes through different development stages, as do germ cells of multicellular organisms. Thus, we consider a more general case, allowing for variation in per cell division mutation rate (e.g., a higher mutation rate in early embryonic development) [27] and describe the accumulation of replication-driven mutations as a piece-wise linear process (following [18]).

For simplicity, we divide germ cell development from fertilization to reproduction into four stages, separated by the settlement of primordial germ cells in the developing gonads (which almost coincides with sexual differentiation), birth, and onset of puberty, respectively. Let  $d_i^s$  and  $\mu_i^s$  be the numbers of cell divisions and replication error rate in the  $i^{\text{th}}$  stage ( $i = 1, 2, 3, 4$ ) in sex  $s$  ( $s \in \{f, m\}$ ). Because there is no sex difference in the first stage,  $d_1^f = d_1^m$  and  $\mu_1^f = \mu_1^m$ , and we replace them by  $d_1$  and  $\mu_1$  (see Table 1 for a list of parameters involved in the model). Previous studies in *Drosophila melanogaster* suggest that the first division of a zygote has an extraordinarily high mutation rate [27,28]. Although the first division in *Drosophila* is quite distinct from that in mammals, it is possible that it would be more mutagenic in mammals as well, so we consider the first division separately as stage 0, of which the mutation rate is  $\mu_0$  for both sexes, and re-define stage 1 as from the second post-zygotic division to sex differentiation. The total number of replication-driven autosomal mutations from one parent to the offspring is then:

$$M_R^s = (\mu_0 + \mu_1 d_1 + \mu_2^s d_2^s + \mu_3^s d_3^s + \mu_4^s d_4^s)H, s \in \{f, m\}$$

where  $H$  is the total number of base pairs in a haploid set of autosomes.

In mammals, all mitotic divisions of female germ cells are completed by birth of the future mother, so  $d_3^f = 0$  and  $d_4^f = 0$ , and the total number of replication-driven mutations inherited from mother is (Fig 2B red line):

$$M_R^f = (\mu_0 + \mu_1 d_1 + \mu_2^f d_2^f)H. \tag{1}$$

**Table 1. A list of parameters used in the model for replication-driven mutations.**

Symbol	Definition
$d_i^s, \mu_i^s$	Number of cell divisions and replication error rate per division in the $i^{\text{th}}$ stage ( $i = 0, 1, 2, 3, 4$ ) in sex $s$ ( $s \in \{f, m\}$ )
	Stage 0: the first post-zygotic division
	Stage 1: from the second post-zygotic division to sex differentiation
	Stage 2: from sex differentiation to birth
	Stage 3: from birth to puberty
	Stage 4: from puberty to reproduction
$t_{sg}$	Duration of spermatogenesis (in years)
$d_{sg}$	number of cell divisions required to complete spermatogenesis from spermatogonial stem cells
$c^m$	Number of cell divisions undergone by spermatogonial stem cells in each year
$P$	Age of puberty (assumed to be the same for both sexes)
$G$	Age of reproduction (assumed to be the same for both sexes)
$H$	Total number of base pairs in a haploid set of autosomes
$M_R^s$	Numbers of autosomal replication-driven mutations inherited from the parent of sex $s$
$M_R$	Total number of autosomal replication-driven mutations inherited by an offspring from both parents
$\alpha_R$	Ratio of male to female replication-driven mutations
$m_{R,y}$	Average yearly mutation rate for replication-driven mutations

doi:10.1371/journal.pbio.1002355.t001

In contrast, male germ cells undergo divisions in all stages outlined above; furthermore, the number of germ cell divisions after puberty ( $d_4^m$ ) is not a fixed number, because after puberty, sperm are continuously produced through asymmetric division of spermatogonial stem cells, at a roughly constant rate. If we assume that males and females have the same ages of onset of puberty and reproduction (denoted by  $P$  and  $G$  respectively), and that a spermatogonial stem cell undergoes  $c^m$  divisions each year, the total number of paternal mutations is a function of reproductive age  $G$  (Fig 2B blue line):

$$M_R^m = [\mu_0 + \mu_1 d_1 + \mu_2^m d_2^m + \mu_3^m d_3^m + \mu_4^m (c^m (G - P - t_{sg}) + d_{sg})]H, \quad (2)$$

where  $t_{sg}$  and  $d_{sg}$  are the time (in years) and the number of cell divisions needed to complete spermatogenesis from spermatogonial stem cells. The two divisions in meiosis are counted as one here, because only one round of DNA replication takes place in meiosis.

Summing Eqs 1 and 2, the total number of autosomal replication-driven mutations inherited by a diploid offspring from both parents is (Fig 2B purple line):

$$M_R = M_R^f + M_R^m = [2\mu_0 + 2\mu_1 d_1 + \mu_2^f d_2^f + \mu_2^m d_2^m + \mu_3^m d_3^m + \mu_4^m (c^m (G - P - t_{sg}) + d_{sg})]H.$$

By dividing Eq 2 by Eq 1, we obtain the ratio of male to female replication-driven mutations:

$$\alpha_R = \frac{M_R^m}{M_R^f} = \frac{\mu_0 + \mu_1 d_1 + \mu_2^m d_2^m + \mu_3^m d_3^m + \mu_4^m d_{sg}}{\mu_0 + \mu_1 d_1 + \mu_2^f d_2^f} + \frac{\mu_4^m c^m}{\mu_0 + \mu_1 d_1 + \mu_2^f d_2^f} \cdot (G - P - t_{sg}),$$

which suggests that, keeping other parameters unchanged, increases in generation time  $G$  will lead to a stronger male bias in mutation, as expected intuitively (Fig 2C).

It follows that the average yearly mutation rate (i.e., the substitution rate if all mutations are neutral) is a function of  $G$ :

$$m_{R,y} = \frac{m_{R,g}}{G} = \frac{2\mu_0 + 2\mu_1 d_1 + \mu_2^f d_2^f + \mu_2^m d_2^m + \mu_3^m d_3^m + \mu_4^m d_{sg} + \mu_4^m c^m (G - P - t_{sg})}{2G}. \quad (3)$$

In order to explore the effect of generation time on the average yearly mutation rate, it is useful to reorganize [Eq 3](#) as:

$$m_{R,y} = \frac{\mu_4^m c^m}{2} + \frac{A^*}{2G}, \quad (4)$$

where  $A^* = 2\mu_0 + 2\mu_1 d_1 + \mu_2^f d_2^f + \mu_2^m d_2^m + \mu_3^m d_3^m - \mu_4^m (c^m P + c^m t_{sg} - d_{sg})$ , which is independent of  $G$ .

[Eq 4](#) suggests that if and only if  $A^* = 0$  will the yearly mutation rate be independent of  $G$ . Otherwise,  $m_{R,y}$  will either increase or decrease monotonically with  $G$ , depending on the sign of  $A^*$ . Changes in the timing of puberty ( $P$ ), in the number of cell divisions ( $d_i^s$ ) and in the replication error rate per cell division in each stage ( $\mu_i^s$ ) will also influence the dependence of  $m_{R,y}$  on  $G$ .

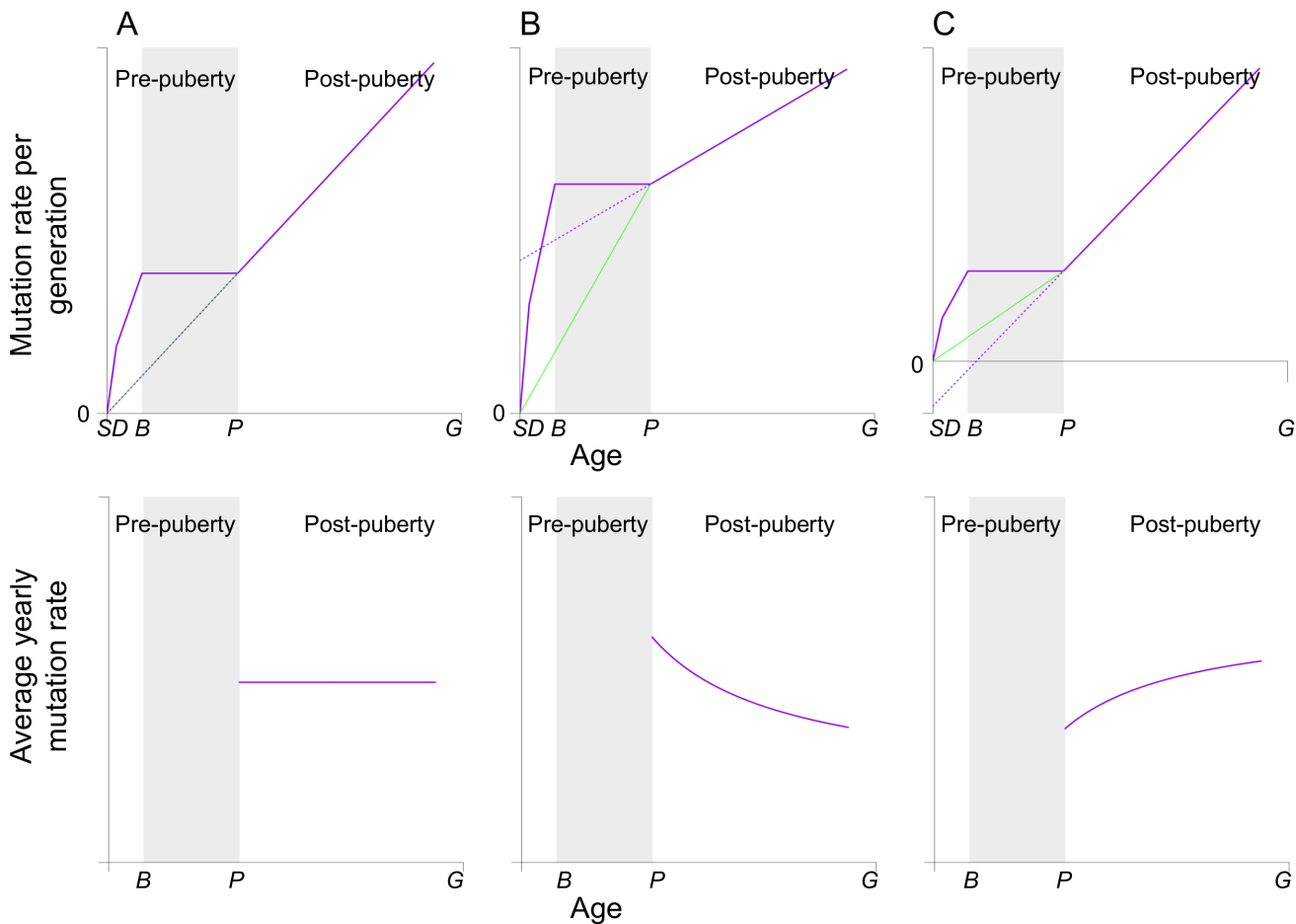
The relationship between  $m_{R,y}$  and  $G$  can also be directly read off the curve in [Fig 3](#). The mutation rate per generation increases linearly with  $G$  after puberty, but this linear relationship does not apply to the period before puberty. If and only if the extended fitted line passes through the origin will the mutation rate per generation be exactly proportional to the generation time, and the average yearly mutation rate unaffected by  $G$ . If the intercept of the extrapolated line at age zero is positive,  $m_{R,y}$  decreases with  $G$ , consistent with the observed “generation time effect” in primates. Conversely, if the intercept is negative,  $m_{R,y}$  increases with  $G$ . In fact, the intercept obtained by extrapolation is exactly  $A^*$  in [Eq 3](#), so interpretation from [Fig 3](#) is equivalent to that suggested by [Eq 4](#).

Although estimates of other parameters exist, little is known about the replication error rate per cell division in germ cells, so it is unclear whether  $A^*$  is positive or negative. However, it seems highly coincidental that an expression that involves multiple variables would happen to equal zero. Therefore, we argue that there is almost certainly an effect of generation time on yearly mutation rate in humans, although the magnitude of the effect could be small. The magnitude of the paternal age effect in pedigree data suggests that there should be generation-time effect in humans (see [S1 Text](#)).

Our model further reveals that, all else being equal, a longer generation time can lead to either an increase or decrease in the average yearly rate at which replicative mutations accrue. Therefore, the general observation that substitution rate in mammals tends to decrease with increasing generation times [\[7,10,16\]](#) is not necessarily expected; in fact, its existence requires very specific conditions on ontogenesis to hold (shown in [Fig 3B](#)). Moreover, given the current understanding of germ cell development in humans, the generation-time effect implies a higher mutation rate per cell division in early embryonic development than in spermatogenesis (see [S1 Text](#) for a discussion of available data in humans and chimpanzees).

Since mammalian species differ drastically in life history traits as well as development and renewal processes of germ cells [\[26,29\]](#), [Eq 4](#) implies that the yearly mutation rate likely varies among species (even if per cell division mutation rates remain constant). As a result, unless life history traits co-vary in certain ways, we should not expect neutral substitution rates to be constant across mammalian species—or even along single evolutionary lineages. An important implication is that changes in life history among hominins [\[30\]](#) introduce uncertainty about dates in human evolution obtained under the assumption of a molecular clock [\[31\]](#).





**Fig 3. The effect of the generation time on the sex-averaged yearly rate of replication-driven mutations.** The sex-averaged mutation rate per generation (solid purple line) increases with the generation time (assumed to be the same for males and females). Depending on the age of puberty ( $P$ ), generation time ( $G$ ), and the per cell division mutation rates, a linear fit to the number of mutations after puberty (dotted purple line) could have a zero, positive, or negative intercept at age zero, and the slope of this linear fit represents the yearly mutation rate after puberty. The slope of the green line represents the average yearly mutation rate prior to puberty. The effect of  $G$  on the overall average yearly mutation rate ( $m_{R,y}$ ) depends on the relative values of the two slopes, which is equivalent to the sign of the intercept of dotted purple line at age zero: (A) If the intercept is zero, the dotted purple and green lines coincide, and the yearly mutation rates before and after puberty are equal, so the  $G$  does not affect  $m_{R,y}$ . (B) If the intercept is positive, the yearly mutation rate after puberty is smaller than that before puberty, so  $m_{R,y}$  decreases with generation time. (C) If the intercept is negative, the yearly mutation rate after puberty is greater than that before puberty, so  $m_{R,y}$  increases with generation time.

doi:10.1371/journal.pbio.1002355.g003

### The Accumulation of Non-replicative Mutations with Cell Divisions

DNA is subject to large numbers of damaging events every day as a result of normal cellular metabolism, and more DNA lesions may be generated by exogenous agents [32]. Typical DNA damage includes depurination and deamination due to DNA hydrolysis; alkylation and oxidation of bases induced by chemicals such as ethylmethane sulfonate or reactive oxygen species; pyrimidine dimers caused by ultraviolet radiation; and single- or double-stranded breaks produced by gamma and X-rays. Most single-stranded lesions cannot pair properly with any regular bases (termed “noncoding lesions”) and thus will block DNA replication if unrepaired (Fig 1). However, a few alterations to nucleotides can pair with bases different from the original Watson-Crick partners; such lesions (termed “miscoding lesions”), if unrepaired before replication, will lead to irreversible replacement of a base pair after cell division (Fig 1) [5].



To model the accrual of non-replicative mutations, we start by considering deamination of methylated CpG sites, which is the best understood example of miscoding lesions, and discuss more complex mutagenesis mechanisms in the [S2 Text](#). This modification turns the methylated cytosine (mC) into a thymine (T); if uncorrected before DNA replication, an adenine instead of a guanine will be incorporated into the nascent strand, which results in a mutation in one of the two daughter cells. While DNA replication and cell division are obviously two distinct events, they are tightly coordinated such that DNA is replicated exactly once before each division (other than in meiosis and under a few unusual conditions). In what follows, we therefore do not distinguish between the two events.

We model the proportion of damaged base pairs at the time of cell division by considering the effects of both damage and repair ([Fig 4A](#)). For simplicity, we assume that single-strand damage occurs at a constant instantaneous rate  $\mu$  throughout cell cycle and that the repair machinery recognizes lesions at a constant rate  $r$  ([Fig 4A](#)). Thus, the proportion of base pairs that carry a lesion at time  $t$  after the last cell division,  $p_1(t)$ , is described by a simple differential equation:

$$\frac{dp_1}{dt} = \mu(1 - p_1) - rp_1,$$

with the initial condition  $p_1(0) = 0$ .

The solution to the differential equation is:

$$p_1(t) = \frac{\mu}{\mu + r} (1 - e^{-(\mu+r)t}).$$

Because each unrepaired single-strand lesion leads to a base pair substitution in one of the two daughter cells, the average mutation rate in one cell division (i.e., the expected fraction of base pairs that differ between a daughter cell and its mother cell) is:

$$M_{NR}(T) = \frac{1}{2}p_1(T) = \frac{\mu}{2(\mu + r)} (1 - e^{-(\mu+r)T}), \quad (5)$$

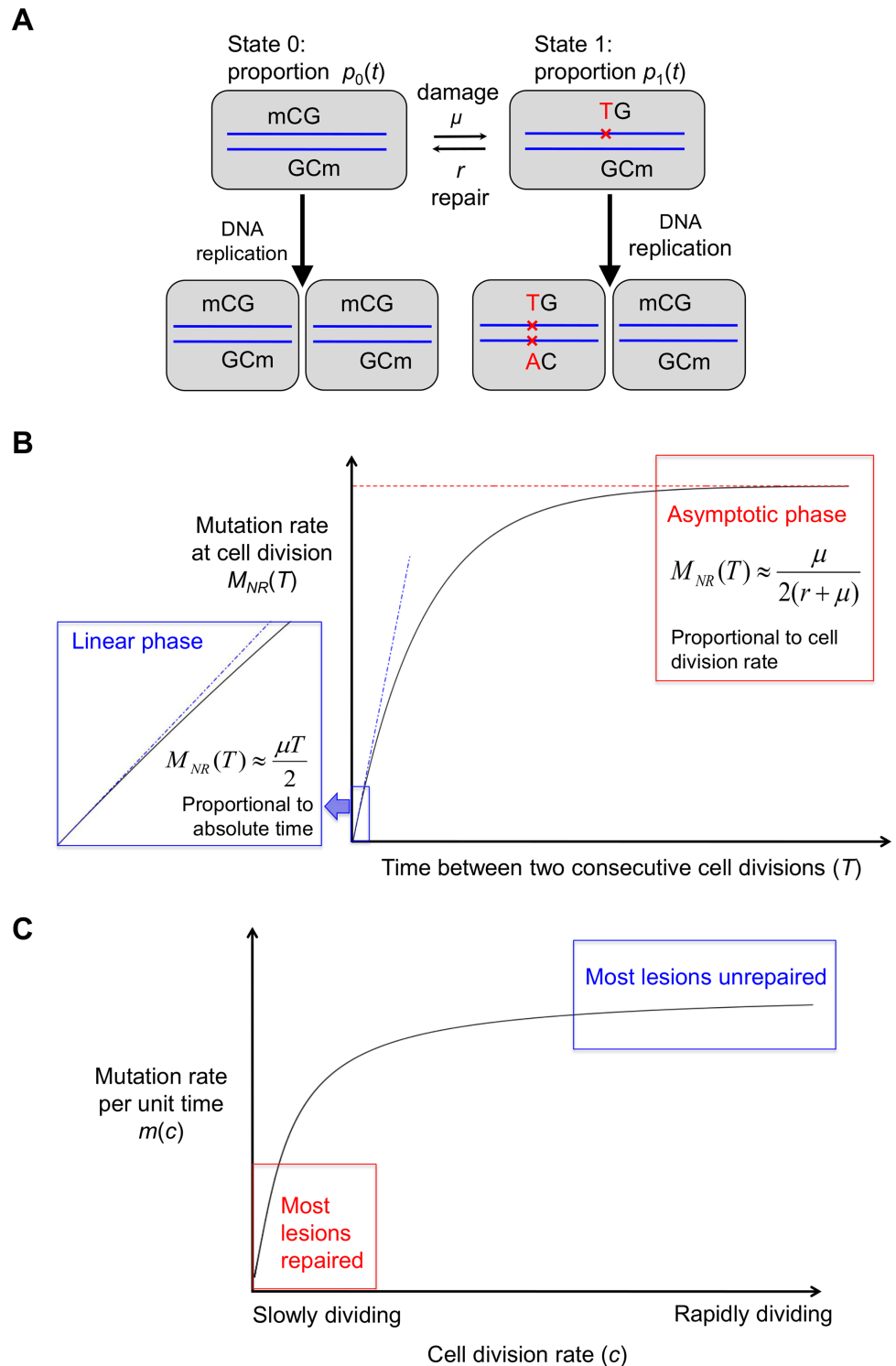
where  $T$  is the time between two consecutive cell divisions ([Fig 4B](#)).

We assume that  $\mu \ll 1/T$  for any biologically reasonable value of  $T$ , so even in the absence of DNA repair, the absolute mutation rate per base pair per cell division ( $\approx \frac{1}{2}\mu T$ ) is very small. In addition, we focus on a single cell lineage and assume an infinite sites model, in which each genomic site can be mutated at most once. Thus, the total mutation rate over many cell divisions is simply the sum of the mutation rates for every division.

A key feature of the result in [Eq 5](#) is that the accumulation of mutations per cell division exhibits two different limiting behaviors, depending on the relative rates of cell division and repair. When the rate at which lesions are repaired is much slower than the rate of cell division ( $rT \ll 1$ ), the number of mutations is approximately proportional to time between two rounds of DNA replication:

$$M_{NR}(T) = \frac{\mu T}{2}. \quad (6)$$

The intuition is that, for a cell under this condition, there is almost no time for the repair machinery to correct lesions, so almost all lesions result in mutations. Consequently, mutations accumulate at a constant rate regardless of the rates of cell division and repair ([Fig 4B](#), red box). In other words, non-replicative mutations that are inefficiently repaired will track absolute time.



**Fig 4. The basic model for non-replicative mutations.** (A) The DNA dynamics before and after cell division. The upper panel shows the DNA states prior to the next cell division, and the lower panel shows the DNA states of the daughter cells after cell division. (B) The per cell division mutation rate increases with the time between two consecutive cell divisions and reaches an asymptote when the cell divides sufficiently slowly. (C) The rate at which non-replicative mutations accumulate per unit time increases with the cell division rate.

doi:10.1371/journal.pbio.1002355.g004

In contrast, in the other limit where the repair is highly efficient relative to the rate of cell division ( $rT \gg 1$ ), the number of mutations approaches an equilibrium level by the time of cell division:

$$M_{NR}(T) = \frac{\mu}{2(\mu + r)}. \tag{7}$$

As a result, mutations accumulate at a rate that is roughly proportional to the number of cell divisions, regardless of absolute time (Fig 4B, blue box). Here, the intuition is that when repair is highly efficient, the few lesions that have not been corrected tend to be those that arose right before the cell division, and therefore the time since the last division has little effect. Importantly, under this scenario, the accrual of mutations that arise from lesions mimics what would be expected from replication errors. We note that the existence of such an equilibrium comes from the assumption of no error in repair; however, even when errors in repair are taken into consideration, there exists a phase in which repair and damage roughly balance out, so the mutation rate is proportional to the cell division rate (see S2 Text).

To understand how the mutation rate of non-replicative mutations depends on absolute time and the rate of cell division in general, we derive the mutation rate per unit time as the product of mutation rate per cell division and the cell division rate ( $c = 1/T > 0$ ):

$$m(c) = cM_{NR}\left(\frac{1}{c}\right) = \frac{c}{2(1 + R)} \left(1 - e^{-\frac{(1+R)\mu}{c}}\right). \tag{8}$$

The mutation rate  $m(c)$  has two limiting behaviors when  $c$  approaches infinity and zero, respectively, which have the same intuitive explanations as Eqs 6 and 7, respectively. Moreover, it can be shown that  $m(c)$  is a concave increasing function of  $c$ . In other words, in a given period of time, faster dividing cell lineages accumulate more non-replicative mutations than slowly dividing lineages, but the increase in the number of mutations is smaller than the increase in the cell division rate. Therefore, when repair is neither inefficient nor extremely efficient, and given fixed damage and repair rates, faster dividing lineages are expected to accumulate non-replicative mutations at a higher rate per year than more slowly dividing ones (Fig 4C and see Table 2 for a list of parameters involved in the model).

This model can be extended readily to incorporate more features, such as other types of non-replicative mutations as well as to understand phenomena such as the strand bias in mutations associated with transcription (see S2 Text) [33,34]. Although the quantitative results

**Table 2. A list of parameters used in the model for non-replicative mutations.**

Symbol	Definition
$\mu$	Instantaneous damage rate
$r$	Instantaneous repair rate
$R = r/\mu$	Relative repair rate compared to damage rate
$p_0(t)$	Proportion of base pairs in the genome that do not carry a lesion at time $t$ since last cell division
$p_1(t)$	Proportion of base pairs in the genome that carry a single-strand lesion at time $t$ since last cell division
$T$	Time between two consecutive divisions of a cell lineage
$M_{NR}(T) = \frac{1}{2} p_1(T)$	Mutation rate per division for a cell that divides every $T$ unit of time
$c = 1/T$	Cell division rate
$m(c) = c * M_{NR}(1/c)$	Mutation rate per unit time for a cell with cell division rate $c$

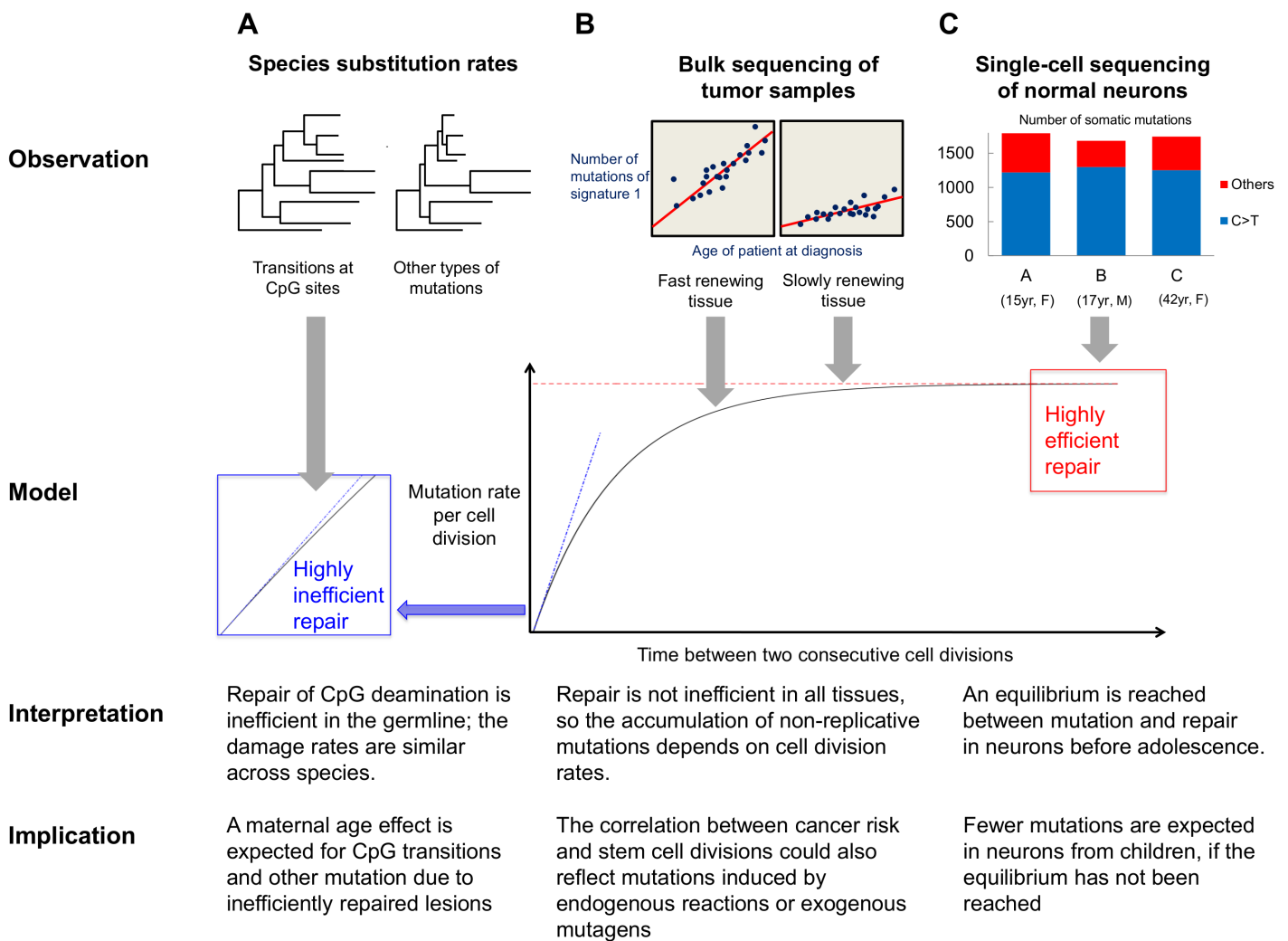
doi:10.1371/journal.pbio.1002355.t002

differ, the main conclusion holds: the accumulation of non-replicative mutations depends critically on the repair efficiency in relation to the cell division rate.

### Discussion

These results demonstrate the fundamental importance of repair efficiency in determining the dependence of mutation rates on age, sex, and cell division rate (Fig 5). When DNA repair is inefficient, we should expect a linear accumulation of damage-induced mutations, partially justifying the expectation that neutral substitution rates of non-replicative mutations should not depend on generation time or other life history traits, and hence may be constant across species. However, our model highlights additional conditions for this expectation to be met: in particular, it reveals that the clock-like behavior of CpG transitions in mammals not only requires a non-replicative origin but also implies both relatively low repair efficiency in germ cells and similar damage rates across mammalian species (Fig 5A).

A further implication is that the number of mutations of maternal origin should increase with the mother's age for CpG transitions and other mutations that arise from inefficiently



**Fig 5. A visual summary of interpretation of existing observations based on our model and further predictions about germline and somatic mutations in humans.**

doi:10.1371/journal.pbio.1002355.g005

repaired lesions. In this regard, we speculate that the current lack of a detectable maternal age effect may be due to underpowered sample sizes (notably because of the strong correlation between maternal and paternal ages). In any case, our model predicts that a maternal age effect should be detectable with sufficient data and reliable identification of parental origin of mutations (e.g., by sequencing of a third generation). Conversely, the detection of a maternal age effect on mutation rate would provide *prima facie* evidence for the existence of non-replicative mutations that are not efficiently repaired (assuming no relationship between the age at which an oocyte is ovulated and the number of cell divisions experienced during oocytogenesis [35]).

Also of note, lesions that have the same damage rate but are recognized by distinct repair mechanisms may differ not only in their absolute mutation rates but also in their time dependencies. Indeed, changes to the repair efficiency (or to the division rate) could alter the sex and time dependence of non-replicative mutations; for example, decreases in repair efficiency could lead mutations that previously tracked cell division rates to depend more on absolute time. Therefore, the phylogenetic molecular clock should not necessarily run at a steady rate even for mutations due to spontaneous DNA damage.

Our modeling results also shed light on studies of somatic mutations. As an illustration, a recent single-cell sequencing study identified mutations in neurons from the cerebral cortex of three healthy individuals [36]. The numbers of mutations in each cell were similar regardless of the donor's sex and age (ranging from 15 to 42 years, Fig 5C) [37]. The genome-wide distribution of the somatic mutations appeared to be associated with transcription, with most identified mutations being C to T transitions at methylated cytosines. These observations led the authors to conclude that the mutations that they observed were due to non-replicative damage that was poorly repaired [36]. However, if mutations are non-replicative in origin and not repaired, more DNA lesions should accrue in older individuals, even in post-mitotic cells. In light of our model, an explanation is that an equilibrium between DNA damage and repair was reached before adolescence, and thus that the number of mutations does not increase further with age (Fig 5C). If this is the case, then there should be fewer somatic mutations in post-mitotic neurons from younger individuals, in which the equilibrium has not been reached.

Similarly, the model helps to interpret patterns observed in tumor samples, in which the total number of somatic mutations increases with the age of patient at diagnosis and grows at higher rates in fast renewing tissues [22]. Deamination at CpG sites make substantial contribution to mutations in almost all cancer types and accumulate at constant yearly rates that appear to be positively correlated with the turnover rates of the corresponding normal tissues (Fig 5B) [23,38]. As we have shown, all else being equal, a positive correlation is expected even for mutations that arise from DNA damage, so long as lesions are not poorly repaired in all somatic tissues.

Importantly, then, the recently reported correlation between number of stem cell divisions and lifetime risk of cancer across tissues is consistent with mutations of both replicative and non-replicative origins, and does not provide any evidence that most mutations are attributable to replication mistakes in stem cell divisions (what the authors referred to as “bad luck” in [9]). Of course, tumorigenesis is a multistep process that depends not only on the accumulation of mutations but also on tissue architecture as well as the order and consequences of specific mutational events, and gaining insight into its causes will likely require consideration of all these facets. What our model makes apparent is that it will also be important to incorporate a realistic model for the source of mutations.

Similar arguments apply to the male bias in mutation found by resequencing pedigrees and the generation time effect in phylogenetics: neither observation provides evidence for a replication-driven mutational process, as they could also reflect mutations arising from residual lesions left after efficient repair. Given these considerations, it becomes clear that, based on

available data, we still do not know if a substantial proportion of human germline and somatic mutations—including those at non-CpG sites—are non-replicative in origin.

In summary, we introduce a model that helps to interpret findings from studies of somatic mutations, human pedigrees, and phylogenies. Although very simple, its behavior appears to be robust. By making explicit the relationship between the genesis of mutations and their accumulation over ontogeny, the model reveals the critical importance of both the source of mutations and the repair efficiency of lesions. Because replicative mutations and non-replicative mutations can display similar properties when repair is efficient, none of the previous observations of correlations between mutation and cell division rates lends strong support to the commonly held belief that most mutations are replicative in origin. Further experimental work is therefore needed to distinguish between different sources of mutation. Notably, fitting models such as this one to growing data from diverse fields should provide a quantitative understanding of how DNA changes accumulate in somatic tissues during a lifetime and in the germline over evolutionary time scales.

## Supporting Information

**S1 Fig. A model for non-replicative mutations with errors in repair.** (A) The DNA dynamics with errors in repair can be described by three states. The upper panel shows the DNA states prior to the next cell division, and the lower panel shows the DNA states of the daughter cells after cell division. (B) The proportion of base pairs without lesion ( $p_0(t)$ ), the proportion of base pairs with single-strand lesions ( $p_1(t)$ ) and the mutation rate per cell division ( $M_{NR}(t)$ ) as functions of the time since the last division. Same values of the damage and repair rates are used for all cases with repair. In the case with no DNA repair, the value of  $r$  is set to zero. (C) Log-log plots for  $p_0(t)$ ,  $p_1(t)$ , and  $M_{NR}(t)$ . The dotted blue lines show the boundaries between the four phases for the case with  $\varepsilon = 0.0001$  (represented by the blue curve). Notice that both axes are on a logarithmic scale, so later phases should be longer than they appear on the plot. (PDF)

**S1 Table. A list of parameters and estimated values used in the model for replication-driven mutations.** See [S1 Text](#) for references behind each parameter value. (DOC)

**S2 Table. A list of parameters used in the model of more complex scenarios of non-replicative mutations.** (DOC)

**S1 Text. The predicted generation time effects in humans and chimpanzees, based on available data.** (DOC)

**S2 Text. More complex scenarios of mutations that arise from DNA lesions.** (DOC)

## Acknowledgments

We thank Phil Green for comments on a previous paper that motivated us to pursue this work; Guy Amster, Priya Moorjani, Eduardo Amorim, Chen Chen, and Laure Ségurel for helpful discussions; and Ludmil Alexandrov, Michael Stratton, and Shamil Sunyaev for sharing unpublished data and valuable suggestions.

## Author Contributions

Conceived and designed the experiments: ZG GS MP. Performed the experiments: ZG MJW. Analyzed the data: ZG MJW. Wrote the paper: ZG GS MP.

## References

1. Muller HJ. Artificial Transmutation of the Gene. *Science*. 1927; 66(1699):84–7. Epub 1927/07/22. doi: [10.1126/science.66.1699.84](https://doi.org/10.1126/science.66.1699.84) PMID: [17802387](https://pubmed.ncbi.nlm.nih.gov/17802387/).
2. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499(7457):214–8. Epub 2013/06/19. doi: [10.1038/nature12213](https://doi.org/10.1038/nature12213) PMID: [23770567](https://pubmed.ncbi.nlm.nih.gov/23770567/); PubMed Central PMCID: PMC3919509.
3. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*. 2014; 46(9):944–50. Epub 2014/08/05. doi: [10.1038/ng.3050](https://doi.org/10.1038/ng.3050) PMID: [25086666](https://pubmed.ncbi.nlm.nih.gov/25086666/); PubMed Central PMCID: PMC4222185.
4. Reijns MA, Kemp H, Ding J, de Procé SM, Jackson AP, Taylor MS. Lagging-strand replication shapes the mutational landscape of the genome. *Nature*. 2015; 518(7540):502–6. doi: [10.1038/nature14183](https://doi.org/10.1038/nature14183) PMID: [25624100](https://pubmed.ncbi.nlm.nih.gov/25624100/); PubMed Central PMCID: PMC4374164.
5. Maki H. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annu Rev Genet*. 2002; 36:279–303. doi: [10.1146/annurev.genet.36.042602.094806](https://doi.org/10.1146/annurev.genet.36.042602.094806) PMID: [12429694](https://pubmed.ncbi.nlm.nih.gov/12429694/).
6. Chang BH, Shimmin LC, Shyue SK, Hewett-Emmett D, Li WH. Weak male-driven molecular evolution in rodents. *Proceedings of the National Academy of Sciences of the United States of America*. 1994; 91(2):827–31. Epub 1994/01/18. PMID: [8290607](https://pubmed.ncbi.nlm.nih.gov/8290607/); PubMed Central PMCID: PMC43042.
7. Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol*. 1996; 5(1):182–7. doi: [10.1006/mpev.1996.0012](https://doi.org/10.1006/mpev.1996.0012) PMID: [8673286](https://pubmed.ncbi.nlm.nih.gov/8673286/).
8. Makova KD, Li WH. Strong male-driven evolution of DNA sequences in humans and apes. *Nature*. 2002; 416(6881):624–6. doi: [10.1038/416624a](https://doi.org/10.1038/416624a) PMID: [11948348](https://pubmed.ncbi.nlm.nih.gov/11948348/).
9. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*. 2015; 347(6217):78–81. doi: [10.1126/science.1260825](https://doi.org/10.1126/science.1260825) PMID: [25554788](https://pubmed.ncbi.nlm.nih.gov/25554788/).
10. Yi S, Ellsworth DL, Li WH. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol Biol Evol*. 2002; 19(12):2191–8. PMID: [12446810](https://pubmed.ncbi.nlm.nih.gov/12446810/).
11. Shimmin LC, Chang BH, Li WH. Male-driven evolution of DNA sequences. *Nature*. 1993; 362(6422):745–7. Epub 1993/04/22. doi: [10.1038/362745a0](https://doi.org/10.1038/362745a0) PMID: [8469284](https://pubmed.ncbi.nlm.nih.gov/8469284/).
12. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. 2012; 488(7412):471–5. doi: [10.1038/nature11396](https://doi.org/10.1038/nature11396) PMID: [22914163](https://pubmed.ncbi.nlm.nih.gov/22914163/); PubMed Central PMCID: PMC3548427.
13. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*. 2015; 47(7):822–6. doi: [10.1038/ng.3292](https://doi.org/10.1038/ng.3292) PMID: [25985141](https://pubmed.ncbi.nlm.nih.gov/25985141/).
14. Crow JF. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet*. 2000; 1(1):40–7. doi: [10.1038/35049558](https://doi.org/10.1038/35049558) PMID: [11262873](https://pubmed.ncbi.nlm.nih.gov/11262873/).
15. Penrose LS. Parental age and mutation. *Lancet*. 1955; 269(6885):312–3. PMID: [13243724](https://pubmed.ncbi.nlm.nih.gov/13243724/).
16. Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*. 2004; 101(39):13994–4001. doi: [10.1073/pnas.0404142101](https://doi.org/10.1073/pnas.0404142101) PMID: [15292512](https://pubmed.ncbi.nlm.nih.gov/15292512/); PubMed Central PMCID: PMC4521089.
17. Kim SH, Elango N, Warden C, Vigoda E, Yi SV. Heterogeneous genomic molecular clocks in primates. *PLoS Genet*. 2006; 2(10):e163. Epub 2006/10/13. doi: [10.1371/journal.pgen.0020163](https://doi.org/10.1371/journal.pgen.0020163) PMID: [17029560](https://pubmed.ncbi.nlm.nih.gov/17029560/); PubMed Central PMCID: PMC1592237.
18. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*. 2014; 15:47–70. doi: [10.1146/annurev-genom-031714-125740](https://doi.org/10.1146/annurev-genom-031714-125740) PMID: [25000986](https://pubmed.ncbi.nlm.nih.gov/25000986/).
19. Bird AP. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 1980; 8(7):1499–504. PMID: [6253938](https://pubmed.ncbi.nlm.nih.gov/6253938/); PubMed Central PMCID: PMC324012.
20. Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Molecular*



- biology and evolution. 2006; 23(3):565–73. Epub 2005/11/11. doi: [10.1093/molbev/msj060](https://doi.org/10.1093/molbev/msj060) PMID: [16280537](https://pubmed.ncbi.nlm.nih.gov/16280537/).
21. Thomas GW, Hahn MW. The human mutation rate is increasing, even as it slows. *Mol Biol Evol*. 2014; 31(2):253–7. doi: [10.1093/molbev/mst218](https://doi.org/10.1093/molbev/mst218) PMID: [24202611](https://pubmed.ncbi.nlm.nih.gov/24202611/).
  22. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110(6):1999–2004. doi: [10.1073/pnas.1221068110](https://doi.org/10.1073/pnas.1221068110) WOS:000315209800019. PMID: [23345422](https://pubmed.ncbi.nlm.nih.gov/23345422/)
  23. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013; 500(7463):415–21. doi: [10.1038/nature12477](https://doi.org/10.1038/nature12477) PMID: [23945592](https://pubmed.ncbi.nlm.nih.gov/23945592/); PubMed Central PMCID: PMC3776390.
  24. Irigaray P, Newby JA, Clapp R, Hardell L, Howard V, Montagnier L, et al. Lifestyle-related factors and environmental agents causing cancer: an overview. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*. 2007; 61(10):640–58. Epub 2007/12/07. doi: [10.1016/j.biopha.2007.10.006](https://doi.org/10.1016/j.biopha.2007.10.006) PMID: [18055160](https://pubmed.ncbi.nlm.nih.gov/18055160/).
  25. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *British journal of cancer*. 2011; 105 Suppl 2:S77–81. Epub 2011/12/14. doi: [10.1038/bjc.2011.489](https://doi.org/10.1038/bjc.2011.489) PMID: [22158327](https://pubmed.ncbi.nlm.nih.gov/22158327/); PubMed Central PMCID: PMC3252065.
  26. Drost JB, Lee WR. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environ Mol Mutagen*. 1995; 25 Suppl 26:48–64. PMID: [7789362](https://pubmed.ncbi.nlm.nih.gov/7789362/).
  27. Gao JJ, Pan XR, Hu J, Ma L, Wu JM, Shao YL, et al. Highly variable recessive lethal or nearly lethal mutation rates during germ-line development of male *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108(38):15914–9. Epub 2011/09/06. doi: [10.1073/pnas.1100233108](https://doi.org/10.1073/pnas.1100233108) PMID: [21890796](https://pubmed.ncbi.nlm.nih.gov/21890796/); PubMed Central PMCID: PMC3179084.
  28. Gao JJ, Pan XR, Hu J, Ma L, Wu JM, Shao YL, et al. Pattern of mutation rates in the germline of *Drosophila melanogaster* males from a large-scale mutation screening experiment. *G3 (Bethesda)*. 2014; 4(8):1503–14. Epub 2014/06/14. doi: [10.1534/g3.114.011056](https://doi.org/10.1534/g3.114.011056) PMID: [24924332](https://pubmed.ncbi.nlm.nih.gov/24924332/); PubMed Central PMCID: PMC4132180.
  29. Hermann BP, Sukhwani M, Hansel MC, Orwig KE. Spermatogonial stem cells in higher primates: are there differences from those in rodents? *Reproduction*. 2010; 139(3):479–93. Epub 2009/11/03. doi: [10.1530/REP-09-0255](https://doi.org/10.1530/REP-09-0255) PMID: [19880674](https://pubmed.ncbi.nlm.nih.gov/19880674/); PubMed Central PMCID: PMC2895987.
  30. Robson SL, Wood B. Hominin life history: reconstruction and evolution. *J Anat*. 2008; 212(4):394–425. doi: [10.1111/j.1469-7580.2008.00867.x](https://doi.org/10.1111/j.1469-7580.2008.00867.x) PMID: [18380863](https://pubmed.ncbi.nlm.nih.gov/18380863/); PubMed Central PMCID: PMC32409099.
  31. Amster G, Sella G. Life history effects on the molecular clock of autosomes and sex chromosomes. *BioRxiv*; 2015. <http://biorxiv.org/content/early/2015/08/10/024281>
  32. Salk JJ, Fox EJ, Loeb LA. Mutational heterogeneity in human cancers: origin and consequences. *Annu Rev Pathol*. 2010; 5:51–75. doi: [10.1146/annurev-pathol-121808-102113](https://doi.org/10.1146/annurev-pathol-121808-102113) PMID: [19743960](https://pubmed.ncbi.nlm.nih.gov/19743960/); PubMed Central PMCID: PMC3375045.
  33. Green P, Ewing B, Miller W, Thomas PJ, Green ED, Program NCS. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*. 2003; 33(4):514–7. doi: [10.1038/ng1103](https://doi.org/10.1038/ng1103) PMID: [12612582](https://pubmed.ncbi.nlm.nih.gov/12612582/).
  34. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*. 2010; 463(7278):191–6. Epub 2009/12/18. doi: [10.1038/nature08658](https://doi.org/10.1038/nature08658) PMID: [20016485](https://pubmed.ncbi.nlm.nih.gov/20016485/); PubMed Central PMCID: PMC3145108.
  35. Rowsey R, Gruhn J, Broman KW, Hunt PA, Hassold T. Examining variation in recombination levels in the human female: a test of the production-line hypothesis. *Am J Hum Genet*. 2014; 95(1):108–12. doi: [10.1016/j.ajhg.2014.06.008](https://doi.org/10.1016/j.ajhg.2014.06.008) PMID: [24995869](https://pubmed.ncbi.nlm.nih.gov/24995869/); PubMed Central PMCID: PMC34085639.
  36. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*. 2015; 350(6256):94–8. doi: [10.1126/science.aab1785](https://doi.org/10.1126/science.aab1785) PMID: [26430121](https://pubmed.ncbi.nlm.nih.gov/26430121/).
  37. Clancy B, Darlington RB, Finlay BL. Translating developmental time across mammalian species. *Neuroscience*. 2001; 105(1):7–17. PMID: [11483296](https://pubmed.ncbi.nlm.nih.gov/11483296/).
  38. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. *Nat Genet*. 2015; 47(12):1402–7. doi: [10.1038/ng.3441](https://doi.org/10.1038/ng.3441) PMID: [26551669](https://pubmed.ncbi.nlm.nih.gov/26551669/).