An integrated model of the effects on fitness of positively selected, conserved, and nearly neutral sites in sexual populations and the population-optimal mutation rate

Gordon Irlam
gordoni@gordoni.com

Los Altos, California, United States

January 1, 2026

Abstract

Positively selected, conserved, neutral, and nearly neutral sites have often been analyzed in isolation. Here, an approximate integrated model of the fitness effects of these classes of sites in sexual populations is developed and analyzed. Within this framework, it is shown that the aggregate fitness effects associated with positively selected sites need not be negligible, even when such sites are rare, and that nearly neutral sites can contribute a substantial component of the total genetic load. The framework defines the latent load as the genetic load associated with positively selected sites that are destined to fix but for which beneficial alleles have not yet established. The germline spontaneous mutation rate that maximizes the geometric mean of population mean fitness is predicted, under the model, to occur when the mean latent load is approximately equal to the mutational load. This population-optimal mutation rate differs from the minimal mutation rate predicted by purely microevolutionary considerations that focus on individual-level selection. It is hypothesized that, over macroevolutionary timescales, species-level processes such as differential extinction may bias the mutation rates of surviving species towards the population-optimal value. The model provides a unified framework for assessing the combined fitness effects of positively selected, conserved, and nearly neutral sites, and highlights the potential macroevolutionary importance of the latent load in shaping the long-term evolution of the mutation rate.

Keywords: positive selection, purifying selection, negative selection, genetic load, latent load, substitutional load, lag load, mutational load, nearly neutral load, optimal mutation rate.

Introduction

The goals of this manuscript are twofold: to develop an approximate equation for the fitness of sexual populations incorporating the effects of positively selected, conserved, and nearly neutral sites; and then to use this equation to understand how aggregate population mean fitness changes with the germline spontaneous mutation rate.

Population mean fitness is important because it reflects the expected number of organisms present in the next generation, and is therefore a key contributor to the success of the population or species. An important result derived here is that the mutation rate per base pair that maximizes the long-term geometric-mean population mean fitness is the one for which the mean fitness losses associated with the latent load and the mutational load are approximately equal. This mutation rate will be termed the population-optimal mutation rate. This finding is significant because it is a mutation rate that is quite different from the as-small-as-possible rate that might be expected from microevolutionary considerations [1, 2]. The finding is also significant because, while the mutational load has been considered to be a primary cause of genetic load [3], the possibility of a similarly sized impact for the latent load appears to have been rarely considered.

Despite the importance of the mutation rate to population mean fitness, and thus the success of species, there are limited microevolutionary mechanisms for keeping the mutation rate near its long-term population-optimal value. Consequently if the mutation rate is to be kept close to its long-term population-optimal value it will likely be necessary to invoke macroevolution and the extinction of species having mutation rates that are significantly too small or too large.

This manuscript builds on the work of Smith on the lag load [4], Haldane on the mutational load [5], and Kimura et al. on the fixation of a beneficial allele, the genetic load in the presence of drift, and the heterozygosity of sites maintained by a flux of mutations [6, 7, 8, 9].

The latent load is defined in this manuscript to be the genetic load associated with sites that could harbor beneficial alleles but at which such alleles have not yet arisen and begun to fix. Kimura makes related arguments to the arguments made here on population-optimal mutation rates, but the latent load is not considered [10, 11]. It is argued in the present manuscript that the latent load is a key component in determining the population-optimal mutation rate.

It is beyond the scope of this manuscript to review the degree to which mutation rates are found to be close to their population-optimal values. This manuscript simply seeks to develop a mathematical model of population mean fitness, and then use it to derive a formula for the population-optimal mutation rate.

Results

Before commencing, it is worth clarifying what is meant here by "fitness", denoted w. Fitness is traditionally defined as the growth rate of a genotype in a population. When organisms are haploid, and the genotype in question is that of the haploid organism, fitness takes on a particularly intuitive form. Fitness is simply the expected relative number of immediate offspring produced by a haploid organism in a given environment. It is an expected value because random stochastic sampling effects

may cause the actual observed number to vary from the expected value, but if it were possible to repeat the same life cycle in the same environment, fitness would equal the mean number of offspring observed. It is relative. An organism with twice the fitness value as some other organism would be expected to have twice as many offspring. In this manuscript, the fitness of diploid organisms is defined in the same manner; as the expected relative number of immediate offspring. Most of the time it is this relative nature of fitness that concerns us, but in order to speak meaningfully of fitness values alone it is necessary to anchor the fitness scale. Fitness here is defined relative to a hypothetical ideal organism that is perfectly adapted to its environment, and which is assigned a fitness value of 1. This means that log fitness, x, is negative, or zero in the one special case. The concept of an ideal organism is a bookkeeping device rather than a realizable genotype.

As used here, genetic load describes the difference between the fitness of an ideal organism at a given point in time and the mean fitness of the current population. Four focal contributors to the genetic load will be considered here:

- what will be termed the "latent load", which results from sites that could harbor beneficial alleles but at which such alleles have not yet arisen and begun to fix;
- the substitutional load, which results when beneficial alleles have become established and the sites are in the process of fixing;
- the mutational load, which results from deleterious mutations to conserved sites;
- and what will be termed the "nearly neutral load", which results from nearly neutral sites in the genome.

The nearly neutral load could in principle be viewed as part of the mutational load. It is however convenient to treat it as a distinct load because unlike the mutational load it contains an effective component that is independent of the mutation rate.

Stochastic tunneling refers to the mechanism by which a beneficial double mutant can occasionally arise and become fixed despite the mutations both individually being deleterious [12]. If relative fitness is defined in relation to a hypothetical globally optimal organism for the current environment, then a very large genetic load relates to the fitness improvements that are not reachable through a monotonically increasing fitness path, but might be reachable through stochastic tunneling. In this manuscript this load will be referred to as the "external load". This load can be defined as the fitness difference between a locally optimal organism that is reachable by a series of site by site monotonically increasing fitness steps when the external environment is held constant, and the globally optimal organism for the current environment. In the case of epistasis there might be several locally optimal organisms with different fitness values. This is resolved in Appendix D.

The external load provides a means for generalizing the idea of the genetic load to allow for interspecies comparisons. Different species can stably coexist with different sums of the latent, substitutional, mutational, nearly neutral loads, and possibly other loads, provided they have offsetting differences in the external load. If different species have different total genetic loads then their populations sizes are expected to change, changing the environment, which changes the loads, and frequently brings the total loads back into balance.

The environment is made up of both external and internal environments. The external environment is the aggregate of all the physical and biological spaces with which organisms making up the population interact. The internal environment is made up of the population's genomes. The latent and substitutional loads are generated as a result of ongoing changes in the external environment. If the external environment was constant the latent and substitutional loads would tend towards zero. The rate of external environmental change may vary over time.

Many alleles are conditionally beneficial, conditionally deleterious, or conditionally neutral, exhibiting different beneficial, deleterious, or neutral effects depending on the allele's environmental context. Such alleles are simply described here as beneficial, deleterious, or neutral depending on their aggregate effects across all organisms in the population. The effective selection coefficient of such an allele is the expected value of the allele's selection coefficients across all organisms.

Figure 1 shows a simplified lifecycle of a site, and the different genetic loads it may generate. In this manuscript, a positively selected site is defined as a site for which there is an allele with a selection coefficient that is beneficial which has not previously been fixed, irrespective of whether the beneficial allele is currently present in the population. This definition is consistent with a conserved, or negatively selected site, not requiring deleterious alleles be present in the population. The definition differs from definitions of positively selected site that require currently segregating beneficial alleles, or that simply require the site underwent positive selection at some point in the past.

A positively selected site is either a latent site or a substituting site. A beneficial allele is considered established if it is not going to die out; something that is only known with certainty in hindsight. A new positively selected site may have a beneficial allele that is already established as a result of standing genetic variation, or more typically, the beneficial allele first needs to arise by mutation and then become established. While waiting for the beneficial allele to establish the positively selected site is a latent site and it generates a latent load. Once the beneficial allele is established the positively selected site is a substituting site and it generates a substitutional load. When the beneficial allele becomes fixed the site is either conserved and generates a mutational load, or nearly neutral and generates a nearly neutral load. Occasional change in the environment causes a site to transition between being a conserved site, a neutral or nearly neutral site, or to become a new positively selected site.

Beneficial mutations that die out rather than become established transiently increase population mean fitness. It is a matter of semantics whether this fitness increase is considered to create a reduction in the latent load, or the substitutional load, but because such mutations fail to establish, the beneficial allele is rare, and so only have a very small effect on the load.

The latent and substitutional loads together comprise what Smith calls the lag load [4]. They exist because the makeup of the population lags behind that of an ideal organism as dictated by the current external environment, which changes over time. Smith's concept of an ideal organism was as the locally optimal organism reached by a series of fitness increasing mutational steps should the external environment be held constant.

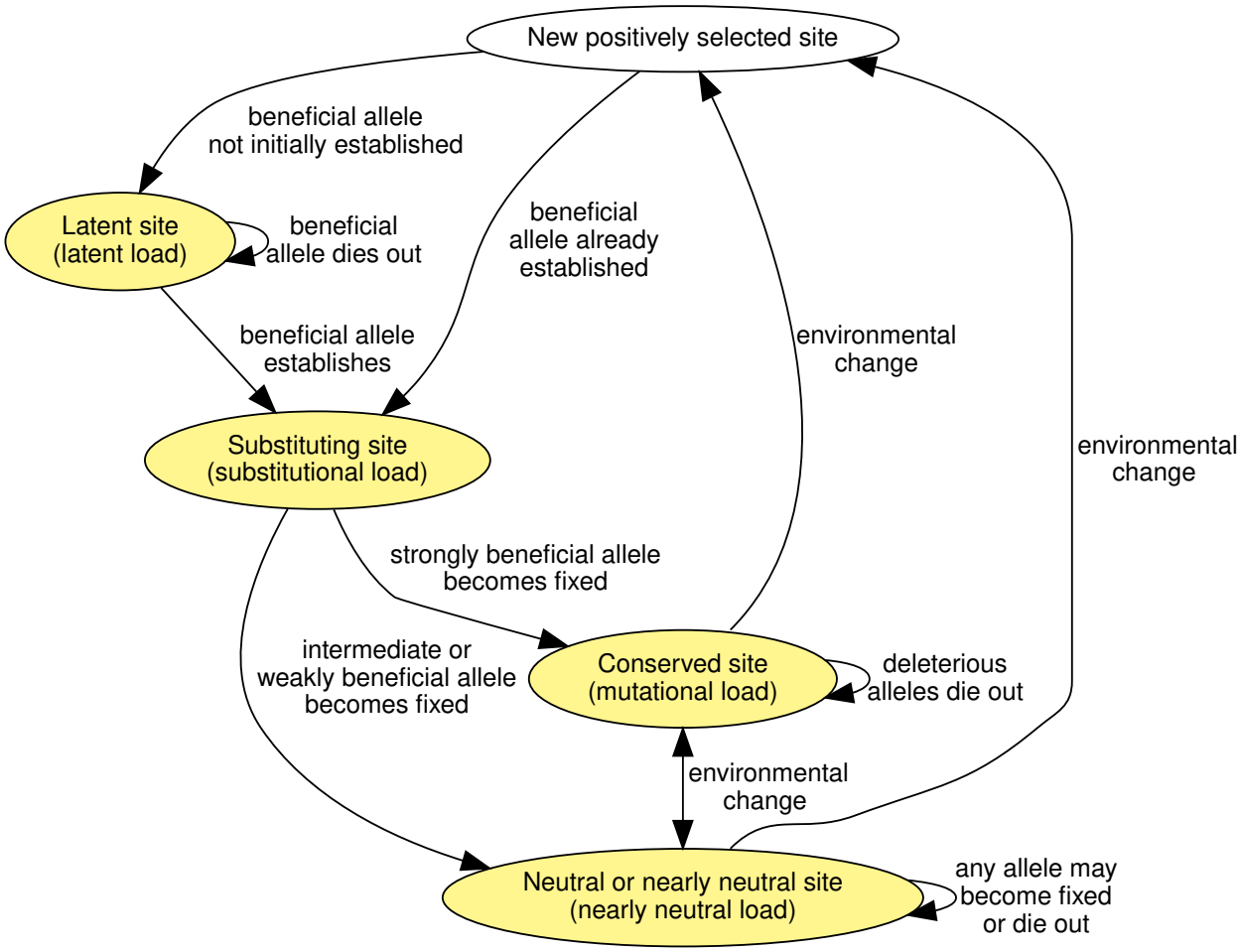

Figure 1: Simplified lifecycle of a site. Transitions between states are the result of mutation, selection, drift, and environmental change. New positively selected site is a transitory state.

w	Fitness of an organism.
x	Log fitness of an organism.
$\overline{x}$	Mean log fitness of the population.
N	Number of organisms in the population.
N_e	Variance effective population size.
P	Ploidy: 1 haploid, or 2 diploid.
Γ_p	Population-wide net rate of new positively selected sites per sexual generation.
μ	Germline mutation rate per base pair per haploid genome per sexual generation.
P_{fix}	Probability of fixation of a beneficial mutation or allele.
s_p	Heterozygous selection coefficient for a beneficial (aka positive) allele.
h	Dominance coefficient. Beneficial homozygotes have selection coefficient $\frac{s_p}{h}$.
f_l	Fraction of new positively selected sites that generate a latent load.
Γ_{ls}	Effective rate of transition from latent to substituting sites per sexual generation.
π_f	Adjustment to P_{fix} on account of unaccounted for interference between sites.
π_e	Adjustment to log fitness effect of selection due to epistasis.
π_M	Adjustment to log fitness effect of latent load due to macroevolutionary selection.
A_s	Aggregate log fitness for fixation of an established allele at a substituting site.
γ	Euler's constant, $0.5772\ldots$.
Γ_s	Per sexual generation rate at which substituting sites become established.
U	Germline deleterious mutation rate per organism per sexual generation.
s_n	Heterozygous selection coefficient for a deleterious (aka negative) allele.
$A_0(s)$	Mean time-aggregated log fitness of mutations at nearly neutral sites that fail to fix.
$x_0(s)$	Log fitness of a nearly neutral site.
ν	Rate of back mutation per base pair per haploid genome per sexual generation.
$D(s)$	Density of sites with beneficial alleles having selection coefficient s.
R	Genome length in Morgans.
L_n	Number of conserved sites per haploid genome.
$\pi_\mu \vert g$	Adjustment to population-optimal mutation rate due to microevolutionary selection.

Table 1: Summary of key notation.

Model

The population is modeled using a Wright-Fisher-like discrete-time model consisting of repeating, non-overlapping, generations, with mating occurring in accordance with individual fitness.

A summary of the key notation used is provided in Table 1. Assume a sexual population of organisms with population size N, and neutral site variance effective population size N_e. Let the ploidy, P, be 1 for a organisms that are counted and experience the force of selection during the haploid state, and 2 for organisms that are counted and experience selective forces while in the diploid state.

Only single nucleotide point mutations are considered. Insertion and deletion (indel) mutations are not included in the analysis. Stochastic tunneling is also not modeled. It would be possible to extend the analysis to incorporate these features.

Epistasis is possible, but epistasis among simultaneously segregating sites is assumed to be minimal.

That is, a site that is either a latent site or a conserved site may affect the fitness of another site, but there should be few interactions between the fitness effects of substituting sites or nearly neutral sites.

The constructed model handles a broad range of selection coefficients. Strong selection occurs when $|PN_e s| \gg 1$, where s is the selection coefficient. Sites with alleles having selection coefficients of intermediate strength behave according to the nearly neutral site model. In fact, the conserved site model is simply a special case of the more general nearly neutral site model.

In the model changes in selection coefficients occur as a result of environmental change, and these changes can be instantaneous or gradual. Latent sites are modeled as coming about as a result of instantaneous changes from neutral, nearly neutral, or conserved sites to sites with alleles having positive selection coefficients. If the new positive selection coefficient is strong this may be followed by further gradual variation in the strong positive selection coefficient of the latent site's allele. The effects of gradual transitions by latent sites over the non-strong selection coefficient range are not modeled. Changes in the selection coefficients at substituting sites are not modeled. This is acceptable as the substitutional load is shown to be much smaller than the latent load.

The selection coefficient of a beneficial allele at a latent site may cease to be beneficial before the site becomes established. The effect of this is handled in the model by dealing with rates of transition from latent to substituting sites. This detail is not shown in Figure 1.

A site may have an associated selection coefficient so small that fixation of the beneficial allele would take longer than the lifetime of the species, but such sites are still relevant. This is because there are likely to be many such sites, and so it is possible to speak in terms of probabilities and rates of fixation.

New positively selected sites

The external environment may undergo change either frequently or infrequently. When it does, such change may reduce the fitness of an organism relative to the ideal organism, which is assumed to have adapted to the change. Such change, however, creates opportunities for beneficial alleles to arise and/or increase in frequency at sites where the allele was previously neutral or deleterious. In addition to new positively selected sites arising as a result of changes in the external environment, new positively selected sites may also arise as a result of changes to the internal environment. The fixation of the beneficial allele at one positively selected site might epistatically lead to the creation of other nearby or distant positively selected sites.

Let Γ_p denote the net rate at which new positively selected sites arise per sexual generation.

Γ_p is a net rate because some positively selected sites may cease to be positively selected as a result of environmental change. It is also a net rate because a reduction in N_e may render some positively selected sites as nearly neutral, and thus reduce Γ_p. Similarly, an increase in N_e may make some nearly neutral sites positively selected sites, increasing Γ_p. As described in Appendix D, epistasis also influences Γ_p. By notational convention, a change in the selection coefficient of an already fixed allele to be beneficial will be excluded from Γ_p. For example, if a site was neutral and fixed at allele B, and then the environment changes and allele B is now beneficial, this does not contribute to Γ_p.

Since base pairs can take on four possible values, this reduces the magnitude of Γ_p by roughly $\frac{1}{4}$.

Probability of a latent site

Let μ denote the mean germline spontaneous mutation rate per base pair per haploid genome per sexual generation.

When a new positively selected site arises, its fitness impact will depend on whether the beneficial allele is already established and will proceed to fixation, or whether it either does not exist at all or existing copies of the allele will die out.

If the positively selected site was previously a neutral site, then there is a chance that the beneficial allele is already present. Assuming $PN_e\mu \ll 1$, Kimura has determined for neutral sites the mean heterozygosity, the nucleotide diversity, as approximately $2PN_e\mu$ [8]. Recall that P is the ploidy. Provided the nucleotide diversity is low, the nucleotide diversity is approximately twice the probability that a given genomic site harbors a mutant allele. Thus the probability of a genomic site harboring a mutant allele is approximately $PN_e\mu$. However, this does not mean that the typical previously neutral genomic site has either zero or one copy of a mutant allele. The majority will have zero, and a few will have a much larger number.

Let s_p denote the heterozygous selection coefficient associated with a beneficial allele. Based on the work of Kimura [6, 13], if the allele has proportion p, the probability of fixation, P_{fix}, is,

$$P_{fix} = \frac{1 - e^{-2PN_e s_p p}}{1 - e^{-2PN_e s_p}} \tag{1}$$

Thus if $2PN_e s_p p \gg 1$ the beneficial allele at the previously neutral site is already established, and so the site will only ever exert a substitutional load. Conversely, if $2PN_e s_p p \ll 1$, the probability of fixation of the beneficial allele is small, and so the substitutional load is preceded by a latent load.

Let f_l denote the probability that a new positively selected site is a latent site. That is, at the time the new positively selected site arises all existing copies of the beneficial allele are destined for extinction. Appendix A shows that f_l will have a value close to 1 (e.g., $f_l = 0.992$ for $PN_e = 2\times10^5$, $\mu = 10^{-8}$, and $s_p = 10^{-3}$ which are values typical of vertebrates).

Latent load

The genetic load associated with a new positively selected site doesn't begin at the time a beneficial allele establishes, but at the time the environment changes to make an allele's selection coefficient positive, thereby creating a new positively selected site. The genetic load from the time the new positively selected site comes into existence until the beneficial allele is established is termed the latent load.

For simplicity, it is assumed that only one of the four possible base pair values will maximize the fitness associated with a new positively selected site, and that the other base pair values are equally

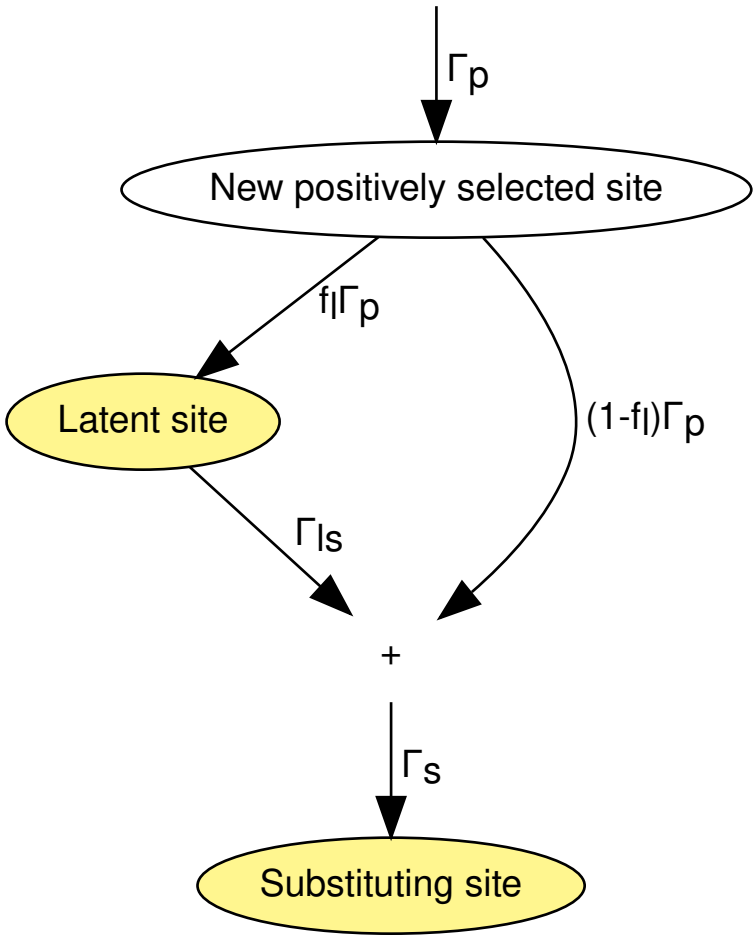

Figure 2: Rates of transition between different types of positively selected sites.

deleterious. Consequently, the mean rate of beneficial mutations per haploid genome at a latent site is approximately $\frac{\mu}{3}$.

For now, assume that there is no epistasis.

Also for now, assume that environmental change causes new positively selected sites to arise in an all-or-nothing manner; a site is initially neutral, nearly neutral, or conserved, and then suddenly becomes a positively selected site with a fixed selection coefficient s_p for the beneficial allele.

Figure 2 shows the rates of transition between different types of positively selected sites.

Let Γ_{ls} denote the rate at which latent positively selected sites transition to being substituting sites.

Let Γ_s denote the per sexual generation rate at which substituting sites become established. Sub-

stituting sites become established either when a beneficial mutation of a latent site establishes, or when the beneficial allele of a new positively selected site is already established.

$$\Gamma_s = \Gamma_{ls} + (1 - f_l)\Gamma_p \tag{2}$$

For a species pair, over macroevolutionary timescales, the temporal mean value of Γ_s is equal to the rate of positively selected adaptive substitution per generation, and can be estimated by,

$$\mathrm{E}[\Gamma_s] \approx \frac{(\alpha_n D_n + \alpha_{nc} D_{nc})g}{2T}$$

where α_n is the fraction of non-synonymous substitutions that are fixed by positive selection, D_n is the number of non-synonymous substitutions, α_{nc} is the fraction of substitutions that are fixed by positive selection in the non-coding region, D_{nc} is the number of substitutions in the non-coding region, g is the generation length, and T is the divergence time. The fixation by drift of sites with alleles having small fitness effect is corrected for the determination of α_n and α_{nc}.

Suppose that genome-wide there are a total of k_l latent sites. Each site exists in every haploid genome. Latent sites arise at a rate of $f_l\Gamma_p$. The difference between these two rates is equal to the rate at which k_l changes per sexual generation, Δk_l,

$$\Delta k_l = f_l\Gamma_p - \Gamma_{ls} \tag{3}$$

Eliminate Γ_p by combining equations 2 and 3,

$$\Gamma_{ls} = f_l\Gamma_s - (1 - f_l)\Delta k_l \tag{4}$$

Fixation of the beneficial allele at a latent site typically starts from a single mutation. Thus, in this case, when $s_p \ll 1$, from equation 1,

$$P_{fix} \approx \frac{2\frac{N_e^{npss}}{N}s_p}{1 - e^{-2PN_e s_p}} \tag{5}$$

where N_e^{npss} is the effective population size at the new positively selected site if it were instead neutral.

The aggregate rate of beneficial mutation at the latent sites is $k_l PN\frac{\mu}{3}$, and the probability of fixation of each beneficial mutation is approximately given by equation 5. The product of these two terms equals the rate at which latent sites transition to substituting sites,

$$\Gamma_{ls} \approx \frac{k_l 2P\pi_f N_e s_p \mu}{3\left(1 - e^{-2PN_e s_p}\right)}$$

where π_f is a factor to account for any change to the probability of fixation on account of a possible difference between the average effective population size of new positively selected sites, N_e^{npss}, and the effective population size of neutral sites, N_e. If, as is likely, new positively selected sites cluster in time and space the two may differ; differences in the effects of linkage between new positively selected sites and neutral sites will need to be taken into account. Appendix B shows that π_f is expected to be close to 1 (e.g., > 0.99 for apes and rodents, and > 0.97 for insects).

Let h denote the dominance coefficient. By convention $h = 1$ for haploid eukaryotic species. Assuming $s_p \ll 1$, the log fitness loss associated with a wild-type homozygote is approximately $\frac{s_p}{h}$. This is the log fitness gained if the organism is a homozygous mutant. The wild-type homozygote contributes to the latent genetic load.

Let $\overline{x_l}$ denote the mean log fitness associated with latent sites,

$$\begin{aligned}\overline{x_l} &\approx -\frac{k_l s_p}{h} \\ &\approx -\frac{3\Gamma_{ls}\left(1 - e^{-2PN_e s_p}\right)}{2P\pi_f N_e \mu h}\end{aligned} \tag{6}$$

In the event that h is not constant but instead a function of the latent sites, h in equation 6 represents the harmonic mean value.

If s_p is strong, the exponential term is approximately zero, equation 6 is effectively independent of s_p, and equally applicable to a distribution of s_p values. When s_p is nearly neutral the exponent term will be close to 1, the log fitness effect will be proportional to s_p, and it will be much smaller than usual. Let $\Gamma_{ls}^{raw}(s)$ denote the density of the rate of transition from latent to substituting sites with beneficial alleles having selection coefficient s. For a range of selection coefficients define the effective Γ_{ls} by,

$$\Gamma_{ls} = \int_0^\infty \Gamma_{ls}^{raw}(s)\left(1 - e^{-2PN_e s}\right) ds \tag{7}$$

Then,

$$\overline{x_l} \approx -\frac{3\Gamma_{ls}}{2P\pi_f N_e \mu h}$$

It is shown in Appendix C that for three animal species that were examined, assuming an exponential distribution of fitness effects at latent sites,

$$\Gamma_{ls} \approx \int_0^\infty \Gamma_{ls}^{raw}(s) ds$$

Epistasis is likely to significantly alter the number of current positively selected sites that actually fix. There are many genotypic paths to the same phenotype. By redefinition of the latent load to

be the fitness effect only of positively selected sites that are destined to fix, Appendix D shows how it is possible to incorporate epistasis into the latent load. To reflect epistasis, let,

$$\overline{x_l} \approx -\frac{3\pi_e \Gamma_{ls}}{2P\pi_f N_e \mu h} \tag{8}$$

where $\pi_e > 0$ is an adjustment to the relative log fitness effect of selection on account of epistasis. Appendix D shows an estimate of π_e is 0.8; approximate range 0.6 to 0.9.

Since f_l is close to 1, barring Γ_p becoming orders of magnitude larger than Γ_s, which is not realistic, $(1 - f_l)\Delta k_l \approx 0$. Thus by equation 4,

$$\Gamma_{ls} \approx f_l \Gamma_s \tag{9}$$

Equation 8 is valid for a population at a given point in time. Also of interest is the mean value of $\overline{x_l}$ over an evolutionary time period, $\mathrm{E}[\overline{x_l}]$. If the generation length is variable then $\mathrm{E}[\overline{x_l}]$ would be taken by averaging over the number of generations.

Equation 8 is true from a microevolutionary perspective, but not necessarily from a macroevolutionary perspective. Over time the number of latent sites for a population slowly fluctuates as new latent sites are created and existing sites transition to substituting sites. If the population is one of many that are in competition, then species with temporarily larger than usual numbers of latent sites for those species will be less fit than usual, and thus more vulnerable to macroevolutionary extinction at that point. This would introduce a downward bias in the number of latent sites in extant species. To reflect this, define $0 < \pi_M \leq 1$, as an adjustment to the log fitness effect of the latent load on account of any macroevolutionary selection. Then,

$$\mathrm{E}[\overline{x_l}] \approx -\,\mathrm{E}\left[\frac{3\pi_M \pi_e f_l \Gamma_s}{2P\pi_f N_e \mu h}\right] \tag{10}$$

Thus far, the only case that has been considered, is one for which alleles at sites that become new positively selected sites instantly go from having a neutral or deleterious effect, to having a fixed positive selection coefficient. It is likely that, more commonly, the selection coefficient changes slowly over time as the environment changes, until it reaches its maximal value. Appendix E demonstrates that provided the allele is strongly beneficial, this gradualism is unlikely to alter the value of $\overline{x_l}$.

Substitutional load

It takes some time for the beneficial allele at a positively selected site to reach fixation. From the moment a beneficial allele is established until it is fixed, it exerts a substitutional load.

Assume $PN_e s_p \gg 1$. Also assume that the selection coefficient of the beneficial allele does not appreciably change between the time it is established and the time it is fixed.

Kimura has determined the number of generations until fixation for a beneficial allele with selective advantage s, $s \gg \frac{1}{N_e}$, $h = \frac{1}{2}$, and initial proportion p, [9] as approximately,

$$\int_p^1 \frac{(e^{2PN_e s\xi} - 1)(e^{-2PN_e s\xi} - e^{-2PN_e s})}{s\xi(1-\xi)} d\xi$$

The integrand denotes the time (measured as the number of generations) spent at each proportion of the population. A second term dealing with the time spent at proportions less than the initial proportion is negligible and is therefore ignored.

$1 - \xi$ is the proportion of the population not containing the beneficial allele. If the log fitness cost of a genomic site not harboring the beneficial allele is C, then multiplying the integrand within the integral by $C(1-\xi)$ gives the time-aggregated sum of the fitness costs associated with the fixation of the beneficial allele,

$$\begin{aligned}
&\int_p^1 \frac{C(e^{2PN_e s\xi} - 1)(e^{-2PN_e s\xi} - e^{-2PN_e s})}{s\xi} d\xi \\
&= C \frac{-\mathrm{Ei}(-2PN_e s\xi) + \log(\xi) - e^{-2PN_e s}(\mathrm{Ei}(-2PN_e s\xi) - \log(\xi))}{s} \Bigg|_{\xi=p}^{1} \\
&\approx C f(s,\xi) \Big|_{\xi=p}^{1} \quad \text{for } s \gg \frac{1}{2PN_e}
\end{aligned}$$

in which,

$$f(s,\xi) = \frac{-\mathrm{Ei}(-2PN_e s\xi) + \log(\xi)}{s}$$

and Ei is the exponential integral,

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt$$

Since h is not guaranteed to equal 0.5 an adjustment to this approach is necessary. Even with this adjustment, the results are only approximate. It is shown below that, under parameter values typical for sexual eukaryotes, the log fitness effect of the substitutional load is at least 50 times smaller than that of the latent load. Thus this approximation will only produce a small error in the overall load.

Let s'_p denote the increase in fitness associated with a homozygous beneficial allele relative to a heterozygous beneficial allele. The log fitness of a homozygous beneficial allele is approximately $\frac{s_p}{h}$, and that of a heterozygous beneficial allele is approximately s_p, so,

$$s'_p \approx \frac{s_p}{h} - s_p = \frac{1-h}{h} s_p \tag{11}$$

Fixation can be modeled as involving two adjacent components, a heuristic approximation chosen for analytical tractability. The first component covers frequencies from p to 50% beneficial allele frequency, with $s = s_p$. The second covers frequencies from 50% to 100% beneficial allele frequency, with $s = s'_p$. Ignoring the rare cases of organisms that are homozygous for the beneficial allele in the first component, and homozygous for the non-beneficial allele in the second component, the aggregate log fitness loss associated with the fixation of a beneficial allele, A_s, can be approximated in two parts. A cost of $\frac{s_p}{h}$ times $f(s_p, \xi)$ with ξ ranging from p to 0.5. And a cost of s'_p times $f(s'_p, \xi)$ for ξ ranging from 0.5 to 1,

$$\begin{aligned} A_s \approx & \frac{s_p}{h} f(s_p, \xi)\Big|_{\xi=p}^{\frac{1}{2}} + s'_p f(s'_p, \xi)\Big|_{\xi=\frac{1}{2}}^{1} \\ = & \frac{-\mathrm{Ei}(-PN_e s_p) + \log(0.5) + \mathrm{Ei}(-2PN_e s_p p) - \log(p)}{h} \\ & - \mathrm{Ei}(-2PN_e s'_p) + \mathrm{Ei}(-PN_e s'_p) - \log(0.5) \end{aligned}$$

Now, $\mathrm{Ei}(x) \approx 0$ for $x \ll -1$. Hence, assuming s_p and $s'_p \gg \frac{1}{PN_e}$,

$$A_s \approx \frac{\mathrm{Ei}(-2PN_e s_p p) - \log(2p)}{h} + \log 2$$

In some cases a simplification is possible. $\mathrm{Ei}(x) \approx \gamma + \log(|x|)$ for $|x| \ll 1$ where γ is Euler's constant, 0.5772.... It is not guaranteed, but if the condition $2PN_e s_p p \ll 1$ holds, then,

$$A_s \approx \frac{\log(PN_e) + \log s_p + \gamma}{h} + \log 2$$

The simplifying condition is the same as the condition for a previously neutral site that is now a positively selected site experiencing a latent load. It thus describes the situation in which fixation derives from a single initial mutation. If fixation derives from multiple initial copies of the beneficial allele, A_s will be less than given above.

In the general case, for strong selection, the mean aggregate log fitness loss, $\overline{A_s}$, is,

$$\overline{A_s} \approx \frac{\overline{\mathrm{Ei}(-2PN_e s_p p)} - \overline{\log(2p)}}{h} + \log 2$$

Let k_s be a value less than or equal to 1 that reflects preexisting established beneficial alleles of new positively selected sites. Then,

$$\overline{A_s} \approx k_s \left(\frac{\log(PN_e) + \overline{\log s_p} + \gamma}{h} + \log 2 \right) \tag{12}$$

This equation applies under strong selection alone.

For non-strong selection, a bound on A_s is provided by combining the maximum fitness loss, $\frac{s_p}{h}$, with the maximum mean number of generations to fixation, $2PN_e$ [9],

$$0 \leq A_s \leq \frac{2PN_e s_p}{h}$$

In both cases, the mean log fitness associated with the substitutional load, $\overline{x_s}$, is given by,

$$\overline{x_s} \approx -\Gamma_s \overline{A_s} \tag{13}$$

Later, the fact that A_s and thus $\overline{x_s}$ are independent of μ will be important.

For strong selection, comparing equations 12 and 13 with equations 4 and 8, and assuming either $\Delta k_l = 0$ or $f_l = 1$, it can be seen that the effect of the substitutional load is smaller than the effect of the latent load when,

$$\log(PN_e s_p) + \gamma + h \log 2 < \frac{3\pi_M \pi_e f_l}{2P\pi_f N_e \mu k_s}$$

For eukaryotes, $N_e\mu$ is around 10^{-3} [14]. Using $N_e\mu = 10^{-3}$, $\pi_M = 1$, $\pi_e = 0.8$, $\pi_f \leq 1$, $f_l = 1$, $k_s \leq 1$, and $P = 2$, gives a value ≥ 600 for the right hand side. Even with $PN_e = 2{\times}10^7$, $s_p = 10^{-3}$, and $h \leq 1$, the left hand side yields a value < 12. Which implies that the log fitness effect of the latent load is more than 50 times as large as the substitutional load in this case.

For non-strong selection, under the same conditions the effect of the substitutional load will be smaller than the latent load if,

$$2PN_e s_p < \frac{3\pi_M \pi_e f_l}{2P\pi_f N_e \mu k_s}$$

Here the left hand side will be small by virtue of the requirement that $PN_e s_p \gg 1$ is false. Thus, again, the log fitness effect of the latent load exceeds that of the substitutional load by more than a factor of 50.

Mutational load

Mutations at conserved sites are deleterious and are removed from the population either by selection or by chance back mutation. Suppose deleterious mutations occur randomly with a mean germline rate of U per organism per sexual generation. Let $-s_n$ denote the heterozygous selection coefficient associated with deleterious alleles, with $s_n > 0$. If $s_n \gg \mu$, and $s_n \gg \frac{1}{PN_e}$, so that selection dominates over drift, then Haldane has shown [5, 15],

$$\overline{x_m} \approx -U \tag{14}$$

Nearly neutral load

The analysis of the mutational load thus far only applies when selection dominates over random drift. When $|PN_e s| \approx 1$, selection and drift are in balance. This is the range of nearly neutral sites [16].

Kimura et al. investigated the mutational load when selection need not dominate over drift [7], and provided formulas for the total mutational load when $|PN_e s_n|$ is small or large, but not for the intermediate range. For $|2N_e s_n| < 5$ in a diploid with $h = \frac{1}{2}$, it was found,

$$\overline{x_m} \approx -\frac{2s_n}{1 + \frac{\nu}{\mu}\exp(4N_e s_n)} \tag{15}$$

where ν is the per site rate of back mutation.

The analysis here extends this finding by deriving a formula for the mutational load for the full range of selection strengths, and importantly partitioning the load into a μ-responsive and a non-μ-responsive component.

Due to random drift, nearly neutral sites can become fixed in the non-beneficial state. This increases the genetic load they generate. Sites that do this might be said to "flap", spending long periods fixed in both beneficial and deleterious states.

Even sites that are not neutral or nearly neutral can find themselves spending some time fixed in a deleterious state [6]. Such sites will be termed intermediate sites. It is convenient to consider the genetic load that results from this propensity to become fixed in a deleterious state as part of an effective nearly neutral load.

Assume $h = \frac{1}{2}$. Suppose there is a single base, base a, that has a selection coefficient of 0, and each of the other bases has a selection coefficient of $-s$; s may be positive or negative, with $|s| \ll 1$. The mutation rate for base a is μ, and the aggregate rate of back mutation from the other bases to base a is ν. This is a simplified model that ignores non-additive dominance and multi-allele dynamics for tractability.

The site will oscillate between state a, in which it was most recently fixed at base a, and state b, in which it was most recently fixed at some other base. Let t_a be the mean time (measured as the number of generations) spent in state a, and t_b be the mean time spent in state b. Let r_a denote the rate of mutation from base a to some other base, and r_b denote the rate of mutation from other bases to base a. Let $A_0(-s) < 0$ denote the mean time-aggregated fitness associated with a mutation that fails to fix when in state a, and $A_0(s) > 0$ denote the mean time-aggregated fitness gain associated with a mutation that fails to fix when in state b. Then the total time-averaged log fitness associated with a site, $x_m^{tot}(s)$, is given by,

$$x_m^{tot}(s) \approx \frac{t_a r_a A_0(-s) + t_b(r_b A_0(s) - Ps)}{t_a + t_b} \tag{16}$$

The term $-Ps$ occurs in the numerator because it is the fitness effect for state b in the absence of

the beneficial allele.

For state a,

$$r_a \approx PN\mu$$

Let P_a denote the probability that a single mutation eventually fixes when in state a. Since initially $p = \frac{1}{PN}$, and $s \ll 1$, from equation 1,

$$P_a \approx \frac{2\frac{N_e}{N}s}{\exp(2PN_e s) - 1}$$

And,

$$t_a = \frac{1}{r_a P_a}$$

Similarly, let P_b denote the probability that a single mutation eventually fixes when in state b,

$$\begin{aligned} P_b &\approx \frac{2\frac{N_e}{N}s}{1 - \exp(-2PN_e s)} \\ &\approx \exp(2PN_e s)P_a \end{aligned}$$

And,

$$\begin{aligned} r_b &= PN\nu \\ &= r_a \frac{\nu}{\mu} \end{aligned}$$

Hence,

$$\begin{aligned} t_b &= \frac{1}{r_b P_b} \\ &\approx \frac{t_a \mu}{\exp(2PN_e s)\nu} \end{aligned}$$

By equation 16,

$$x_m^{tot}(s) \approx \frac{PN\mu A_0(-s) + \frac{1}{\exp(2PN_e s)}\left(PN\mu A_0(s) - \frac{\mu}{\nu}Ps\right)}{1 + \frac{\mu}{\exp(2PN_e s)\nu}} \tag{17}$$

$A_0(s)$ can be determined in a manner similar to the substitutional load. Kimura has determined the expected number of generations until extinction of an allele with selective advantage s and proportion p as [9],

$$\frac{u(p)}{1-u(p)}\int_p^1 \psi(\xi)(1-u(\xi))^2 d\xi + \int_0^p \psi(\xi)(1-u(\xi))u(\xi)d\xi \tag{18}$$

where, using Kimura's notation,

$$V_{\delta p} = \frac{p(1-p)}{PN_e}$$
$$M_{\delta p} = sp(1-p)$$

so that,

$$u(\xi) = \frac{1-\exp(-2PN_e s\xi)}{1-\exp(-2PN_e s)}$$

and,

$$\psi(\xi) = \frac{1-\exp(-2PN_e s)}{s\xi(1-\xi)}\exp(2PN_e s\xi)$$

In equation 18, ξ is the proportion of haploid genomes with the allele. Multiplying the integrands by $Ps\xi$ therefore gives the time-aggregated fitness,

$$\begin{aligned} A_0(s) =& \frac{P(1-\exp(-2PN_e sp))}{\exp(-2PN_e sp)-\exp(-2PN_e s)}\int_p^1 \frac{\exp(2PN_e s\xi)(\exp(-2PN_e s\xi)-\exp(-2PN_e s))^2}{(1-\exp(-2PN_e s))(1-\xi)}d\xi \\ &+ P\int_0^p \frac{\exp(2PN_e s\xi)(\exp(-2PN_e s\xi)-\exp(-2PN_e s))(1-\exp(-2PN_e s\xi)}{(1-\exp(-2PN_e s))(1-\xi)}d\xi \end{aligned}$$

Assuming $N \gg 1$, since $p = \frac{1}{PN}$, p will be small and the second term is negligible. Replacing ξ by $1-\xi$ in the first integral then gives,

$$
\begin{aligned}
A_0(s) \approx & \frac{P(1-\exp(-2PN_e sp))}{\exp(-2PN_e sp)-\exp(-2PN_e s)} \\
& \frac{1}{\exp(4PN_e s)-\exp(2PN_e s)} \int_0^{1-p} \frac{\exp(2PN_e s(1-\xi))(\exp(2PN_e s\xi)-1)^2}{\xi} d\xi \\
\approx & \frac{2P\frac{N_e}{N}s}{(\exp(2PN_e s)-1)(1-\exp(-2PN_e s))} \\
& \times \begin{cases} (2PN_e s)^2 \frac{\xi}{2}\Big|_{\xi=0}^{1-p}, & |2PN_e s| \ll 1 \\ (\mathrm{Ei}(2PN_e s\xi)+\mathrm{Ei}(-2PN_e s\xi)-2log\xi)\Big|_{\xi=0}^{1-p}, & s \neq 0 \end{cases}
\end{aligned}
$$

Now $\mathrm{Ei}(x) \approx \gamma + \log|x|$ when $|x| \ll 1$, $\mathrm{Ei}(x) \approx 0$ when $x \ll -1$, and $\mathrm{Ei}(x) \approx \frac{\exp(x)}{x}$ when $x \gg 1$, so,

$$
A_0(s) \approx \begin{cases} P\frac{N_e}{N}s, & |2PN_e s| \ll 1 \\ \frac{1}{N}\frac{s}{|s|}, & |2PN_e s| \gg 1 \\ \frac{2P\frac{N_e}{N}s}{(\exp(2PN_e s)-1)(1-\exp(-2PN_e s))} & \\ \times(\mathrm{Ei}(2PN_e s)+\mathrm{Ei}(-2PN_e s)-2(\gamma+\log|2PN_e s|)), & s \neq 0 \end{cases}
$$

From this it follows that $A_0(-s) \approx -A_0(s)$. Hence, by equation 17,

$$
x_m^{tot}(s) \approx -\frac{PN\mu(\exp(2PN_e s)-1)A_0(s)+\frac{\mu}{\nu}Ps}{\exp(2PN_e s)+\frac{\mu}{\nu}} \tag{19}
$$

Assuming the Jukes-Cantor model ($\frac{\mu}{\nu}=3$), in several special cases this gives,

$$
x_m^{tot}(s) \approx \begin{cases} -\frac{3}{4}Ps, & |2PN_e s| \ll 1 \\ -P\mu, & 2PN_e s \gg 1 \\ P|s|-P\nu, & 2PN_e s \ll -1 \end{cases}
$$

which agrees with intuition and with equation 14.

For the general case, with a Jukes-Cantor model, Figure 3 shows the two components of $x_m^{tot}(s)$; a μ-responsive component and a non-μ-responsive component, which are summed to give $x_m^{tot}(s)$. See File S1 for the plotting code [17]. The non-μ-responsive component will be termed the s component, since for a given PN_e value, its value is dependent on s. In the figure the fitness effect of the μ component is divided by $P\mu$, and the s component is multiplied by N_e. The s component starts at zero, reaches a maximum of around $-\frac{0.3}{N_e}$ at $PN_e s \approx 0.8$, and then slowly falls back to almost zero by the time $PN_e s = 4$. The μ component starts at zero, interestingly increases to a maximum value larger than $-P\mu$ at $PN_e s \approx 2.9$, and then asymptotically falls back to the value $-P\mu$.

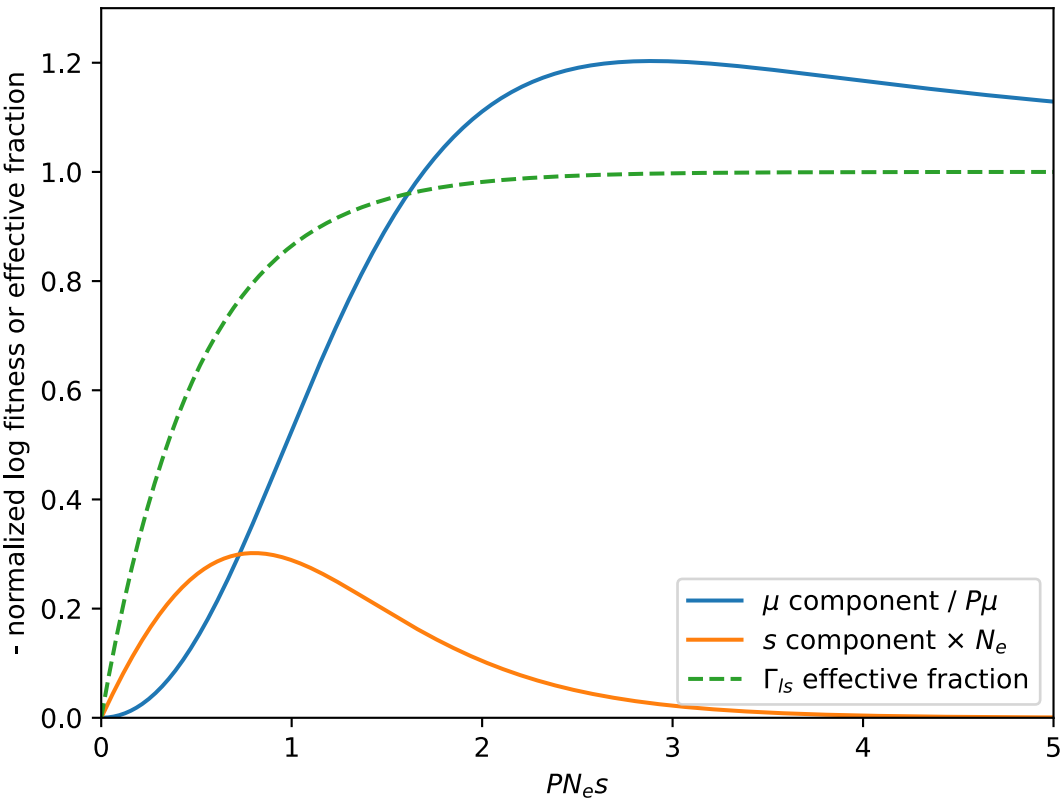

Figure 3: Time-averaged log fitness components for a site. Mutation rate μ, log fitness effect s, and coefficient of dominance, $h = \frac{1}{2}$, assuming a Jukes-Cantor model. Also shown is the Γ_{ls} effective fraction. Ploidy P. Effective population N_e.

Also shown for reference in Figure 3 is the Γ_{ls} effective fraction, $1 - e^{-2PN_e s}$. For small $PN_e s$, the Γ_{ls} effective fraction falls off to a lesser extent than the μ component.

Figure 3 shows that the s component of a nearly neutral site can be much larger than the standard $P\mu$ associated with the mutational load when $PN_e\mu \ll 1$. Consequently, depending on the number of nearly neutral sites, and their selection coefficients, the fitness loss associated with them could be substantial. They are nearly neutral with respect to individual selection, but not necessarily with respect to aggregate fitness.

An illustrative example, typical of vertebrates, helps clarify this point. Consider an exponential distribution of fitness effects for new positively selected sites with mean effect size $\overline{s_p} = 10^{-3}$. Numerical integration of the s component of equation 19 weighted by the distribution of fitness effects for a Jukes-Cantor model with $P = 2$ and $N_e = 10^5$ yields a mean log fitness of -2.4×10^{-8} per site. See File S1 for the numerical integration code. This is the mean value of the s component of nearly neutral sites. Meanwhile, the mean log fitness from the mutational load of those new sites that will become conserved is approximately $-P\mu$. All non-neutral sites were once new positively selected sites. Therefore, assuming their selection coefficients haven't changed, for $\mu = 10^{-8}$ the nearly neutral load would be larger than the mutational load.

Let $D_0(s)$ denote the genomic density of nearly neutral sites with selection coefficient s for some definition of nearly neutral. Then the log fitness loss associated with nearly neutral sites, $\overline{x_0}$, is given by,

$$\overline{x_0} = \int_0^\infty x_m^{tot}(s) D_0(s) ds$$

At times it may be useful to split $\overline{x_m^{tot}}$ into its μ and s components, and speak of an effective mutational load having fitness effect $\overline{x_m^{eff}}$ which combines the μ component of nearly neutral sites with the mutational load of conserved sites, and to speak of the effective nearly neutral load having fitness effect $\overline{x_0^{eff}}$ which consists of just the s component of nearly neutral and intermediate sites.

$$\begin{aligned}\overline{x_m^{eff}} &= \int_0^\infty x_m^{eff}(s) D(s) ds \\ &= -U^{eff}; \text{ this is the definition of } U^{eff} \text{ assuming } h = \frac{1}{2}\end{aligned} \tag{20}$$

$$\overline{x_0^{eff}} = \int_0^\infty x_0^{eff}(s) D(s) ds \tag{21}$$

where $D(s)$ is the density of sites with selection coefficient s in the genome, and for $h = \frac{1}{2}$,

$$x_m^{eff}(s) \approx -\frac{PN\mu(\exp(2PN_e s) - 1)A_0(s)}{\exp(2PN_e s) + \frac{\mu}{\nu}}$$

$$x_0^{eff}(s) \approx -\frac{Ps}{1 + \frac{\nu}{\mu}\exp(2PN_e s)}$$

This last equation is in agreement with equation 15 derived by Kimura et al. [7].

Interference between positively selected, conserved, and nearly neutral sites

When combining multiple site types, it is important to consider possible interference between their fitness effects. Interference occurs when the presence of one site affects the fitness effect of another.

There is no interference from the latent load on the other loads because, apart from rare copies of beneficial alleles that die out, there is no variation at the latent site. Interference by the other loads on the latent load is primarily captured by their effects on the effective population size. Appendix B discusses the clustering of new positively selected sites, an effect not captured by the effective population size.

Because the substitutional load is much smaller than the latent load, interference from conserved sites or other substituting sites, on the substitutional load is expected to only have a second-order effect on total fitness. The rarity of substituting sites also means that interference by substituting sites on the mutational load is unlikely to exert a significant overall fitness effect.

Interference by nearly neutral sites on substituting sites and conserved sites is expected to have a small effect. Nearly neutral sites may be numerous, but their individual selection effects are minimal. As a result, the inter-organism variance and the resulting aggregate selection effects of such sites will be minor. Finally, interference from substituting sites, conserved sites, and other nearly neutral sites on the nearly neutral load is possible; however, these effects are again taken into account through the effective population size.

Fitness of positively selected, conserved, and nearly neutral sites

The fitness effects of positively selected sites are captured by $\overline{x_l}$ and $\overline{x_s}$. The μ-responsive fitness effects of conserved sites, intermediate sites, and nearly neutral sites are captured by $\overline{x_m^{eff}}$. And the non-μ-responsive fitness effects of nearly neutral sites and intermediate sites are captured by $\overline{x_0^{eff}}$.

Let $\overline{x}$ denote the mean log fitness of a population. Combining equations 8, 13, 20, and 21,

$$\begin{aligned}\overline{x} &= \overline{x_l} + \overline{x_s} + \overline{x_m^{eff}} + \overline{x_0^{eff}} + \cdots \\ &\approx -\frac{3\pi_e\Gamma_{ls}}{2P\pi_f N_e \mu h} - \Gamma_s A_s - U^{eff} + \int_0^\infty x_0^{eff}(s)D(s)ds + \cdots \end{aligned} \tag{22}$$

where "$\cdots$" denotes the effect of other genetic loads, not considered here, such as the external load and the load associated with heterozygote advantage.

For multiple sites with different dominance coefficients, equation 22 is still applicable, provided that h in the first term and in A_s is replaced by appropriately weighted average values.

Also note, equation 19 for $x_0^{eff}(s)$ is only applicable for $h = \frac{1}{2}$ and a single beneficial allele. If this is not the case, a more complex formula will be necessary.

The population-optimal mutation rate

The likelihood of a mutation being harmful is far greater than the likelihood that it is beneficial. Further, the magnitude of the fitness effect of harmful mutations is typically greater than that of beneficial mutations. Consequently, it has been hypothesized that natural selection drives the mutation rate down to a lower bound set by the balance between selection for lower mutation rates and random genetic drift [1, 2].

Here it is proposed that, in the long run, evolution biases the germline spontaneous mutation rate per sexual generation so that it is as large as possible to maximize the ability to adapt, and hence to minimize the fitness cost associated with the latent load, but not so large that the mutational load incurs a heavy cost. That is, evolution results in a mutation rate that approaches the species or population-optimal value. This is distinct from the optimal mutation rate for an individual organism or its genes. This process may occur through macroevolution. Species that evolve mutation rates that are too low fail to adapt, incur a larger latent load, and are more likely to become extinct. Species with mutation rates that are too high incur a heavy mutational load, produce fewer successful offspring, and are likewise more likely to become extinct.

At first glance, it may seem implausible that species go extinct due to suboptimal mutation rates. But this is proposed to be only a contributing factor. Other stochastic factors also contribute to the success of a species. But in the same way that the selective advantage of a new genotype does not determine the realized number of copies in the next generation, but only the odds of copies, the optimality of the mutation rate affects the odds of species survival. Over macroevolutionary time scales these odds can be expected to accumulate, leading to mutation rates in the surviving species that are roughly appropriate for the relevant lineages.

An analogy might be drawn between the proposed macroevolution of sexual species with variable mutation rates, and the observed microevolution of asexual organisms with variable mutation rates. A sexual species is viewed as analogous to a single asexual organism. In bacteria when there is a significant selective challenge, and a higher mutation rate is desirable, mutator alleles that increase the mutation rate frequently arise and occasionally spread by hitchhiking with beneficial mutations [18]. Once the challenge has been met, the mutation rates subsequently decline [19]. For this analogy to be valid there needs to be some mechanism for species to occasionally acquire higher mutation rates simultaneously with increased population mean fitness. Whether this is possible through stochastic effects in small species founder populations is not certain. Another mechanism through which a higher mutation rate might be possible is if the species evolves a longer lifespan. Other things being equal, a longer lifespan might frequently be expected to increase the per generation mutation rate, while at the same time increasing the number of offspring and thus offering a possible fitness boost. This mechanism also is not certain as there are complications in so far as a longer lifespan might also reduce the population mean fitness of the progeny.

Let L_n denote the number of sites under the control of negative selection per haploid genome, so that,

$$U = PL_n\mu \tag{23}$$

Strictly speaking, in what follows, L_n and U should be considered effective quantities, reflecting not just conserved sites but also the partial μ log fitness component of nearly neutral sites.

Substituting equation 23 into equation 22, and assuming any other loads are either small or only weakly dependent on μ, gives,

$$\overline{x} \approx -\frac{3\pi_e\Gamma_{ls}}{2P\pi_f N_e\mu h} - PL_n\mu + K$$

where K consists of terms that are either independent of, or only weakly dependent on, μ.

Suppose P, L_n, and μ are fixed, but that the other parameters are functions of time.

A population is most likely to be successful if the product of successive $\overline{w}$ terms is as large as possible, or equivalently their generationally averaged geometric-mean value is maximized. Let s be the aggregate relative selective advantage of an organism. It is unusual for the total population to grow by a large relative amount in a single generation. It is also highly unlikely that some genotype in the population is fitter than the mean log fitness of the population by anything other than a small amount. Remembering $\bar{\cdot}$ denotes the mean over the population, while Geomean$[\cdot]$ and E$[\cdot]$ are averages over generations.

$$\begin{aligned}
\text{Geomean}[\overline{w}] &= \text{Geomean}[\overline{1+s}] \\
&= \text{Geomean}[1+\overline{s}] \\
&\approx \text{Geomean}[e^{\overline{s}}], \text{ since } \overline{s} \text{ is usually small} \\
&\approx \text{Geomean}[e^{\overline{\log(1+s)}}], \text{ since } s \text{ is usually small} \\
&= \text{Geomean}[e^{\overline{x}}] \\
&= e^{\text{E}[\overline{x}]}
\end{aligned}$$

The population is therefore most likely to be successful if E$[\overline{x}]$ is maximized. Using equation 10,

$$\text{E}[\overline{x}] \approx -\text{E}\left[\frac{3\pi_M\pi_e f_l\Gamma_s}{2P\pi_f N_e\mu h}\right] - PL_n\mu + K \tag{24}$$

Differentiating equation 24 with respect to μ shows that the geometric-mean population mean fitness is maximized when,

$$\text{E}\left[\frac{3\pi_M\pi_e f_l\Gamma_s}{2P\pi_f N_e\mu^2 h}\right] \approx PL_n$$

So the maximal population fitness is given by $\hat{\mu}|g$, where,

$$\hat{\mu}|g \approx \sqrt{\mathrm{E}\left[\frac{3\pi_M \pi_e f_l \Gamma_s}{2P^2 \pi_f N_e L_n h}\right]} \tag{25}$$

The given g qualifier signifies this value is for a given generation length. It is not necessarily the optimal value if the generation length, and thus Γ_p and Γ_s, are free to also change.

Equation 25 gives the fixed mutation rate that would be near optimal for the species over the time frame (measured in generations) for which the expectation was taken. Care must be taken in applying equation 25 if the parameter values vary over time; it might not be possible to accurately drop the expectation operator.

It should be obvious that species do not know in advance the parameter values for the optimal mutation rate; rather the equation for population mean fitness acts as a mutation rate "filter" through which competing species must pass.

The maximal value log fitness is given by,

$$\overline{x} \approx -2U + K$$

At the population-optimal mutation rate, mean fitness losses from the latent load and mutational load are approximately equal.

Microevolution might drive μ to a value below than the population-optimal value. Let $\pi_\mu|g$ denote the multiplicative factor by which the mutation rate is smaller or larger than its optimal value. Then,

$$\mu \approx (\pi_\mu|g)(\hat{\mu}|g) \approx \sqrt{\mathrm{E}\left[\frac{3(\pi_\mu|g)^2 \pi_M \pi_e f_l \Gamma_s}{2P^2 \pi_f N_e L_n h}\right]} \tag{26}$$

and

$$\overline{x_m} \approx -U \approx (\pi_\mu|g)^2\, \mathrm{E}[\overline{x_l}]$$

Most of the terms in the right hand side of equation 26 are expected to have similar values within very broad clades (Linnaean kingdoms). The exceptions are L_n, which might show some N_e dependence but should still be largely constant within slightly less broad clades (Linnaean phyla), N_e, and lastly Γ_s, which might like Γ_p exhibit considerable stochastic variation.

Ignoring the temporal effects of the expectation operator, equation 26 then suggests an inverse (square root) relationship between population sizes and mutation rates for sexual species. Species with large population sizes such as invertebrates, and especially unicellular eukaryotes, are expected to have smaller mutation rates per sexual generation than vertebrates. This appears to agree with

empirical observations [14]. Assuming other parameters are held constant, equation 26 predicts the log mutation rate equals $-\frac{1}{2}$ times the log effective population size, plus a constant, plus noise from the variation in Γ_s. This also approximately agrees with empirical observations both within and between clades [2, 14].

Discussion

This manuscript has considered sexual species. For the optimal mutation rate of asexual populations analogous considerations are expected to apply, though the underlying dynamics will differ. Here the concept of a species could be replaced by lineages of organisms having identical-by-descent mutation rates. The log fitness loss associated with the asexual analogue of the latent load is inversely proportional to the product of the population size and the mutation rate, while the mutational load is proportional to the mutation rate. This again leads to a population-optimal mutation rate in which the mean latent load and the mutational load are approximately equal. Assuming other factors are held constant, the population-optimal mutation rate is predicted to be inversely proportional to the square root of the population size.

A model of the fitness effects of positively selected, conserved, and nearly neutral sites was constructed. A new load, the latent load, was carved out from the lag load and analyzed. The mean latent load is comparable in size to the mutational load, and much larger than the substitutional load. The nearly neutral load was also defined and analyzed.

The constructed model implies the existence of a spontaneous mutation rate per base pair that maximizes the aggregate mean fitness of a population. If the mutation rate is too low, beneficial alleles take a long time to establish, increasing the latent load. If it is too high, negative selection incurs a large mutational load. The population-optimal mutation rate is one at which the mean fitness losses from the latent and mutational loads are approximately equal. The extent to which evolution approaches the population-optimal mutation rate is a question for further study.

Appendix A: The latent fraction of new positively selected sites is close to one

The latent load depends on the value of f_l, the proportion of new positively selected sites for which no existing copies of the beneficial allele are already established.

For a new positively selected site that arises at a previously neutral site, f_l can be estimated as follows.

Assume the beneficial allele suddenly has selection coefficient, s_p, and all other possible alleles have a selection coefficient of zero.

From equation 1, assuming $2PN_e s_p \gg 1$, the probability of fixation of an allele with proportion x is given by,

$$f(x) \approx 1 - e^{-2PN_e s_p x}$$

The work of Kimura provides a method for determining the mean value of $f(x)$ [8]. Adopting Kimura's notation, let,

$$V_{\delta p} = \frac{p(1-p)}{PN_e}$$
$$M_{\delta p} = sp(1-p)$$

for some heterozygous selection coefficient s, with dominance coefficient, $h = \frac{1}{2}$. In a moment, the limit as $s \to 0$ to describe a previously neutral site will be considered. Let,

$$\begin{aligned} G(x) &= \exp\left(-2\int_0^x \frac{M_{\delta p}}{V_{\delta p}} dp\right) \\ &= \exp\left(-2PN_e sx\right) \end{aligned}$$

Let,

$$\begin{aligned} u(p) &= \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} \\ &= p \text{ in the limit as } s \to 0 \end{aligned}$$

If the mutation rate is μ, and one third of mutations produce the currently neutral, but soon-to-be beneficial allele, and $\psi_f(\xi)$ is defined by,

$$\begin{aligned} \psi_f(\xi) &= \frac{2PN\mu f(\xi)}{3V_{\delta\xi}u'(\xi)} \\ &\approx \frac{2}{3}P^2 N_e N\mu \frac{1 - e^{-2PN_e s_p \xi}}{\xi(1-\xi)} \end{aligned}$$

and $I_f(p)$ by,

$$I_f(p) = (1 - u(p))\int_0^p \psi_f(\xi)u(\xi)d\xi + u(p)\int_p^1 \psi_f(\xi)(1 - u(\xi))d\xi$$

then

$$\overline{f(x)} = I_f\left(\frac{1}{PN}\right)$$
$$\approx \frac{2}{3}PN_e\mu\left(\log(\xi) - \mathrm{Ei}(-2PN_e s_p\xi)\right)\Big|_{\xi=\frac{1}{PN}}^{1}$$

because the first term of $I_f\left(\frac{1}{PN}\right) \approx 0$ when $PN \gg 1$, where Ei is the exponential integral,

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt$$

Now, $\mathrm{Ei}(x) \approx 0$ for $x \ll -1$. In addition, $\mathrm{Ei}(x) \approx \gamma + \log(|x|)$ for $|x| \ll 1$ where γ is Euler's constant, $0.5772\ldots$. Thus, assuming $2PN_e s_p \gg 1$ and $2\frac{N_e}{N}s_p \ll 1$,

$$\overline{f(x)} \approx \frac{2}{3}PN_e\mu(\log(2PN_e s_p) + \gamma)$$

This is the probability of fixation of the now-beneficial allele at a site that was neutral and suddenly became a new positively selected site. The latent fraction, f_l is $1 - \overline{f(x)}$, or,

$$f_l \approx 1 - \frac{2}{3}PN_e\mu(\log(2PN_e s_p) + \gamma) \tag{27}$$

Now consider a range of selection coefficients. Suppose an exponentially distributed distribution of fitness effects of new positively selected sites having mean selection coefficient $\overline{s_p}$. The exponential distribution is chosen here for analytical convenience. From equation 7 the latent load is likely to be significant when $2PN_e s_p \gtrsim 1$. Evaluating the latent fraction for just those sites,

$$f_l \approx 1 - \frac{2}{3}PN_e\mu\frac{\int_{\frac{1}{2PN_e}}^{\infty} \frac{1}{\overline{s_p}}\exp\left(\frac{-s}{\overline{s_p}}\right)(\log(2PN_e s) + \gamma)\,ds}{\int_{\frac{1}{2PN_e}}^{\infty} \frac{1}{\overline{s_p}}\exp\left(\frac{-s}{\overline{s_p}}\right)ds}$$
$$= 1 - \frac{2}{3}PN_e\mu\left(\frac{-\exp\left(\frac{-s}{\overline{s_p}}\right)\log(2PN_e s) + \mathrm{Ei}\left(\frac{-s}{\overline{s_p}}\right)\Big|_{s=\frac{1}{2PN_e}}^{\infty}}{\exp\left(\frac{-1}{2PN_e\overline{s_p}}\right)} + \gamma\right)$$
$$= 1 - \frac{2}{3}PN_e\mu\left(\frac{-\mathrm{Ei}\left(\frac{-1}{2PN_e\overline{s_p}}\right)}{\exp\left(\frac{-1}{2PN_e\overline{s_p}}\right)} + \gamma\right)$$

For values typical of vertebrates, $PN_e = 2\times10^5$, $\mu = 10^{-8}$, and $\overline{s_p} = 10^{-3}$, this results in $f_l = 0.992$.

This analysis assumes $h = \frac{1}{2}$; other dominance coefficients may change the scaling but are unlikely to alter the qualitative result.

This result applies to a new positively selected site that arises at a previously neutral site. If the new positively selected site arises at a site where the beneficial allele was previously deleterious, the initially expected proportion of the beneficial allele is smaller, making f_l even closer to 1.

Appendix B: Clustering of new positively selected sites

Adaptation to an environmental change frequently appears to consist of an adaptive walk of three to six steps [20], rather than a long series of steps where each step is triggered by the preceding step. It does not appear to be known whether the ordering of the steps is largely arbitrary, or whether the ordering is fixed due to sign epistasis between adaptive sites. It also does not appear to be known whether the adaptive sites are located close to each other on the genome. If the ordering of the steps is arbitrary and the sites are located close to each other, then there is a risk that a beneficial allele undergoing a selective sweep at one site may interfere with and reduce the probability of fixation of a beneficial allele at another.

Most interference between sites is captured by N_e. However, the effect of the clustering of new positively selected sites will not be captured by N_e because N_e is the neutral site effective population size averaged across the entire genome, rather than the neutral effective population size of positively selected sites clustered close to other positively selected sites. This reduction in the probability of fixation is denoted π_f. Since $\pi_f \leq 1$, equation 10 in the absence of π_f may underestimate the mean latent load.

There is some concern that unobserved new positively selected sites of smaller fitness effect than those detected in the studies of adaptive walks cited above might increase the mean length of an adaptive walk. However, as will be shown, if they are of smaller fitness effect they are unlikely to overlap in time with the detected sites. Conceivably, such sites could then overlap in time with each other reducing their probability of fixation. This effect is not modeled here, but could contribute to a reduction in the estimate for π_f.

The rate of adaptive amino acid substitution at a given location in the genome has been reported to be positively correlated with the mutation rate and the recombination rate, and negatively correlated with gene density in *Drosophila* [21]. Based on this it was estimated that genome-wide Hill-Robertson interference is responsible for the loss of about 27% of all adaptive substitutions. However, the meaning of this is not clear for the model used in this manuscript. If a beneficial mutation fails to establish and ultimately fix, then the latent site will remain, beneficial mutations will reoccur, and ultimately establish and fix. In the model used, adaptive substitutions are delayed rather than permanently lost.

Barton has calculated the probability of fixation of a strongly selected beneficial mutation at a site that is tightly linked to another site that is undergoing a selective sweep [22]. For a beneficial mutation with an effect size much smaller than that of the allele that is sweeping, and occurring prior to the sweeping allele reaching a frequency of 50%, the reduction can be substantial.

Consider the probability that two would be selective sweeps at two nearby new positively selected

sites overlap critically in time.

Suppose there are two sites. A first site site harboring a beneficial allele with selection coefficient s_p, for which $PN_e s_p \gg 1$, and a second site with a beneficial allele with selection coefficient S_p, $PN_e S_p \gg 1$ that is going to undergo an selective sweep. Assume the two sites are linked. Also assume the two sites were created as a result of the same environmental change, and that initially both sites have a frequency of zero for their respective beneficial alleles. Assume the second site is first establish; otherwise swap the two sites.

Suppose $h = \frac{1}{2}$. The time from having a single copy of the second beneficial allele until it reaches the critical value of 50%, T_S, is approximately given by [23],

$$T_S \approx \frac{\log PN_e}{S_p}$$

Let T_l denote the time, measured from the establishment of the second site, until the first site would have become established in the absence of the second site. Beneficial mutations at the first site are occurring at rate $PN\frac{\mu}{3}$, and the probability of fixation is normally $2\frac{N_e}{N}s_p$. Consequently T_l has an exponential distribution with rate parameter $\lambda_s = 2PN_e s_p \frac{\mu}{3}$.

Let P_{crit} denote the probability that the first site would have been established after the second site, but prior to the second site reaching a frequency of 50%; this is given by,

$$\begin{aligned} P_{crit} &\approx 1 - e^{-\lambda_s T_S} \\ &= 1 - (PN_e)^{-2PN_e \frac{s_p}{S_p} \frac{\mu}{3}} \end{aligned}$$

For $s_p = S_p$ and $P = 2$, this gives $P_{crit} \approx 0.28\%$ when $N_e = 2{\times}10^4$ and $\mu = 1{\times}10^{-8}$, $P_{crit} \approx 1.7\%$ when $N_e = 2{\times}10^5$ and $\mu = 5{\times}10^{-9}$, and $P_{crit} \approx 9.2\%$ when $N_e = 1{\times}10^6$, and $\mu = 5{\times}10^{-9}$, values intended to be roughly typical for apes, rodents, and insects respectively. These values are likely overestimates. The fact that the second site established first typically means $s_p < S_p$.

P_{crit} gives the probability of a potentially critical interaction between the two sites. The potentially critical interaction only applies to the first site. The second site is already established, and certain to fix. Selecting one of these two sites at random, the probability of a potentially critical interaction is $\frac{P_{crit}}{2}$. The extent to which this then reduces the probability of fixation of the first site will depend on the ratio of the selection coefficients, the degree of linkage between the sites, and the time at which the beneficial mutation which would have been establishing occurs. Consequently, the mean reduction in the probability of fixation, consequent reduction in the effective population size, and thus increase in the latent load, might be at most a few percent.

Thus the tight linkage of two positively selected sites will infrequently affect their probability of fixation. The mean number of sites actually interacting is limited by adaptive walks that typically involve three to six steps, not all of which may potentially occur concurrently, and not all of which are at sites located close to each other on the genome. Hence the clustering of new positively selected sites appears unlikely to appreciably increase the latent load, and π_f is expected to be close to 1 (e.g., > 0.99 for apes and rodents, and > 0.97 for insects).

The population-optimal mutation rate is obtained by taking the square root of an expression involving the effective population size; equation 25. Thus the population-optimal mutation rate will be increased on account of clustering of new positively selected sites by an even smaller proportion than the latent load.

Appendix C: The effective rate of latent to substituting sites is approximately equal to the raw rate

Let $\rho_p(s)$ denote the distribution of fitness effects for the beneficial alleles of sites transitioning from latent to substituting sites having selection coefficient s. By equation 9 this is approximately the same as the distribution of fitness effects for new substituting sites. Theory suggests $\rho_p(s)$ might be modeled as an exponential distribution with mean fitness effect size $\overline{s_p}$ [24],

$$\rho_p(s) \approx \frac{1}{\overline{s_p}} \exp\left(-\frac{s}{\overline{s_p}}\right)$$

Note that $\rho_p(s)$ is not the same as the distribution of fitness effects for beneficial mutations that occur. Beneficial mutations of smaller effect size are less likely to fix, and so the sites that harbor them will persist as positively selected sites for longer, allowing the same beneficial mutation to recur more often prior to fixation.

$\Gamma_{ls}^{raw}(s)$ is the distribution of fitness effects for sites transitioning from latent to substituting sites per sexual generation.

Let,

$$\Gamma_{ls}^{raw} = \int_0^\infty \Gamma_{ls}^{raw}(s) ds$$

Then by definition,

$$\Gamma_{ls}^{raw}(s) = \rho_p(s)\Gamma_{ls}^{raw}$$

The effective rate per sexual generation of transition from latent to substituting sites, Γ_{ls}, is defined by equation 7 as,

$$\begin{aligned}
\Gamma_{ls} &= \int_0^\infty \Gamma_{ls}^{raw}(s)\left(1 - \exp\left(-2PN_e s\right)\right) ds \\
&\approx \Gamma_{ls}^{raw} \frac{1}{\overline{s_p}} \int_0^\infty \exp\left(-\frac{s}{\overline{s_p}}\right) - \exp\left(-\left(2PN_e + \frac{1}{\overline{s_p}}\right) s\right) ds \\
&\approx \left(\frac{2PN_e\overline{s_p}}{2PN_e\overline{s_p} + 1}\right) \Gamma_{ls}^{raw}
\end{aligned}$$

For humans, it has been estimated that $\overline{s_p} \approx 6.4\times10^{-3}$ with a 95% confidence interval of 7×10^{-4} to 1×10^{-1} [25], and N_e is of the order of 10,000. The distribution of fitness effects for positive selection in *Mus musculus* and in *Drosophila melanogaster* have previously been modeled using a single fitness class with PN_es_p equal to 15 and 23 respectively [26]. Consequently,

$$\Gamma_{ls} \approx \Gamma_{ls}^{raw}$$

with an error of 3% or less for the three species examined.

Appendix D: Epistasis and the latent load

Epistasis occurs when the fixation of one positively selected site changes the selection coefficients of alleles at other sites. There are several possibilities. An affected site may be a latent site, in which case the selection coefficient of the beneficial allele may increase, which is termed positive epistasis. Or the selection coefficient of the beneficial allele may decrease, which is termed negative epistasis, or it may even become negative, which is termed sign epistasis. Or an affected site may be a conserved site. A change in the selection coefficient of an allele of a conserved site normally will not have any appreciable effect on population mean fitness. The one exception is sign epistasis, in which the selection coefficient of a previously deleterious allele becomes positive, and which creates the potential for future changes in fitness.

Epistasis involving substituting sites and nearly neutral sites is not modeled. Epistasis between latent sites and sign epistasis at conserved sites is considered here.

Sign epistasis is taken care of by the definition of Γ_p as a net rate, and by the definition of Γ_{ls} as the rate of transition from latent sites to substituting sites. This takes into account positively selected sites disappearing and also emerging due to sign epistasis.

In the presence of positive or negative epistasis, the latent load is defined in terms of the change in fitness that results from the fixation of all current positively selected sites that are destined to fix were the environment held constant. Sites of final sign epistasis are excluded from this definition. It is impossible to hold the environment constant, life may not be deterministic, and it is only known retrospectively whether a positively selected site is destined to fix or, alternatively, to undergo sign epistasis and cease being a positively selected site. This might cause empirical difficulties in precisely determining the latent load, but it introduces no conceptual challenges. Similarly, the external load can be defined in terms of the difference between the organism that results from the fixation of all current positively selected sites that are destined to fix and the globally optimal organism.

Let π_e denote the mean multiplicative adjustment to the log fitness effect of the beneficial alleles of latent sites that are destined to fix so as to incorporate the fitness effects of the positive and negative epistasis experienced by all the other latent sites that are destined to fix. It follows that π_e is also the multiplicative adjustment to the log fitness effect of the latent load that takes into account epistasis.

Mean negative epistasis between beneficial mutations appears to commonly occur [27, 28, 29, 30, 31].

Using colony size as a proxy for fitness, Schoustra et al. investigated the epistatic effects of single and double mutants for the loss of an unused expensive cellular function [31]. For mutations of similar effect size, they found double mutants had a mean fitness epistatic effect size relative to the mean of the two single mutants fitness effects of -0.18 in a rich physical environment and -0.66 in a poor physical environment when likely cases of sign epistasis are removed [31][Calculated from data supplement]. This is suggestive of an mean epistatic effect size relative to the sum of the two log fitness effects of around -0.2.

At any point in time there might be many latent sites. However, not all are destined to fix; some undergo reciprocal sign epistasis in response to those that fix. As discussed in Appendix B, adaptive walks in response to an environmental change typically consist of three to six steps. Not all of these steps may be capable of occurring concurrently; some may only arise in response to others. Thus the latent positively selected can probably be arranged into very small groups corresponding to concurrently feasible sites of adaptive walks for particular phenotypic effects, with relatively little overall epistasis occurring between groups. If the mean size of these groups is around two then based on the data of Schoustra et al. an estimate for π_e is 0.8; approximate range 0.6 to 0.9.

Appendix E: The emergence of positively selected sites and the latent load

Until this point, the model has assumed an allele is either neutral or deleterious, and then instantaneously becomes beneficial with a constant selection coefficient. The real world is unlikely to operate like this. More likely the beneficial nature of the allele increases gradually over time.

Let the selection coefficient, $s(i)$, of the beneficial allele at the site be a function of the number of generations, i, since the allele first started having a beneficial effect. Suppose the selection coefficient is strong, $PN_e s(i) \gg 1$ for all i, so that the probability of fixation of a single copy of the allele arising at generation j is $2\frac{N_e}{N}s(j)$.

The expected log fitness associated with the latent load of a single latent site with age i is given by the latent load for one period times the product of the probabilities that the beneficial allele was not destined to fix in any prior period,

$$\text{expected log fitness of a latent site with age } i \approx -\frac{s(i)}{h}\prod_{j=0}^{i-1}\left(1 - PN\frac{\mu}{3} \times 2\frac{N_e}{N}s(j)\right)$$

Then, the mean log fitness associated with latent sites, $\overline{x_l}$, is,

$$\begin{aligned}\overline{x_l} &= \text{latent to substitutional rate} \times \text{the sum of the expected log fitness for each age}\\ &\approx -\Gamma_{ls} \times \sum_{i=0}^{\infty}\frac{s(i)}{h}\prod_{j=0}^{i-1}\left(1 - \frac{2}{3}PN_e\mu s(j)\right)\end{aligned}$$

As is to be expected, for $s(i) = s_p$, this produces,

$$\overline{x_l} \approx -\frac{3\Gamma_{ls}}{2PN_e\mu h}$$

Let r denote the ratio of $\overline{x_l}$ for arbitrary $s(i)$ compared to that for $s(i) = s_p$,

$$r = \frac{2}{3}PN_e\mu \sum_{i=0}^{\infty} s(i) \prod_{j=0}^{i-1} \left(1 - \frac{2}{3}PN_e\mu s(j)\right)$$

Let,

$$a_i = \frac{2}{3}PN_e\mu s(i)$$

then,

$$\begin{aligned}
r &= \sum_{i=0}^{\infty} a_i \prod_{j=0}^{i-1} (1 - a_j) \\
&= a_0 + a_1(1 - a_0) + a_2(1 - a_0)(1 - a_1) + \cdots \\
&= 1 - (1 - a_0) + a_1(1 - a_0) + a_2(1 - a_0)(1 - a_1) + \cdots \\
&= 1 - (1 - a_0)(1 - a_1) + a_2(1 - a_0)(1 - a_1) + \cdots \\
&= 1 - (1 - a_0)(1 - a_1)(1 - a_2) + \cdots \\
&= 1 - (1 - a_0)(1 - a_1)(1 - a_2) \cdots \\
&= 1, \text{ provided } \lim_{i \to \infty} s(i) > \epsilon \text{ for some } \epsilon > 0
\end{aligned}$$

Thus provided the beneficial allele is always strong, the gradual emergence of latent sites does not alter the formula for the fitness effect of the latent load, $\overline{x_l}$.

It is possible that gradual emergence may alter f_l, the proportion of new positively selected sites that are latent sites. However, equation 27 shows f_l is a decreasing function of s_p, and so the effect will be to push f_l even closer to 1.

Data availability

The code and data used to produce Figure 3 are publicly available as File S1.

File S1 - Nearly neutral log fitness software and results.
`https://doi.org/10.5281/zenodo.18102709`

Acknowledgments

I am grateful to Michael Lynch for the time he spent reviewing this manuscript and for providing valuable feedback. I am also grateful to Steven Greidinger, Philip Gerrish, and Benjamin Good for the time they spent reviewing an earlier related manuscript.

Conflict of interest statement

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

References

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

[2] 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.

[3] Edward J Klekowski. Genetic load and its causes in long-lived plants. *Trees*, 2:195–203, 1988.

[4] J Maynard Smith. What determines the rate of evolution? *The American Naturalist*, 110(973):331–338, 1976.

[5] JBS Haldane. The effect of variation of fitness. *The American Naturalist*, 71(735):337–349, 1937.

[6] Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.

[7] Motoo Kimura, Takeo Maruyama, and James F Crow. The mutation load in small populations. *Genetics*, 48(10):1303, 1963.

[8] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[9] Motoo Kimura and Tomoko Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763, 1969.

[10] Motoo Kimura. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *Journal of Genetics*, 57(1):21–34, 1960.

[11] Motoo Kimura. On the evolutionary adjustment of spontaneous mutation rates. *Genetics Research*, 9(1):23–34, 1967.

[12] Yoh Iwasa, Franziska Michor, and Martin A Nowak. Stochastic tunnels in evolutionary dynamics. *Genetics*, 166(3):1571–1579, 2004.

[13] Zaheerabbas Patwa and Lindi M Wahl. The fixation probability of beneficial mutations. *Journal of The Royal Society Interface*, 5(28):1279–1289, 2008.

[14] Michael Lynch, Farhan Ali, Tongtong Lin, Yaohai Wang, Jiahao Ni, and Hongan Long. The divergence of mutation rates and spectra across the Tree of Life. *EMBO Reports*, 24(10):e57561, 2023.

[15] Aneil F Agrawal and Michael C Whitlock. Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):115–135, 2012.

[16] Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, pages 263–286, 1992.

[17] Gordon Irlam. An integrated model of the effects on fitness of positively selected, conserved, and nearly neutral sites in sexual populations and the population-optimal mutation rate: Supplement 1, 2025. doi:10.5281/zenodo.18102709.

[18] Paul D Sniegowski, Philip J Gerrish, and Richard E Lenski. Evolution of high mutation rates in experimental populations of E. coli. *Nature*, 387(6634):703–705, 1997.

[19] Sébastien Wielgoss, Jeffrey E Barrick, Olivier Tenaillon, Michael J Wiser, W James Dittmar, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E Lenski, and Dominique Schneider. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences*, 110(1):222–227, 2013.

[20] Sijmen E Schoustra, Thomas Bataillon, Danna R Gifford, and Rees Kassen. The properties of adaptive walks in evolving populations of fungus. *PLOS Biology*, 7(11):e1000250, 2009.

[21] David Castellano, Marta Coronado-Zamora, Jose L Campos, Antonio Barbadilla, and Adam Eyre-Walker. Adaptive evolution is substantially impeded by Hill–Robertson interference in Drosophila. *Molecular Biology and Evolution*, 33(2):442–455, 2016.

[22] Nick H Barton. Linkage and the limits to natural selection. *Genetics*, 140(2):821–841, 1995.

[23] Sarah P Otto and Michael C Whitlock. Fixation probabilities and times. *Encyclopedia of Life Sciences*, pages 1–5, 2005.

[24] H Allen Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526, 2003.

[25] David Castellano, Moisès Coll Macià, Paula Tataru, Thomas Bataillon, and Kasper Munch. Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. *Genetics*, 213(3):953–966, 2019.

[26] Tom R Booker. Inferring parameters of the distribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare. *G3: Genes, Genomes, Genetics*, 10(7):2317–2326, 2020.

[27] Darin R Rokyta, Paul Joyce, S Brian Caudle, Craig Miller, Craig J Beisel, and Holly A Wichman. Epistasis between beneficial mutations and the phenotype-to-fitness map for a ssDNA virus. *PLOS Genetics*, 7(6):e1002075, 2011.

[28] Martijn F Schenk, Ivan G Szendro, Merijn LM Salverda, Joachim Krug, and J Arjan GM De Visser. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Molecular Biology and Evolution*, 30(8):1779–1787, 2013.

[29] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034):1190–1192, 2011.

[30] Xinzhu Wei and Jianzhi Zhang. Patterns and mechanisms of diminishing returns from beneficial mutations. *Molecular Biology and Evolution*, 36(5):1008–1021, 2019.

[31] Sijmen Schoustra, Sungmin Hwang, Joachim Krug, and J Arjan GM de Visser. Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. *Proceedings of the Royal Society B: Biological Sciences*, 283(1837):20161376, 2016.