Supporting Information for: "The Rate of Observable Molecular Evolution When Mutation May Not Be Weak"

Jason de Koning and Bianca De Sanctis

February 2, 2018

1 SI Methods

1.1 Introducing an initial configuration

Especially as population-scaled mutation rates grow large, it is important to relax the third weak mutation assumption that the population starts with only one mutant allele, $p = 1$. To do this, we allow an initial frequency distribution across the population, $f(p)$, such that a fraction of the population initially exists in each state, p, according to the probabilities of generating p mutations in one generation under the Wright-Fisher Markov model. This can be accomplished with

$$\begin{aligned} k(\mathbf{p}) &= \frac{\sum_{p=1}^{2N-1} w_p P^*_{\text{Fix}}(p)}{\sum_{p=1}^{2N-1} w_p T_\mu(p) + \sum_{p=1}^{2N-1} w_p T^*_{\text{Abs}}(p)} \\ &= \frac{P^*_{\text{Fix}}(\mathbf{p})}{T_\mu(\mathbf{p}) + T^*_{\text{Abs}}(\mathbf{p})} \\ &= \frac{P^*_{\text{Fix}}(\mathbf{p})}{\frac{1}{1-P(0,0)} + T^*_{\text{Abs}}(\mathbf{p})} \end{aligned} \tag{1}$$

where,

$$w_p = \frac{P(0,p)}{1-P(0,0)} \tag{2}$$

and

$$P(0,p) = \binom{2N}{p} \mu^p (1-\mu)^{2N-p} \tag{3}$$

with a forward mutation rate of μ. Each integration over p in equation 1 can be truncated once the probability $P(0,p)$ becomes small enough to be considered negligible. When using direct, sparse matrix methods(1), these computations take trivially longer than when assuming $p = 1$, since the vast majority of computational time is spent computing the transition matrix and performing an LU decomposition, which do not change with p. Numerical results from this approach match very closely to an exact computational approach, with which the substitution rate can be calculated directly from the Wright-Fisher transition matrix (Extended Data Fig. 2; see Methods, main text).

It is interesting to note that Kimura's formulation uses the approximation $T_\mu = 1/(2N\mu)$ in place of the correct value under the Wright-Fisher model, which when $p = 1$ is $(1 - P(0,0))/P(0,1)$. While these quantities are close to each other for small mutation rates, they become different as mutation rates are increased.

1.2 Recovering the weak mutation model as a special case of the rate of observable evolution

As noted above, Kimura assumed that $T_\mu = 1/(2N\mu)$ generations. By introducing this assumption, equation 3 in the main text can be rewritten as

$$k = \frac{2N\mu P^*_{\text{Fix}}}{1 + 2N\mu T^*_{\text{Abs}}}. \tag{4}$$

Here we see the weak-mutation substitution rate in the numerator (with the addition of the effect of mutation incorporated in P^*_{Fix}), and a retardation factor in the denominator. Compatible with Kimura's assumptions, when $2N\mu T^*_{\text{Abs}}$ is small, the contribution of the retardation factor will be negligible and the weak-mutation rate of evolution will agree with the rate of observable evolution. We can now reintroduce Kimura's weak mutation assumptions that: 1) no additional mutations can be generated during the segregation of an initial mutant by substituting P_{Fix} for P^*_{Fix}; and 2) that absorption times are negligible compared to mutation times by making absorptions instantaneous and taking T^*_{Abs} to 0. Doing this, we see the substitution rate simplifies to Kimura's rate,

$$\begin{aligned} k &= \frac{2N\mu P_{\text{Fix}}}{1 + 2N\mu \cdot 0} \\ &= 2N\mu P_{\text{Fix}} \end{aligned}$$

1.3 Alternative derivation of the rate of evolution

Let M be the time until the first fixation, and F_i be the event where the ith mutation-absorption cycle is the first one to fix. We wish to derive an expression for $E(M)$, since the expected substitution rate will then be $k = 1/E(M)$. To begin, we condition on when the fixation occurs.

$$E(M) = \sum_{i=1}^{\infty} E(M|F_i)P(F_i)$$

where $E(M|F_i)$ is the expected time to fixation, given that fixation occurs on the cycle i. In writing an expression for $E(M|F_i)$, we need to account for the time it takes i mutations to arise, the time it takes $i-1$ mutations to go extinct once arisen, and the time it takes for 1 mutation to fix once arisen. Let T_μ be the mean number of generations for a mutation to arise. This yields

$$E(M|F_i) = iT_\mu + (i-1)T^*_{\text{Ext}} + T^*_{\text{Fix}}$$

Since a single mutation fixes with probability P^*_{Fix},

$$P(F_i) = P^*_{\text{Fix}}(1 - P^*_{\text{Fix}})^{i-1}$$

Using both of the above formulas in our earlier expression for $E(M)$, we get

$$\begin{aligned} E(M) &= \sum_{i=1}^{\infty} \left(iT_\mu + (i-1)T^*_{\text{Ext}} + T^*_{\text{Fix}}\right) P^*_{\text{Fix}}(1 - P^*_{\text{Fix}})^{i-1} \\ &= (T_\mu + T^*_{\text{Ext}}) \left(\sum_{i=1}^{\infty} iP^*_{\text{Fix}}(1 - P^*_{\text{Fix}})^{i-1} \right) + (T^*_{\text{Fix}} - T^*_{\text{Ext}}) \left(\sum_{i=1}^{\infty} P^*_{\text{Fix}}(1 - P^*_{\text{Fix}})^{i-1} \right) \end{aligned}$$

Notice that the first summation is the expected value of a geometric random variable with success probability P^*_{Fix}, which simplifies to $1/P^*_{\text{Fix}}$. The second summation is the sum over all probabilities of the same geometric random variable, which simplifies to 1. Therefore,

$$E(M) = (T_\mu + T^*_{\text{Ext}}) \frac{1}{P^*_{\text{Fix}}} + (T^*_{\text{Fix}} - T^*_{\text{Ext}})$$

and the expected substitution rate is $k = 1/E(M)$.

Rearranging gives a more intuitive formulation.

$$E(M) = (T_\mu + T^*_{\text{Ext}}) \left(\frac{1}{P^*_{\text{Fix}}} - 1 \right) + T_\mu + T^*_{\text{Fix}}$$

Let T^*_{Abs} be the mean time to absorption, unconditional on the absorbing state. Since the only absorbing states are extinction and fixation we have

$$T^*_{\text{Abs}} = P^*_{\text{Fix}} T^*_{\text{Fix}} + (1 - P^*_{\text{Fix}}) T^*_{\text{Ext}}$$

and thus we can rewrite k, the expected substitution rate, as

$$k = \frac{1}{E(M)} = \frac{T_\mu^{-1} P^*_{\text{Fix}}}{1 + T_\mu^{-1} T^*_{\text{Ext}} + T_\mu^{-1} P^*_{\text{Fix}} (T^*_{\text{Fix}} - T^*_{\text{Ext}})} = \frac{P^*_{\text{Fix}}}{T_\mu + T^*_{\text{Abs}}}.$$

1.4 SLiM simulation code

```
initialize() {
        // mu is the mutation rate, s is the selection coefficient
        initializeMutationRate(mu);
        initializeMutationType("m1", 0.5, "f", s);
        m1.convertToSubstitution = F;
        m1.mutationStackPolicy = F;

        // single locus
        initializeGenomicElementType("g1", m1, 1.0);
        initializeGenomicElement(g1, 0, 0);
        initializeRecombinationRate(0);
}

1 {
        // N is the population size
        sim.addSubpop("p1", N);
        writeFile(..., append=T);
}

1:5000000 late() {
        muts = sim.mutations;
        freqs = sim.mutationFrequencies(p1,muts);
        if( sum(freqs) == 1.0 ) {
                // Return process
                p1.genomes.removeMutations(sim.mutationsOfType(m1),T);
                writeFile(...,append=T);
        }
}
```

2 SI Figures

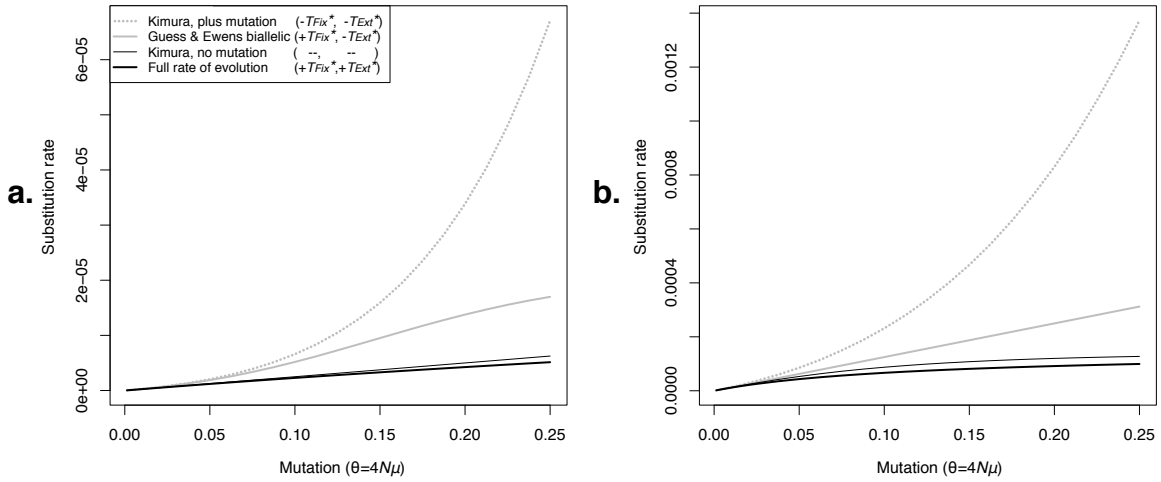

Figure S1: **Effect of mutation on different rates of evolution.**
Conditions are the same as in Fig. 1C. Modifying Kimura's weak-mutation rate of evolution to include mutation in the probability of fixation predicts a large acceleration in the rate of evolution compared to Kimura's model ("Kimura, plus mutation") for larger values of θ. Similarly, including mutation and the expected time to fixation, as in Guess and Ewens ("Guess Ewens biallelic"), predicts an acceleration in the rate of evolution. However, a deceleration is predicted (and observed in simulation; Fig. 1C) when the absorption times are included as in the rate of observed evolution. a. $S = 0$ (neutral). b. $S = 50$ (strong positive selection).

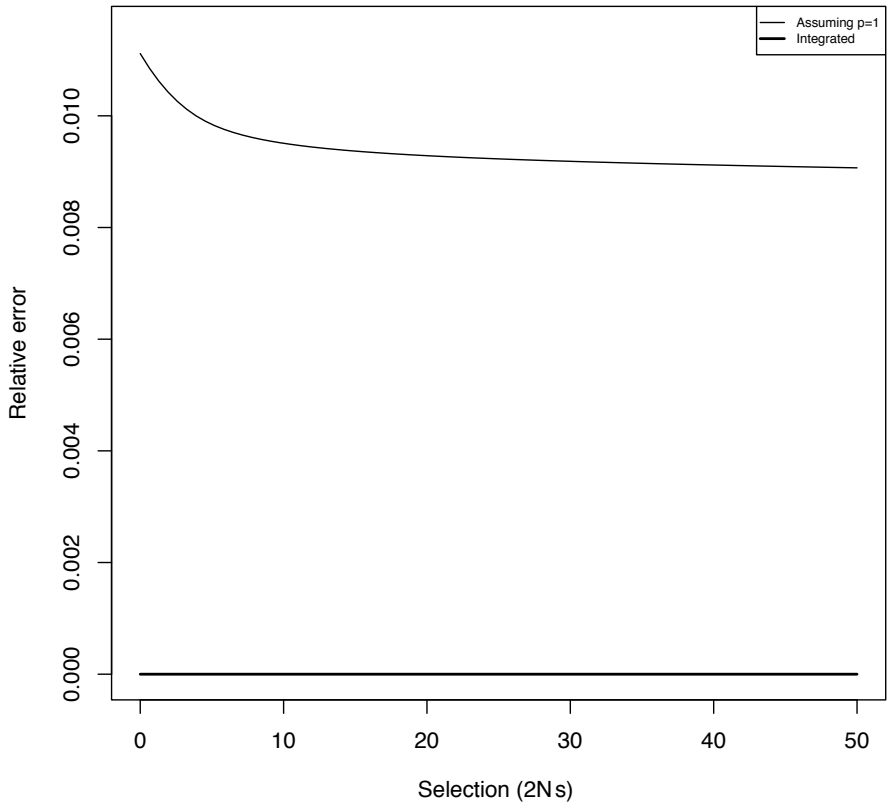

Figure S2: **Relative error when assuming $p = 1$ and including an initial configuration, $f(p)$.**
Relative error between direct computation of the rate of evolution and equation 3 (main text). When integrating over $f(p)$, the two calculations match closely, differing by only about 10^{-13}. As expected, when $p = 1$ is assumed, the error is much greater.

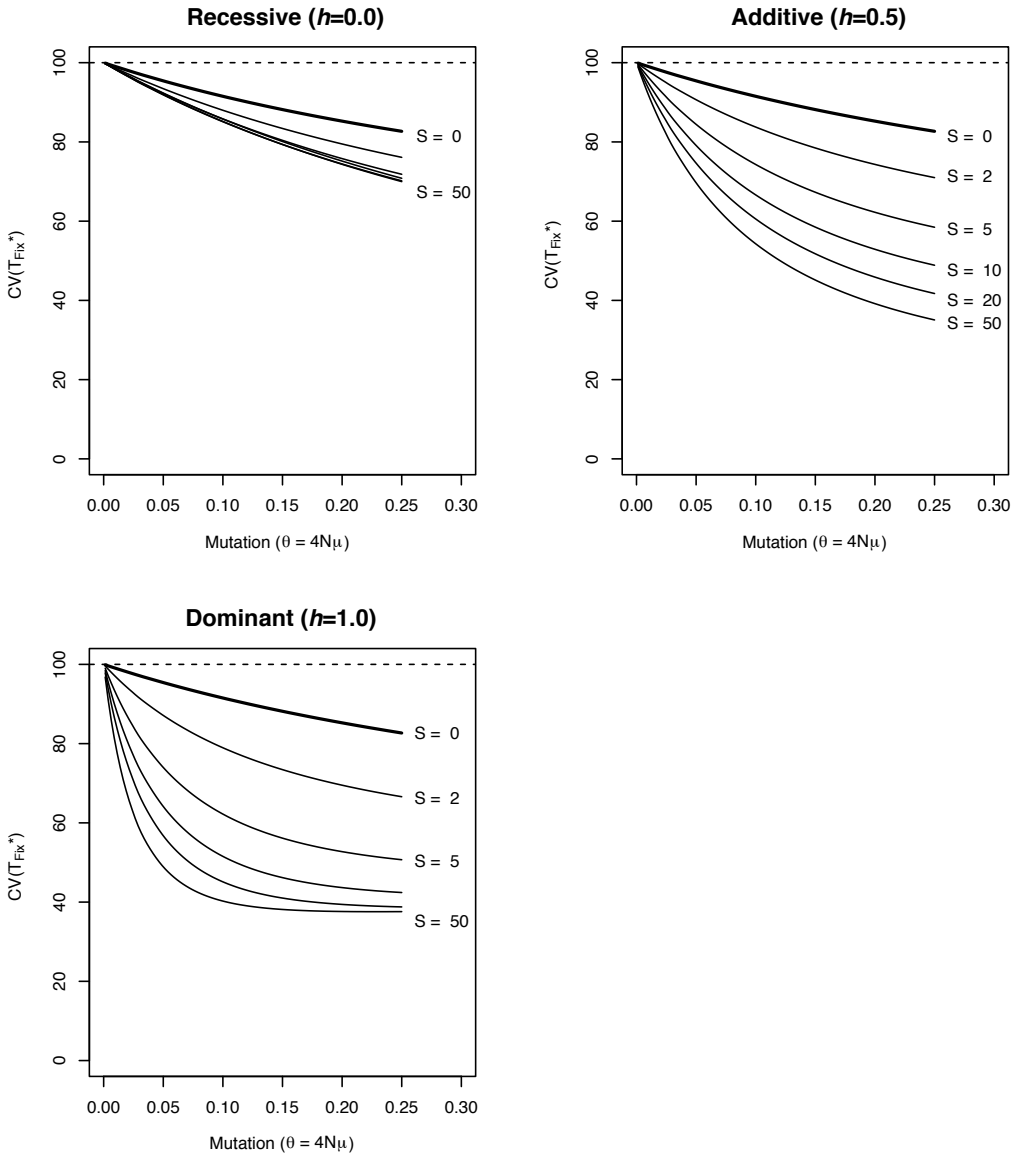

Figure S3: **The molecular clock is not a Poisson process.**
A Poisson process has often been used to model substitution processes, which requires that the mean of the substitution rate be equal to its variance. Although we cannot easily calculate the variance of the substitution rate, we can calculate the variance or coefficient of variation (CV) of the time between fixations. For a Poisson substitution process, $CV(T^*_{\mathrm{Fix}}) = 100\%$. Values other than 100% imply that the molecular clock is over- or under-dispersed. Here we show that not only do mutation and selection cause substantial over/under dispersion, the neutral molecular clock ($S = 0$) is not a Poisson process, in general, in the Wright-Fisher model including mutation.

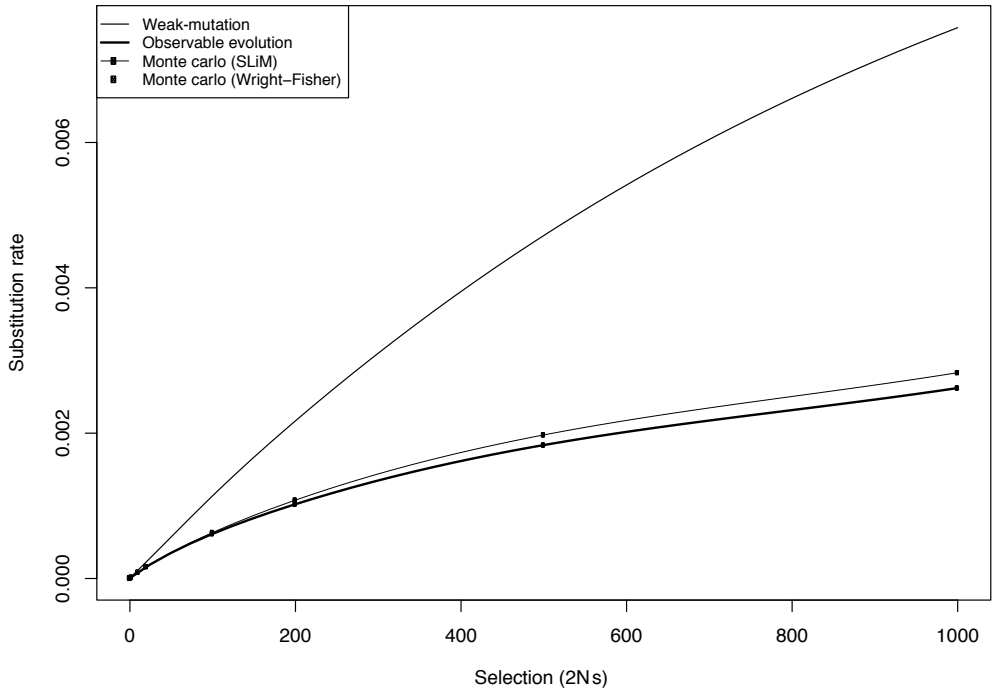

Figure S4: **Simulation versus theory for a more extreme parameter range.** Calculations and simulations were made assuming a modest value of $\theta = 0.1$. For extreme values of the population-scaled selection coefficient, SLiM simulations quantitatively diverged but qualitatively agreed with standard Wright-Fisher simulations and theory. This variation is expected to some extent, as SLiM implements an individual-based simulation under Wright-Fisher like conditions, not the Wright-Fisher model itself.

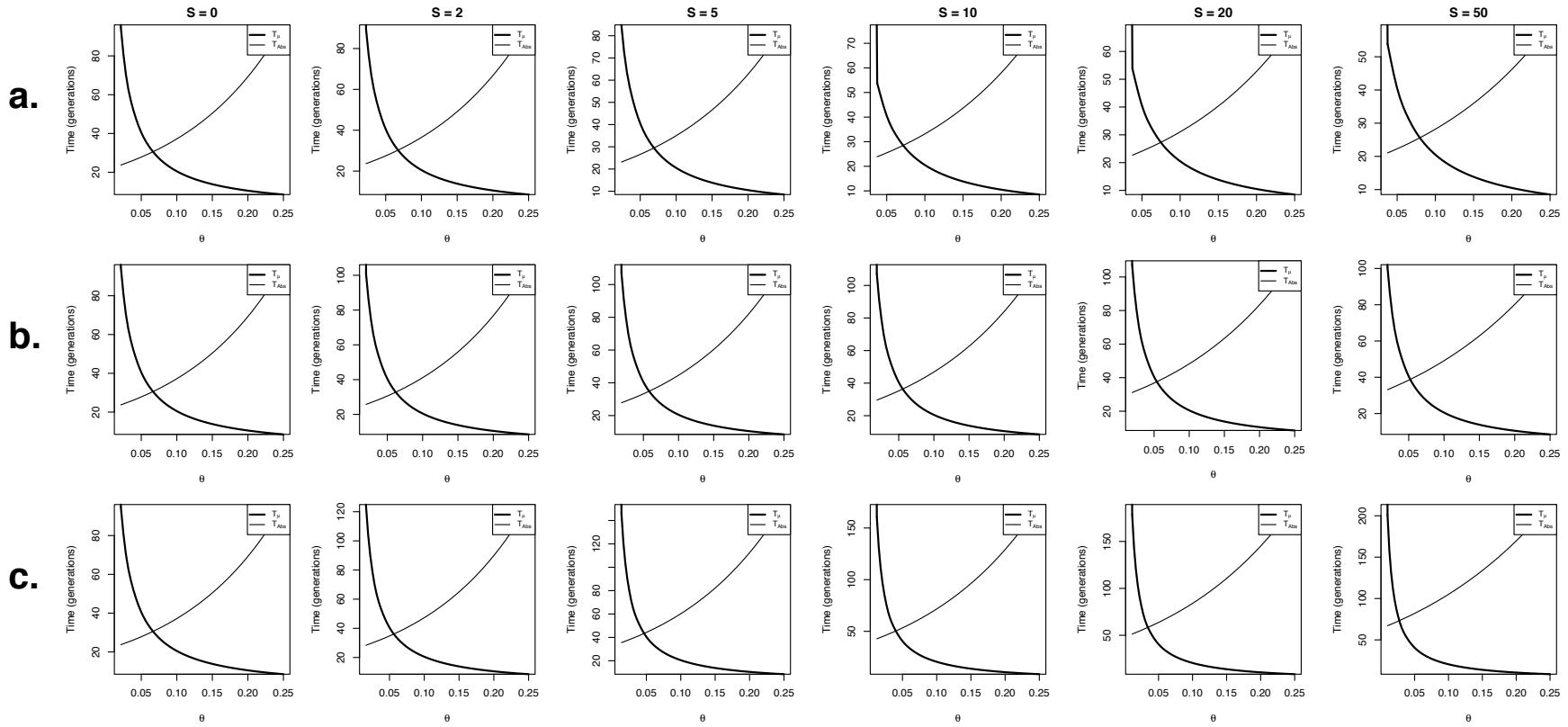

Figure S5: **Time to absorption and time to mutation when mutation may not be weak.**
The weak-mutation rate of evolution assumes $T_\mu \gg T^*_{\text{Abs}}$. Here it is shown that this assumption fails badly when θ is as small as 0.05. All quantities were integrated over $f(p)$ as explained in the Supplementary Methods. a. $h = 0$ (recessive). b. $h = 0.5$ (additive). c. $h = 1.0$ (dominant).

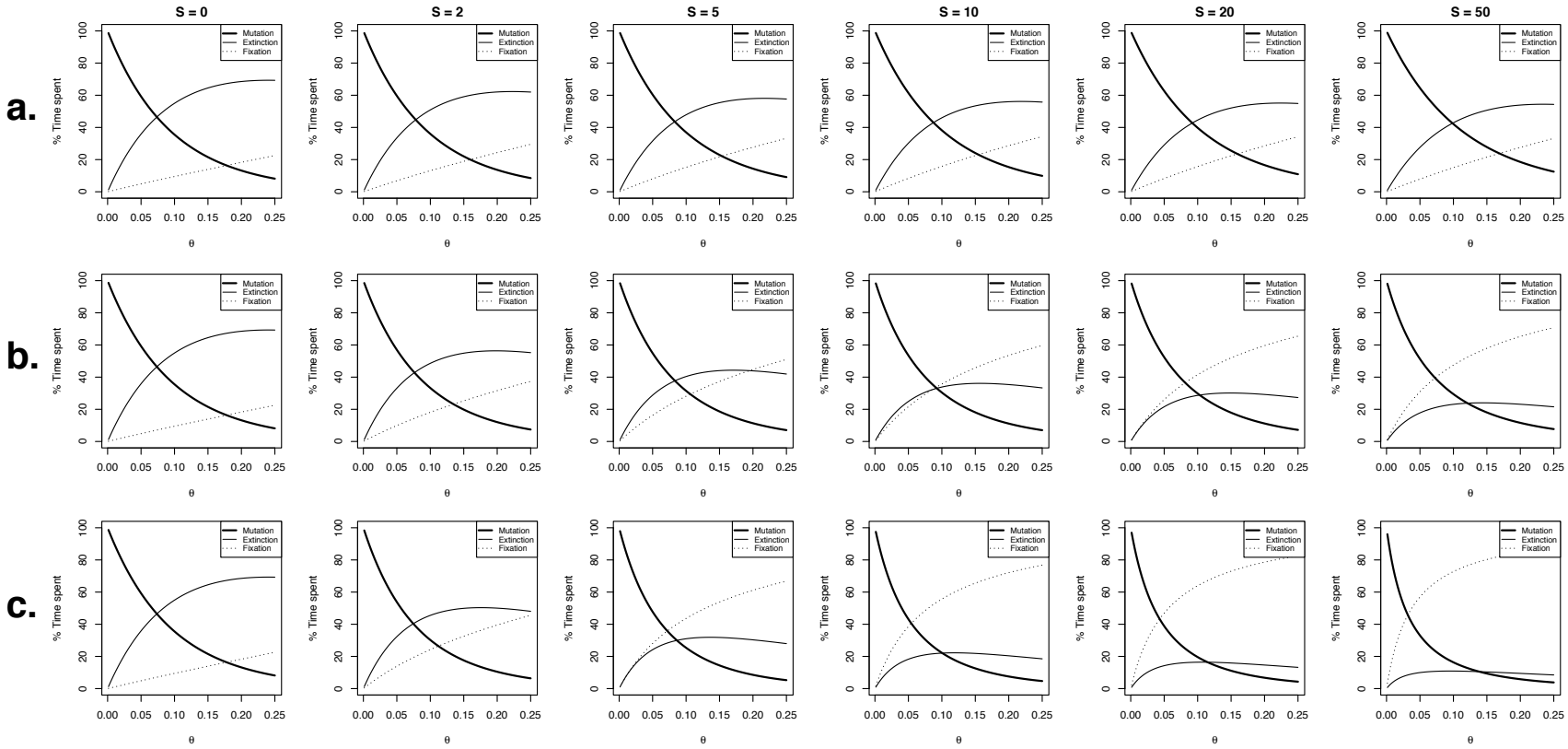

Figure S6: **Fraction of time in the between fixations cycle spent waiting for different processes.**
Following the logic of the main text, the mean cumulative time between fixations spent generating mutations was computed as $T_\mu \cdot 1/P_{\mathrm{Fix}}$, the mean time spent segregating prior to extinctions was $T^*_{\mathrm{Ext}} \cdot \left((1/P^*_{\mathrm{Fix}}) - 1\right)$, and the mean time spent in the fixation phase was T^*_{Fix}. All quantities were integrated over $f(p)$ as explained in the Supplementary Methods, and were normalized by the mean time between fixations. a. $h = 0$ (recessive). b. $h = 0.5$ (additive). c. $h = 1.0$ (dominant).

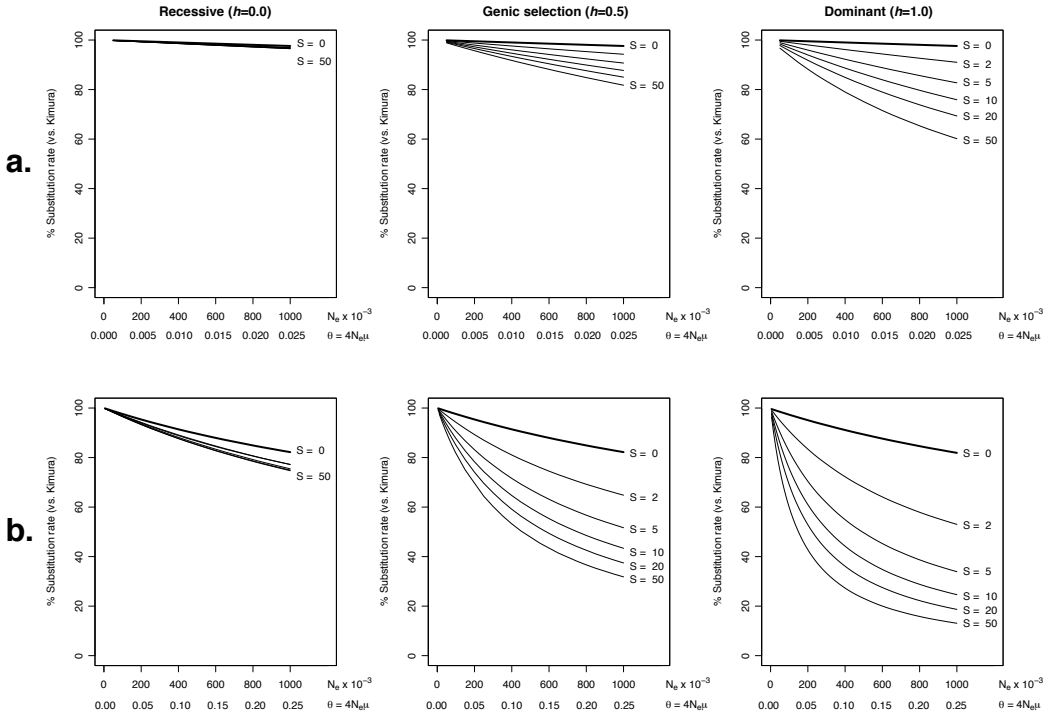

Figure S7: **Predicted deceleration for CpG and non-CpG transitions in large rodent populations.**
A. The non-CpG forward mutation rate from Uchimura et al.(2) (5.6×10^{-9}) was used over a range of effective population sizes relevant to rodents. Ancestral effective population sizes of mice have been estimated to be $277,000$(4), while other studies have suggested that *Mus musculus castaneus* has a current effective population size over $700,000$(3). B. Using the CpG-transition forward mutation rate(2) (6.73×10^{-8}). For reasonable values of the effective population size (N_e), these results predict a slowdown of CpG transitions that may be detectable in natural populations.

References

[1] Krukov, I., De Sanctis, B.D., de Koning, A.P.J.. Wright-Fisher exact solver (WFES): scalable analysis of population genetic models without simulation or diffusion theory. *Bioinformatics* **33,** 1416-1417 (2017)

[2] Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J., Yagi, T.. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Research* **25,** 1-10 (2015)

[3] Phifer-Rixey, M., Bonhomme, F., Boursot, P., Churchill, G.A., Pilek, J., Tucker, P.K., Nachman, M.W., Adaptive evolution and effective population size in wild house mice. *Molecular Biology and Evolution* **29,** 2949-2955 (2012)

[4] Geraldes, A., Basset, P., Smith, K.L., Nachman, M.W., Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination. *Molecular Ecology* **20,** 4722-4736 (2011)