Accumulation and maintenance of information in evolution

Michal Hledík*[1], Nick Barton†[1], and Gašper Tkačik†[1]

[1] *Institute of Science and Technology Austria, AT-3400 Klosterneuburg, Austria*

Abstract

Selection accumulates information in the genome — it guides stochastically evolving populations towards states (genotype frequencies) that would be unlikely under neutrality. This can be quantified as the Kullback-Leibler (KL) divergence between the actual distribution of genotype frequencies and the corresponding neutral distribution. First, we show that this population-level information sets an upper bound on the information at the level of genotype and phenotype, limiting how precisely they can be specified by selection. Next, we study how the accumulation and maintenance of information is limited by the cost of selection, measured as the genetic load or the relative fitness variance, both of which we connect to the control-theoretic KL cost of control. The information accumulation rate is upper bounded by the population size times the cost of selection. This bound is very general, and applies across models (Wright-Fisher, Moran, diffusion) and to arbitrary forms of selection, mutation and recombination. Finally, the cost of maintaining information depends on how it is encoded: specifying a single allele out of two is expensive, but one bit encoded among many weakly specified loci (as in a polygenic trait) is cheap.

1 Introduction

Throughout evolution, selection accumulates information in the genome. It guides evolving populations towards fitter phenotypes, genotypes and genotype frequencies, which would be highly unlikely to arise by chance. This information – the degree to which selection can control the stochastic process of evolution – has been a long-standing subject of research [1]–[7], and relates to basic questions in evolutionary biology and genetics.

How well can selection specify the genotype and the phenotype? The degree to which within- and between-species genetic variation are shaped by selection has been the subject of the neutralist-selectionist debate [8]–[11]. Today, we know that much of the human genome is involved in various biochemical processes [12], [13], but this does not mean that it is strongly shaped by selection [14]–[16]. Here we ask a related question in information-theoretic terms: how much information can selection accumulate and maintain in the genome? Much of the sequence is to some degree random, and given its size $l \approx 3 \times 10^9$ base pairs, it likely contains far less information than the maximum conceivable 6×10^9 bits of information. A similar question has been raised in the context of origin of life: given high mutation rates, how much information could be maintained in the genome of early organisms [2]?

Analogous questions can be asked about the phenotype. How many traits can selection optimize? It is easy to list a large number of potentially relevant traits: take the expression of all genes in all cell types and conditions, or regulatory interactions between pairs of genes. For a fit organism, these traits need to be specified with some precision, and this precision is likely limited (even if it is to some degree facilitated by correlations among traits). For example, a study of selective constraint on human gene expression [17] gave evidence of constraint, but overall, this seems weak. Given the large number of possibly important phenotypes, how precisely can selection specify them?

Quantifying genetic information. An established method in bioinformatics quantifies the information content of a short genomic motif, such as a binding site, by comparing an alignment of its instances across the genome to the genomic background [18], [19]. Our definition of genetic information is mathematically similar, but aims to apply more generally (to large regions without multiple instances available). It is therefore based in theoretical population genetics rather than sequence data analysis. A key related concept is the repeatability of evolution [20], [21]. Evolution is stochastic due to genetic drift and mutation, but selection can reduce the space of possible outcomes. For example, suppose that in a sequence of length l, n sites are under strong selection for specific nucleotides. By fixing those nucleotides, selection will accumulate $2n$ bits of information. Meanwhile, the remaining $l-n$ sites will be occupied by random nucleotides, and if a replicate population evolves under identical conditions, the $l-n$ nucleotides will likely be different. Therefore our concept of information in a sequence is inversely related to how differently it could have evolved under identical conditions.

In general, however, the information content of the genome cannot be quantified by simply counting the sites that are under selection. A single bit of information can be spread across many loci under weak selection – a phenomenon particularly relevant when selection acts on polygenic traits, long recognized in quantitative genetics and described by the infinitesimal model [22], [23]. Polygenicity and weak selection also resolve the apparent contradiction between the variety of phenotypes, or biochemical processes involving the DNA, and the lack of strong selective constraint on all of them. Selection might act on a small number of high level traits, which are influenced by large numbers of loci spread across the genome (described by the omnigenic model [24]), which experience only weak selection individually.

In Section 2, we define information on three levels – the population state (genotype frequencies), the genotype, and the phenotype. There are simple inequalities between the three levels. This means that the upper bound on information accumulation rate, which we prove at the population level, also implies a bound at the genotype and phenotype levels. We use the KL divergence, a central quantity in information theory [25] to quantify the difference between their actual distribution and their corresponding neutral distribution.

Notably, the neutral phenotype distribution corresponds approximately to the phenotype distribution among random DNA sequences. Recent work with random mutant libraries suggests that for some phenotypes, this distribution is accessible experimentally (gene expression driven by random promoters [26]–[28] or enhancers [29]). Any departure from this neutral distribution amounts to accumulation of information.

*E-mail: michal.hledik@ist.ac.at.

†Contributed equally.

Uploaded on July 3, 2022.

Cost of information. After defining what genetic information means, we ask how quickly it can accumulate and how much of it can be maintained. We look for answers in terms of the cost of selection – the amount of relative fitness variation in a population. This cost, traditionally measured as the relative fitness variance or the genetic load, is itself limited. In a population with constant size, relative fitness is proportional to the expected number of offspring, and the number of offspring can only vary between zero and the reproductive capacity of the organism.

We rely on an information-theoretic measure of cost of selection, which is itself upper bounded by the relative fitness variance and genetic load, but has favorable mathematical properties. It relates the cost of selection to the KL cost of control [30]–[32], or the thermodynamic power [33].

The relationship between information accumulation rate and the cost of selection has been studied by Kimura [1] and later Worden [3], MacKay [4] and Barton [7]. In Sec. 3 we discuss these works in more detail and present a new, more general bound. The problem of maintenance has been studied by Eigen [2], Watkins [5] and Peck and Waxman [6]. We discuss these in Sec. 4 and present example calculations that suggest general trends in the amount of information that can be maintained per unit cost.

2 Quantifying genetic information

The measures of information studied in this paper are based on comparisons between the distributions of various variables under selection versus neutrality. The focus on probability distributions accounts for the stochasticity of evolution, and the difference between the distributions with and without selection corresponds to the control that selection exerts on evolution. We quantify this difference in bits, using the KL divergence [25]

$$D(U) = \sum_u \psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)} \tag{1}$$

where U is a variable that takes values u with probabilities $\psi^U(u)$ with selection and $\varphi^U(u)$ under neutrality. Below we focus on three variables – genotype frequencies (which describe population states), genotypes and phenotypes.

For a pair of variables U, V, statistical dependencies are reflected in their joint and conditional KL divergence, $D(U, V)$ and $D(U|V)$ (see SI Sec. S1 for the definitions). Both are nonnegative quantities, and they follow the chain rule

$$D(U, V) = D(U) + D(V|U) = D(V) + D(U|V). \tag{2}$$

The chain rule allows a comparison of the effects of selection on different variables, as well as on the same variable at different times.

2.1 Population-level information

Evolution is a stochastic process happening to populations, and genotype frequencies form the state space. We use X to denote the genotype frequencies as a random variable, with each value x being a vector with an element x_g for each genotype g, normalized as $\sum_g x_g = 1$. As an example, Fig. 1A shows a common evolutionary scenario where a single locus, two allele system starts from a single copy of a beneficial allele A, and later the frequency evolves stochastically.

X takes values x with probabilities $\psi^X(x)$ under selection and $\varphi^X(x)$ under neutrality. Fig. 1B shows examples of these distributions for the single locus system at three different times. In general, these distributions are shaped

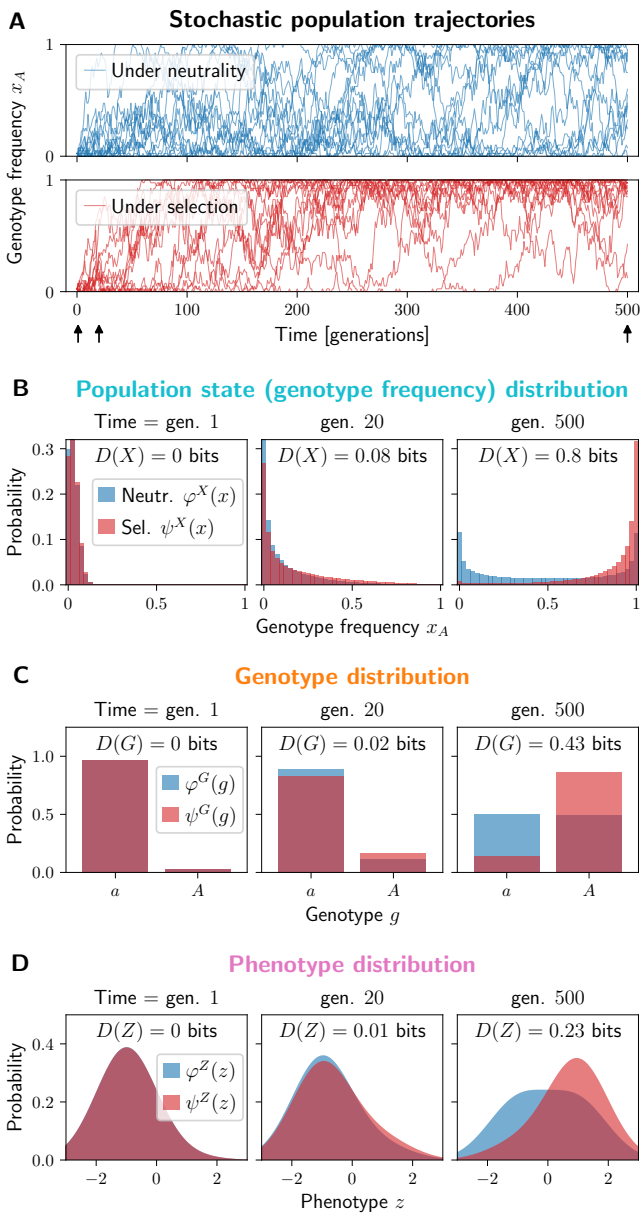

Figure 1: Selection controls the evolution of a single locus, two allele system and drives the distribution of the population states, genotype and phenotype away from neutrality. (A) Stochastic trajectories of the frequency x_A of the beneficial allele A, under neutrality and under selection (blue and red). The allele A starts at a single copy, and under selection it tends to increase in frequency. Black arrows indicate the times when the distributions are plotted in (B-D). At time = 500 generations, the system is approximately stationary. (B-D) The probability distributions of the genotype frequency x_A (B), genotype g (C) and a noisy phenotype z (D) under neutrality (blue) and under selection (red) after a varying number of generations of evolution. The associated measures of information $D(X)$, $D(G)$ and $D(Z)$ are indicated in each panel. (B) The neutral distribution φ^X converges to a symmetric U shape, while the distribution under selection is biased towards high frequencies of the beneficial allele A. The information $D(X)$ increases over time. (C) The neutral genotype distribution φ^G converges to a uniform distribution, due to symmetry between alleles a and A. Under selection (ψ^G), the beneficial allele A has a higher probability, but it does not dominate completely, so the genotype-level information $D(G)$ is less than the maximum 1 bit. $D(G)$ is also upper bounded by $D(X)$. (D) A phenotype with different means and a Gaussian noise for each allele, $\zeta(z|g) = \mathcal{N}(z; \mu_g, \sigma)$ with $\mu_a = -1$, $\mu_A = +1$ and $\sigma = 1$. The information $D(Z)$ is upper bounded by $D(G)$, with a gap due to the partially overlapping distributions $\zeta(z|a)$ and $\zeta(z|A)$. Generated using a haploid Wright-Fisher model (SI Sec. S4) with population size $N = 40$, mutation rate $\mu = 0.005$ and fitness 1 (allele a) and 1.05 (allele A).

by various evolutionary forces – mutation, drift, recombination, selection (ψ^X only), and others. We refer to $D(X)$, the KL divergence between ψ^X and φ^X, as the population-level information.

The example in Fig. 1 illustrates two important phenomena we discuss in the rest of the paper. The first phenomenon is the accumulation of information. A population evolves from an initial distribution (in the simplest case, $\psi^X = \varphi^X$ and $D(X) = 0$, but this is not necessary). For example, the initial state x may be completely specified as in Fig. 1A, or both ψ^X and φ^X may start at the neutral stationary distribution. Over time, selection causes ψ^X to diverge from φ^X and the information $D(X)$ accumulates (Fig. 1B). We study this in detail in Sec. 3. The second phenomenon is the maintenance of information, and it takes place when both $\psi^X(x)$ and $\varphi^X(x)$ are stationary, and the information $D(X)$ is constant. In Sec. 4 we study how much information can be maintained at a given cost of selection.

The population-level information $D(X)$ has been studied under different names and in different roles [7], [34]–[36]. It captures any departure of the genotype frequency distribution ψ^X from its neutral counterpart φ^X – notably, selection can favor not only high frequencies of fit genotypes, but also higher or (more typically) lower amounts of genetic variation within populations. Note that $D(X)$ refers to the effects of selection on the genotype frequencies, rather than allele frequencies. It therefore includes effects of selection on correlations between loci (linkage disequilibrium), which are generated by physical linkage, by chance in finite populations, or due to functional interactions (epistasis) – see also SI Sec. S2.

Notably, $D(X)$ (or $D(G)$ introduced below) appears as a term in free fitness – a quantity analogous to free energy which, under some assumptions, increases over time [35], [37], [38]. This implies that evolution maximizes the expected log-fitness while constraining $D(X)$ – see SI Sec. S8.

2.2 Genotype-level information

If we sample a random genotype from a population in a given state x, we find the genotype g with a probability given simply by its frequency $\psi^{G|X}(g|x) = \varphi^{G|X}(g|x) = x_g$. Taking into account evolutionary stochasticity, we average over all population states x with their probabilities $\varphi^X(x)$ or $\psi^X(x)$,

$$\varphi^G(g) = \sum_x \varphi^X(x)\, x_g, \qquad \psi^G(g) = \sum_x \psi^X(x)\, x_g. \tag{3}$$

Under symmetric point mutations, the neutral distribution φ^G converges to a uniform distribution over all genotypes, while selection typically concentrates ψ^G among a smaller number of fit genotypes. This is also the case for the single locus system in Fig. 1C. The divergence between ψ^G and φ^G is the genotype-level information $D(G)$.

If selection precisely specifies n out of l nucleotides in the genome – i.e. $\psi^G(g)$ is uniform over a fraction $1/4^n$ out of 4^l possible genotypes – this implies $D(G) = 2n$ bits. This corresponds to the intuition of $2n$ bits of information encoded in the genome. More typically, selection will specify many sites only weakly (biasing the probability towards some alleles, see also Fig. 1C), and may contribute to $D(G)$ through LD – correlations between linked or epistatically interacting sites. Without linkage or epistasis, $D(G)$ is approximately additive across loci (see Fig. S1).

$D(G)$ generalizes some previous definitions of genetic information [1], [3], [6] which focused on strong selection or uniform distributions, and coincides with others in important special cases [4], [5].

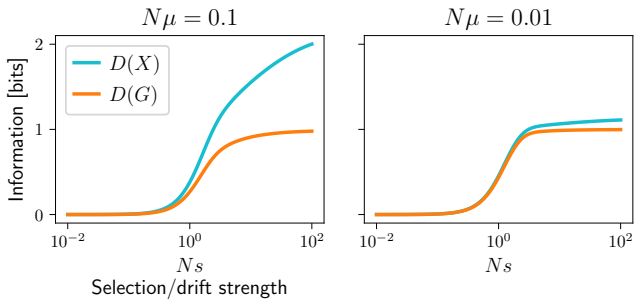

Figure 2: Illustration of $D(X)$ (cyan) and $D(G)$ (orange) for a single locus, two allele system at stationary distributions ψ^X, φ^X as a function of selection strength Ns for two different mutation strengths $N\mu$. The genotype-level information $D(G)$ grows with Ns, from 0 up to 1 bit one out of the two alleles dominates, with the steepest increase around $Ns = 1$. The population-level information $D(X)$ can be much greater than $D(G)$ when mutation is strong, and generates diversity within the population that selection can shape (or suppress). When mutation is weak, $D(X)$ and $D(G)$ are similar, since the population state can be specified by the allele that is currently fixed and $D(X|G) = 0$. Computed using a Wright-Fisher model as in Fig. 1, with population size $N = 100$.

2.3 Phenotype-level information

Finally, selection controls evolution on the level of the phenotype Z. Z could be a categorical trait such as the presence/absence of a disease or the correct/incorrect protein fold, a quantitative trait, a comprehensive characterization of an individual, or its fitness. Given a genotype g, the probability of the phenotype z will be given by the possibly noisy genotype-phenotype relationship $\psi^{Z|G}(z|g) = \varphi^{Z|G}(z|g) = \zeta(z|g)$. When there are no environmental effects or intrinsic noise, $\zeta(z|g)$ will be concentrated at a single value z for each genotype g. Taking into account the variation within populations, as well as the evolutionary stochasticity, the marginal probability of z is

$$\psi^Z(z) = \sum_g \psi^G(g)\, \zeta(z|g), \qquad \varphi^Z(z) = \sum_g \varphi^G(g)\, \zeta(z|g). \tag{4}$$

We show the distributions ψ^Z, φ^Z for the single locus system in Fig. 1D, where the trait has a genotype-dependent mean and a Gaussian noise. While under neutrality, φ^Z tends to spread out over time, selection causes ψ^Z to be more concentrated. The divergence between ψ^Z and φ^Z is the phenotype-level information $D(Z)$.

If we can take the genotype distribution φ^G to be uniform over all possible DNA sequences of some length, then φ^Z is the phenotype distribution among such random sequences. Examples of this distribution have recently been measured experimentally, for gene expression generated by random promoter sequences in *S. cerevisiae* and *E. coli* [26], [28]. If a healthy cell requires the gene expression to be in some narrow range, this translates to a requirement on the phenotype-level information $D(Z)$, and this requirement will increase if the expression needs to be specified across cell states.

2.4 The relationship between the three levels

The definitions above, combined with the chain rule (Eq. (2)) lead to a hierarchy among the three levels,

$$D(X) \geq D(G) \geq D(Z). \tag{5}$$

This inequality can be observed across columns of panels in Fig. 1B-D.

Intuitively, the phenotype-level information $D(Z)$ is bounded by the genotype-level information $D(G)$, since the information about the phenotype has to be encoded in the genome. A special case of this relationship has been noted by Worden [3], who however, worked in a deterministic setting (see SI Sec. S3). The difference between the two, $D(G) - D(Z) = D(G|Z)$, can have two sources. First, the phenotype distribution $\zeta(z|g)$ may overlap between genotypes, causing the phenotype to be specified less precisely than the genotype (as in Fig. 1D). Second, selection may favor genotypes based on criteria other than the phenotype Z, such as other phenotypes or robustness.

Similarly, $D(G)$ can only be as large as the population-level information $D(X)$. To increase the probability of a genotype g, selection must increase the probability of population states with a high frequency of g. However, selection can also shape the patterns of genetic diversity in populations, without impacting the average genotype frequencies, therefore contributing to the difference $D(X) - D(G) = D(X|G)$. In populations with weak mutation, which tend to have little diversity, this difference is small – see Fig. 2.

We rely on the inequalities in Eq. (5) in two ways. First, an upper bound on the population-level information $D(X)$ which we prove in Sec. 3 also implies an upper bound on the genotype and phenotype-level information $D(G)$ and $D(Z)$. In other words, selection can only fine-tune the phenotype to a degree to which it can control the population state.

Second, $D(X)$ and $D(G)$ can be difficult to estimate directly for systems with multiple loci, due to the high dimensionality (see Fig. S1). In such situations, $D(Z)$ for fitness or a low-dimensional phenotype Z can serve as a lower bound on $D(G)$ and $D(X)$. If Z is the trait under selection, or fitness itself, this lower bound can be tight. This approach is applicable even for essentially black box genotype-phenotype models, such as models of gene regulation or protein folding.

3 Accumulation of information

In this section we show how the rate at which $D(X)$, the population-level information, increases over time, is limited by the population size and the variation in fitness. We start by pointing out a connection between population genetics and control theory.

3.1 Accumulation of information and the cost of control

We consider a population evolving over time, with a trajectory $X^0, X^1, \ldots, X^T$ forming a Markov chain between generations 0 and T (such as in Fig. 1A). The divergence of the trajectories' distribution from neutrality, $D(X^0, X^1, \ldots, X^T)$, has been proposed as a measure of predictability of evolution [21]. Using the chain rule (Eq. (2)), we can decompose it in two ways,

$$\begin{aligned} &D(X^0, X^1, \ldots, X^T) \\ &= \underbrace{D(X^0)}_{\substack{\text{Initial}\\ \text{information}}} + \underbrace{\sum_{t=0}^{T-1} D(X^{t+1}|X^t)}_{\text{KL cost of control}} \qquad (6) \\ &= \underbrace{D(X^T)}_{\substack{\text{Final}\\ \text{information}}} + \underbrace{\sum_{t=0}^{T-1} D(X^t|X^{t+1})}_{\substack{\text{Effect of selection on}\\ \text{trajectories reaching } X^T}} . \qquad (7) \end{aligned}$$

In Eq. (6), we distinguish between the divergence of the initial states X^0 and the additional conditional divergence in each generation, $D(X^{t+1}|X^t)$. The latter can be recognized as the KL cost of control, averaged over the initial states x^t [30], [31]. In the context of population genetics, selection takes the role of control.

Eq. (7) makes the distinction between the distribution of endpoints X^T, and the conditional distribution of the states that precede those endpoints. Selection can shape the full trajectories, but only the effects on X^T constitute the final population-level information.

Together, Eq. (6) and (7) imply a bound on the information accumulated between times 0 and T in terms of the KL cost of control,

$$D(X^T) - D(X^0) \leq \sum_{t=0}^{T-1} D(X^{t+1}|X^t). \qquad (8)$$

Specifically, the information accumulated over a single generation, $\Delta D(X^t) = D(X^{t+1}) - D(X^t)$, is upper bounded as

$$\Delta D(X^t) \leq D(X^{t+1}|X^t). \qquad (9)$$

Analogous bounds for continuous time Markov chains and the diffusion approximation are in SI Sec. S6 and S7.

Note that control theory is concerned with computing optimal control policies, which maximize an imposed objective while minimizing the cost $\sum_{t=0}^{T-1} D(X^{t+1}|X^t)$. This is analogous to computing the optimal artificial selection – in fact, the KL divergence control theory framework has recently been used to study artificial selection on quantitative traits [32].

In contrast, natural selection is typically given by the biological or ecological circumstances, and not necessarily optimized in this sense. Still, the KL cost of control provides bounds on the rate at which selection accumulates information (Eq. (8,9)) and it has a meaning in population genetics, which we discuss in the next section.

We also note that Eq. (9) is related to the proof that free fitness increases over time [37], [38], see SI Sec. S8.

3.2 Variation in fitness as cost of control

To compute $D(X^{t+1}|X^t)$ in population genetics, we need to specify a model. We analyze multiple general model classes in the SI: Wright-Fisher and discrete Moran models in SI Sec. S5, continuous time Moran model in SI Sec. S6 and the diffusion approximation in SI Sec. S7. In summary, the bound in Eq. (9) always takes the form

$$\Delta D(X^t) \leq kN \sum_{x^t} \psi^{X^t}(x^t)\, C(x^t) = kN \langle C \rangle_t, \qquad (10)$$

where N is the population size, kN is the number of individuals that are sampled with selection in each generation ($k = 1$ under asexual reproduction and $k = 2$ under sexual reproduction when 2 parents are sampled with selection for each individual). $C(x^t)$ is the cost of selection at the population state x^t (see below), and $\langle C \rangle_t$ is the expected cost at time t. To upper bound information accumulated over multiple generations, we need to sum over them,

$$D(X^T) - D(X^0) \leq kN \sum_{t=0}^{T-1} \sum_{x^t} \psi^{X^t}(x^t)\, C(x^t) = kN \langle C \rangle_{0,T}. \qquad (11)$$

The cost $C(x)$ is a measure of fitness variation in a population in the state x,

$$C(x) = \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_g(x). \qquad (12)$$

where $\hat{w}_g(x)$ is the (frequency dependent) relative fitness of genotype g. When sampling genotypes as parents for

the next generation, g is picked with probability x_g under neutrality and $x_g \hat{w}_g(x)$ under selection – $C(x)$ is the KL divergence between these two distributions.

$C(x)$ is related to two more established measures of cost in population genetics – the relative fitness variance $V(x)$ and the genetic load $L(x)$, which have been studied under a number of circumstances – e.g. mutation-selection balance [39], genetic drift [40], [41], certain types of epistasis and the evolution of sex [42], [43], ongoing substitutions [44]–[46] or stabilizing selection on quantitative traits [47]. They are defined as

$$V(x) = \sum_g x_g \left(\hat{w}_g(x) - 1\right)^2 \tag{13}$$

$$L(x) = 1 - \frac{1}{\hat{w}_{\max}(x)}, \tag{14}$$

where $\hat{w}_{\max}(x)$ is the maximum relative fitness present in the population x, $\hat{w}_{\max}(x) = \max_{g;\, x_g > 0} \hat{w}_g(x)$. We derive the relationships between $C(x)$, $V(x)$ and $L(x)$ in the SI Sec. S9. $V(x)$ and $L(x)$ satisfy the inequality $V(x) \leq \frac{L(x)}{1-L(x)}$ (see also [48]) and both provide an upper bound on $C(x)$,

$$C(x) \leq \frac{V(x)}{\ln 2}, \qquad C(x) \leq \log_2 \frac{1}{1 - L(x)}. \tag{15}$$

In addition, under weak selection and in the diffusion approximation, $C(x) = V(x)/(2 \log 2)$. The bounds in Eq. (10,11) can therefore also be rewritten in terms of $V(x)$ or $L(x)$ using Eq. (15).

Assuming constant population size, relative fitness is proportional to the expected number of offspring, and therefore limited by the species' reproductive capacity. The quantities $\hat{w}_{\max}(x)$, $L(x)$, $V(x)$ and $C(x)$ and as a consequence $\Delta D(X)$, are therefore all limited in realistic settings (SI Sec. S9).

In the context of artificial selection or genetic algorithms, an alternative measure of cost is the population size N, which is the number of cultivated plants or animals, or fitness function evaluations [49], [50]. We note that according to the bounds in Eq. (10,11), the maximal accumulation rate is also proportional to N. Furthermore, increasing the strength of selection (and therefore $C(x)$) beyond an optimal value may increase the immediate response to selection, but reduces the long term response, due to loss of genetic diversity [49], [50]. Therefore in practice, $C(x)$ will be limited even in this context.

3.3 Example 1: the fates of a beneficial allele

The bounds in Eq. (10,11) hold in genetically diverse populations with clonal interference or recombination. Still, it is interesting to consider the case of sequential fixation/loss of mutations, as was done previously [1], [7], [44].

Suppose that a beneficial allele A appears in one copy at time $t = 0$, and is guaranteed to be fixed or lost before another mutation appears that could interfere with it. The population and genotype-level information, $D(X^t)$ and $D(G^t)$, start at 0 and accumulate over time as selection tends to increase the frequency of A (Fig. 3A). The cumulative cost of selection $N\langle C\rangle_{0,t}$ serves as the upper bound on both $D(X^t)$ and $D(G^t)$.

Note that under relatively strong selection ($Ns = 3$, Fig. 3A right), A increases in frequency considerably faster than under neutrality, leading to high $D(X^t)$. But some of these gains are later lost as A is fixed or lost. This is an example of how only the probabilities of endpoints, and not the shape of the trajectories, matters for the information that is ultimately accumulated (the two terms in Eq. (7)).

The increments in $D(X^t)$ and $D(G^t)$ in each generation are plotted in Fig. 3B, along with the bound by $N\langle C\rangle_t$, Eq. (10). The bound on $\Delta D(X^t)$ is relatively tight. $\Delta D(G^t)$ can temporarily exceed $N\langle C\rangle_t$, since the accumulation bound in Eq. (10) does not directly apply to the genotype level, but this is only a transient phenomenon due to the inequality between the cumulative genotype- and population-level information $D(G^t) \leq D(X^t)$.

Both $D(X^t)$ and $D(G^t)$ saturate at the same value $D(X^\infty) = D(G^\infty)$, since the ultimate fate of the population is given simply by whether the allele A is fixed or lost. The fixation probability is $1/N$ under neutrality and $\psi^{\text{fix}} = \psi^{X^\infty}(1) = \psi^{G^\infty}(A)$ under selection, and the accumulated information is a function of this probability,

$$D(X^\infty) = D(G^\infty) = \tag{16}$$

$$= \psi^{\text{fix}} \log_2(N\psi^{\text{fix}}) + (1 - \psi^{\text{fix}}) \log_2 \frac{N(1 - \psi^{\text{fix}})}{N - 1} \tag{17}$$

This function is plotted in cyan in Fig. 3C. According to Eq. (11), it provides a lower bound on the total cost, $N\langle C\rangle_{0,\infty} \geq D(X^\infty)$, given a fixation probability. This holds when the allele A has a constant, frequency independent selective advantage, as in the three examples in Fig. 3A,B (full black line and black points in Fig. 3C). By computing a suitable frequency dependent selection, which optimizes the fixation probability while constraining the total cost $N\langle C\rangle_{0,\infty}$, we can reduce the cost considerably (dash-dotted black line in Fig. 3C, see SI Sec. S11 and Fig. S4 for details). This is achieved by making selection weaker at high frequencies, where the risk of losing A is low. Still, the cost stays above $D(X^\infty)$, as it has to under arbitrary frequency and time-dependent selection.

Under both forms of selection, the bound is only tight when selection is weak. To emphasize this, we plot the information accumulated per unit cost, $D(X^\infty)/\langle C\rangle_{0,\infty}$, as function of the fixation probability ψ^{fix} in Fig. 3D. At weak selection, ψ^{fix} is only perturbed a little from its neutral value $1/N$, but up to N bits can be accumulated per unit cost. A special case of this was shown by Barton [7]. Similar scaling with N was also found in a different setting by Kimura [45].

Stronger selection accumulates more information, but at a disproportionately higher cost, since a large part of it is spent on shaping trajectories rather than outcomes. In the extreme case, to achieve $\psi^{\text{fix}} = 1$, only individuals carrying the A allele can be allowed to reproduce and A gets fixed in only one generation – a highly unlikely way to fixation under neutrality. In this case, selection has the same effect on each genotype sampled as a parent in the first generation, as on the allele that is ultimately fixed (both are A with probability $1/N$ under neutrality and 1 under selection). As a result, the cost is equal to the accumulated information, $\langle C\rangle_{0,\infty} = D(G^\infty) = D(X^\infty)$ and only 1 bit per unit cost is accumulated (Fig. 3D). This is why previous results derived in deterministic settings [1], [3] claimed much more stringent limits on accumulation of information.

3.4 Example 2: accumulation of information under mutation

Unlike the example above, real systems experience ongoing mutation. On the one hand, mutation is necessary to supply beneficial alleles for adaptation, but on the other hand, mutation can disrupt existing adaptation. In this section, we assume that the single locus, two allele system starts at the neutral stationary distribution with $D(X^0) = D(G^0) = 0$, and then selection is turned on. Adaptation exploits copies of the allele A that either segregate in the population by chance at time 0, or arise later by mutation.

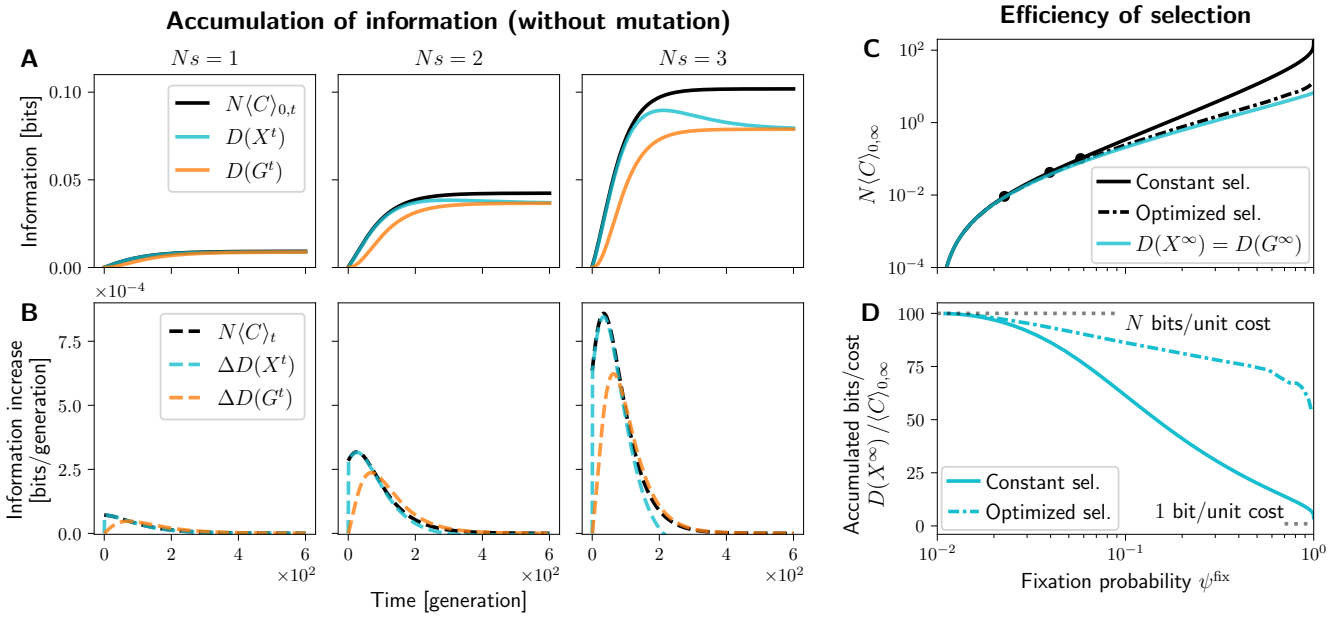

Figure 3: Information accumulation associated with the fixation or loss of a beneficial allele in a haploid single locus, two allele system. The beneficial allele starts at a single copy and evolves under drift and selection, but no mutation. (A) The population-level information ($D(X^t)$, cyan) and genotype-level information ($D(G^t)$, orange) over time, for three different strengths of selection (Ns). Both $D(X^t)$ and $D(G^t)$ start at zero, accumulate over time as selection tends to increase the frequency of the beneficial allele, and saturate as the allele is fixed or lost. The black line is the upper bound according to Eq. (11) with $k = 1$. (B) The increments in $D(X^t)$ and $D(G^t)$ per generation (cyan and orange dashed), and the upper bound according to Eq. (10) with $k = 1$ (black dashed). (C) The cyan line shows the total information accumulated, $D(X^\infty) = D(G^\infty)$, as function of the fixation probability ψ^{fix}. $D(X^\infty)$ serves as a lower bound on N times the total cost of selection, plotted in black, regardless of the form selection takes. The full black line corresponds to constant selection coefficient, with black points showing the three cases in (A,B). The dash-dotted black line shows frequency dependent selection that maximizes ψ^{fix} (and therefore also $D(X^\infty)$) while constraining $N\langle C\rangle_{0,\infty}$. (D) Same data as in (C), but the vertical axis now shows the ratio of the information $D(X^\infty)$ and the total cost of selection $\langle C\rangle_{0,\infty}$ for constant selection (full black) and optimized frequency dependent selection (dash-dotted black). At most N bits can be accumulated per unit cost, and this is achieved at weak selection. At strong selection, this reduces to as low as 1 bit per unit cost. Figure computed using the Wright-Fisher model as in Fig. 1, with population size $N = 100$.

Fig. 4A shows the information $D(X^t)$ and $D(G^t)$ over time. Accumulation take place on the time scale of $1/\mu$. Note that the bound Eq. (11) is not very tight. This is even more apparent in Fig. 4B, where the average cost per generation $N\langle C\rangle_t$ remains positive even after the system has reached the new stationary state, while the increments in $D(X^t)$ and $D(G^t)$ are zero. This corresponds to the cost of maintaining information, which we discuss in Sec. 4.

In summary, the accumulation of information is upper bounded by the KL cost of control, which in turn corresponds to the population size times the variation in fitness. However, if selection changes not only the probabilities of the final states, but also the paths that lead there (because it is strong, because adaptation is maintained for a long time, or because adaptation is reversed by time-dependent selection), then the information accumulated is less than the total cost.

3.5 Comparison with the fitness flux bound

The fitness flux theorem [35] implies another upper bound on information accumulation rate, $\Delta D(X^t) \leq 2N\langle\phi\rangle_t$, where $\langle\phi\rangle_t$ is the expected fitness flux per generation. It is plotted in gray in Fig. 4. It differs from the cost of selection bound both quantitatively and in terms of interpretation.

Quantitatively, neither bound is tighter in general. In Fig. 4B, the cost of selection bound is tighter in early stages of adaptation, and the fitness flux bound is tighter in the late stages. This is consistent with the interpretation of fitness flux as the rate of ongoing adaptation, or the rate of ascent in the mean fitness landscape/seascape [35]. This rate is high in the early stages of adaptation, when the population is far from the fitness peak and tends to climb up quickly. Later, when the population approaches a stationary distribution, there is no more adaptation on average, and $2N\langle\phi\rangle_t$ as well as $\Delta D(X^t)$ vanish. Meanwhile, the cost of selection bound $kN\langle C\rangle_t$ is tighter in the earlier stages when most of the cost is spent on new adaptation, but it remains positive under stationarity, due to maintenance costs.

Technically, the fitness flux theorem was originally derived in [35] under the diffusion approximation, and requires an additional assumption that the neutral process is at a stationary distribution with detailed balance. We derive and discuss the technical aspects of the fitness flux bound in SI Sec. S10 and Fig. S2,S3.

4 Maintenance of information

In this section, we ask how much information can be maintained in the genome for a given cost of selection. A general bound analogous to Eq. (10) seems to be out of reach for now, but we can study how the information maintained depends on key evolutionary parameters. We start by analyzing the single locus, two allele system, and then proceed to systems with large numbers of loci.

4.1 Single locus: weak selection is most efficient

Fig. 5A shows the information, $D(X)$ and $D(G)$, maintained by the single locus, two allele system at the stationary state under various strengths of selection. Stronger selection maintains more information – up to 1 bit at the genotype level, and more on the population level. However, it comes

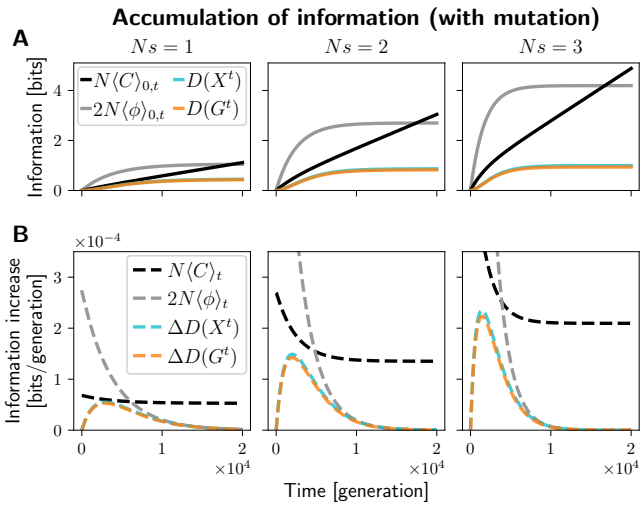

Figure 4: Information accumulation in a single locus, two allele system and the associated upper bounds. The system starts from a neutral stationary distribution over allele frequencies, where $D(X^0) = D(G^0) = 0$. Then it evolves under selection with varying strengths (Ns, left to right) for 2×10^4 generations. (A) The cumulative information at the population level ($D(X^0)$, cyan) and genotype level ($D(G^0)$, orange) over time. Due to the weak mutation $N\mu = 0.01$, the two measures of information are similar. The black and gray lines show upper bounds by the cumulative cost of selection and the cumulative fitness flux. (B) The increments in information per generation, $\Delta D(X^t)$ (cyan dashed) and $\Delta D(G^t)$ (orange dashed) and the upper bounds on these increments in terms of the cost of selection $kN\langle C\rangle_t$ (black, in this case $k = 1$) and the expected fitness flux $2N\langle\psi\rangle_t$ (gray). Note that the cost of selection bound is briefly nearly tight under weak selection ($Ns = 1$, left), and the fitness flux bound is tight near stationarity, when both the accumulation rate and the fitness flux approach 0. Figure computed using the Wright-Fisher model as in Fig. 1. The population size is fixed at $N = 100$. For technical reasons, the expected fitness flux curves were computed using an equivalent Moran model, see SI Sec. S10 and Fig. S2.

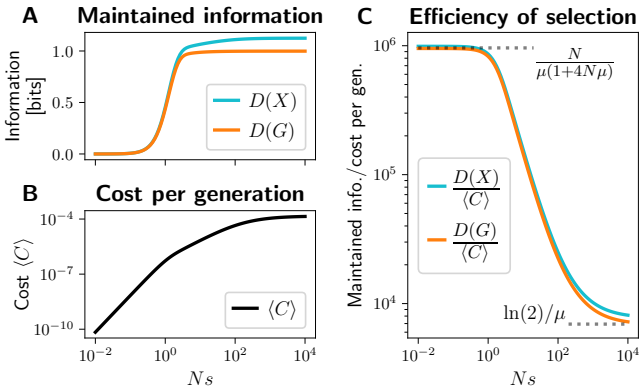

Figure 5: Maintenance of information in the single locus, two allele system. (A) The main plot shows the stationary values of information, $D(X)$ (cyan) and $D(G)$ (orange), as function of selection strength Ns. Stronger selection keeps the beneficial allele at higher frequencies, but this is associated with higher average cost of selection $\langle C\rangle$, shown in (B). Note that much of the time, one of the alleles is fixed and the cost C is zero. $\langle C\rangle$ is the average cost per generation over the stationary distribution of allele frequencies. (C) The ratio of the maintained information and the average cost of selection, $D(X)/\langle C\rangle$ (cyan) and $D(G)/\langle C\rangle$ (orange). Selection is most efficient when it is relatively weak ($Ns \ll 1$), maintaining up to $\frac{N}{\mu(1+4N\mu)}$ bits per unit cost at the genotype level, and inefficient when strong ($Ns \gg 1$), maintaining only about $\ln(2)/\mu$ bits per unit cost (dotted horizontal lines). The population size is $N = 100$ and mutation rate $\mu = 10^{-4}$.

with a higher cost of selection $\langle C\rangle$, Fig. 5B. Notably, the cost increases faster then the maintained information. As a result, the amount of information maintained per unit cost decreases with selection strength, Fig. 5C.

There are two important asymptotic regimes. When selection is very strong, $Ns \gg 1$, deleterious mutations are purged as soon as they arise, and $D(G) \approx 1$ bit. Mutations arise with a probability $N\mu$ per generation, and purging each costs $C \approx 1/(N\ln(2))$ (assuming truncation selection with $\alpha = 1 - 1/N$, see SI Sec. S9). In this regime,

$$\text{Strong selection:} \qquad \frac{D(G)}{\langle C\rangle} \approx \frac{\ln 2}{\mu}, \tag{18}$$

bits can be maintained per unit cost (see Fig. 5C). Similar arguments apply when $N\mu > 1$. The inverse scaling with μ is expected based on the deterministic mutation load [39] or Eigen's error catastrophe [2] which occurs when selection cannot maintain sequences without error, and it was also derived by Watkins [5].

Selection is much more efficient when it is weak, $Ns \ll 1$. Both the cost and the maintained information can be calculated under the diffusion approximation (see SI Sec. S4B for details). If mutation is also weak, $N\mu \ll 1$, the amount of genetic variation (pairwise diversity) scales with $2N\mu$, and the cost (variation in fitness) is approximately $\langle C\rangle \approx N\mu s^2/(2\ln 2)$. Meanwhile, selection shifts the mean frequency of A away from $1/2$ by about $Ns/2$, and this is associated with genotype-level information $D(G) \approx N^2 s^2/(2\ln 2)$ bits. In this regime, up to N/μ bits per unit cost are maintained. When mutation $N\mu$ is not negligible, a more accurate result is

$$\text{Weak selection:} \qquad \frac{D(G)}{\langle C\rangle} \approx \frac{N}{\mu(1+4N\mu)}, \tag{19}$$

see SI Sec. S4. This limit is also highlighted in Fig. 5C. The special case when $N\mu \gg 1$, $\frac{D(G)}{\langle C\rangle} \approx 1/(4\mu^2)$ was previously derived by Watkins [5].

By itself, a single locus under weak selection cannot contribute much to biological function. However, selection can act on a polygenic trait influenced by many loci. If they are unlinked, we expect both the maintained information and the cost of selection to be approximately additive, and the ratio $D(G)/\langle C\rangle$ to scale according to Eq. (19). To confirm this, we next study a polygenic system.

4.2 Information stored among many loci

We use an individual-based model to study a population of N haploids with $l = 1000$ biallelic loci, mutation and free recombination. Offspring are produced by sampling pairs of parents with selection, shuffling their genomes (at each locus, the allele from either parent is inherited with probability $1/2$), and flipping each allele with probability μ. Selection acts on a fully heritable, additive trait with equal effects, $z_g =$ (the number of A alleles in g), with fitness being $w_g = (1+s)^{z_g}$.

The results are shown in Fig. 6. The panel 6A shows an example of a stochastic population trajectory, indicating the phenotypes present in the population over time. The system is initialized with random genomes that contain the beneficial allele at each locus with probability $1/2$, with z taking values around $l/2 = 500$ with binomial noise. Selection with $s = 0.01$ makes the beneficial alleles more frequent over time. The stationary distribution over phenotypes is in Fig. 6B. Under neutrality, $\varphi^Z = \text{Binom}(l, 1/2)$ by symmetry. The distribution ψ^Z under selection is shifted relatively far from φ^Z, leading to $D(Z) = 88.0$ bits of information on the phenotype level.

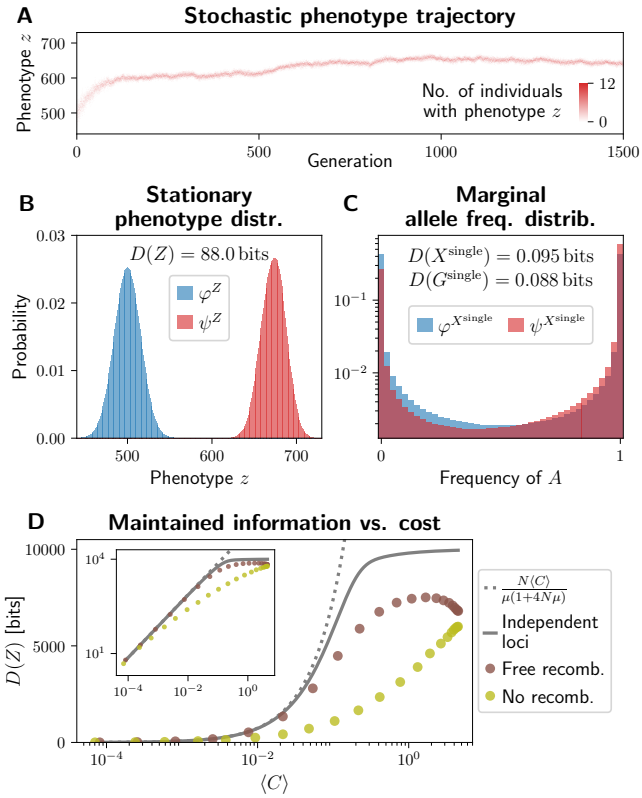

Figure 6: Maintenance of information in a system with $l = 1000$ biallelic loci. Selection is directional on an additive trait Z (= the number of beneficial alleles). (A) A heatmap showing the number of individuals in a population occupying each value of the phenotype z at each generation. The population is initialized as a collection of random genomes, each containing the beneficial allele at around $l/2 = 500$ loci. Over time, this number stochastically increases due to selection. Only the first 1500 generations of the trajectory are shown, the full trajectory was 5×10^3 generations of burn-in and 2×10^5 to estimate the stationary distributions in (B,C). (B) The stationary distribution over the phenotype Z, under neutrality (φ^Z, blue) and selection (ψ^Z, red), along with the phenotype-level information $D(Z)$. Due to symmetry between loci and alleles, $\varphi^Z(z) = \text{Binom}(z; l, 0.5)$ is binomial. Under selection, ψ^Z is obtained as the histogram over individuals and over 2×10^5 generations at stationarity. (C) The marginal distribution over allele frequencies at individual loci, under neutrality ($\varphi^{X^{\text{single}}}$, blue, computed using a transition matrix for the single locus system) and under selection ($\psi^{X^{\text{single}}}$, red, computed as a histogram over all loci and 2×10^5 generations at stationarity). The associated $D(X^{\text{single}})$ and $D(G^{\text{single}})$ correspond to information maintained at one locus, and because the loci are approximately independent, the total information is about $l = 1000$ times more. The population size is $N = 40$, mutation strength $N\mu = 0.02$ and selection strength $Ns = 0.4$. (D) The relationship between the maintained information $D(Z)$ and the cost of selection $\langle C \rangle$, with recombination (brown points), without recombination (olive points). This is compared with predictions under the assumption of independent loci (gray line; computed using single locus diffusion approximation and multiplying both information and cost by the number of loci) and the linear scaling with $\langle C \rangle$ based on Eq. (19) (dotted gray). Computed for a system with $l = 10^4$ loci, population size $N = 40$, mutation strength $N\mu = 0.02$ and variable Ns. Distributions estimated from a stochastic trajectory over 5×10^4 generations, after 5×10^3 generations of burn-in. The inset shows identical data with a log vertical scale.

The population state distribution and the genotype distribution are inaccessible due to their dimensionality (see Fig. S1). However, we know that they are lower bounded by $D(Z)$, which is easy to compute, and $D(Z) \approx D(G)$ since Z is the only trait under selection. Since the loci are unlinked and have equal effects, the information $D(Z)$ can be divided evenly among them. The marginal distribution over allele frequencies is only slightly different from neutrality (Fig. 6C), by about $D(X^{\text{single}}) = 0.095$ bits in terms of allele frequency distribution and $D(G^{\text{single}}) = 0.088$ in terms of allele probabilities. The 1000 loci, however, combine to produce a large shift in the phenotype distribution, $D(Z) \approx 1000 D(G^{\text{single}})$.

This information is maintained at a very low cost of selection, $\langle C \rangle = 0.0012$ bits per generation, or relative fitness variance $\langle V \rangle = 0.0017$. This amounts to $D(Z)/\langle C \rangle = 7.1 \times 10^4$ bits per unit cost, only a little below the single locus limit $N/\mu/(1 + 4N\mu) = 7.4 \times 10^4$ under weak selection.

4.3 Interference between loci

In practice, the selection on different loci might interfere, and this can hinder the maintenance of information. The interaction may be due to Hill-Robertson interference, linkage, or epistasis.

In Fig. 6D we vary the selection coefficient s on individual alleles in a $l = 10^4$ locus system, and plot the maintained $D(Z)$ against the cost $\langle C \rangle$. We use the individual-based model to compute these with free recombination (as in 6A-C) and with zero recombination (offspring genotypes are identical to those of single parents, up to mutation). We compare the results with the weak selection scaling according to Eq. (19), and results for 10^4 loci that evolve independently (cost and information are summed over 10^4 single locus systems).

With free recombination, weak selection maintains about as much information as if the loci were independent (brown points and gray line in Fig. 6D, inset), approximately according to Eq. (19) (gray dotted line). However, when selection is strong ($\langle C \rangle \approx 0.1$ or more), individual alleles experience additional fluctuations in frequency due to random associations with alleles at other loci in a finite population [51], [52], reducing the efficiency of selection. As a result, the freely recombining loci maintain less information than if they were independent. This is in addition to the fact that under strong selection, maintenance is more costly even for independent loci (full gray line departs from dotted gray line, Fig. 6D). Extremely strong selection, which removes potentially adaptive variation at other loci, maintains even less information than more moderate selection, and it makes recombination ineffective (brown points at high $\langle C \rangle$ in Fig. 6D).

Without recombination, less information is maintained at any given cost (olive points in Fig. 6D). In fact, Watkins [5] has shown that due to clonal interference, organisms with no recombination cannot maintain more than the order of $\ln(N)/\mu$ bits of information even if the cost is unlimited, making Haldane's and Eigen's results [2], [39] pertinent to asexual populations.

The advantage of recombination has also been recognized in a similar context by MacKay [4] and Peck and Waxman [6], and relates to the evolution of sex and epistasis. Recombination is advantageous when facing unconditionally deleterious or beneficial alleles [43], but can be disadvantageous when adaptation depends on beneficial combinations of alleles [53]. However, it is not clear if any form of selection can maintain more information at a given cost than $\frac{N}{\mu(1+4N\mu)}$ achieved by weak directional selection with recombination.

5 Discussion

Selection exerts control on evolving populations, but its capacity is limited. The limits to selection have been approached from various angles. Here we build upon previous work that had developed the idea that selection accumulates and maintains information in the genome [1], [2], and that this is associated with a cost in terms of variation in fitness, such as genetic load or fitness variance [39], [44]. The early work has suggested remarkably simple limits to selection: that the maximal rate of accumulation is bounded by the cost itself [1], [3], and that maintenance is limited to about $1/\mu$ functional sites in the genome [2], [39].

Later work has pointed out that both accumulation [4], [7] and maintenance [5], [6] can exceed these limits, notably when recombination is involved. However, the general bounds remained unclear, possibly in part due to the difficulty of defining genetic information in general.

The measures of information that we have introduced in Sec. 2 coincide with or generalize previous definitions, and offer two advantages. First, they facilitate connections between different levels – e.g. between the abstract population-level information that has been studied theoretically in different contexts [34]–[36] and the effect that selection has on the distribution of phenotypes.

Second, the generality of our definition allows proving a general bound on information accumulation rate. This turns out to be a factor N faster than the early bounds, but depends on selection on individual loci being weak. The bound relies on a measure of cost of selection that connects the genetic load and fitness variance [48] with the KL cost in control theory [30], [31], recently used in the context of artificial selection [32].

How much information can be maintained in the genome at a given cost remains an open problem, but we have discussed how this might scale with the population size and the mutation rate. The scaling in Eq. (19) generalizes a result by Watkins [5] to realistic populations with $N\mu < 1$. Still, more work is needed to make claims about the information content of any real organism's genome. Typical populations have N_e/μ much greater than the genome size, suggesting that the genome size or other factors are more limiting than Eq. (19). The maintenance can be made more difficult by linkage or epistasis, and parts of the genome are likely under strong selection which is more costly. Still, Eq. (19) suggests that in theory, the genome could contain a substantial amounts of information among weakly selected loci, e.g. coding for polygenic traits. This is consistent with recent work [54] pointing out that mutation load does not pose severe limitations to the functional fraction of the human genome.

Similarly, the bound on accumulation rate in Eq. (10) hypothetically allows accumulation of information amounting to 10% of the human genome in about 10^6 generations (6×10^8 bits, assuming effective population size $N_e \approx 10^4$, $k = 2$ and meager cost $\langle C \rangle \approx 0.03$ or relative fitness variance $\langle V \rangle \approx 0.018$ devoted to accumulation). But this is unlikely to have happened. Some selection was likely strong and more costly, and selection could have fluctuated, reversing previous adaptation. However, under the right conditions, information can accumulate very fast.

Our findings are complementary to the point raised by Kondrashov [41], that the survival of populations could be threatened by large numbers of weakly deleterious mutations ($Ns < 1$). While selection cannot purge them, it can perturb the allele frequency distribution of each by a small amount, and thus shift the distribution of higher level traits very far from neutrality. This is similar to the resolution by Charlesworth [55]. In fact, information accumulation and maintenance are most cost-efficient in this regime. This does not mean that a genomic architecture, where most mutations operate at $Ns < 1$ and information is encoded among many weakly specified sites, would evolve as an adaptation to maximise information gain. Nevertheless, such an architecture might arise in multicellular organisms as a side effect of their small effective population sizes and long genomes [56], [57].

Focus on the information content of genomes, rather than their fraction under selection, could help better frame the controversy sparked by some publications from the ENCODE project [12]–[16], [54], [58]. On the one hand, genomic regions under detectable selection (less than 15% in humans [59]) likely contain less than 2 bits per base pair, because their current function could be achieved by a number of alternative sequences (e.g. due to synonymous mutations in coding regions, or flexibility of transcription factor binding site sequence and location). On the other hand, regions without detectable selection could contain a considerable amount of information in the aggregate, at a low cost, encoding polygenic traits.

In bioinformatics, there already is a measure of information content applicable to short regulatory motifs [18], [19]. Future work could examine the precise relationship between this measure and our theoretical definitions. The generality of our framework also opens new directions for future research. One is to predict the maximal amount of information that can be maintained in genomes and populations with realistic parameters. Another is to study the information content of genomic elements with well-described genotype-phenotype maps (e.g. promoters [26], [27]), under different hypotheses about selection on the phenotype.

Acknowledgements

We thank Ksenia Khudiakova, Wiktor Młynarski and Sean Stankowski for discussions. G.T. and M.H. acknowledge funding from the Human Frontier Science Program (HFSP RGP0032/2018). N.B. acknowledges funding from ERC grant 250152 "Information and Evolution".

References

[1] M. Kimura, "Natural selection as the process of accumulating genetic information in adaptive evolution," *Genetics Research*, vol. 2, no. 1, pp. 127–140, 1961.

[2] M. Eigen, "Selforganization of matter and the evolution of biological macromolecules," *Naturwissenschaften*, vol. 58, no. 10, pp. 465–523, 1971.

[3] R. P. Worden, "A speed limit for evolution," *Journal of Theoretical Biology*, vol. 176, no. 1, pp. 137–152, 1995.

[4] D. J. C. MacKay, "Why have sex? Information acquisition and evolution," in *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003, pp. 269–280, ISBN: 978-0-521-64298-9.

[5] C. Watkins, "Selective breeding analysed as a communication channel: Channel capacity as a fundamental limit on adaptive complexity," in *2008 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2008, pp. 514–518.

[6] J. R. Peck and D. Waxman, "Is life impossible? Information, sex, and the origin of complex organisms," *Evolution*, vol. 64, no. 11, pp. 3300–3309, 2010.

[7] N. H. Barton, "How does epistasis influence the response to selection?" *Heredity*, vol. 118, no. 1, pp. 96–109, 2017.

[8] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, vol. 217, no. 5129, pp. 624–626, 1968.

[9] J. Hey, "The neutralist, the fly and the selectionist," *Trends in Ecology & Evolution*, vol. 14, no. 1, pp. 35–38, 1999.

[10] A. D. Kern and M. W. Hahn, "The Neutral Theory in light of natural selection," *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1366–1371, 2018.

[11] J. D. Jensen, B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch, D. Charlesworth, and B. Charlesworth, "The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018," *Evolution*, vol. 73, no. 1, pp. 111–114, 2019.

[12] The ENCODE Project Consortium *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[13] M. Kellis *et al.*, "Defining functional DNA elements in the human genome," *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6131–6138, 2014.

[14] W. F. Doolittle, "Is junk DNA bunk? A critique of ENCODE," *Proceedings of the National Academy of Sciences*, vol. 110, no. 14, pp. 5294–5300, 2013.

[15] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik, "On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE," *Genome Biology and Evolution*, vol. 5, no. 3, pp. 578–590, 2013.

[16] T. D. P. Brunet and W. F. Doolittle, "Getting "function" right," *Proceedings of the National Academy of Sciences*, vol. 111, no. 33, E3365–E3365, 2014.

[17] E. C. Glassberg, Z. Gao, A. Harpak, X. Lan, and J. K. Pritchard, "Evidence for weak selective constraint on human gene expression," *Genetics*, vol. 211, no. 2, pp. 757–772, 2019.

[18] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *Journal of Molecular Biology*, vol. 188, no. 3, pp. 415–431, 1986.

[19] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[20] A. E. Lobkovsky and E. V. Koonin, "Replaying the tape of life: Quantification of the predictability of evolution," *Frontiers in Genetics*, vol. 3, 2012.

[21] M. Lässig, V. Mustonen, and A. M. Walczak, "Predicting evolution," *Nature Ecology & Evolution*, vol. 1, no. 3, pp. 1–9, 2017.

[22] R. A. Fisher, "The correlation between relatives on the supposition of Mendelian inheritance," *Transactions of the Royal Society of Edinburgh*, vol. 52, no. 02, pp. 399–433, 1918.

[23] N. H. Barton, A. M. Etheridge, and A. Véber, "The infinitesimal model: Definition, derivation, and implications," *Theoretical Population Biology*, vol. 118, pp. 50–73, 2017.

[24] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: From polygenic to omnigenic," *Cell*, vol. 169, no. 7, pp. 1177–1186, 2017.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second. John Wiley & Sons, 2006, ISBN: 978-0-471-24195-9.

[26] C. G. de Boer, E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev, "Deciphering eukaryotic gene-regulatory logic with 100 million random promoters," *Nature Biotechnology*, vol. 38, no. 1, pp. 56–65, 2020.

[27] E. D. Vaishnav *et al.*, "The evolution, evolvability and engineering of gene regulatory DNA," *Nature*, vol. 603, no. 7901, pp. 455–463, 2022.

[28] M. Lagator, S. Sarikas, M. Steinrueck, D. Toledo-Aparicio, J. P. Bollback, C. C. Guet, and G. Tkačik, "Predicting bacterial promoter function and evolution from random sequences," *eLife*, vol. 11, S. Krishna, A. M. Walczak, E. van Nimwegen, and T. Einav, Eds., e64543, 2022.

[29] T. Fuqua *et al.*, "Dense and pleiotropic regulatory information in a developmental enhancer," *Nature*, vol. 587, no. 7833, pp. 235–239, 2020.

[30] E. Todorov, "Linearly-solvable Markov decision problems," in *Advances in Neural Information Processing Systems*, vol. 19, 2006, pp. 1369–1376.

[31] E. A. Theodorou, "Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations," *Entropy*, vol. 17, no. 5, pp. 3352–3375, 2015.

[32] A. Nourmohammad and C. Eksin, "Optimal evolutionary control for artificial selection on molecular phenotypes," *Physical Review X*, vol. 11, no. 1, p. 011 044, 2021.

[33] D. S. Pavlichin, Y. Quek, and T. Weissman, "Minimum power to maintain a nonequilibrium distribution of a Markov chain," Preprint, arXiv:1907.01582 [Preprint], 2019. arXiv: `1907.01582`.

[34] N. H. Barton and H. P. de Vladar, "Statistical mechanics and the evolution of polygenic quantitative traits," *Genetics*, vol. 181, no. 3, pp. 997–1011, 2009.

[35] V. Mustonen and M. Lässig, "Fitness flux and ubiquity of adaptive evolution," *Proceedings of the National Academy of Sciences*, vol. 107, no. 9, pp. 4248–4253, 2010.

[36] K. Bod'ová, G. Tkačik, and N. H. Barton, "A general approximation for the dynamics of quantitative traits," *Genetics*, vol. 202, no. 4, pp. 1523–1548, 2016.

[37] Y. Iwasa, "Free fitness that always increases in evolution," *Journal of Theoretical Biology*, vol. 135, no. 3, pp. 265–281, 1988.

[38] G. Sella and A. E. Hirsh, "The application of statistical physics to evolutionary biology," *Proceedings of the National Academy of Sciences*, vol. 102, no. 27, pp. 9541–9546, 2005.

[39] J. B. S. Haldane, "The effect of variation on fitness," *The American Naturalist*, vol. 71, no. 735, pp. 337–349, 1937.

[40] M. Kimura, T. Maruyama, and J. F. Crow, "The mutation load in small populations," *Genetics*, vol. 48, no. 10, pp. 1303–1312, 1963.

[41] A. S. Kondrashov, "Contamination of the genome by very slightly deleterious mutations: Why have we not died 100 times over?" *Journal of Theoretical Biology*, vol. 175, no. 4, pp. 583–594, 1995.

[42] M. Kimura and T. Maruyama, "The mutational load with epistatic gene interactions in fitness," *Genetics*, vol. 54, no. 6, pp. 1337–1351, 1966.

[43] A. S. Kondrashov, "Deleterious mutations and the evolution of sexual reproduction," *Nature*, vol. 336, no. 6198, pp. 435–440, 1988.

[44] J. B. S. Haldane, "The cost of natural selection," *Journal of Genetics*, vol. 55, no. 3, p. 511, 1957.

[45] M. Kimura, "Limitations of Darwinian selection in a finite population," *Proceedings of the National Academy of Sciences*, vol. 92, no. 6, pp. 2343–2344, 1995.

[46] W. J. Ewens, "Remarks on the substitutional load," *Theoretical Population Biology*, vol. 1, no. 2, pp. 129–139, 1970.

[47] R. Lande, "Genetic variation and phenotypic evolution during allopatric speciation," *The American Naturalist*, vol. 116, no. 4, pp. 463–479, 1980.

[48] E. E. Shnol, E. A. Ermakova, and A. S. Kondrashov, "On the relationship between the load and the variance of relative fitness," *Biology Direct*, vol. 6, no. 1, p. 20, 2011.

[49] A. Robertson, "Some optimum problems in individual selection," *Theoretical Population Biology*, vol. 1, no. 1, pp. 120–127, 1970.

[50] N. Barton and T. Paixão, "Can quantitative and population genetics help us understand evolutionary computation?" In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '13, New York, NY, USA: Association for Computing Machinery, 2013, pp. 1573–1580, ISBN: 978-1-4503-1963-8.

[51] W. G. Hill and A. Robertson, "The effect of linkage on limits to artificial selection," *Genetics Research*, vol. 8, pp. 269–294, 1966.

[52] N. H. Barton, "Linkage and the limits to natural selection," *Genetics*, vol. 140, no. 2, pp. 821–841, 1995.

[53] F. A. Kondrashov and A. S. Kondrashov, "Multidimensional epistasis and the disadvantage of sex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12 089–12 092, 2001.

[54] B. Galeota-Sprung, P. Sniegowski, and W. Ewens, "Mutational load and the functional fraction of the human genome," *Genome Biology and Evolution*, vol. 12, no. 4, pp. 273–281, 2020.

[55] B. Charlesworth, "Why we are not dead one hundred times over," *Evolution*, vol. 67, no. 11, pp. 3354–3361, 2013.

[56] M. Lynch and J. S. Conery, "The origins of genome complexity," *Science*, vol. 302, no. 5649, pp. 1401–1404, 2003.

[57] M. Lynch, *The Origins of Genome Architecture*. Sinauer Associates, 2007, ISBN: 978-0-87893-484-3.

[58] D. Graur, "An upper limit on the functional fraction of the human genome," *Genome Biology and Evolution*, vol. 9, no. 7, pp. 1880–1885, 2017.

[59] C. M. Rands, S. Meader, C. P. Ponting, and G. Lunter, "8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage," *PLOS Genetics*, vol. 10, no. 7, e1004525, 2014.

Supplementary information: Accumulation and maintenance of information in evolution

Michal Hledík*[1], Nick Barton†[1], and Gašper Tkačik†[1]

[1] *Institute of Science and Technology Austria, AT-3400 Klosterneuburg, Austria*

Contents

S1 Joint and conditional KL divergence, chain rule **2**

S2 Population-level information: allele frequencies and LD **2**

S3 Violation of the bound by Worden 1995 by drift **2**

S4 The single locus, two allele system used for figures **3**
S4A Wright-Fisher model . . . 3
S4B The diffusion approximation . . . 3

S5 The bound on information accumulation rate – Markov chains **4**
S5A Wright-Fisher model . . . 4
S5A.1 Selection in an infinite offspring pool . . . 5
S5A.2 Selection among parents . . . 6
S5B Discrete-time Moran model . . . 6

S6 The bound on information accumulation rate – continuous-time Markov chains **7**
S6A Continuous-time Moran model . . . 7

S7 The bound on information accumulation rate – diffusion approximation **8**
S7A Application to population genetics . . . 8

S8 Relationship with free fitness and statistical physics **9**
S8A Boltzmann form of stationary distributions . . . 9
S8B Free fitness . . . 10
S8C Monotonic convergence of stochastic processes to their stationary distributions . . . 10
S8D Free fitness as a trade-off between fitness and information . . . 11

S9 Properties of measures of cost of selection **11**

S10 Fitness flux theorem **12**
S10A Discrete-time Markov chains . . . 13
S10B Diffusion approximation . . . 13
S10C Comparisons of the discrete and the diffusion formulas . . . 14
S10D Interpretation of the bounds under diffusion . . . 14

S11 Frequency dependent selection that maximizes fixation probability **15**

List of Figures

S1 Distributions on the three levels with multiple loci . . . 17
S2 Fitness flux formulas comparison . . . 18
S3 Bounds from an intermediate frequency . . . 19
S4 Frequency dependent selection . . . 20

*Electronic address: michal.hledik@ist.ac.at.
†Contributed equally.
Uploaded on July 3, 2022.

S1 Joint and conditional KL divergence, chain rule

For a single variable U, the KL divergence [1] between its distributions with and without selection is

$$D(U) = \sum_u \psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)} \tag{S1}$$

where U takes values u with probabilities $\psi^U(u)$ under selection and $\varphi^U(u)$ under neutrality.

To be well defined, the KL divergence requires that the support of ψ^U is a subset of the support of φ^U. In other words, if for some u we have $\varphi^U(u) = 0$, then also $\psi^U(u) = 0$ – outcomes impossible under neutrality are also impossible under selection. This condition needs to be respected when setting the initial conditions (ψ^U and ϕ^U at time zero). Over time, selection increases or decreases the probability of population states, genotypes or phenotypes that arise by reproduction with mutation, a by finite factor proportional to fitness. But selection cannot create entirely new states. On the other hand, some genotypes that arise by mutation can have zero fitness, and therefore be impossible under selection ($\psi^U(u) > 0$ but $\psi^U(u) = 0$). When this happens, the corresponding term $\psi^U(u) \log_2 \frac{\psi^U(u)}{\varphi^U(u)}$ is set to zero. Therefore this is a very natural assumption, and analogous arguments apply to joint/conditional distributions which we discuss next.

For a pair of variables U, V we can write their joint and conditional KL divergence [1],

$$D(U,V) = \sum_{u,v} \psi^{U,V}(u,v) \log_2 \frac{\psi^{U,V}(u,v)}{\varphi^{U,V}(u,v)}, \tag{S2}$$

$$D(U|V) = \sum_v \psi^V(v) \sum_u \psi^{U|V}(u|v) \log_2 \frac{\psi^{U|V}(u|v)}{\varphi^{U|V}(u|v)} \tag{S3}$$

where $\psi^{U,V}(u,v)$ and $\psi^{U|V}(u|v)$ are the joint and conditional probabilities under selection, and $\varphi^{U,V}(u,v)$ and $\varphi^{U|V}(u|v)$ under neutrality. With these definitions, the chain rule states two possible decompositions of the joint KL divergence,

$$D(U,V) = D(U) + D(V|U) = D(V) + D(U|V). \tag{S4}$$

S2 Population-level information: allele frequencies and LD

When there is a fixed number of loci, instead of genotype frequencies, an alternative way to describe a population state is in terms of allele frequencies. Allele frequencies by themselves, however, do not capture correlations between loci and therefore can miss some of the information that selection can accumulate on the population level. This can be expressed using the chain rule,

$$D(X) = D(\text{Allele freq.}) + D(X|\text{Allele freq.}), \tag{S5}$$

where the term $D(\text{Allele freq.}|X) = 0$ because regardless of selection, allele frequencies are fully determined by the genotype frequencies X. The term $D(X|\text{Allele freq.})$ quantifies how different from neutrality are the correlations between loci.

S3 Violation of the bound by Worden 1995 by drift

The genotype-level information introduced by Worden [2] (see Eq. (11) there) is the KL divergence between the genotype frequencies and a uniform distribution,

$$I = \sum_g x_g \log_2(M x_g), \tag{S6}$$

where M is the number of possible genotypes and x_g the frequency of genotype g (denoted q_j in [2]). Worden also introduced a similar genotype-level measure (Eq. (8) in [2], which was upper bounded by I. These measures of information can be seen as special cases of $D(G)$ and $D(Z)$ when there is no evolutionary stochasticity – ψ^X is concentrated at a single value x, and φ^X is concentrated at a single value of x that is uniform over all possible genotypes.

Worden proposes a bound on the rate of increase of I starting from a uniform x, and the maximal rate is proportional to a quantity similar to the genetic load, i.e. roughly a factor N (population size) times more stringent than the bound presented here.

The proof relies on the assumption that the population is large and x_g evolve deterministically, but later, validity in finite populations is claimed (Sec. 2.6 in [2]). This is mistaken: in a realistic population, I can hardly be zero to start with, as there will be more possible genotypes than individuals and x_g cannot be uniform. Starting from near uniform x, random drift will tend to remove variability from the population and concentrate all genotypes around some random ancestral genotype, and I will increase even without selection. This can also be seen as a consequence of the convexity of KL divergence: random fluctuations in x will, on average, increase I. This highlights the need to consider stochasticity as well as population variation when quantifying the intuitive notion of genetic information.

Worden's stringent bound does hold if the genotype frequencies evolve deterministically and there is no recombination. This is consistent with our observation that selection accumulates information less cost-efficiently when $Ns \gg 1$ and the fixation probability of a beneficial mutation is close to 1 (Main Text Fig. 3C,D).

S4 The single locus, two allele system used for figures

We use a haploid single locus, two allele system to produce Figures 1-5. The figures are produced with a Wright-Fisher model, Moran model (only the fitness flux in Fig. 4, S2B and S3), and some intuition can be gained by approximating it as diffusion under weak selection. Note that this is only an illustration, more general classes of models are discussed in sections S5, S6 and S7.

The system has two alleles, a and A, where the latter is beneficial under selection. It is parametrized by the population size N, mutation rate μ and selection coefficient $s(x)$ (which is frequency dependent only in Fig. 3C,D and S4).

S4A Wright-Fisher model

Under the Wright-Fisher model, the state space is a set of discrete frequencies of the A allele, $x_A = 0, 1/N, 2/N, \ldots, 1$, while the a allele always has the complementary frequency $x_a = 1 - x_A$. The two alleles have the following properties.

Allele g	frequency x_g	fitness $w_g(x)$	relative fitness $\hat{w}_g(x)$
a	$x_a = 1 - x_A$	$w_a(x) = 1$	$\hat{w}_a(x) = \frac{1}{1+s\,x_A}$
A	x_A	$w_a(x) = 1 + s$	$\hat{w}_A(x) = \frac{1+s}{1+s\,x_A}$

The probability of sampling allele A as a parent is $x_A \hat{w}_A(x)$, and the probability of sampling it as offspring is

$$q_A(x) = x_A(1-\mu) + x_a\mu \tag{S7}$$

$$p_A(x) = x_A\hat{w}_A(x)(1-\mu) + x_a\hat{w}_a(x)\mu = \frac{x_A(1+s)(1-\mu) + (1-x_A)\mu}{1+sx_A}, \tag{S8}$$

under neutrality and under selection respectively. The Wright-Fisher transition probabilities are given by the binomial distribution,

$$Q(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} q_A(x)^{Nx_A^{t+1}} (1-q_A(x))^{N-Nx_A^{t+1}}. \tag{S9}$$

$$P(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} p_A(x)^{Nx_A^{t+1}} (1-p_A(x))^{N-Nx_A^{t+1}}. \tag{S10}$$

This is a case of the Wright-Fisher model with selection among parents (see Eq. (S27,S28) and SI Sec. S5A.2). An analogous discrete-time Moran model can be written by plugging Eq. (S7,S8) into Eq. (S43,S44).

All calculations were done with $N \leq 200$. Given the small size of the system, we can compute the full matrix $P(x^{t+1}|x^t)$, and calculate the distribution over genotype frequencies over time by iterating $\psi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} \psi^{X^t}(x^t)P(x^{t+1}|x^t)$.

S4B The diffusion approximation

We write the diffusion approximation for the evolution of the frequency x_A. Following the notation in SI Sec. S7, the first two moments of change of x_A are given by

$$a_A = \mu(1 - 2\,x_A) \qquad \text{expected change due to mutation,} \tag{S11}$$

$$a_A^s = \frac{s\,x_A(1-x_A)}{1+sx_A} \qquad \text{expected change due to selection,} \tag{S12}$$

$$b_{AA} = \frac{x_A(1-x_A)}{N} \qquad \text{drift covariance.} \tag{S13}$$

The diffusion equation for this system is a special case of Eq. (S57,S58).

Maintenance of information under weak selection. Since the diffusion process takes place along only one dimension, the stationary distribution can be determined by equating the probability flux to zero. For simplicity, we neglect the mean fitness $1 + sx_A \approx 1$ in the denominator in Eq. (S12), by assuming that selection is weak, $s \ll 1$.

The stationary distributions under selection and under neutrality are

$$\tilde{\psi} = \frac{(x_A(1-x_A))^{2N\mu-1}e^{2Nsx_A}}{Z(N,\mu,s)} \qquad \tilde{\varphi} = \frac{(x_A(1-x_A))^{2N\mu-1}}{Z(N,\mu,0)} \tag{S14}$$

with the normalization constant

$$Z(N,\mu,s) = \int_0^1 (x_A(1-x_A))^{2N\mu-1}e^{2Nsx_A}dx_A = \Gamma(2N\mu)^2\,{}_1\tilde{F}_1(2N\mu;\,4N\mu;\,2Ns), \tag{S15}$$

where Γ is the Gamma function and ${}_1\tilde{F}_1$ is the regularized confluent hypergeometric function.

Similar integrals yield results for the maintained information $D(X)$, $D(G)$ and the associated expected cost at the stationary state. We calculate the expectation of the cost $C(x) = \frac{V(x)}{2\ln 2} = \frac{s^2 x_A(1-x_A)}{2\ln 2}$ (see SI Sec. S9 and S7A). For the

genotype-level information $D(G)$, we also need the expected frequency of A which is equal to its marginal probability, $\langle x_A \rangle = \tilde{\psi}^G(A)$. For brevity, we only write the leading terms in s for each quantity.

$$\langle C \rangle = \int_0^1 \tilde{\psi}\, C(x)\, dx_A \qquad = \frac{N\mu\, s^2}{(4N\mu+1)\, 2\ln 2} + O\left(s^4\right) \tag{S16}$$

$$\langle x_A \rangle = \tilde{\psi}^G(A) = \int_0^1 \tilde{\psi}\, x_A\, dx_A \qquad = \frac{1}{2} + \frac{Ns}{8N\mu+2} + O\left(s^3\right) \tag{S17}$$

$$D(G) = \langle x_A \rangle \log_2 \frac{\langle x_A \rangle}{1/2} + (1-\langle x_A \rangle) \log_2 \frac{1-\langle x_A \rangle}{1/2} = \frac{(Ns)^2}{(4N\mu+1)^2\, 2\ln 2} + O\left(s^4\right) \tag{S18}$$

$$D(X) = \int_0^1 \tilde{\psi} \log_2 \frac{\tilde{\psi}}{\tilde{\varphi}}\, dx_A \qquad = \frac{(Ns)^2}{(4N\mu+1)\, 2\ln 2} + O\left(s^4\right) \tag{S19}$$

Notably, at weak selection, both the cost $\langle C \rangle$ and the information $D(X)$, $D(G)$ scale with s^2. Their ratio is therefore given by the population size and the mutation rate,

$$\frac{D(G)}{\langle C \rangle} = \frac{N}{\mu\,(4N\mu+1)} + O(s^2), \tag{S20}$$

$$\frac{D(X)}{\langle C \rangle} = \frac{N}{\mu} + O(s^2). \tag{S21}$$

The ratio $\frac{N}{\mu\,(4N\mu+1)}$ is shown in Fig. 5C.

S5 The bound on information accumulation rate – Markov chains

The bound on information accumulation rate, as stated in Main Text Eq. (10,11) holds across several different model classes. Here we derive it for models that are Markov chain, in particular the Wright-Fisher model and the discrete Moran model. The two following sections contain similar derivations for continuous time Markov chains and the diffusion approximation. Note that all of the model parameters, such as those that describe selection, mutation or population size, can be time dependent, but we do not write it explicitly as we only need to focus on a single time step.

In the Markov chains class of models, the population state X^t takes discrete values x^t at discrete time steps t. The distribution over states is governed by

$$\varphi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} Q(x^{t+1}|x^t)\, \varphi^{X^t}(x^t) \qquad \text{under neutrality,} \tag{S22}$$

$$\psi^{X^{t+1}}(x^{t+1}) = \sum_{x^t} P(x^{t+1}|x^t)\, \psi^{X^t}(x^t) \qquad \text{under selection,} \tag{S23}$$

where $\varphi^{X^t}(x^t)$ and $\psi^{X^t}(x^t)$ are the marginal probabilities of population states at time t, and $Q(x^{t+1}|x^t) = \varphi^{X^{t+1}|X^t}(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t) = \psi^{X^{t+1}|X^t}(x^{t+1}|x^t)$ are the transition probabilities. $Q(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t)$, as well as all the parameters that we later introduce to specify them, can be time-dependent, but we do not write it explicitly. The population-level information at time t is

$$D(X^t) = \sum_{x^t} \psi^{X^t}(x^t) \log_2 \frac{\psi^{X^t}(x^t)}{\varphi^{X^t}(x^t)}. \tag{S24}$$

In general, the chain rule Eq. (S4) yields a bound

$$\Delta D(X^t) = D(X^{t+1}) - D(X^t) = D(X^{t+1}|X^t) - D(X^t|X^{t+1}) \tag{S25}$$

$$\leq D(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)}{Q(x^{t+1}|x^t)}. \tag{S26}$$

The expression $\leq D(X^{t+1}|X^t)$ corresponds to the expected KL cost of control [3], [4]. In the special case when $Q(x^{t+1}|x^t)$ and $P(x^{t+1}|x^t)$ are independent of time $\psi^{X^t}(x^t)$ is the stationary distribution of $P(x^{t+1}|x^t)$, $D(X^{t+1}|X^t)$ is also the KL divergence rate between $P(x^{t+1}|x^t)$ and $Q(x^{t+1}|x^t)$. We now examine specific forms of the transition probabilities given by the Wright-Fisher model and the Moran model.

S5A Wright-Fisher model

In this general model, each time step t represents a generation, and consists of sampling a new population of N offspring genotypes that constitute the population at $t+1$. The basic assumption is that the offspring genotypes are sampled independently with with probabilities $q_g(x^t)$ without selection or $p_g(x^t)$ with selection, leading to multinomial probability distributions over the frequencies x^{t+1} in the next generation,

$$Q(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} \prod_g q_g(x^t)^{Nx_g^{t+1}}, \tag{S27}$$

$$P(x^{t+1}|x^t) = \binom{N}{Nx^{t+1}} \prod_g p_g(x^t)^{Nx_g^{t+1}}, \tag{S28}$$

where $\binom{N}{Nx^{t+1}} = \frac{N!}{\prod_g (Nx_g^{t+1})!}$ is the multinomial coefficient. We can use these expressions to write down the general bound Eq. (S26) as

$$\Delta D(X^t) \leq D(X^{t+1}|X^t) = N \sum_{x^t} \psi^{X^t}(x^t) \sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)} \tag{S29}$$

The probabilities $q_g(x^t)$ capture arbitrary mutation and recombination, and we show examples of the form $q_g(x^t)$ can take below. Selection is modelled by the relationship between $q_g(x^t)$ and $p_g(x^t)$, and this can be done in two ways, which we discuss in the following subsections.

Asexual reproduction. In an asexual population with mutation, $q_g(x)$ will have the form

$$q_g(x) = \sum_{g'} x_{g'} \, \rho^{\text{asex}}_{g'g}, \tag{S30}$$

where $\rho^{\text{asex}}_{g'g}$ is the probability that a parent with genotype g' produces offspring with genotype g, and includes arbitrary mutation.

Sexual reproduction, random mating. Provided that recombination happens always between two parental genotypes, we sum over possible pairs of parental genotypes,

$$q_g(x) = \sum_{g_1 g_2} x_{g_1} x_{g_2} \, \rho^{\text{rec}}_{g_1 g_2 g} \tag{S31}$$

with $x_{g_1} x_{g_2}$ being the probability of a parental pair g_1, g_2 and $\rho^{\text{rec}}_{g_1 g_2 g}$ the probability that this pair produces offspring with genotype g. This includes arbitrary mutation and recombination.

Alternatively, we can distinguish between two sexes, classifying each genotype as male ($g \in \mathcal{G}^M$) of female ($g \in \mathcal{G}^F$). We sum over all male-female pairs,

$$q_g(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} \frac{x_{g_m} x_{g_f}}{Z(x)} \, \rho^{\text{sex}}_{g_m g_f g} \tag{S32}$$

with $\rho^{\text{sex}}_{g_m g_f g}$ being the probability that parents g_m, g_f give rise to offspring g and

$$Z(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} x_{g_m} x_{g_f} \tag{S33}$$

is a normalization factor such that $x_{g_m} x_{g_f}/Z(x)$ is the probability of a parental pair g_m, g_f.

Nonrandom mating. In the case of sexual reproduction, we can replace the expression $x_{g_m} x_{g_f}/Z(x)$ by a different expression for the probability of sampling a mating pair g_m, g_f. For example, we can include a factor $0 \leq \sigma_{g_m g_f} \leq 1$ corresponding to the probability that individuals with this pair of genotypes will mate. Then $q_g(x)$ will have the form

$$q_g(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} \frac{x_{g_m} x_{g_f} \sigma_{g_m g_f}}{\tilde{Z}(x)} \, \rho^{\text{sex}}_{g_m g_f g}, \tag{S34}$$

with normalization

$$\tilde{Z}(x) = \sum_{g_m \in \mathcal{G}^M} \sum_{g_f \in \mathcal{G}^F} x_{g_m} x_{g_f} \sigma_{g_m g_f}. \tag{S35}$$

S5A.1 Selection in an infinite offspring pool

Here we assume that all individuals (under asexual reproduction) or all pairs of individuals (under sexual reproduction) in the population at time t contribute a large number of offspring to a common pool. The next generation then consists of N individuals sampled from this pool, with or without selection.

The genotype frequencies in the pool will be equal to $q_g(x^t)$, since this is the probability that a random invididual (or pair of individuals) from the population x^t has offspring with genotype g. Under neutrality, the genotypes that survive and constitute the next generation are sampled at random. Under selection, genotypes from the pool are sampled with probabilities proportional to fitness $w_g(x^t)$, leading to

$$p_g(x^t) = \frac{q_g(x^t) w_g(x^t)}{\sum_{g'} q_{g'}(x^t) \, w_{g'}(x^t)} = q_g(x^t) \, \tilde{w}_g(x^t). \tag{S36}$$

where $\tilde{w}_g(x^t) = \frac{w_g(x^t)}{\sum_{g'} q_{g'}(x^t) \, w_{g'}(x)}$ is the relative fitness of g calculated within the offspring pool where selection takes place. Combining it with Eq. (S29), we obtain the bound

$$\Delta D(X^t) \leq D(X^{t+1}|X^t) = N \sum_{x^t} \psi^{X^t}(x^t) \sum_g q_g(x^t) \, \tilde{w}_g(x^t) \log_2 \tilde{w}_g(x^t) = N \langle C^t_{\text{pool}} \rangle, \tag{S37}$$

where the last expression coincides with the definition of the cost of selection in Main Text Eq. (S89), calculated at time t within the offspring pool where selection takes place, and averaged over possible population states x^t.

S5A.2 Selection among parents

Here we assume that selection takes place before reproduction. From the population at time t, we first sample N genotypes (under asexual reproduction) or N pairs of genotypes (under sexual reproduction) as parents. These are sampled independently, with probabilities proportional to fitness. Then we sample one offspring genotype for each parent/pair of parents, with mutation and recombination, and these constitute the population at time $t+1$.

When sampling genotypes as parents, g gets picked with probability given by its frequency x_g^t under neutrality, and $x_g^t \hat{w}_g(x^t)$ under selection where $\hat{w}_g(x^t) = \frac{w_g(x^t)}{\sum_{g'} x_{g'} w_{g'}(x)}$ is the relative fitness of genotype g, now computed within the adult population at time t.

We can rewrite Eq. (S29) with

$$p_g(x^t) = q_g\left(\hat{w}(x^t) \circ x^t\right), \tag{S38}$$

where $(\hat{w}(x^t) \circ x^t)_g = \hat{w}_g(x^t)\, x_g^t$ is the vector of genotype frequencies that are weighted by their relative fitness. However, Eq. (S38) is difficult to analyse. An easier approach is to introduce an intermediate variable Y^t, which signifies the genotype frequencies among parents (or pairs of parents) at generation t. The Wright-Fisher model is a Markov chain of the form

$$\cdots \to X^t \to Y^t \to X^{t+1} \to Y^{t+1} \to \cdots. \tag{S39}$$

Selection operates in the step $X^t \to Y^t$ when parents are sampled from the existing population, but not in the step $Y^t \to X^{t+1}$ where reproduction takes place with mutation and recombination. This can be expressed by writing the chain rule

$$D(X^{t+1}|X^t) = D(Y^t|X^t) + D(X^{t+1}|Y^t, X^t) - D(Y^t|X^{t+1}, X^t) \leq D(Y^t|X^t) \tag{S40}$$

where the term $D(X^{t+1}|Y^t, X^t) = D(X^{t+1}|Y^t) = 0$ because $Y^t \to X^{t+1}$ is the reproduction step with no selection. The term $D(Y^t|X^{t+1}, X^t)$ is nonnegative and can be dropped since an upper bound on $D(X^{t+1}|X^t)$ is sufficient for our purposes, but deserves some attention as it could make the bound loose. $D(Y^t|X^{t+1}, X^t)$ can only be large when, given the genotype frequencies X^t, X^{t+1} at two subsequent generations, there is uncertainty about the genotype frequencies among the parents Y^t sampled from X^t that gave rise to X^{t+1}. In diverse populations with low mutation rates or large genomes, parents tend to be easy to identify, and $D(Y^t|X^{t+1}, X^t)$ will be small.

Finally, to calculate $D(Y^t|X^t)$ we note that conditionally on X^t, Y^t is a multinomial variable with kN trials ($k = 1$ under asexual reproduction or $k = 2$ under sexual reproduction) and probabilities x_g^t under neutrality or $x_g^t \hat{w}_g(x^t)$ under selection. Then the bound can be written as

$$\Delta D(X^t) \leq D(Y^t|X^t) = kN \sum_{x^t} \psi^{X^t}(x^t) \sum_g x_g^t\, \hat{w}_g(x^t) \log_2 \hat{w}_g(x^t) = kN\langle C^t\rangle. \tag{S41}$$

Here we have assumed no distinction between sexes, but we can extend to that case by sampling N parents of each sex separately and find

$$\Delta D(X^t) \leq D(Y^t|X^t) = N\langle C^t_{\text{male}}\rangle + N\langle C^t_{\text{female}}\rangle. \tag{S42}$$

S5B Discrete-time Moran model

This model can be defined similarly to the Wright-Fisher model, but under the Moran model each time step consists of only one birth and one death, and there are N such time steps per generation.

The genotype that dies is chosen at random from the population x^t, and the probability that it will be g is equal to its frequency x_g^t. The probability that the genotype born is g' is $q_{g'}(x^t)$ under neutrality and $p_{g'}(x^t)$ under selection. These can take the same form as under the Wright-Fisher model, with selection within an infinite offspring pool or among parents. This gives rise to the transition probabilities

$$Q(x^{t+1}|x^t) = \begin{cases} x_g^t q_{g'}(x^t); & x^{t+1} = x^t + \frac{e^{g'}}{N} - \frac{e^g}{N},\ g' \neq g \quad (g \text{ dies, } g' \text{ born}), \\ \sum_g x_g^t q_g(x^t); & x^{t+1} = x^t, \\ 0; & \text{otherwise}, \end{cases} \tag{S43}$$

$$P(x^{t+1}|x^t) = \begin{cases} x_g^t p_{g'}(x^t); & x^{t+1} = x^t + \frac{e^{g'}}{N} - \frac{e^g}{N},\ g' \neq g \quad (g \text{ dies, } g' \text{ born}), \\ \sum_g x_g^t p_g(x^t); & x^{t+1} = x^t, \\ 0; & \text{otherwise}, \end{cases} \tag{S44}$$

where e^g is a vector of genotype frequencies with one at element g and zeros elsewhere. Now the bound on information accumulation rate, Eq. (S26), simplifies to

$$\Delta D(X^t) \leq D(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \left(\sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)} - \sum_g \beta_g(x^t) \log_2 \frac{\beta_g(x^t)}{\alpha_g(x^t)} \right) \tag{S45}$$

$$\leq \sum_{x^t} \psi^{X^t}(x^t) \sum_g p_g(x^t) \log_2 \frac{p_g(x^t)}{q_g(x^t)}, \tag{S46}$$

$$\leq kN\langle C^t\rangle \tag{S47}$$

where we have used $\alpha_g(x^t) = \frac{x_g^t q_g(x^t)}{\sum_{g'} x_{g'}^t q_{g'}(x^t)}$ and $\beta_g(x^t) = \frac{x_g^t p_g(x^t)}{\sum_{g'} x_{g'}^t p_{g'}(x^t)}$ to denote the probability that if the genotype that is born and dies is the same, it is the genotype g. Even though selection makes $\beta_g(x^t)$ different from $\alpha_g(x^t)$, such replacements

leave the population unchanged regardless of g, and this is associated with the nonpositive second term inside the brackets in Eq. (S45), which reduces the amount of information that can be accumulated.

Eq. (S46) is almost identical to the bound on information accumulation rate under the Wright-Fisher model, Eq. (S29). This means that the discussion of the two models of selection in Sec. S5A.1 and S5A.2 also applies to the Moran model, which allows us to write Eq. (S47). The only difference is that in each time step, we sample only one genotype from the offspring pool, or one parent (or pair of parents). This is reflected in the missing factor N. As there are N time steps per generation, the bound on information accumulated per generation again scales with $kN\langle C^t\rangle$.

Note, however, that the effective population size of the Moran model (i.e. the population size of a Wright-Fisher model with the same covariance in allele frequency change per generation) is $N_e = N/2$. This is because in the Moran model, both the births and deaths are random events, whereas the Wright-Fisher model only has random births. We can therefore expect the tighter bound $\Delta D(X^t) \lesssim kN_e\langle C^t\rangle$ to hold for large enough populations, as both models approach the same diffusion limit. We also prove the bound under the diffusion approximation separately in Sec. S7.

S6 The bound on information accumulation rate – continuous-time Markov chains

In this class of models, the genotype frequencies X^t are discrete, but the time t is continuous. The distributions $\varphi^{X^t}(x)$, $\psi^{X^t}(x)$ over X^t are governed by the master equations

$$\frac{d}{dt}\varphi^{X^t}(x) = \sum_{x'} \bar{Q}(x,x')\,\varphi^{X^t}(x') \qquad \text{under neutrality,} \tag{S48}$$

$$\frac{d}{dt}\psi^{X^t}(x) = \sum_{x'} \bar{P}(x,x')\,\psi^{X^t}(x') \qquad \text{under selection,} \tag{S49}$$

where $\bar{Q}(x,x')$ and $\bar{P}(x,x')$ are transition rates from x' to x under neutrality and selection respectively. These can be time dependent. The population-level information $D(X^t)$ is defined as in Eq. (S24), but now changes continuously in time. The rate of this change is upper bounded as

$$\frac{d}{dt}D(X^t) \leq \sum_x \psi^{X^t}(x) \sum_{x,\,x'\neq x} \left(\bar{P}(x',x)\log\frac{\bar{P}(x',x)}{\bar{Q}(x',x)} + \bar{Q}(x',x) - \bar{P}(x',x)\right). \tag{S50}$$

When $\bar{P}$ and $\bar{Q}$ are independent of time and $\psi^{X^t}(x)$ is the stationary distribution associated with $\bar{P}$, the right hand side corresponds to the KL divergence rate between $\bar{P}$ and $\bar{Q}$, as derived in [5]. The bound Eq. (S50) can be verified algebraically, or derived from the discrete bound Eq. (S26) by taking $Q = e^{\epsilon\bar{Q}}$, $P = e^{\epsilon\bar{P}}$ and the limit $\epsilon \to 0$. We now show an example of the form that $\bar{P}$ and $\bar{Q}$ can take.

S6A Continuous-time Moran model

This model is based on its discrete-time counterpart in Sec. S5B. Only transitions consisting of replacing one genotype (g, death) by another (g', birth) are allowed, and the transition rates have the form

$$\bar{Q}(x',x) = \begin{cases} Nx_g q_{g'}(x); & x' = x - \frac{e^g}{N} + \frac{e^{g'}}{N},\ g' \neq g, \\ -\sum_{g,g'\neq g} Nx_g q_{g'}(x); & x' = x, \\ 0; & \text{otherwise,} \end{cases} \tag{S51}$$

$$\bar{P}(x',x) = \begin{cases} Nx_g p_{g'}(x); & x' = x - \frac{e^g}{N} + \frac{e^{g'}}{N},\ g' \neq g, \\ -\sum_{g,g'\neq g} Nx_g p_{g'}(x); & x' = x, \\ 0; & \text{otherwise,} \end{cases} \tag{S52}$$

where time is measured in generations, i.e. there are on average N replacement events per unit time. With this form of the transition rates, we can rewrite the general bound Eq. (S50) as

$$\frac{d}{dt}D(X^t) \leq N\sum_x \psi^{X^t}(x) \sum_{g,g'\neq g} x_g \left(p_{g'}(x)\log\frac{p_{g'}(x)}{q_{g'}(x)} + q_{g'}(x) - p_{g'}(x)\right) \tag{S53}$$

$$= N\sum_x \psi^{X^t}(x) \sum_g \left(p_g(x)\log\frac{p_g(x)}{q_g(x)} - x_g q_g(x)\left(\frac{p_g(x)}{q_g(x)}\log\frac{p_g(x)}{q_g(x)} + 1 - \frac{p_g(x)}{q_g(x)}\right)\right) \tag{S54}$$

$$\leq N\sum_x \psi^{X^t}(x) \sum_g p_g(x)\log\frac{p_g(x)}{q_g(x)}. \tag{S55}$$

This is the same bound as for the discrete Moran model in Eq. (S47), up to the factor N, which is due to the different unit of time.

S7 The bound on information accumulation rate – diffusion approximation

Here we show an upper bound on the rate of accumulation of information under the diffusion approximation. The approach is similar to Iwasa [6] (who assumed detailed balance) and Hasegawa [7] (who did not), but here we distinguish between the processes with and without selection. We start by deriving a general bound for a pair of diffusion processes, and then apply it to the population genetics context in Sec. S7A.

For brevity, we write the probability density over population states x at time t as $\varphi = \varphi(x,t)$ under neutrality and $\psi = \psi(x,t)$ under selection. The population-level information is now determined by integration,

$$D(X) = \int \psi \log_2 \frac{\psi}{\varphi}\, dx = \frac{1}{\ln 2} \int \psi \ln \frac{\psi}{\varphi}\, dx. \tag{S56}$$

Note that while we stick to measuring information in bits, it is more convenient to use the natural logarithm during the derivation.

The diffusion equation is parametrized by the first and second moment of change in x_g. Selection is assumed to only exert control through the first moment, which we label as a_g under neutrality and $a_g + a_g^s$ under selection. The second moment is $b_{gg'}$, both under neutrality and under selection. All these are functions of x and t, e.g. $a_g = a_g(x,t)$, but we do not write this for brevity. We sum over any index that appears twice in a term, e.g. $\partial_g b_{g'g} = \sum_{g=1}^{\mathcal{G}-1} \partial_g b_{g'g}$. Note that the diffusion is described in the subspace of $\mathcal{G}-1$ genotype frequencies, where $\mathcal{G}$ is the number of genotypes – the last frequency is determined by normalization, $x_{\mathcal{G}} = 1 - \sum_{g=1}^{\mathcal{G}-1} x_g$. The diffusion equation is

$$\partial_t \varphi = -\partial_g \left(u_g \varphi\right), \qquad u_g = a_g \qquad - \frac{1}{2}\partial_{g'} b_{gg'} - \frac{1}{2} b_{gg'} \partial_{g'} \ln \varphi \qquad \text{under neutrality,} \tag{S57}$$

$$\partial_t \psi = -\partial_g \left(v_g \psi\right), \qquad v_g = a_g + a_g^s - \frac{1}{2}\partial_{g'} b_{gg'} - \frac{1}{2} b_{gg'} \partial_{g'} \ln \psi \qquad \text{under selection,} \tag{S58}$$

where we introduced the velocity fields $u_g = u_g(x,t)$ and $v_g = v_g(x,t)$ such that $u_g\varphi$ and $v_g\psi$ are the probability fluxes under neutrality and under selection respectively. From their definition, it follows that

$$\partial_g \ln \frac{\psi}{\varphi} = -2 b_{gg'}^{-1} \left(v_{g'} - u_{g'} - a_{g'}^s\right). \tag{S59}$$

The rate of change of $D(X)$ can be written as

$$\frac{d}{dt} D(X) = \frac{1}{\ln 2} \int \partial_t \left(\psi \ln \frac{\psi}{\varphi}\right) dx, \tag{S60}$$

and the integrand can be written as

$$\partial_t \left(\psi \ln \frac{\psi}{\varphi}\right) = \ln \frac{\psi}{\varphi}\, \partial_t \psi + \partial_t \psi - \frac{\psi}{\varphi}\, \partial_t \varphi \tag{S61}$$

$$= -\ln \frac{\psi}{\varphi}\, \partial_g \left(v_g \psi\right) - \partial_g \left(v_g \psi\right) + \frac{\psi}{\varphi}\, \partial_g \left(u_g \varphi\right) \tag{S62}$$

$$= -\partial_g \left(v_g \psi \ln \frac{\psi}{\varphi} + \left(v_g - u_g\right)\psi\right) + \psi \left(v_g - u_g\right) \partial_g \ln \frac{\psi}{\varphi}. \tag{S63}$$

The first term is a divergence and vanishes after integration in Eq. (S60) because $u_g\varphi$ and $v_g\psi$ cannot cross the domain boundary (assuming $\psi/\varphi < \infty$). Therefore

$$\frac{d}{dt} D(X) = \frac{1}{\ln 2} \int \psi \left(v_g - u_g\right) \partial_g \ln \frac{\psi}{\varphi}\, dx \tag{S64}$$

$$= -\frac{2}{\ln 2} \int \psi \left(v_g - u_g\right) b_{gg'}^{-1} \left(v_{g'} - u_{g'} - a_{g'}^s\right) dx \tag{S65}$$

$$= -\frac{2}{\ln 2} \int \psi \left(v_g - u_g - \frac{1}{2} a_g^s\right) b_{gg'}^{-1} \left(v_{g'} - u_{g'} - \frac{1}{2} a_{g'}^s\right) dx + \frac{1}{2 \ln 2} \int \psi\, a_g^s\, b_{gg'}^{-1}\, a_{g'}^s\, dx. \tag{S66}$$

The last expression has two terms, both of which are quadratic forms. The first one makes a nonpositive contribution, leading to the upper bound in information accumulation rate,

$$\frac{d}{dt} D(X) \leq \frac{1}{2 \ln 2} \int \psi\, a_g^s\, b_{gg'}^{-1}\, a_{g'}^s\, dx. \tag{S67}$$

On the right hand side we can identify the KL cost of control [4] in bits. This bound holds for any pair of diffusion processes with the same fluctuations covariance $b_{gg'}$. We will now discuss it in the context of population genetics.

S7A Application to population genetics

The bound Eq. (S67) does not depend on the form of a_g, and therefore it can be used to model arbitrary mutation and recombination, for example by taking

$$a_g = q_g - x_g \tag{S68}$$

with $q_g = q_g(x)$ as introduced in the discrete models above. Notably, a_g does not need to be the gradient of a scalar potential.

The fluctuations in genotype frequencies can modeled based on multinomial sampling in the Wright-Fisher model (Eq. (S27,S28)),

$$b_{gg'} = \frac{\delta_{gg'} x_g - x_g x_{g'}}{N}, \tag{S69}$$

with no summation over g, and where the (effective) population size N can be time-dependent. This has the inverse [8]

$$b^{-1}_{gg'} = N \left(\frac{\delta_{gg'}}{x_g} + \frac{1}{x_{\mathcal{G}}} \right). \tag{S70}$$

The control term a^s_g imposed by selection can be written as

$$a^s_g = (\hat{w}_g - 1)\, x_g \qquad \text{where} \qquad \hat{w}_g = \frac{w_g}{\sum_g w_g x_g} \tag{S71}$$

where w_g is the fitness of genotype g, possibly time and frequency dependent, and $\hat{w}_g$ is the relative fitness. With these definitions, we find that

$$\sum_{g,g'=1}^{\mathcal{G}-1} a^s_g\, b^{-1}_{gg'}\, a^s_{g'} = N \sum_{g=1}^{\mathcal{G}} \frac{(a^s_g)^2}{x_g} = N \sum_{g=1}^{\mathcal{G}} (\hat{w}_g - 1)^2\, x_g = NV(x) \tag{S72}$$

and the bound on information accumulation rate is

$$\frac{d}{dt} D(X) \leq \frac{N\langle V \rangle}{2 \ln 2} = N \langle C \rangle. \tag{S73}$$

In the last equation, we equate $\frac{\langle V \rangle}{2 \ln 2} = \langle C \rangle$ – we show that this holds under weak selection in Sec. S9.

We can get intuition about the tightness of this bound by analyzing the first, nonpositive term in Eq. (S66). The bound is only tight when $v_g - u_g - \frac{1}{2} a^s_g = 0$ for all x with nonzero ψ. An interesting specific case is when the neutral process is at an equilibrium with detailed balance, such that $u_g = 0$. We note that v_g can be decomposed as $v_g = v'_g + a^s_g$ into the contribution from selection a^s_g and all other evolutionary forces, v'_g. Our bound is then tight when $a^s_g = -2v'_g$, i.e. when selection induces a probability flux in exactly the opposite direction and exactly twice the magnitude as the all the other evolutionary forces combined. While this might occasionally and approximately be the case, it will only be a transient phenomenon.

Suppose that the population starts at the neutral equilibrium with $v'_g = 0$ and then selection starts to act, e.g. after a change in the environment. For any nonzero a^s_g, the bound cannot be tight as $a^s_g \neq -2v'_g = 0$. After some time a new equilibrium might be reached, with $v_g = v'_g + a^s_g = 0$, where again, the bound is not tight $a^s_g \neq -2v'_g = 2a^s_g$. In this case, maintenance costs are incurred but no further adaptation takes place. The bound can only be tight for a moment when adaptation is taking place, selection pulls the population in the opposite direction as the other evolutionary forces combined, but selection is twice as strong, $a^s_g = -2v'_g$.

S8 Relationship with free fitness and statistical physics

Stochastic models in population genetics show some mathematical properties analogous to statistical physics. In particular, a quantity called free fitness, analogous to free energy in physics, can be defined and shown to monotonically increase over time. In this section we provide some background about free fitness and discuss two connections with our work. First, the increasing property of free fitness can be proved by a method similar to the proof of our bound on information accumulation rate. Second, free fitness can be written as the difference between mean log fitness and genetic information as defined in this paper (e.g. on the genotype or population level), implying that evolution tends to maximize mean log fitness at a given amount of information.

S8A Boltzmann form of stationary distributions

Under suitable conditions, the stationary distributions in population genetics models take a form similar to the Boltzmann distribution. Models on both the population level and the genotype level display this property.

- If mutation is weak ($NU \ll 1$ where U is the total mutation rate across the studied genomic region), populations are mostly monomorphic, with only occasional fixations of a different genotype. The system can then be described with the most recently fixed genotype g and the distribution $\psi^G(g)$. The stationary distribution $\tilde{\psi}^G(g)$ takes the form [9], [10]

$$\tilde{\psi}^G(g) = \frac{1}{Z^G} e^{2N \ln w_g}, \tag{S74}$$

  where $2N$ is again analogous to inverse temperature, log fitness $\ln w_g$ is analogous to negative energy, and Z^G is normalization constant.

- Assuming many biallelic loci under linkage equilibrium, we can describe the system with the vector of allele frequencies p and the joint distribution $\psi^P(p)$ over them. The stationary distribution $\tilde{\psi}^P(p)$ can be derived from the diffusion approximation and takes the form [11], [12]

$$\tilde{\psi}^P(p) = \frac{1}{Z^P} \prod_i (p_i q_i)^{2N\mu - 1}\, e^{2N \ln \bar{w}(p)}, \tag{S75}$$

  where p_i and $q_i = 1 - p_i$ are the allele frequencies of the two alleles at locus i and Z^P is a normalization constant. Twice the population size $2N$ takes the role of inverse temperature and log mean fitness $\ln \bar{w}(p)$ takes the role of negative energy. The factors $(p_i(1-p_i))^{2N\mu-1}$ correspond to mutation and drift potential (similar to e.g. chemical potential), which will be made clearer in the next subsection.

The formulas apply to haploids, but similar formulas apply to diploids or when mutation coefficients vary across loci or alleles (see e.g. the SI of [10]). Importantly, they depend on the assumption of detailed balance. At stationarity, net probability flux between any two allele frequency vectors or genotypes must be zero, making $\tilde{\psi}^G(g)$ and $\tilde{\psi}^P(p)$ equilibrium distributions. This can be violated under certain forms of mutation, recombination and strong selection, leading to additional terms related to robustness [13].

S8B Free fitness

When a system starts from an arbitrary initial distribution and approaches the stationary distribution in Eq. (S75) or Eq. (S74), we can track this progress using free fitness [6], [10], which increases monotonically in time as was shown previously and as we can also prove in more generality here.

- On the genotype level, we can define free fitness F^G at any distribution $\psi^G(g)$ away from equilibrium as a sum of expected log fitness and entropy terms,

$$F^G = \underbrace{\langle \ln w_g \rangle_{\psi^G}}_{\text{Selection}} + \frac{1}{2N} \underbrace{\left\langle \ln \frac{1}{\psi^G(g)} \right\rangle_{\psi^G}}_{\text{Entropy}} \tag{S76}$$

$$= \underbrace{\frac{\ln Z^G}{2N}}_{\substack{\text{Equilibrium}\\ \text{free fitness}}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^G(g)}{\tilde{\psi}^G(g)} \right\rangle_{\psi^G}}_{\substack{\text{KL divergence}\\ \text{from equilibrium}}} = \tilde{F}^G - \frac{1}{2N} D_{KL}(\psi^G || \tilde{\psi}^G) \tag{S77}$$

  where $\langle \cdot \rangle_{\psi^G}$ denotes an expectation over $g \sim \psi^G(g)$. In the special case of equilibrium $\psi^G = \tilde{\psi}^G$, we obtain $F^G = \tilde{F}^G = \frac{\ln Z^G}{2N}$, and away from equilibrium, free fitness is reduced by an amount proportional to the KL divergence between the actual distribution ψ^G and the equilibrium $\tilde{\psi}^G$.

- On the population level with linkage equilibrium, free fitness at a distribution ψ^P can be defined as a sum of three terms – a negative potential for selection, mutation and drift, and entropy:

$$F^P = \underbrace{\langle \ln \bar{w}(p) \rangle_{\psi^P}}_{\text{Selection}} + \underbrace{\frac{2N\mu - 1}{2N} \left\langle \sum_i \ln(p_i q_i) \right\rangle_{\psi^P}}_{\text{Mutation and drift}} + \frac{1}{2N} \underbrace{\left\langle \ln \frac{1}{\psi^P(p)} \right\rangle_{\psi^P}}_{\text{Entropy}} \tag{S78}$$

$$= \underbrace{\frac{\ln Z^P}{2N}}_{\substack{\text{Equilibrium}\\ \text{free fitness}}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^P(p)}{\tilde{\psi}^P(p)} \right\rangle_{\psi^P}}_{\substack{\text{KL divergence}\\ \text{from equilibrium}}} = \tilde{F}^P - \frac{1}{2N} D_{KL}(\psi^P || \tilde{\psi}^P) \tag{S79}$$

  where the expectations $\langle \cdot \rangle$ are taken over $p \sim \psi^P(p)$. While mutation and drift now appear as additional terms in free fitness, free fitness can again be decomposed into its value at equilibrium and a difference proportional to the KL divergence away from it.

The key property of the free fitness is that it is a non-decreasing function of time, until it is maximized at equilibrium. This was proved by Iwasa [6] for the case of F^P and Sella and Hirsh [10] for the case of F^G. In the next section we show that both results can also be derived by the same method as our bound on information accumulation rate.

S8C Monotonic convergence of stochastic processes to their stationary distributions

The general bounds on information accumulation rate (Eq. (S26) for Markov chains, Eq. (S50) for continuous time Markov chains and Eq. (S67) for the diffusion approximation) apply for any pair of stochastic processes, provided that they have compatible support such that the KL divergence is well defined. To make this explicit, we focus on the case of discrete Markov chains, consider some general process ξ instead of the neutral process φ, and introduce notation that generalizes $D(X^t)$,

$$D_{\psi||\xi}(X^t) = \sum_{x^t} \psi^{X^t}(x^t) \log_2 \frac{\psi^{X^t}(x^t)}{\xi^{X^t}(x^t)} \tag{S80}$$

$$D_{\psi||\xi}(X^{t+1}|X^t) = \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} \psi^{X^{t+1}|X^t}(x^{t+1}|x^t) \log_2 \frac{\psi^{X^{t+1}|X^t}(x^{t+1}|x^t)}{\xi^{X^{t+1}|X^t}(x^{t+1}|x^t)}. \tag{S81}$$

The KL divergence chain rule now yields the inequality

$$\Delta D_{\psi||\xi}(X^t) = D_{\psi||\xi}(X^{t+1}) - D_{\psi||\xi}(X^t) \leq D_{\psi||\xi}(X^{t+1}|X^t). \tag{S82}$$

If $\xi = \varphi$ is the neutral process, this is the KL cost of selection bound on the information accumulation rate (Eq. (S26) and Main Text Eq. (9). But we can also choose ξ such that it does contain selection and has the same transition probabilities as ψ, but starts from a different initial condition, i.e. $\xi^{X^{t+1}|X^t} = \psi^{X^{t+1}|X^t}$ and $\xi^{X^0} \neq \psi^{X^0}$. Then we find that

$D_{\psi||\xi}(X^{t+1}|X^t) = 0$ and $\Delta D_{\psi||\xi}(X^t) \leq 0$, i.e. the divergence between ψ^{X^t} and ξ^{X^t} is non-increasing over time, because relative to ξ, there is no control exerted on ψ.

If, in addition, the system has a unique stationary distribution $\tilde{\psi}^X$ and ξ^{X^0} is initialized there (and therefore stays there indefinitely, $\xi^{X^0} = \xi^{X^t} = \tilde{\psi}^X$ for any t), we find that ψ^{X^t} converges to this stationary distribution $\tilde{\psi}^X$ monotonically in terms of the KL divergence $D_{\psi||\xi}(X^t)$. Similar proofs apply to continuous time Markov chains and diffusion, since we only need to replace φ by ξ and repeat the derivation leading to Eq. (S50) and Eq. (S67).

In the two regimes discussed above in Sec. S8A, the population state X corresponds to some fixed genotype G or a vector of allele frequencies P. Therefore $D_{\psi||\xi}(X^t) = D_{KL}(\psi^{G^t}||\tilde{\psi}^G)$ or $D_{\psi||\xi}(X^t) = D_{KL}(\psi^{P^t}||\tilde{\psi}^P)$ are non-increasing functions of time. Together with Eq. (S77,S79), this implies that F^G or F^P are non-decreasing functions of time.

Iwasa [6] and Sella and Hirsh [10] proved the same result by different methods. In our framework it emerges as a special case of the information accumulation bound with zero control. The key part of our proof, stating that $D_{\psi||\xi}(X^t)$ is non-increasing, is also more general (regarding the state space, the form of the stationary distribution, and detailed balance – although free fitness is not defined so generally). A similarly general proof for continuous time Markov chains, as well as several related results for replicator dynamics and reaction networks, is reviewed in reference [14].

S8D Free fitness as a trade-off between fitness and information

We can rewrite the expressions for free fitness using the genotype and population-level information respectively. We first write down the neutral stationary distributions. On the genotype level, we assume that it is uniform over 4^l possible sequences of length l,

$$\tilde{\varphi}^G(g) = \frac{1}{4^l}. \tag{S83}$$

On the population level,

$$\tilde{\varphi}^P(p) = \frac{1}{Z^{\varphi,P}} \prod_i (p_i q_i)^{2N\mu-1}, \tag{S84}$$

where $Z^{\varphi,P}$ is the normalization constant, to be distinguished from Z^P in Eq. (S75) which includes selection.

Using $\tilde{\varphi}^G(g)$ and $\tilde{\varphi}^P(p)$, we can rewrite free fitness, Eq. (S77,S79), as

$$F^G = \underbrace{\langle \ln w_g \rangle_{\psi^G}}_{\text{Selection}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^G(g)}{\tilde{\varphi}^G(g)} \right\rangle_{\psi^G}}_{\substack{\text{Genotype-level}\\ \text{information}}} - \underbrace{\frac{1}{2N} \ln(4^l)}_{\substack{\text{Independent}\\ \text{of } \psi^G}} = \langle \ln w_g \rangle_{\psi^G} - \frac{1}{2N} D(G) + \text{const.} \tag{S85}$$

on the genotype level and

$$F^P = \underbrace{\langle \ln \bar{w}(p) \rangle_{\psi^P}}_{\text{Selection}} - \frac{1}{2N} \underbrace{\left\langle \ln \frac{\psi^P(p)}{\tilde{\varphi}^P(p)} \right\rangle_{\psi^P}}_{\substack{\text{Population-level}\\ \text{information}}} + \underbrace{\frac{\ln Z^{\varphi,P}}{2N}}_{\substack{\text{Independent}\\ \text{of } \psi^P}} = \langle \ln \bar{w}(p) \rangle_{\psi^P} - \frac{1}{2N} D(P) + \text{const.} \tag{S86}$$

on the population level. In both cases we have emphasized that terms independent of ψ^G or ψ^P are constant in time and therefore not important for the dynamics of free fitness. Up to the constant, this formula for free fitness has also been used in the paper on fitness flux [15]. The fitness flux theorem (ref. [15] and Sec. S10) then provides perhaps the most elegant proof that free fitness is a non decreasing function of time, as it relates changes in $D(P)$ to changes in expected fitness.

Free fitness tends to increase over time until it is maximized at the Boltzmann-like equilibrium distribution $\tilde{\psi}^G$ or $\tilde{\psi}^P$. In other words, evolution maximizes the expected log fitness while constraining the amount of genetic information, with $1/(2N)$ serving as a Lagrange multiplier that controls the trade-off.

S9 Properties of measures of cost of selection

Here we prove general inequalities between the genetic load $L(x)$, relative fitness variance $V(x)$, and the information theoretic cost $C(x)$. We also derive the form of $C(x)$ for the special cases of weak selection and truncation selection. The three measures are defined as

$$L(x) = 1 - \frac{1}{\hat{w}_{\max}(x)} \tag{S87}$$

$$V(x) = \sum_g x_g \left(\hat{w}_g(x) - 1\right)^2, \tag{S88}$$

$$C(x) = \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_g(x), \tag{S89}$$

where x_g is the frequency of genotype g in the population and $\hat{w}_{\max}(x) = \max_{g;\, x_g>0} \hat{w}_g(x)$ is the relative fitness of the fittest individual that is present in the population ($x_g > 0$).

We note that some previous work has defined $\hat{w}_{\max}(x)$ to be the maximum fitness possible, i.e. the fitness of an ideal genotype with no deleterious mutations regardless of whether such an individual exists. Load computed with such a definition is higher, and this has led to claims of severe restrictions on the rate of adaptive substitutions [16] and the functional fraction of the human genome [17]. However, load under this definition has been criticized as irrelevant, since the ideal genotype has a vanishing probability of existing in the population, and if only the fitness values likely to be present in the population are considered, load-based restrictions are more permissive [18], [19]. Our definitions of $L(x)$, $V(x)$ and $C(x)$ all focus on the existing variation of fitness in the population x. $L(x)$ is also related to the concept of lead, which was defined as the difference between the maximum and the mean log fitness in a traveling wave [20].

Truncation selection and limitations by reproductive capacity. Under truncation selection, a fraction α of individuals in the population has constant relative fitness, equal to the maximum $\hat{w}_g(x) = \hat{w}_{\max}(x)$ and the remaining fraction $1-\alpha$ has relative fitness zero $\hat{w}_g(x) = 0$. By definition, the mean relative fitness must be $\sum_g x_g \hat{w}_g(x) = 1$, which requires $\hat{w}_{\max}(x) = 1/\alpha$. The three measures of cost of selection then are

$$L^{\text{trunc}}(x) = 1 - \alpha, \tag{S90}$$

$$V^{\text{trunc}}(x) = \alpha \left(\frac{1}{\alpha} - 1\right)^2 + (1-\alpha)(0-1)^2 = \frac{1}{\alpha} - 1, \tag{S91}$$

$$C^{\text{trunc}}(x) = -\log_2 \alpha. \tag{S92}$$

At a constant population size, the expected number of offspring of an individual is equal to their relative fitness $\hat{w}_g(x)$ (or $2\hat{w}_g(x)$ under sexual reproduction, with two parents per offspring). In a species with a reproductive capacity R, we have $\hat{w}_{\max}(x) \leq R$ and the load is limited as $L \leq 1 - 1/R$. $V(x)$ and $C(x)$ at given R are maximized under truncation selection, when only the most extreme relative fitness values available are occupied ($N\alpha$ individuals have relative fitness $\hat{w}_g(x) = 1/\alpha = \hat{w}_{\max}(x) = R$ and $N - N\alpha$ individuals have fitness 0). This implies upper bounds $V(x) \leq R - 1$ and $C(x) \leq \log_2 R$.

General inequality between $L(x)$ and $V(x)$. From Eq. (S87), the genetic load $L(x)$ determines the maximum relative fitness in the population, $\hat{w}_{\max}(x) = 1/(1 - L(x))$. Given that relative fitness of all individuals in the population must lie between 0 and $\hat{w}_{\max}(x)$, its variance $V(x)$ is maximized when only these extreme values are occupied, i.e. under truncation selection. In that case we have $V^{\text{trunc}}(x) = 1/\alpha - 1$ with $\alpha = 1/\hat{w}_{\max}(x)$. This implies a general bound,

$$V(x) \leq V^{\text{trunc}}(x) = \frac{L(x)}{1 - L(x)}, \tag{S93}$$

with equality under truncation selection. The same inequality was derived in ref. [21] by other means.

General inequality between $L(x)$ and $C(x)$. Since logarithm is an increasing function and $x_g \hat{w}_g(x) \leq 0$, we can upper bound each term of the form $x_g \hat{w}_g(x) \log_2 \hat{w}_g(x)$ in Eq. (S89) by $x_g \hat{w}_g(x) \log_2 \hat{w}_{\max}(x)$. Summing over g, we obtain

$$C(x) \leq \sum_g x_g \hat{w}_g(x) \log_2 \hat{w}_{\max}(x) = \log_2 \hat{w}_{\max}(x) = \log_2 \frac{1}{1 - L(x)}. \tag{S94}$$

Equality is again achieved under truncation selection.

General inequality between $V(x)$ and $C(x)$. We use the inequality $\log_2 u \leq \frac{u-1}{\ln 2}$ in Eq. (S89) to obtain

$$C(x) \leq \sum_g x_g \hat{w}_g(x) \frac{\hat{w}_g(x) - 1}{\ln 2} = \frac{1}{\ln 2} \left(\sum_g x_g \hat{w}_g(x)^2 - 1 \right) = \frac{V(x)}{\ln 2}. \tag{S95}$$

Equality is approached under truncation selection when $\alpha \to 1$.

$C(x)$ under weak selection. Here we assume that for all genotypes g present in the population ($x_g > 0$), the relative fitness $\hat{w}_g(x)$ is close to 1. We can then use the Taylor expansion

$$\hat{w}_g(x) \log_2 \hat{w}_g(x) = \frac{1}{\ln 2} \left(\hat{w}_g(x) - 1 + \frac{1}{2} (\hat{w}_g(x) - 1)^2 + O\left((\hat{w}_g(x) - 1)^3\right) \right). \tag{S96}$$

Combining this with Main Text Eq. (S89), we find

$$C(x) = \frac{1}{\ln 2} \sum_g x_g \left(\hat{w}_g(x) - 1 + \frac{1}{2} (\hat{w}_g(x) - 1)^2 + O\left((\hat{w}_g(x) - 1)^3\right) \right) \tag{S97}$$

$$= \frac{V(x)}{2 \ln 2} + \frac{1}{\ln 2} \sum_g x_g \, O\left((\hat{w}_g(x) - 1)^3\right), \tag{S98}$$

or in short, $C(x) \approx V(x)/(2 \ln 2)$ under weak selection. This is particularly relevant for the diffusion limit, when the population size is sent to infinity, and the selection strength is rescaled inversely to the population size.

S10 Fitness flux theorem

In this section we compare the newly introduced bound on information accumulation rate (the *cost of selection bound*) and a similar bound implied by the fitness flux theorem [15] (the *fitness flux bound*). The fitness flux theorem was originally derived under the diffusion approximation. For better comparison, we also derive an analogous result for discrete-time Markov chains. We then discuss the distinct interpretation of the two bounds, and illustrate them (using both the discrete and diffusion expressions) in Fig. S2.

S10A Discrete-time Markov chains

The fitness flux theorem, like its counterparts in statistical physics (e.g. [22]), is based on the comparison of forward and reverse path probabilities. For simplicity, we will not derive the fitness flux theorem in its general form, but rather the form that allows a direct comparison with the cost of selection bound.

We focus on short paths consisting of only one step, (X^t, X^{t+1}). The probability of the forward path (x^t, x^{t+1}) is $\psi^{X^t}(x^t)P(x^{t+1}|x^t)$. We consider a probability distribution over reverse paths, $\psi^{X^{t+1}}(x^{t+1})P(x^t|x^{t+1})$, which is normalized to 1 – we can write this as

$$1 = \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \exp\ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)} \tag{S99}$$

By Jensen's inequality,

$$0 \geq \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)}, \tag{S100}$$

Next, inside the logarithm, we divide and multiply by the neutral probabilities,

$$0 \geq \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})\, \frac{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}}{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)\, \frac{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}} \tag{S101}$$

$$= \sum_{x^t\, x^{t+1}} \psi^{X^t}(x^t)\, P(x^{t+1}|x^t) \ln \frac{\frac{\psi^{X^{t+1}}(x^{t+1})\, P(x^t|x^{t+1})}{\varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})}}{\frac{\psi^{X^t}(x^t)\, P(x^{t+1}|x^t)}{\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t)}}, \tag{S102}$$

where we assumed that the neutral process is at a stationary distribution with detailed balance, i.e. $\varphi^{X^t}(x^t)\, Q(x^{t+1}|x^t) = \varphi^{X^{t+1}}(x^{t+1})\, Q(x^t|x^{t+1})$. Finally, we rearrange terms and divide by $\ln 2$ to get an expression in bits,

$$\text{Fitness flux bound:} \qquad \Delta D(X^t) \leq \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)\, Q(x^t|x^{t+1})}{Q(x^{t+1}|x^t)\, P(x^t|x^{t+1})} = 2N\langle\phi\rangle_t, \tag{S103}$$

$$\text{Cost of selection bound:} \qquad \Delta D(X^t) \leq \sum_{x^t} \psi^{X^t}(x^t) \sum_{x^{t+1}} P(x^{t+1}|x^t) \log_2 \frac{P(x^{t+1}|x^t)}{Q(x^{t+1}|x^t)} \leq kN\langle C\rangle_t, \tag{S104}$$

where $\langle\phi\rangle_t$ is the discrete analog of the fitness flux, averaged over the possible transitions (x^t, x^{t+1}),

$$\phi(x^t, x^{t+1}) = \frac{1}{2N} \log_2 \frac{P(x^{t+1}|x^t)\, Q(x^t|x^{t+1})}{Q(x^{t+1}|x^t)\, P(x^t|x^{t+1})}. \tag{S105}$$

The interpretation of this expression and the relationship to fitness accumulation is most clear in the diffusion approximation, see below. To help compare the fitness flux bound with the cost of selection bound, we take the expectation over the final state x^{t+1} and compute the expected fitness flux from any initial state x^t, $\phi(x^t) = \sum_{x^{t+1}} P(x^{t+1}|x^t)\phi(x^t, x^{t+1})$. An example plot of $\phi(x^t)$ is in Fig. S2A. For comparison, we also included in Eq. (S104) the cost of selection bound.

While the cost of control is a non-negative conditional KL divergence, the fitness flux bound contains an additional term related to the reverse transition probabilities, and can be negative (this is more easily interpretable as the mutation term in the diffusion approximation). The fitness flux bound relies on the additional assumption that the neutral process is at a stationary distribution with detailed balance. This can be satisfied in the single locus, two allele system when using the Moran model, but it is violated by the Wright-Fisher model which we use throughout most of the paper.

S10B Diffusion approximation

Mustonen and Lässig [15] derive the fitness flux theorem using a similar method but under the diffusion approximation, where the fitness flux is related to the rate accumulation of fitness. We include here an informal account of how that relates to the formula in Eq. (S104).

In continuous time, we can generalize the definition of fitness flux in Eq. (S105) to an arbitrary time interval Δt and the transition $(x^t, x^{t+\Delta t})$.

$$\phi(x^t, x^{t+\Delta t}) = \frac{1}{\Delta t}\frac{1}{2N} \log_2 \frac{P(x^{t+\Delta t}|x^t)\, Q(x^t|x^{t+\Delta t})}{Q(x^{t+\Delta t}|x^t)\, P(x^t|x^{t+\Delta t})}, \tag{S106}$$

where we also divided by Δt to get the fitness flux per generation. Under the diffusion approximation, if Δt is small, the transition probabilities P, Q will be approximately normal with parameters given by $a(x^t)$, $a^s(x^t)$ and $b(x^t)$ (see Sec. S7; we drop the dependence on x^t for brevity),

$$P(x^{t+\Delta t}|x^t) \approx \mathcal{N}(x^{t+\Delta t};\, x^t + a + a^s,\, b), \tag{S107}$$

$$Q(x^{t+\Delta t}|x^t) \approx \mathcal{N}(x^{t+\Delta t};\, x^t + a,\, b). \tag{S108}$$

Then in Eq. (S106) we recover the definition of fitness flux from [15],

$$\phi(x^t, x^{t+\Delta t}) \approx \frac{1}{N\Delta t \ln 2}\, (x^{t+\Delta t} - x^t)_g\, b^{-1}_{gg'}\, a^s_{g'} \tag{S109}$$

where we sum over repeated indices as in Sec. S7. Note that the factor $1/\ln 2$ appears because we use base 2 logarithms throughout. If the process takes place in a fitness landscape/seascape F, the vector $b^{-1}_{gg'}\, a^s_{g'}/N = \partial_g F$ is its gradient, and $\phi(x^t, x^{t+\Delta t})$ is the rate at which the system climbs it up and therefore accumulates fitness (Fig. 1 and Eq. S8-S10 in [15]). This interpretation of fitness flux is exact in the diffusion approximation as $\Delta t \to 0$, but only approximate for the discrete formulas in Eq. (S105) and Eq. (S103).

We can take the expectation over $x^{t+\Delta t}$ to obtain the expected fitness flux per generation from any starting position x^t,

$$\phi(x^t) \approx \frac{1}{N \ln 2}(a + a^s)_g b^{-1}_{gg'}\, a^s_{g'}, \tag{S110}$$

which is also plotted in Fig. S2A. Finally, we can take the expectation over x^t to obtain the fitness flux bound in the diffusion approximation. It compares with the cost of selection bound as follows,

$$\text{Fitness flux bound:} \qquad \frac{d}{dt} D(X^t) \leq \frac{2}{\ln 2} \int \psi \left(a_g + a^s_g\right) b^{-1}_{gg'}\, a^s_{g'}\, dx = 2N\langle \phi \rangle_t, \tag{S111}$$

$$\text{cost of selection bound:} \qquad \frac{d}{dt} D(X^t) \leq \frac{1}{2 \ln 2} \int \psi\, a^s_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx \qquad \leq N\langle C \rangle_t. \tag{S112}$$

Note that in Eq. (S111) we added the factor 2 which was missing in [15], as pointed out in [23].

Again, the fitness flux bound requires the neutral process to be at a stationary distribution with detailed balance, which means zero neutral flux $u_g = 0$ (see Eq. (S57)). (There is a more general flux theorem, which does not require detailed balance [15], but it does not provide a bound on $\Delta D(X^t)$.) The single locus, two allele system satisfies the detailed balance, since diffusion only takes place along a single dimension. However, detailed balance is rare in systems with multiple loci with recombination and general forms of mutation.

S10C Comparisons of the discrete and the diffusion formulas

The two bounds, computed using both the Markov chain and the diffusion formulas, are compared in Fig. S2BC for the single locus, two allele system. Note that the bounds are obtained by averaging the functions $2N_e\phi(x_A)$ and $N_e C(x_A)$, such as those plotted in Fig. S2A, with respect to the distribution $\psi^X(x_A)$.

In Fig. S2B, we use the Wright-Fisher model to compute the distribution over allele frequencies and the information increments. The population size is $N_{\text{WF}} = N_e = 100$, mutation strength $N\mu = 0.01$ and selection strength varies across columns. The cost of selection bound is in black, and the diffusion formula (full line, based on Eq. (S112)) is in an agreement with the discrete formula (dashed line, based on Eq. (S104)). The discrete version of the fitness flux bound is violated, as the Wright-Fisher model does not satisfy detailed balance (grey dashed, based on Eq. (S103)). This is because cycles such as $0 \to 1 \to 2 \to 0$ copies of the A allele take place more often than the reverse cycle, since two alleles can get lost by drift in a single generation with a high probability, but are unlikely to arise by mutation. However, detailed balance holds in the diffusion approximation. If fitness flux is computed according to the diffusion formula, the bound holds under weak selection when the diffusion approximation is close to the Wright-Fisher model, but fails when selection is stronger (full grey line, based on Eq. (S111)).

In Fig. S2C, we use the Moran model of the same system to compute the distributions and information increment over time. Note that in order to have the same magnitude of genetic drift as the Wright-Fisher model, the Moran model needs twice as large a census population size, $N_{\text{Moran}} = 2N_e$. Time is measured in generations ($N_{\text{Moran}} = 2N_e$ replacements each), and the increments in information are also computed per generation.

The discrete and diffusion formulas are again in agreement for the cost of selection bound. Note that we computed it as $N_e\langle C \rangle_t$ rather than $N_{\text{Moran}}\langle C \rangle_t$, to account for the additional stochasticity due to random deaths, even though $N_{\text{Moran}} = 2N_e$ parents are sampled with selection in each generation. The fitness flux bound now holds in its discrete version – the Moran model only allows allele frequency changes by $\pm 1/N_{\text{Moran}}$ and therefore satisfies detailed balance. If fitness flux is computed using the diffusion formula in Eq. (S111), it upper bounds the Moran model accumulation of information when selection is weak, but again fails when selection is strong and diffusion departs from the discrete model.

In conclusion, care is needed when applying the fitness flux bound to discrete models. In Main Text Fig. 4, we plot the information accumulation and the cost of selection bound based on the Wright-Fisher model, and the fitness flux bound based on the Moran model.

S10D Interpretation of the bounds under diffusion

The cost of control is non-negative and determined solely by the magnitude of the selection term a^s_g, with the inverse drift covariance $b^{-1}_{gg'}$ acting as the metric. As shown in Sec. S7 and Fig. S2A, this is proportional to the variance in fitness,

$$\frac{d}{dt} D(X^t) \leq N\langle C \rangle_t \approx \frac{N}{2 \ln 2}\langle V \rangle_t. \tag{S113}$$

In Fig. S2BC, this bound starts off large and reduces slightly as selection removes variation from the population. It remains positive at the stationary state, where it represents the cost of maintenance.

In contrast, the fitness flux bound in Eq. (S111) contains the sum $a^s_g + a_g$, corresponding to selection and mutation contributions to the fitness flux. As a result, the fitness flux bound can be written as a sum of a selection term and a mutation term,

$$\frac{d}{dt} D(X^t) \leq 2N\langle \phi \rangle_t = \underbrace{\frac{2N}{\ln 2}\langle V \rangle_t}_{\text{Selection term}} + \underbrace{\frac{2}{\ln 2} \int \psi\, a_g\, b^{-1}_{gg'}\, a^s_{g'}\, dx}_{\text{Mutation term}}. \tag{S114}$$

The selection term is proportional to the fitness variance – like the cost of selection bound, but with a 4 times higher numerical coefficient. When the selection term in Eq. (S114) dominates (a regime also discussed in [15]), the two bounds are proportional to each other, but the cost of selection bound is tighter. In general, the mutation term in Eq. (S114) causes the two bounds to behave in qualitatively different ways. Notably, the mutation term can be substantial and comparable to the selection term even when the parameters N, s, μ suggest that selection is strong and mutation is weak (i.e. under any or all of the conditions $Ns \gg 1$, $N\mu \ll 1$ and $s \gg \mu$). We illustrate and explain this in the following paragraphs.

Fig. S2A shows that the selection term in Eq. (S114) dominates especially at intermediate allele frequencies (around $x_A = 0.5$), where the fitness variance is high and mutation in opposing directions cancels out. Near $x_A = 0$, fitness variance is low and mutation towards the fitter allele A leads to a positive expected fitness flux. Near $x_A = 1$, mutation towards the deleterious allele a dominates and leads to a negative expected fitness flux (Fig. S2A).

This enables a more detailed understanding of the fitness flux bound in the scenario in Main Text Fig. 4 and Fig. S2C. The system is initialized at the neutral stationary distribution, which is symmetric. Therefore the mutation term in fitness flux vanishes, because the positive and negative contributions (at $x_A < 0.5$ and $x_A > 0.5$ respectively) cancel out. The bound is therefore proportional to the average fitness variance. Over time, as selection shifts the distribution towards higher x_A, mutation contributes more and more negatively, until the mutation and selection terms exactly cancel at stationarity. This happens regardless of N, μ and s. At stationarity, populations fluctuate around frequencies where mutation and selection are balanced, typically close to $x_A = 0$ or $x_A = 1$.

The only regime when the selection term dominates fitness flux for an extended period of time is when not only mutation is weak overall, but also the population is initialized at an intermediate frequency. This is shown in Fig. S3, which uses the same parameters as Fig. S2C, but the population is initialized at the frequency $x_A = 0.5$.

Note that only the process with selection can be initialized at $x_A = 0.5$. The neutral process must always be at the neutral stationary distribution to satisfy the assumptions of the fitness flux theorem. As a result, the information $D(X^t)$ is very high initially and decreases until drift spreads out the distribution ψ^{X^t} towards $x_A = 0$ and 1. Later, $D(X^t)$ slightly increases over a much longer (mutation-limited) time scale, see Fig. S3A. The early phase is useful for illustrating the behavior of the fitness flux and the cost of selection, but neither bound is very informative there, as information is being lost.

In Fig. S3B we plot the increments in information, the fitness flux bound and the cost of selection bound per generation. Both bounds are proportional to the relative fitness variance in the early phase at intermediate frequencies, albeit with different proportionality constants (full and dotted purple lines in Fig. S3B). In the later phase, when the population is mostly fixed for one of the alleles, the mutation term in Eq. (S114) becomes important and negative, and the fitness flux bound departs from the fitness variance approximation.

S11 Frequency dependent selection that maximizes fixation probability

In Main Text Fig. 3C,D, we compare the efficiency of selection (accumulated information per unit cost of selection) under constant selection, and under a specific form of frequency dependent selection. This frequency dependence is optimal in the sense that it maximizes the fixation probability of A at a given cumulative cost of selection. Below we describe the optimization procedure.

The calculation is done using a single locus, two allele Wright-Fisher model as described in Sec. S4, but instead of a single selection coefficient, we have a vector $\boldsymbol{s} = \left(s\left(0\right), s\left(\frac{1}{N}\right), s\left(\frac{2}{N}\right), \ldots, s\left(1\right)\right)^{\intercal}$ of selection coefficients $s(x_A)$ for each possible allele frequency x_A. Given $\boldsymbol{s}$, we can compute

- The (right stochastic) transition matrix $P(\boldsymbol{s})$ according to Eq. (S10), using the respective selection coefficient for each starting frequency (rows of $P(\boldsymbol{s})$).
- The vector $\boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s})$ of fixation probabilities for each possible starting frequency. It is given by the last column of the matrix power $P(\boldsymbol{s})^t$ at infinite time $t \to \infty$. Numerically, we keep doubling t, until the total probability that neither allele is fixed is less than 10^{-6} for all starting frequencies.
- The vector $\boldsymbol{c}(\boldsymbol{s})$ of cost of selection at each frequency, $\boldsymbol{c}(\boldsymbol{s}) = \left(C\left(0\right), C\left(\frac{1}{N}\right), C\left(\frac{2}{N}\right), \ldots, C\left(1\right)\right)^{\intercal}$. Note that $C(0) = C(1) = 0$ since there is no fitness variation when one of the alleles is fixed.
- The vector $\boldsymbol{\gamma}(\boldsymbol{s})$ of expected total cost of selection until either allele is fixed, for each starting frequency. The first and last elements are again equal to zero, since one of the alleles is fixed and no more cost is incurred. The remaining elements can be computed using the recurrence relation

$$\boldsymbol{\gamma}(\boldsymbol{s}) = \boldsymbol{c}(\boldsymbol{s}) + P(\boldsymbol{s})\boldsymbol{\gamma}(\boldsymbol{s}), \tag{S115}$$

  where $\boldsymbol{c}(\boldsymbol{s})$ is the immediate cost and $P(\boldsymbol{s})\boldsymbol{\gamma}(\boldsymbol{s})$ is the expected future cost. Eq. (S115) can be solved for $\boldsymbol{\gamma}(\boldsymbol{s})$ as a system of linear equations.
- The value vector $\boldsymbol{v}(\boldsymbol{s}) = \boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s}) - \lambda\boldsymbol{\gamma}(\boldsymbol{s})$, where λ is the Lagrange multiplier which quantifies the constraint on the total cost.

We look for $\boldsymbol{s}$ which maximizes value $\boldsymbol{v}(\boldsymbol{s})$. This is an instance of a Markov decision process [24] similar to the pursuit/first passage problem [25] with a small modification to include an unwanted absorbing state (loss of the A allele). The frequency-dependent selection $\boldsymbol{s}$ corresponds to the decision policy. The optimal policy does not depend on time or the initial state, and maximizes all elements of $\boldsymbol{v}(\boldsymbol{s})$ simultaneously [25].

We optimize $\boldsymbol{s}$ iteratively. It is initialized at all zeros, and we alternate between a value update and a policy (selection) update,

$$\text{Value update:} \qquad \boldsymbol{v} := \boldsymbol{v}(\boldsymbol{s}) = \boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s}) - \lambda \boldsymbol{\gamma}(\boldsymbol{s}) \tag{S116}$$

$$\text{Policy update:} \qquad \boldsymbol{s} := \text{argmax}_{\boldsymbol{s}} \left(P(\boldsymbol{s})\boldsymbol{v} - \lambda \boldsymbol{c}(\boldsymbol{s}) \right) \tag{S117}$$

The value update uses the current estimate of $\boldsymbol{s}$. The policy update uses the current estimate of $\boldsymbol{v}$ to compute the selection coefficient at each frequency, which maximizes the expected value in the next step, minus the immediate cost. The maximization is independent for each element of $\boldsymbol{s}$ and is done by binary searching for zero gradient ($\log_{10} s$ in range from 10^{-5} to 10^{2}, binary search depth 10).

To produce Fig. 3C,D, we vary the cost constraint λ and compute $\boldsymbol{s}$ by 60 value-policy update iterations. Examples of the frequency dependent $\boldsymbol{s}$ are shown in Fig. S4A. Notably, selection is strongest at low frequencies of x_A when A is at the greatest risk of being lost, but weak at higher frequencies to reduce costs. Fig. S4B,C show the fixation probability $\boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s})$ and the expected total cost $\boldsymbol{\gamma}(\boldsymbol{s})$ for each starting frequency, from which the Main Text Fig. 3C,D uses only the values for the initial frequency $1/N$.

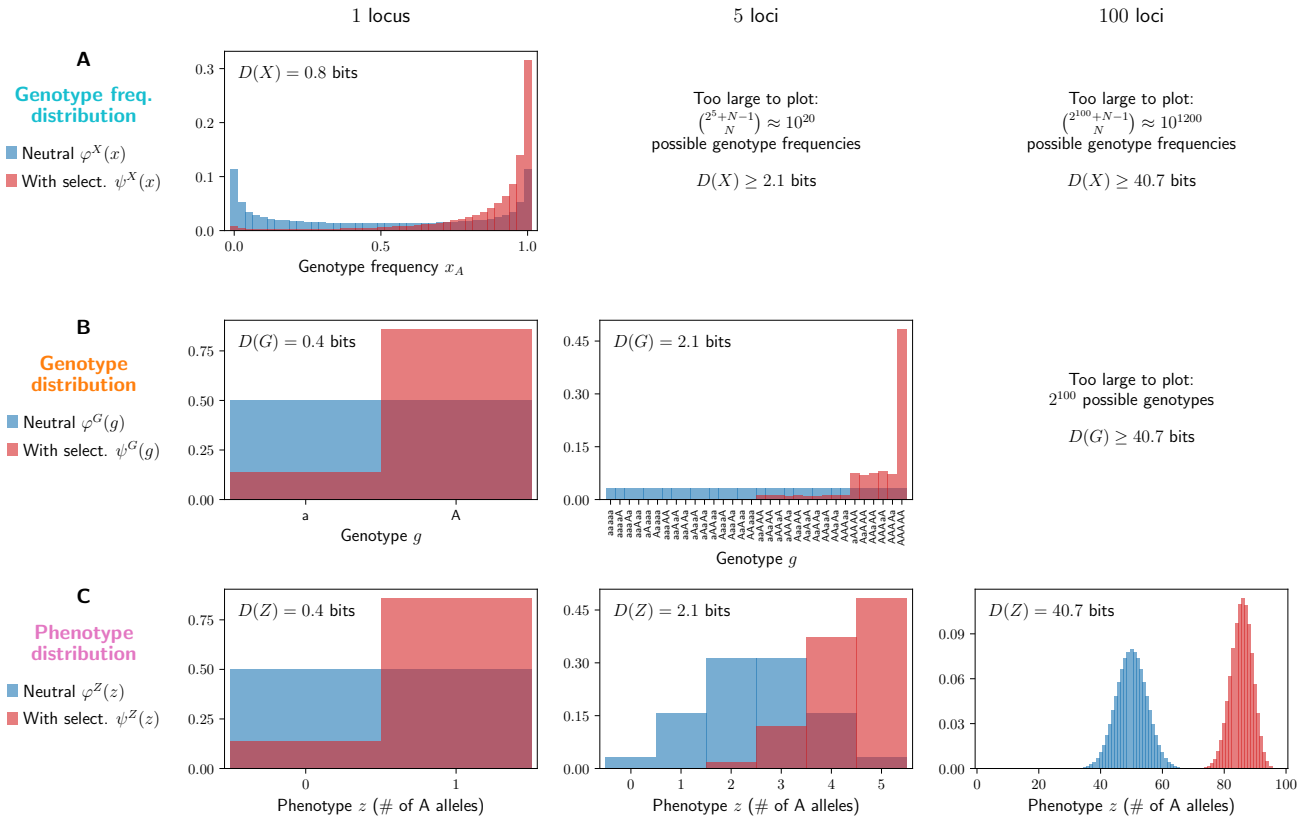

Figure S1: An example of the distributions over genotype frequencies, genotypes and phenotypes, with and without selection and for a varying number of loci, along with the corresponding measures of information. Based on a Wright-Fisher model with parameters $N = 40$, $\mu = 0.005$, $s = 0.05$ (as in Main Text Fig. 1). Fitness is multiplicative across loci, $w = (1 + s)^z$ where z is the number of beneficial A alleles carried – i.e. there is directional selection for an additive, fully heritable phenotype z (as in Main Text Fig. 6). All distributions are at the stationary state, computed using a transition matrix-based model (1 locus) or a long run (10^5 generations after 200 generations of burn-in) of an individual-based model (> 1 locus).
(A) Distributions over genotype frequencies ($\psi^X(x)$ with selection and $\varphi^X(x)$ without, red and blue) and the population-level information, $D(X)$. The distributions cannot be plotted and $D(X)$ cannot be directly estimated for 5 and 100 loci, due to the large number of possible genotype frequencies x, but we can still lower bound $D(X)$ by $D(G)$ or $D(Z)$.
(B) Distributions over genotypes ($\psi^G(g)$ with selection and $\varphi^G(g)$ without, red and blue) and the genotype-level information, $D(G)$. This information is less than $D(X)$, because selection not only gives preference to the fitter alleles, but also reduces the genetic variation within populations. The number of possible genotypes becomes too large for 100 loci, but we can still lower bound $D(G)$ by $D(Z)$.
(C) Distributions over the phenotype ($\psi^Z(z)$ with selection and $\varphi^Z(z)$ without, red and blue) and the phenotype-level information, $D(Z)$. The phenotype is simply the number of A alleles across all loci in an individual – an additive trait of varying polygenicity. In this example, selection favors the A allele at each locus, making fitness a function of z. In such cases, $D(G) \approx D(Z)$, since grouping genotypes into bins of equal z reduces the state space but preserves selective differences. In general, the trait might be unrelated to fitness or form only a component of it, leading to $D(G) > D(Z)$. Note that $D(Z)$ is approximately proportional to the number of loci, with each locus encoding about 0.4 bits. This is because loci evolve approximately independently, as there is zero epistasis, free recombination and little Hill-Robertson interference (see also Main Text Sec. 4B-C).

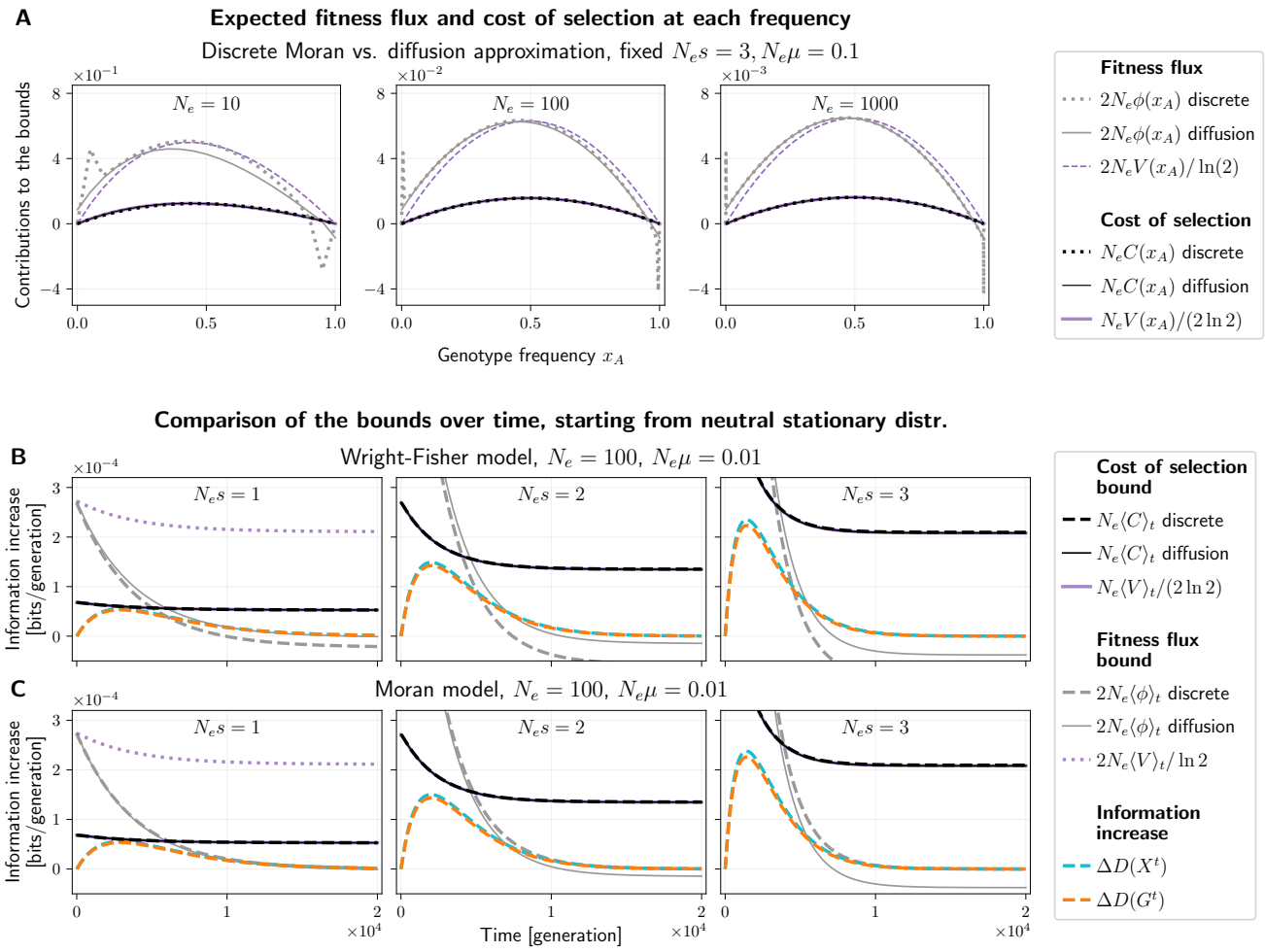

Figure S2: Comparisons between the fitness flux, the information-theoretic cost of selection and fitness variance, using the single locus, two allele systems under the Wright-Fisher model, the equivalent discrete-time Moran model and the diffusion approximation.
(A) The expected fitness flux per generation ($\phi(x_A)$, gray dotted) and the cost of selection ($C(x_A)$, black dotted) for each frequency x_A under the discrete Moran-model for three different effective population sizes N_e (left to right; census population size $N_{\text{Moran}} = 2N_e$ to account for the additional stochasticity in the Moran model compared to the Wright-Fisher model). We fixed $N_es = 3$ and $N_e\mu = 0.1$ to obtain models with a similar behavior but different time scales and granularity. Both $\phi(x_A)$ and $C(x_A)$ are multiplied by N_e and the same numerical factor as in the bounds on information accumulation rate (the bounds are obtained by averaging these values with respect to $\psi^X(x_A)$). The discrete formulas are compared with their diffusion approximation (full gray and black lines). As expected, the diffusion approximation is closer to the discrete model at higher N_e. The fitness flux $\phi(x_A)$ converges to the diffusion approximation non-uniformly across x_A, due to the spikes next to the domain boundaries. These spikes do not disappear but get "squeezed out" at high N_e. We also plot multiples of the relative fitness variance $V(x_A)$ (purple dashed and full lines), which approximate the fitness flux and the cost of selection. The cost $C(x_A)$ can be approximated very closely by $V(x_A)/(2\ln 2)$ as long as $s \ll 1$ (here, the largest value is $s = 0.3$ for $N_e = 10$). Fitness flux is the sum of a selection term proportional to $V(x_A)$ which is largest at intermediate x_A, and a mutation term which dominates near $x_A = 0$ or 1 and causes the discrepancy between $\phi(x_A)$ and $2V(x_A)/\ln 2$. Even when mutation rate is small compared to selection, $\mu \ll s$, mutation is important as the system approaches the stationary distribution concentrated near $x_A = 0$ and 1.
(B) The Wright-Fisher model uses the same parameters as Main Text Fig. 4, namely $N = N_e = 100$; $N_e\mu = 0.01$ and N_es, varied across columns. We plot the increase of information per generation (population level $\Delta D(X^t)$, blue dashed; genotype level $\Delta D(G^t)$, orange dashed), the upper bound in terms of cost of selection $N_e\langle C\rangle_t$, computed using the discrete formula (Eq. (S104), black dashed) and the diffusion formula (Eq. (S112), black full), and the upper bound in terms of the fitness flux computed using the discrete formula (Eq. (S103), gray dashed) and the diffusion formula (Eq. (S111), gray full). We also show the fitness variance approximations of the two bounds (Eq. (S113), purple solid and the fitness variance term in Eq. (S114), purple dotted, outside the plot range for $N_es = 2$ and 3). The discrete fitness flux bound is violated, since the Wright-Fisher model does not satisfy detailed balance under neutrality. The continuous fitness flux bound holds under weak selection, but fails when selection gets stronger, as differences grow between the discrete system and the diffusion approximation.
(C) The Moran model has the same effective parameters but double the census population size, $N_{\text{Moran}} = 2N_e$. The curve descriptions are the same as for the Wright-Fisher model, but note that each generation consists of $2N_e$ replacements, and the information increase as well as the upper bounds are rescaled accordingly. Also note that we plot the cost of selection bound $N_e\langle C\rangle_t$ using the effective population size N_e rather than the census population size $N_{\text{Moran}} = 2N_e$, which would lead to a twice as large (and less tight) bound. The Moran morel satisfies detailed balance under neutrality, and the discrete fitness flux bound holds under arbitrary selection. The continuous fitness flux bound again fails under strong selection.
Note that the diffusion formula for the fitness flux bound would correctly upper bound the accumulation of information in a system modeled fully using the diffusion approximation. These figures show that it does not always upper bound the accumulation of information in discrete models, especially not when selection is strong. The Main Text Fig. 4 uses the information accumulation curves and cost of selection bounds based on the Wright-Fisher model, and the fitness flux bound based on the Moran model.

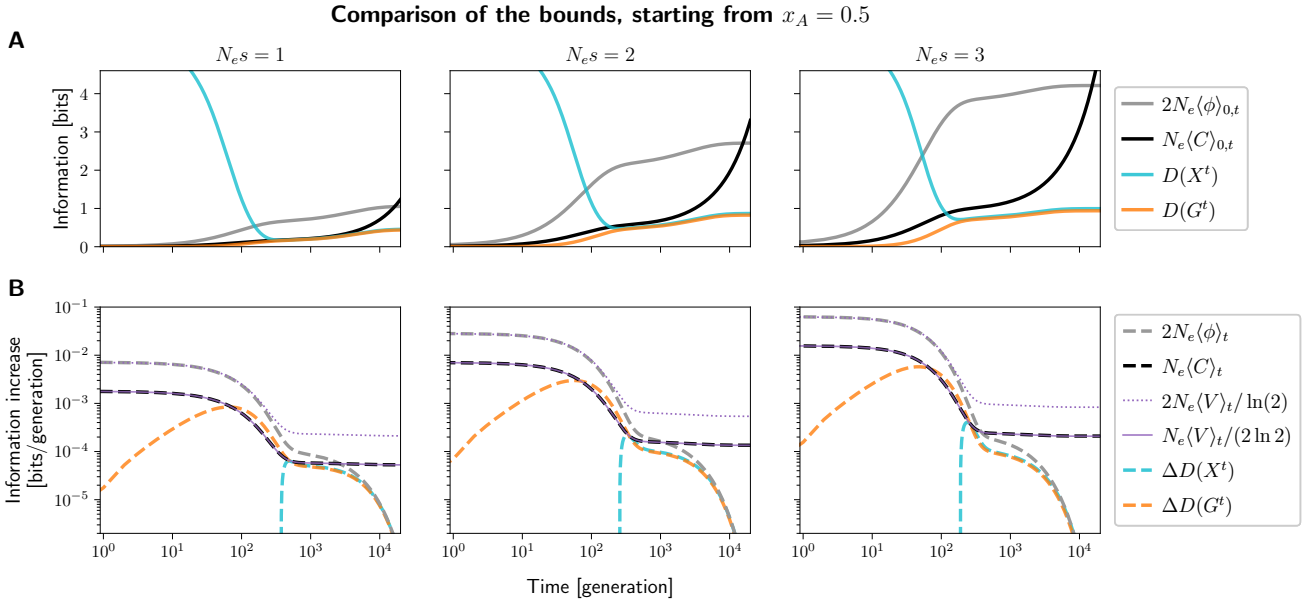

Figure S3: Demonstration of the fitness flux and the cost of selection bounds in the single locus, two allele system initialized at an intermediate frequency $x_A = 0.5$. Note that only the process with selection is initialized at $x_A = 0.5$; the neutral process must always be at the neutral stationary distribution to satisfy the assumptions behind the fitness flux theorem. This leads to a high initial value of $D(X^t)$ at $t = 0$. Calculated using the discrete Moran model with the same parameters as in Fig. S2C.
(A) The cumulative information at the population level ($D(X^t)$, blue) and genotype level ($D(G^t)$, orange), as well the cumulative fitness flux ($2N_e\langle\phi\rangle_{0,t}$, gray) and cumulative cost of selection ($N_e\langle C\rangle_{0,t}$, black). $D(X^t)$ starts out high due to the initial distribution being different under selection than under neutrality ($D(X^0) \approx 12.3$ bits, outside the plot range, in all three cases). At first, drift causes the distribution under selection to spread out towards extreme frequencies, $D(X^t)$ decreases on the time scale of $\sim N_e = 100$ generations. Meanwhile, $D(G^t)$ accumulates as the mean frequency x_A increases from 0.5 due to selection. The two measures of information eventually reach similar values, since mutation is low and the population is mostly fixed for one of the alleles. The cumulative fitness flux and cost of selection upper bound the cumulative information increase $D(X^t) - D(X^0)$, but are not very informative in this case, because despite selection acting, $D(X^t)$ started at a value that is higher than can be maintained and is lost rather than accumulated.
(B) The increase of information per generation on the population and genotype levels ($\Delta D(X^t)$, blue dashed; $\Delta D(G^t)$, orange dashed), and the upper bounds in terms of fitness flux ($2N_e\langle\phi\rangle_t$, gray dashed) and the cost of selection ($N_e\langle C\rangle_t$, black dashed). Note the log scales on both axes. Initially, $\Delta D(X^t)$ is negative and falls outside of the plot as drift spreads out the distribution over allele frequencies. Meanwhile, $\Delta D(G^t)$ is positive as the mean frequency x_A increases due to selection – this is associated with a positive fitness flux and cost of selection. After about $2N_e = 200$ generations, one of the alleles is likely to be fixed by drift, but the mean frequency continues to slowly increase at a mutation-limited rate (time scale $1/\mu = 10^4$ generations), which also causes modestly positive $\Delta D(X^t)$ and $\Delta D(G^t)$. In this phase, the cost of selection remains nearly constant, while fitness flux slowly decays, providing a fairly tight bound on $\Delta D(X^t)$. The cost of selection bound can be very well approximated with the relative fitness variance ($N_e\langle V\rangle_t/(2\ln 2)$, purple solid line). The fitness flux is proportional to the fitness variance in the first phase, when x_A is near 0.5 ($2N_e\langle V\rangle_t/\ln 2$, purple dotted line), but departs from it later as x_A tends to take values near 0 or 1 where mutation is important (see Fig. S2A). We note that similar behavior can also observed for different values of $N_e\mu, N_e s$.

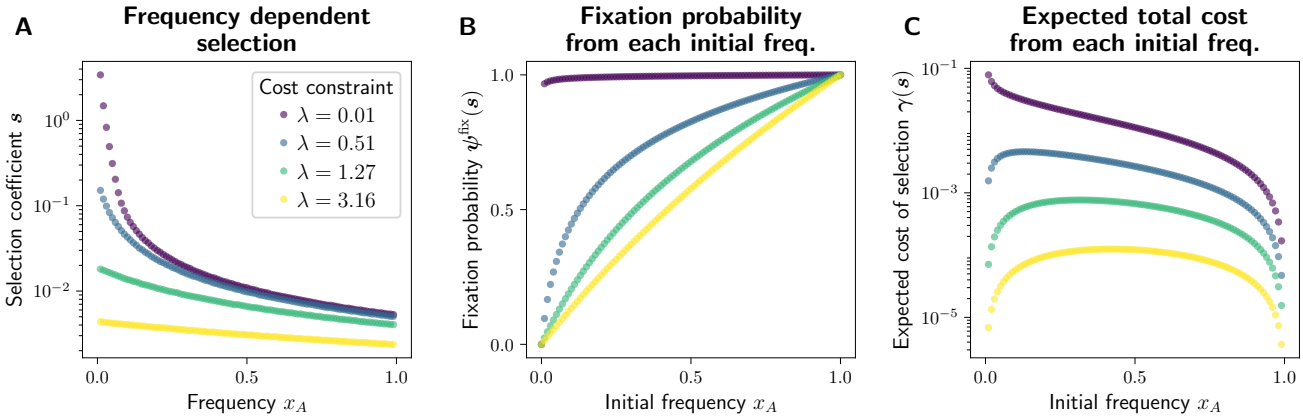

Figure S4: Frequency dependent selection that optimizes fixation probability of an allele at a constrained total cost.
(A) The frequency dependent selection coefficient $\boldsymbol{s}$, computed as described in Sec. S11, under various cost constraints λ. When this constraint is greater, selection is weaker overall.
(B) Fixation probability $\boldsymbol{\psi}^{\text{fix}}(\boldsymbol{s})$ of the allele A for each starting frequency x_A. The fixation probability is close to the frequency itself under weak selection (large λ), and higher when selection is overall stronger.
(C) The expected total cost of selection, $\boldsymbol{\gamma}(\boldsymbol{s})$ associated with trajectories starting from each initial frequency x_A. It is low for high frequencies, where the allele A is expected to be fixed soon, and for low frequencies under weak selection, where it is expected to be lost soon.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second. John Wiley & Sons, 2006, ISBN: 978-0-471-24195-9.

[2] R. P. Worden, "A speed limit for evolution," *Journal of Theoretical Biology*, vol. 176, no. 1, pp. 137–152, 1995.

[3] E. Todorov, "Linearly-solvable Markov decision problems," in *Advances in Neural Information Processing Systems*, vol. 19, 2006, pp. 1369–1376.

[4] E. A. Theodorou, "Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations," *Entropy*, vol. 17, no. 5, pp. 3352–3375, 2015.

[5] G. Kesidis and J. Walrand, "Relative entropy between Markov transition rate matrices," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 1056–1057, 1993.

[6] Y. Iwasa, "Free fitness that always increases in evolution," *Journal of Theoretical Biology*, vol. 135, no. 3, pp. 265–281, 1988.

[7] H. Hasegawa, "Thermodynamic properties of non-equilibrium states subject to Fokker-Planck equations," *Progress of Theoretical Physics*, vol. 57, no. 5, pp. 1523–1537, 1977.

[8] C. S. Withers and S. Nadarajah, "The spectral decomposition and inverse of multinomial and negative multinomial covariances," *Brazilian Journal of Probability and Statistics*, vol. 28, no. 3, pp. 376–380, 2014.

[9] J. Berg, S. Willmann, and M. Lässig, "Adaptive evolution of transcription factor binding sites," *BMC Evolutionary Biology*, vol. 4, no. 1, p. 42, 2004.

[10] G. Sella and A. E. Hirsh, "The application of statistical physics to evolutionary biology," *Proceedings of the National Academy of Sciences*, vol. 102, no. 27, pp. 9541–9546, 2005.

[11] S. Wright, "The distribution of gene frequencies in populations," *Proceedings of the National Academy of Sciences*, vol. 23, no. 6, pp. 307–320, 1937.

[12] H. P. de Vladar and N. H. Barton, "The contribution of statistical physics to evolutionary biology," *Trends in Ecology & Evolution*, vol. 26, no. 8, pp. 424–432, 2011.

[13] R. Rao and S. Leibler, "Evolutionary dynamics, evolutionary forces, and robustness: A nonequilibrium statistical mechanics perspective," *Proceedings of the National Academy of Sciences*, vol. 119, no. 13, e2112083119, 2022.

[14] J. C. Baez and B. S. Pollard, "Relative entropy in biological systems," *Entropy*, vol. 18, no. 2, p. 46, 2016.

[15] V. Mustonen and M. Lässig, "Fitness flux and ubiquity of adaptive evolution," *Proceedings of the National Academy of Sciences*, vol. 107, no. 9, pp. 4248–4253, 2010.

[16] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, vol. 217, no. 5129, pp. 624–626, 1968.

[17] D. Graur, "An upper limit on the functional fraction of the human genome," *Genome Biology and Evolution*, vol. 9, no. 7, pp. 1880–1885, 2017.

[18] W. J. Ewens, "Remarks on the substitutional load," *Theoretical Population Biology*, vol. 1, no. 2, pp. 129–139, 1970.

[19] B. Galeota-Sprung, P. Sniegowski, and W. Ewens, "Mutational load and the functional fraction of the human genome," *Genome Biology and Evolution*, vol. 12, no. 4, pp. 273–281, 2020.

[20] M. M. Desai and D. S. Fisher, "Beneficial mutation–selection balance and the effect of linkage on positive selection," *Genetics*, vol. 176, no. 3, pp. 1759–1798, 2007.

[21] E. E. Shnol, E. A. Ermakova, and A. S. Kondrashov, "On the relationship between the load and the variance of relative fitness," *Biology Direct*, vol. 6, no. 1, p. 20, 2011.

[22] G. E. Crooks, "Path-ensemble averages in systems driven far from equilibrium," *Physical Review E*, vol. 61, no. 3, pp. 2361–2366, 2000.

[23] N. H. Barton, "How does epistasis influence the response to selection?" *Heredity*, vol. 118, no. 1, pp. 96–109, 2017.

[24] R. Bellman, "A Markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[25] J. H. Eaton and L. A. Zadeh, "Optimal pursuit strategies in discrete-state probabilistic systems," *Journal of Basic Engineering*, vol. 84, no. 1, pp. 23–29, 1962.