Evidence for the macroevolutionarily optimal mutation rate

Gordon Irlam
gordoni@gordoni.com

Los Altos, California, United States

October 19, 2024

Abstract

It has been hypothesized that evolution tunes spontaneous mutation rates in sexual species towards population optimal values. By comparing substitution rates at nonsynonymous and synonymous sites in the genomes of a number of related eukaryotic species pairs the approximate range of some key mutation related parameters is determined. For eukaryotes, the rate of occurrence of beneficial, or adaptive, mutational prospects that can subsequently mutate and fix is found to typically be in the range 10^{-3} to 10^{-2} population wide prospective sites per sexual generation. Deleterious, or negative, sites are typically created at the rate 10^{-1} to 10^{0} sites per haploid genome per sexual generation. And the spontaneous mutation rate is typically in the range 10^{-9} to 10^{-8} mutations per base pair per sexual generation. Using the values of these parameters, implied population sizes are computed based on the assumption of population optimal spontaneous mutation rates. Implied population sizes appear highly reasonable. This adds evidence to the hypothesis that macroevolution tunes spontaneous mutation rates towards population optimal values.

Keywords: beneficial mutation rate, rate of adaptive substitution, deleterious mutation rate, spontaneous mutation rate, optimal mutation rate.

Introduction

This manuscript is concerned with the determination of mutation rates in sexual eukaryotes.

It has been hypothesized that the spontaneous mutation rate is driven to the lowest value possible, taking into account the cost of producing such a low mutation rate [1, 2]. The argument is simple. There is a far greater likelihood of a mutation being harmful than beneficial, and so adaptations that reduce the mutation rate will be favored by microevolution. A contrast can be drawn between this and the drift-barrier hypothesis. The drift-barrier hypothesis argues that genetic drift acts to create a barrier that impedes the fixation of mutations that contribute only small improvements in replication fidelity [3, 4]. In both cases, the mutation rate is as small as possible, the difference is what determines the final rate.

What is good for the organism or the gene, isn't necessarily good for the population or the species. If the mutation rate is driven too low, the species won't be able to adapt, and will go extinct. Too high, and the species will bear a large deleterious mutational load. There thus exists a population optimal mutation rate between these two extremes. For sexual species it is possible to calculate this population optimal mutation rate [5]. Since microevolution favors the as-small-as-possible mutation rate, the only way species could achieve the population optimal mutation rate is through macroevolution. Those species with mutation rates that are too high or too low go extinct, leaving behind those species that are close to the population optimal mutation rate. The macroevolutionarily optimal mutation rate is the population optimal mutation rate.

The primary aim of this manuscript is to evaluate whether sexual species appear to favor the population optimal mutation rate. In order to do this, several mutation related parameters are estimated: the population wide rate of occurrence of beneficial mutational prospects that can subsequently mutate and fix, Γ_p, the per haploid genome rate of deleterious mutation, Γ_n, and the germline spontaneous mutation rate per base pair, μ_s. These rates are each expressed as rates per sexual generation. For a given genome, these values are not fully independent; Γ_n is given by the product of the length of the genome under the control of negative selection, L_n, and the spontaneous germline mutation rate, μ_s. Assuming all prospective beneficial sites eventually fix, Γ_p is the same as the rate of adaptive substitution. Γ_n differs from the genome wide rate of spontaneous mutation, U, by the exclusion of neutral and beneficial mutations.

By comparing substitution rates at nonsynonymous and synonymous sites in the genomes of relatively recently diverged sexual species with a known divergence time and a known time between sexual generations, the three mutation related rates will be estimated. It is then possible to compute implied population sizes: the population sizes which have the same optimal mutation rates as the rates estimated. The implied population sizes are found to be highly reasonable, adding support to the hypothesis that evolution tunes the mutation rate towards the population optimal mutation rate and towards maximal mean population fitness.

A secondary motivation for estimating the mutation related rates is with these rates in hand it should be possible to verify that they correspond to regions where sex offers a significant fitness advantage, and so validating our understanding of eukaryogenesis and the evolution of sex.

Results

This study seeks to determine the approximate values of the key mutation related parameters, not their precise values. For particular species others may have come up with more precise estimates of some of the parameter values, but this is not known to have been previously done with a consistent framework that spans species and parameter values. To simplify things it is assumed spontaneous mutations occur at random across the full length of the genome.

Method of parameter estimation

Consider two related species. Let A and S denote the number of nonsynonymous and synonymous sites that are either in common between two orthologous genes, or alternatively the coding portions of the genomes, of the two species. Let D_n and D_s be the number of nonsynonymous and synonymous substitutions occurring between the two genes or genomes. Let $K_a = \frac{D_n}{A}$ (also known as d_N) and $K_s = \frac{D_s}{S}$ (also known as d_S).

The spontaneous mutation rate

Let g be the generation time, and T be the time since the two species diverged. μ_s can be estimated from the genome wide synonymous mutation rate,

$$\mu_s = \frac{K_s g}{2T} \tag{1}$$

Positive selection

Let α be the fraction of nonsynonymous substitutions that are positively selected; as opposed to being neutral nonsynonymous substitutions. In the long run, assuming all prospective beneficial sites are eventually fixed, the rate at which beneficial mutational prospects occur is equal to the rate of adaptive substitution. Consequently, the per generation rate of beneficial mutational prospects is,

$$\Gamma_p = \frac{\alpha K_a A' g}{2T} \tag{2}$$

where A' is the value of A adjusted to take into account nonsynonymous sites in the genome that weren't analyzed.

Estimating alpha

It is possible to estimate α, if, in addition to estimates of the number of substitutions, data on polymorphisms are available [6, 7]. Table 1 shows a sampling of estimates for α. As can be seen

Species	Estimate of α	Year	Source
Homo sapiens and old world monkeys	0.35	2001	[8]
Homo sapiens and *Pan troglodytes*	0.10 - 0.13	2007	[9]
Homo sapiens and *Pan troglodytes*	0.10 - 0.20	2008	[10]
Homo sapiens and macaques	0 or 0.31 - 0.40	2009	[11]
Mus musculus and *Mus famulus* or *Rattus*	0.57	2010	[12]
Drosophila simulans and *Drosophila yakuba*	0.45	2002	[6]
Drosophila melanogaster and *Drosophila simulans*	0.29	2007	[13]
Drosophila americana and *Drosophila ezoana*	0.57	2007	[14]
Drosophila miranda and *Drosophila pseudoobscura*	0.44 - 0.61	2008	[15]
Drosophila melanogaster and *Drosophila simulans*	0.52	2009	[11]
44 animal species pairs	0.58 ± 0.20	2016	[16][Supplement 1]
9 out of 10 plant species	0.0	2010	[17]

Table 1: Some estimates of α the fraction of positively selected nonsynonymous substitutions.

there is considerable variability regarding the value of α. For animals, α has an approximate mean value of 0.6 and a standard deviation of 0.2. For plants it appears indistinguishable from zero.

Negative selection

Random deleterious mutations to the genome can be corrected either by negative, or purifying selection, or by back mutation. Back mutation is likely to be rare, so we focus on purifying selection. We model deleterious mutations as occurring at a rate Γ_n per haploid genome per generation.

Let G be the number of genes in the genome, L be the length of the genome in base pairs, and l_n be the fraction of L under the control of negative selection. Define, n_G, the average number of negatively selected base pairs per gene by,

$$n_G = \frac{l_n L}{G} \tag{3}$$

Since deleterious mutations can occur outside of coding regions, n_G isn't particularly physically meaningful. It is however useful.

The per generation rate at which deleterious mutations are occurring is given by,

$$\begin{aligned} \Gamma_n &= l_n L \mu_s \\ &= \frac{n_G G K_s g}{2T} \text{ by equation 1 and 3} \end{aligned} \tag{4}$$

l_n isn't always known, so equation 4 is in terms of n_G, which is assumed to be relatively constant for relatively similar species.

Implied population sizes assuming population optimal mutation rates

Let N be the organism population size. Assuming evolution drives the mutation rate towards the value that produces the maximum population mean fitness, the spontaneous mutation rate will be the population optimal mutation rate [5],

$$\mu_s \approx \frac{3\Gamma_p}{2P^2 N \Gamma_n h\Pi}$$

where P is the ploidy (haploids 1, diploids 2) and $h\Pi$ is the geometric mean value of the beneficial mutation dominance coefficient multiplied by a factor describing the reduction in the probability of beneficial fixation due to linked and unlinked sites in the genome.

This equation can be trivially rearranged to give N_i, the population size implied by μ_s, Γ_p, and Γ_n, assuming the population optimal mutation rate,

$$N_i \approx \frac{3\Gamma_p}{2P^2 \Gamma_n \mu_s h\Pi} \tag{5}$$

Alternatively, using equations 1, 2, and 4,

$$N_i \approx \frac{3\alpha K_a A' T}{P^2 n_G G K_s^2 g h\Pi} \tag{6}$$

The behavior of N_i when N isn't fixed, but varies over time, is interesting. The establishment time is the average time from when a beneficial mutation first occur, until it has a high likelihood of fixing. If the time period over which variation takes place is short in comparison to the establishment time, then the genetic variability of the population is reduced. This reduces the probability of fixation. The resulting effect will depend on the nature of the changes in population size. For a cyclically varying population, N_i will be the population harmonic mean [18]. If the time period over which variation takes place is long in comparison to the establishment time, then the aggregate rate of beneficial fixation is proportional to the sum of the population values. N_i will be the population arithmetic mean.

Parameter estimation

Comparison of genomes

To estimate Γ_p, Γ_n, and μ_s, a number of relatively recently diverged species pairs were chosen. Species were selected based on the availability of sequenced genomes, availability of estimates of divergence times, and availability of organism generation times. The pairs need to have diverged relatively recently so that the average generation time is meaningful. It is assumed that the number of within species polymorphisms is small in comparison to the number of between species substitutions, so that all differences between genomes can be considered to represent substitutions. Many

Species	G	Genes aligned	A	K_a	K_s	$\frac{K_a}{K_s}$	A'
Homo sapiens and *Pan troglodytes*	19,932	17,678	2.2×10^7	0.0034	0.014	0.25	2.7×10^7
Mus musculus and *Rattus norvegicus*	22,517	18,162	2.2×10^7	0.031	0.19	0.16	2.7×10^7
Gallus gallus and *Phasianus colchicus*	16,248	14,312	1.9×10^7	0.017	0.11	0.16	2.1×10^7
Xenopus laevis and *Xenopus tropicalis*	21,885	17,864	2.3×10^7	0.049	0.29	0.17	2.8×10^7
Oryzias latipes and *Nothobranchius furzeri*	22,145	17,031	2.2×10^7	0.11	0.98	0.11	2.9×10^7
Drosophila simulans and *Drosophila yakuba*	14,217	12,854	1.6×10^7	0.035	0.29	0.12	1.8×10^7
Plasmodium vivax and *Plasmodium gonderi*	5,389	3,343	3.1×10^6	0.15	4.17	0.035	8.3×10^6
Arabidopsis thaliana and *Camelina sativa*	27,271	21,557	2.0×10^7	0.055	0.27	0.20	2.5×10^7
Elaeis guineensis and *Cocos nucifera*	26,295	16,665	1.6×10^7	0.033	0.11	0.30	2.4×10^7
Populus trichocarpa and *Hevea brasiliensis*	31,543	15,327	1.5×10^7	0.13	0.74	0.17	3.3×10^7

Table 2: Genome wide estimates of coding sites and substitution rates. G is the lesser number of genes of the two species. A is the numbers of nonsynonymous sites. K_a and K_s are the number of substitutions per nonsynonymous and synonymous site respectively. A' is the number of nonsynonymous sites adjusted for genes and segments of genes that weren't analyzed.

of the selected species are model organisms. Model organisms are often proposed on the basis of their short generation times. This might introduce a slight bias leading to underestimates of typical values for Γ_p, Γ_n, and μ_s.

For each species pair orthologous genes were identified using protein-protein BLAST to determine reciprocal best hits. Protein sequences of orthologous genes were then aligned using the Needleman-Wunsch algorithm. Aligned protein sequences were mapped back to aligned nucleotide sequences. Genes containing tandem repeat regions were excluded. Estimates of A, S, K_a, and K_s for each gene pair were made. Genome wide estimates of A and S were computed as the sums of individual gene pairs, and genome wide estimates of K_a and K_s computed as A and S weighted averages of the values of the individual gene pairs. Estimates of Γ_p, Γ_n, and μ_s are likely underestimates on account of regions of the genome with the greatest variability not aligning and being excluded from the analysis. This becomes increasingly significant for longer divergence times. See Materials and methods for a more detailed description of the methodology. The results are shown in Table 2.

The K_a and K_s values for humans and chimps of 0.0034 and 0.014 compare reasonably well to previously reported values of 0.0029 and 0.013 respectively [19][Supplement S23, site weighted K_a and K_s values divided by $2T$].

To account for unanalyzed genes the value of A was multiplied by the smaller of the coding sequence sizes for the two species divided by the analyzed coding sequence size giving A'. Use of the smaller coding sequence size is important because while most analyzed species are diploid, *Xenopus laevis* is tetraploid [20], and *Camelina sativa* is hexaploid [21]. Polyploidy is also the reason for using the lesser of the two gene counts for G in preference to the mean value.

Estimating the number of negatively selected base pairs per gene

To compute Γ_n we need to estimate n_G, the average number of base pairs under the control of negative selection per gene:

Species	T	g	α	n_G	Γ_p	Γ_n	μ_s
Homo sapiens and *Pan troglodytes*	6.7×10^6	25	0.15	8,400	0.026	4.3	2.5×10^{-8}
Mus musculus and *Rattus norvegicus*	20.9×10^6	0.5	0.57	8,400	0.0057	0.44	2.3×10^{-9}
Gallus gallus and *Phasianus colchicus*	34.1×10^6	1	0.58	8,400	0.0030	0.22	1.6×10^{-9}
Xenopus laevis and *Xenopus tropicalis*	64.0×10^6	3	0.58	8,400	0.019	1.3	6.9×10^{-9}
Oryzias latipes and *Nothobranchius furzeri*	93.0×10^6	1.5	0.58	8,400	0.015	1.5	7.9×10^{-9}
Drosophila simulans and *Drosophila yakuba*	11.4×10^6	0.1	0.45	4,400	0.0012	0.080	1.3×10^{-9}
Plasmodium vivax and *Plasmodium gonderi*	9.5×10^6	0.18	0.58	1,300	0.0068	0.28	4.0×10^{-8}
Arabidopsis thaliana and *Camelina sativa*	9.4×10^6	0.2	0.10	780	0.0015	0.062	2.9×10^{-9}
Elaeis guineensis and *Cocos nucifera*	43.0×10^6	50	0.10	780	0.046	1.3	6.5×10^{-8}
Populus trichocarpa and *Hevea brasiliensis*	80.0×10^6	25	0.10	780	0.066	2.9	1.2×10^{-7}

Table 3: Estimates of mutation related rates. T is the time since the species diverged in years. g is the estimated generation time in years. α is the proportion of substitutions that are adaptive. n_G is the mean number of base pairs under negative selection per gene. Γ_p is the rate of positive site creation per generation. Γ_n is the rate of negative site creation per haploid genome per generation. μ_s is the germline rate of mutation per base pair per generation.

- For humans $l_n = 0.054$ [22], $L = 3.1\times10^9$, and from Table 2, G = 19,932. By equation 3 this results in a value for n_G of 8,400. For many other species n_G is unknown. Not knowing any better, we assume that the same value of n_G applies to other vertebrate species.
- For *Drosophila* spp., $l_n = 0.45$ is the mid-point estimate of 37-53% [23], $L = 1.4\times10^8$, and G = 14,217, giving n_G = 4,400.
- For *Arabidopsis thaliana*, $l_n = 0.177$ [24], $L = 1.2\times10^8$, and G = 27,271, giving n_G = 780. The same value is assumed to apply to other plant species.
- Estimating n_G for *Plasmodium* spp. is more difficult. Like *A. thaliana*, *Plasmodium* spp. have a compact genome, hence a rough approximation can be obtained by taking the n_G value for *A. thaliana*, and scaling it by the relative mean gene coding sequence lengths; 1,230 for *A. thaliana*, and 2,090 for *Plasmodium vivax*. This gives n_G = 1,300.

Calculation of the mutation related rates

In computing Γ_p, estimates of α are based on Table 1, except that for plants we assume an α of 0.1 rather than the somewhat implausible value of 0.0. An α of 0.1 appears within the 95% confidence interval of 8 of the 10 plants reported by Gossmann et al. [17]. The correct value of α for *Plasmodium* spp. is particularly uncertain. Using equations 1, 2, and 4, it is possible to compute estimates for Γ_p, Γ_n, and μ_s. The results are presented in Table 3.

For the considered species pairs excluding plants, Γ_p is in the range 0.0012 to 0.026. Γ_n is in the range 0.080 to 4.3. And, also excluding *Plasmodium* spp., μ_s ranges from 1.3×10^{-9} to 2.5×10^{-8}.

Comparison of mutation related rates

The estimate $\Gamma_p = 0.0012$ for *Drosophila simulans* and *Drosophila yakuba* can be compared to a published estimate of a rate of adaptive substitution of 0.0022 for the same species based on an

Species	μ_s	Source
Homo sapiens	2.0×10^{-8}	[25][Table 4, $C\mu_b$]
Homo sapiens	2.5×10^{-8}	[26]
Mus musculus	3.8×10^{-9}	[27]
Mus musculus	5.7×10^{-9}	[28][mean value]
Mus musculus	1.1×10^{-8}	[25][Table 4, $C\mu_b$]
Drosophila melanogaster	2.8×10^{-9}	[29]
Drosophila melanogaster	3.5×10^{-9}	[30]
Drosophila melanogaster	8.5×10^{-9}	[25][Table 4, $C\mu_b$]
Arabidopsis thaliana	7.0×10^{-9}	[31]

Table 4: Estimates of the spontaneous mutation rate, μ_s, by various authors. μ_s is the rate of mutation per base pair per sexual generation.

analysis of 35 genes [6].

Some of the reported values for μ_s can be compared to estimates by others shown in Table 4. While similar in magnitude, a discrepancy of more than a factor of two exists between the estimate made here and the lowest value reported by others for *Drosophila* spp. and *A. thaliana*.

Plasmodium has a high spontaneous mutation rate per sexual generation. The average mutation rate is reported elsewhere to be 1.7×10^{-9} per generation [32]. The *Plasmodium* life cycle involves at least 200 generations per year [33] but only one sexual generation every 65 days [34]. This reported value would imply a spontaneous mutation rate per sexual generation of at least 6.1×10^{-8}. This compares to the estimate made here of 4.0×10^{-8}. This high rate of spontaneous mutation may go some way to explaining why *Plasmodium* is able to rapidly evolve drug resistance [35].

Implied population sizes assuming population optimal mutation rates

The mean value of h for deleterious mutations isn't known with great precision, but it might be around 0.2 or 0.3 [36], suggesting a value of h for beneficial mutations equal to 1 minus this value, or 0.7 to 0.8. Π is likely close to 1 [5]. The resulting values for N_i, computed using equation 6 and assuming $h\Pi = 0.7$ are shown in Table 5.

If the hypothesis that spontaneous mutation rates are evolutionarily tuned towards producing the population optimal fitness was false we could get all sorts of random values out of equation 6. We don't. Most of the values are around 10^5 to 10^6, which seems quite reasonable. Moreover, the implied population size for hominids is on the small side, and that for *Drosophila* spp. is on the high side, agreeing with an intuitive inverse correlation between organism size and population density and thus population size [37]. Thus the hypothesis appears to be a reasonably good hypothesis. Not to much weight should be placed on the results for plants, due to the use of a synthetic value for α,

By considering the heterozygosity of synonymous sites Lynch et al. have estimated effective population sizes for a number of species as shown in Table 6. The values reported here are roughly 10 times larger than this. It must however be remembered that effective populations are perhaps

Species	N_i
Homo sapiens and *Pan troglodytes*	1.3×10^5
Mus musculus and *Rattus norvegicus*	3.0×10^6
Gallus gallus and *Phasianus colchicus*	4.8×10^6
Xenopus laevis and *Xenopus tropicalis*	1.2×10^6
Oryzias latipes and *Nothobranchius furzeri*	7.0×10^5
Drosophila simulans and *Drosophila yakuba*	6.4×10^6
Plasmodium vivax and *Plasmodium gonderi*	1.3×10^6
Arabidopsis thaliana and *Camelina sativa*	4.4×10^6
Elaeis guineensis and *Cocos nucifera*	2.9×10^5
Populus trichocarpa and *Hevea brasiliensis*	1.1×10^5

Table 5: Estimates of N_i, the implied population size assuming the population optimal mutation rate.

Species	N_e	Source
Homo sapiens	2.1×10^4	[4][Supplement S1]
Pan troglodytes	2.9×10^4	[4][Supplement S1]
Mus musculus	1.8×10^5	[4][Supplement S1]
Drosophila melanogaster	8.6×10^5	[4][Supplement S1]
Plasmodium falciparum	3.5×10^5	[4][Supplement S1]
Arabidopsis thaliana	2.9×10^5	[4][Supplement S1]
vertebrates	$\sim10^4$	[38]
trees	$\sim10^4$	[38]

Table 6: Estimates of the effective population size, N_e, from the heterozygosity of synonymous sites.

10 times smaller than mean actual populations on account of population bottlenecks, expansions, and other factors that reduce genetic variability [38, 39]. By way of contrast, the implied population is immune to long term population fluctuations. Population sizes are thus consistent with the mutation rate being tuned towards the population optimal mutation rate.

Because tuning of the mutation rate likely only occurs over macroevolutionary time frames, equation 5 might only be expected to be valid in the big macroevolutionary picture. It probably can't be used to predict the average population of an individual species with great accuracy.

The implied population for the predominantly haploid *Plasmodium* spp. is 1.3×10^6, which may on first glance appear smaller than expected, and likely much smaller than the actual population. It should be remembered that *Plasmodium* undergoes severe constraints on random mating coupled with frequent severe population bottlenecks by virtue of it being constrained to individual host organisms. Frequent population bottlenecks are likely to significantly reduce the implied population size. A small implied population size and a small size of the portion of the genome under the control of negative selection might explain why *Plasmodium* has such a high rate of spontaneous mutation. Along with the ploidy, these two terms appear as the denominator in the formula for the optimal mutation rate [5],

$$\mu_s \approx \sqrt{\frac{3\Gamma_p}{2P^2 N L_n h \Pi}} \tag{7}$$

Discussion

Genomic analysis suggests that for sexual species Γ_p is typically in the range 10^{-3} to 10^{-2} population wide beneficial mutational prospects per sexual generation. Γ_n is typically in the range 10^{-1} to 10^{0} deleterious sites per haploid genome per sexual generation. And μ_s is typically in the range 10^{-9} to 10^{-8} spontaneous mutations per base pair site per sexual generation. This is the area of the parameter space of interest when seeking to understand the evolution of sex.

Holding the other parameters constant, equation 7 predicts an inverse relationship between the population size and the mutation rate. The empirical evidence of a sub-unity negative relationship between the log effective population size and the log mutation rate thus supports the hypothesis that evolution tunes mutation rates towards population optimal values [4].

In this manuscript, the estimated values of the mutation related parameters were used to compute implied population sizes under the assumption of optimal mutation rates. The implied population sizes appear highly reasonable. This suggests that macroevolution has tuned the spontaneous mutation rate towards the population optimal mutation rate, that is the macroevolutionarily optimal mutation rate. This is the mutation rate which produces the optimal population mean fitness.

Materials and methods

See File S1 for the bioinformatics analysis code and results of the gene by gene analysis of each species pair [40].

The coding sequences (CDS) of genes from sequenced species genomes were obtained from NCBI. Genomes were filtered to remove any non-nuclear genes, documented pseudogenes, or duplicate protein isoforms. At this stage multiple distinct protein isoforms of each gene were potentially present. This is important. When seeking to assess the divergence of closely related genomes, filtering out shorter isoforms of each gene might have resulted in the longest isoform of orthologous genes in each species having few exons directly in common but still being related. This would have led to undercounting of orthologous sites and misestimating substitution rates.

Reciprocal best hits were computed using protein-protein BLAST [41] using the task blastp-fast, the BLOSUM90 matrix, soft masking of low complexity regions, a minimum expect value of 10^{-6}, and reporting the single best high-scoring segment pair. For each gene of the first species having multiple isoforms with reciprocal best hits, the reciprocal best hit of the isoform with the highest number of matching residues was then selected. This resulted in orthologous gene pairs.

Protein sequences were aligned to each other using the Needleman-Wunsch algorithm [42] implemented by Biopython PairwiseAligner [43]. Scores of match 5, mismatch -8, gap open -50, end gap open -25, gap extend -2, and end gap extend -1, were found to do a good job of conservatively pre-

dicting the alignments for most orthologous gene pairs examined. Nucleotide sequences were then aligned based on the protein sequences. Stop residues '*', unknown nucleotides 'X', and leading gaps '-' were not aligned, but the remainder of the protein sequence was aligned.

Mutation rates may vary over the genome. For each orthologous gene pair PAML's CODEML [44] was used to count the number of synonymous and nonsynonymous sites and estimate the number of substitutions that have occurred. Rarely CODEML failed to deliver a result within its search bounds, producing an estimate of 50.0 for the separation time. Such results were excluded from the analysis.

Genes containing tandem repeats may bias the analysis due to tandem repeat copy number polymorphism and random gene conversion within such genes. So Tandem Repeat Finder [45] was used on the nucleotide sequence to exclude such genes from the analysis. Default recommended parameter values were used, match 2, mismatch 7, delta (indels) 7, PM 80, PI 10, minscore 50, except maxperiod was increased from 500 to 2,000 to ensure identifying the human FLG gene which has repeats of period 972.

Estimates of K_a and K_s are highly variable for small numbers of residues. Only genes having an aligned protein length of 20 residues or more and with a Tandem Repeat Finder score of 100 or less were selected for inclusion in the resulting analysis. A score of more than 100 corresponds to more than 16 residues making up any discovered total tandem repeat region length.

Genome wide estimates of A and S were produced by summing the individual gene values. K_a and K_s for the genome were computed as the A and S weighted averages of the individual gene values.

Divergence times were estimated using TimeTree [46].

Noting that laboratory lifespans are typically longer than wild lifespans, generation times were estimated as follows. *Homo sapiens* and *Pan troglodytes*, 25 years [47]. *Mus musculus* and *Rattus norvegicus*, 6 months based on reproductive life spans of 7 to 8 months and 12 to 15 months respectively. *Gallus gallus* and *Phasianus colchicus*, 1 year for *Gallus* [48]. *Xenopus laevis* and *Xenopus tropicalis*, 3 years based on sexual maturity of 12 to 18 months and 4 to 6 months and laboratory lifespans of 10 to 15 years and 10 years respectively [49, 50, 51]. *Oryzias latipes* and *Nothobranchius furzeri*, 1.5 years, based on 50% laboratory mortality after 3 years and exactly 1 year respectively [52, 53]. *Drosophila simulans* and *Drosophila yakuba*, 0.1 years [6]. *Plasmodium vivax* and *Plasmodium gonderi* 65 days, based on *Plasmodium falciparum* and *Plasmodium reichenowi* [34]. *Arabidopsis thaliana* and *Camelina sativa* 0.2 year, based on a life cycle of as little as 6 weeks and a crop season of 85-100 days respectively. *Elaeis guineensis* and *Cocos nucifera* 50 years, based on economic lifespan of 30 years and lifespan of 100 years or more and a lifespan of 80 to 100 years respectively. *Populus trichocarpa* and *Hevea brasiliensis* 25 years, based on being suitable for timber production after 25 years and an economic life of 25 years respectively [54]. It should be noted that *P. trichocarpa* and *H. brasiliensis* are both Malpighailes, but so are annuals of the genus *Viola*. If the ancestors of these trees were short lived, the generation time would be much smaller.

Accession numbers of genomes used in this study are shown in Table 7.

Species	Accession
Homo sapiens	GCF_000001405.39
Pan troglodytes	GCF_002880755.1
Mus musculus	GCF_000001635.27
Rattus norvegicus	GCF_015227675.2
Gallus gallus	GCF_000002315.5
Phasianus colchicus	GCF_004143745.1
Xenopus tropicalis	GCF_000004195.4
Xenopus laevis	GCF_017654675.1
Oryzias latipes	GCF_002234675.1
Nothobranchius furzeri	GCF_001465895.1
Drosophila simulans	GCF_016746395.2
Drosophila yakuba	GCF_016746365.2
Plasmodium vivax	GCF_000002415.2
Plasmodium gonderi	GCF_002157705.1
Arabidopsis thaliana	GCF_000001735.4
Camelina sativa	GCF_000633955.1
Elaeis guineensis	GCF_000442705.1
Cocos nucifera	GCA_008124465.1
Populus trichocarpa	GCF_000002775.4
Hevea brasiliensis	GCF_001654055.1

Table 7: Accession numbers of genomes used in this study.

Data availability

The code and results from this study are publicly available.

File S1 - Mutation rate analysis software and results.
`https://doi.org/10.5281/zenodo.13956219`

Conflict of interest statement

The author declares they have no conflicts of interest in relation to the content of this manuscript.

References

[1] Egbert Giles Leigh Jr. Natural selection and mutability. *The American Naturalist*, 104(937):301–305, 1970.

[2] Paul D Sniegowski, Philip J Gerrish, Toby Johnson, and Aaron Shaver. The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066, 2000.

[3] Michael Lynch. The lower bound to the evolution of mutation rates. *Genome Biology and Evolution*, 3:1107–1118, 2011.

[4] Michael Lynch, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, 2016.

[5] Gordon Irlam. An integrated model of the effects on fitness of beneficial, detrimental, and neutral mutations in sexual populations. *BioRxiv*, 2023. `doi:10.1101/2023.06.26.546616`.

[6] Nick GC Smith and Adam Eyre-Walker. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–1024, 2002.

[7] Ana Filipa Moutinho, Thomas Bataillon, and Julien Y Dutheil. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, 34(3):315–338, 2020.

[8] Justin C Fay, Gerald J Wyckoff, and Chung-I Wu. Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234, 2001.

[9] Jun Gojobori, Hua Tang, Joshua M Akey, and Chung-I Wu. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings of the National Academy of Sciences*, 104(10):3907–3912, 2007.

[10] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLOS Genetics*, 4(5):e1000083, 2008.

[11] Adam Eyre-Walker and Peter D Keightley. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, 26(9):2097–2108, 2009.

[12] Daniel L Halligan, Fiona Oliver, Adam Eyre-Walker, Bettina Harr, and Peter D Keightley. Evidence for pervasive adaptive protein evolution in wild mice. *PLOS Genetics*, 6(1):e1000825, 2010.

[13] Joshua A Shapiro, Wei Huang, Chenhui Zhang, Melissa J Hubisz, Jian Lu, David A Turissini, Shu Fang, Hurng-Yi Wang, Richard R Hudson, Rasmus Nielsen, et al. Adaptive genic evolution in the Drosophila genomes. *Proceedings of the National Academy of Sciences*, 104(7):2271–2276, 2007.

[14] Xulio Maside and Brian Charlesworth. Patterns of molecular variation and evolution in Drosophila americana and its relatives. *Genetics*, 176(4):2293–2305, 2007.

[15] Doris Bachtrog. Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. *BMC Evolutionary Biology*, 8(1):1–14, 2008.

[16] Nicolas Galtier. Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics*, 12(1):e1005774, 2016.

[17] Toni I Gossmann, Bao-Hua Song, Aaron J Windsor, Thomas Mitchell-Olds, Christopher J Dixon, Maxim V Kapralov, Dmitry A Filatov, and Adam Eyre-Walker. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, 27(8):1822–1832, 2010.

[18] Sarah P Otto and Michael C Whitlock. The probability of fixation in populations of changing size. *Genetics*, 146(2):723–733, 1997.

[19] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[20] Adam M Session, Yoshinobu Uno, Taejoon Kwon, Jarrod A Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, Akira Hikosaka, Atsushi Suzuki, Mariko Kondo, et al. Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, 538(7625):336–343, 2016.

[21] Sateesh Kagale, Chushin Koh, John Nixon, Venkatesh Bollina, Wayne E Clarke, Reetu Tuteja, Charles Spillane, Stephen J Robinson, Matthew G Links, Carling Clarke, et al. The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, 5(1):3706, 2014.

[22] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[23] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[24] Annabelle Haudry, Adrian E Platts, Emilio Vello, Douglas R Hoen, Mickael Leclercq, Robert J Williamson, Ewa Forczek, Zoé Joly-Lopez, Joshua G Steffen, Khaled M Hazzouri, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8):891–898, 2013.

[25] John W Drake. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences*, 870(1):100–107, 1999.

[26] Michael W Nachman and Susan L Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[27] Adeolu B Adewoye, Sarah J Lindsay, Yuri E Dubrova, and Matthew E Hurles. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nature Communications*, 6(1):6684, 2015.

[28] Brandon Milholland, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8(1):1–8, 2017.

[29] Peter D Keightley, Rob W Ness, Daniel L Halligan, and Penelope R Haddrill. Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics*, 196(1):313–320, 2014.

[30] Peter D Keightley, Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L Blaxter. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19(7):1195–1201, 2009.

[31] Mao-Lun Weng, Claude Becker, Julia Hildebrandt, Manuela Neumann, Matthew T Rutter, Ruth G Shaw, Detlef Weigel, and Charles B Fenster. Fine-grained analysis of spontaneous mutation spectrum and frequency in Arabidopsis thaliana. *Genetics*, 211(2):703–714, 2019.

[32] Selina ER Bopp, Micah J Manary, A Taylor Bright, Geoffrey L Johnston, Neekesh V Dharia, Fabio L Luna, Susan McCormack, David Plouffe, Case W McNamara, John R Walker, et al. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLOS Genetics*, 9(2):e1003293, 2013.

[33] Dorothy E Loy, Weimin Liu, Yingying Li, Gerald H Learn, Lindsey J Plenderleith, Sesh A Sundararaman, Paul M Sharp, and Beatrice H Hahn. Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *International Journal for Parasitology*, 47(2-3):87–97, 2017.

[34] Gavin G Rutledge, Ulrike Böhme, Mandy Sanders, Adam J Reid, James A Cotton, Oumou Maiga-Ascofare, Abdoulaye A Djimdé, Tobias O Apinjoh, Lucas Amenga-Etego, Magnus Manske, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature*, 542(7639):101–104, 2017.

[35] Toshihiro Mita and Kazuyuki Tanabe. Evolution of Plasmodium falciparum drug resistance: implications for the development and containment of artemisinin resistance. *Japanese Journal of Infectious Diseases*, 65(6):465–475, 2012.

[36] Chenlu Di and Kirk E Lohmueller. Revisiting dominance in population genetics. *Genome Biology and Evolution*, 16(8):evae147, 2024.

[37] David J Currie. What shape is the relationship between body size and population density? *Oikos*, pages 353–358, 1993.

[38] Michael Lynch. The origins of eukaryotic gene structure. *Molecular Biology and Evolution*, 23(2):450–468, 2006.

[39] Richard Frankham. Effective population size/adult population size ratios in wildlife: a review. *Genetics Research*, 66(2):95–107, 1995.

[40] Gordon Irlam. Evidence for the macroevolutionarily optimal mutation rate: Supplement 1, 2024. `doi:10.5281/zenodo.13956219`.

[41] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[42] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[43] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[44] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[45] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.

[46] Sudhir Kumar, Glen Stecher, Michael Suleski, and S Blair Hedges. Timetree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819, 2017.

[47] Adam Eyre-Walker and Peter D Keightley. High genomic deleterious mutation rates in hominids. *Nature*, 397(6717):344–347, 1999.

[48] Mahendra Mariadassou, Marie Suez, Sanbadam Sathyakumar, Alain Vignal, Mariangela Arca, Pierre Nicolas, Thomas Faraut, Diane Esquerré, Masahide Nishibori, Agathe Vieaud, et al. Unraveling the history of the genus Gallus through whole genome sequencing. *Molecular Phylogenetics and Evolution*, 158:107044, 2021.

[49] Souqi Liao, Wenyan Dong, Hui Zhao, Ruijin Huang, Xufeng Qi, and Dongqing Cai. Cardiac regeneration in Xenopus tropicalis and Xenopus laevis: discrepancies and problems. *Cell & Bioscience*, 8(1):1–3, 2018.

[50] Barney Reed. Good practice guidance for the housing and care of Xenopus iaevis. *Animal Technology and Welfare*, 8(3):137, 2009.

[51] Thomas Naert, Dionysia Dimitrakopoulou, Dieter Tulkens, Suzan Demuynck, Marjolein Carron, Rivka Noelanders, Liza Eeckhout, Gert Van Isterdael, Dieter Deforce, Christian Vanhove, et al. RBL1 (p107) functions as tumor suppressor in glioblastoma and small-cell pancreatic neuroendocrine carcinoma in Xenopus tropicalis. *Oncogene*, 39(13):2692–2706, 2020.

[52] Nobuo Egami. Further notes on the life span of the teleost, Oryzias latipes. *Aging in Cold Blooded Animals*, page 118, 1974.

[53] Eva Terzibasi Tozzini, Alexander Dorn, Enoch Ng'oma, Matej Polačik, Radim Blažek, Kathrin Reichwald, Andreas Petzold, Brian Watters, Martin Reichard, and Alessandro Cellerino. Parallel evolution of senescence in annual fishes in response to extrinsic mortality. *BMC Evolutionary Biology*, 13(1):1–12, 2013.

[54] Yi Peng Teoh, Mashitah Mat Don, and Salmiah Ujang. Assessment of the properties, utilization, and preservation of rubberwood (hevea brasiliensis): a case study in malaysia. *Journal of Wood Science*, 57(4):255–266, 2011.