The evolutionary origin of the universal distribution of fitness effect

Ayuna Barlukova, Gabriele Pedruzzi, and Igor M. Rouzine*

Sorbonne Université, Institute de Biologie Paris-Seine

Laboratoire de Biologie Computationnelle et Quantitative, LCQB, F-75004 Paris, France

*Correspondence to: igor.rouzine@sorbonne-universite.fr

Abstract

An intriguing fact long defying explanation is the observation of a universal exponential distribution of beneficial mutations in fitness effect for different microorganisms. Here we use a general and straightforward analytic model to demonstrate that, regardless of the inherent distribution of mutation fitness effect across genomic sites, an observed exponential distribution of fitness effects emerges naturally, as a consequence of the evolutionary process. Using this result, we develop a technique to measure the mutation fitness effects for specific genomic sites from a single-time sequence set and apply it to influenza A H1N1 hemagglutinin protein. Our results demonstrate the difference between the distribution of fitness effects experimentally observed for naturally occurring mutations and the inherent distribution obtained in directed-mutagenesis experiments. The technique will enable researchers to measure fitness effects of mutations across the genome from a single DNA sample, which is important for predicting the evolution of a population.

Introduction

Evolutionary dynamics of a population of nucleic acid sequences is controlled by several acting forces, including random mutation, natural selection, genetic drift, and linkage decreased by recombination. Of central interest is the adaptation of an organism to a new environment, which occurs due to fixation in a population of rare mutations that confer a benefit to the fitness of the organism (Imhof and Schlotterer 2001; Kassen and Bataillon 2006; Acevedo, et al. 2014; Stern, et

al. 2014; Wrenbeck, et al. 2017). The advantage of each favorable mutation is measured by the relative change it causes in genome fitness (average progeny number). Thus, the knowledge of fitness effects for different mutations is essential for predicting the evolutionary trajectory of a population, such as occurs, for example, during the development of resistance of a pathogen to treatment or the immune response.

Recent advancements in theoretical population genetics provide accurate and general expressions for the speed of adaptation of an asexual population, its genetic diversity, mutation fixation probability, and phylogenetic properties within the framework of the traveling wave theory (Tsimring, et al. 1996; Rouzine, et al. 2003; Rouzine and Coffin 2005; Desai and Fisher 2007; Rouzine and Coffin 2007; Brunet, et al. 2008; Rouzine, et al. 2008; Neher, et al. 2010; Rouzine and Coffin 2010; Hallatschek 2011; Good, et al. 2012; Walczak, et al. 2012; Neher and Hallatschek 2013). In all these models, the distribution of fitness effects among mutation sites DFE serves as an important input parameter.

The average-over-genome fitness effect of a beneficial mutation in HIV genome was estimated using genetic samples from HIV infected patients (Rouzine and Coffin 1999). Finding out the distribution of the fitness effect over sites (DFE) over genomic sites in several viruses and bacteria required specially designed and rather elaborate experiments (Imhof and Schlotterer 2001; Kassen and Bataillon 2006; Acevedo, et al. 2014; Stern, et al. 2014; Wrenbeck, et al. 2017). Recently, selection coefficients across the sites of the hemagglutinin gene of human influenza A/H3N2 were estimated by fitting the deterministic one-locus model and its approximate extension for two-loci (Illingworth and Mustonen 2012). The authors fit the model to time-series data on allele frequencies of hemagglutinin (HA) gene of human influenza A H3N2. (Keightley and Eyre-Walker 2007) proposed a method of DFE estimation in mutation-selection-drift equilibrium based on the assumption that DFE has the shape of the gamma distribution. They estimate parameters of gamma distribution from maximization of the likelihood under the assumption that the derived sites are binomially distributed. Thus, the two modeling papers used strong assumptions about the dynamics of the system.

(Eyre-Walker and Keightley 2007) reviewed different types of experiments to estimate DFE. They noted that there is a lack of understanding in DFE. In particular, during evolution, DFE is likely to change in time, which is not in agreement with the assumption of Gillespie-Orr theory of constant DFE (Gillespie 1982; Orr 2003).

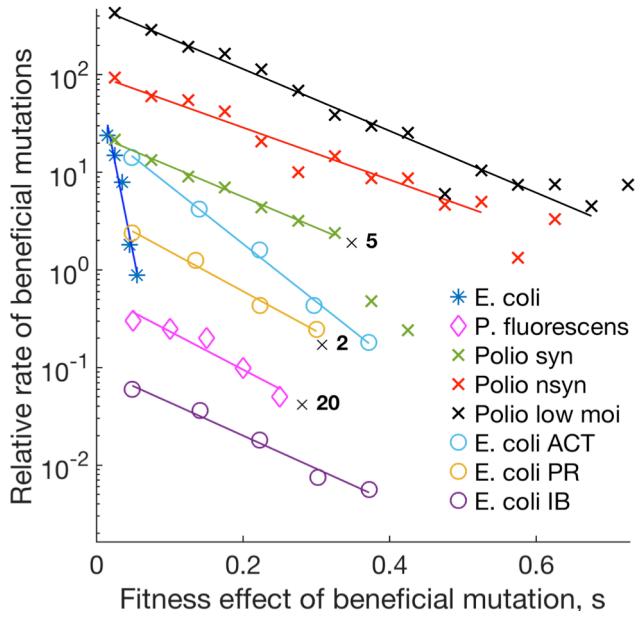

Fig. 1: Different studies on distribution of fitness effects of beneficial mutations demonstrate an exponential form. Y-axis: Frequency of beneficial alleles (arbitrary units). X-axis: Mutation gain in fitness due to a beneficial mutation (selection coefficient). Symbols represent results obtained for different sites of the genome in experiments on *Escherichia coli* (Imhof and Schlotterer 2001), *Pseudomonas fluorescens* (Kassen and Bataillon 2006), poliovirus synonymous mutations, poliovirus non-synonymous mutations (Acevedo, et al. 2014), poliovirus low MOI (Stern, et al. 2014), *E. coli* acetamide (ACT), propionamide (PR), and isobutyramide (IB) (Wrenbeck, et al. 2017).

To fill this gap of knowledge, we use a different approach based on mechanistic description. We study the dynamics of a system in the state of adaptation (non-equilibrium). Based on our results, we propose a more general method of measuring selection coefficients for specific sites not restricted to the one-site model approximation. In fact, this approximation usually does not work for highly diverse and rapidly changing RNA viruses. The reason is linkage of many evolving loci causing clonal interference and genetic hitchhiking effects complicated by stochastic effects in finite population. Linkage effects greatly modify the speed of evolution and other parameters. Our method takes these effects into consideration automatically by taking advantage of the prediction of a narrow solitary wave in fitness space (Rouzine, et al. 2003; Rouzine and Coffin 2005, 2010; Good, et al. 2012).

The key to the method is revealed by the intriguing fact long defying explanation: the frequent occurrence of a universal form of DFE of beneficial mutations. Previous studies in *E. coli, Pseudomonas aeruginosa, Pseudomonas fluorescence*, poliovirus show that the rate of beneficial mutation often decreases with their fitness effect exponentially (Fig. 1) (Imhof and Schlotterer 2001; Kassen and Bataillon 2006; Acevedo, et al. 2014; Stern, et al. 2014; Wrenbeck, et al. 2017).

In the present work, we offer a simple interpretation of this phenomenon. We demonstrate that, regardless of the initial distribution of fitness effects across genomic sites, an exponential DFE emerges naturally, as a consequence of an evolutionary process. Further, we apply these findings to a nucleic acid sequence set of HA protein of influenza virus A strain H1N1 to obtain the relative value of the selection coefficient for each variable, non-synonymous site in the protein.

Studying the existing literature on DFE, we found out that two different distributions, which are described below, were both referred as DFE. We note that there is an inherent constant distribution of selection coefficients of a genome, which represents the number of genome sites in the given small interval of the values of selection coefficient. This distribution can be measured directly only by a site-directed mutagenesis experiment. We will refer to the first as "intrinsic DFE" to emphasize the fact that it is the property of the pathogen/environment and does not depend on the state of population. Another distribution is the distribution of new beneficial mutations arising naturally, which depends on the state of adapting population. In our work, we will use term "DFE" to denote the second distribution, which is the relative rate of mutations naturally occurring during experimental evolution (see results shown in Fig. 1). We show, in our work, by mathematical analysis of a simple and general population model and by direct comparison with data on influenza A, that these two distributions turn out to be quite different from each other. We will focus on beneficial mutations.

Results

In order to explain the exponential shape of DFE observed in the experiments, we start by noting that beneficial mutations can emerge only at the sites currently occupied with less-fit alleles. Here we assume bi-allelic approximation, when two alleles are considered: the best-fit and the next less-fit. Although each position, in principle, can have four nucleotides A, C, T, G, in real viral data, on moderate time scales 1-10,000 generations, most variable sites display only two alleles in a sample. In this case, if a genomic site is occupied by the less-fit allele, it can become only the best-fit by mutation, and, vice versa, a genomic site occupied by the best-fit allele, it can only lose in fitness. If a population is well adapted during the process of evolution, most of genome sites, in each genome, already carry best-fit alleles and cannot experience beneficial mutations.

Therefore, the observed DFE will be affected by the occupation number distribution of less-fit alleles among sites with different s, i.e., by the state of population.

Let denote the average frequency of less-fit alleles at a site with fitness effect s by $f(s)$. We note that $f(s)$ can also be viewed as the frequency of sites available for beneficial mutations. For example, consider a sequence of the form 1000001, where 1 stands for the less-fit allele and 0 for the best-fit allele. Then, only the first and the last positions in the sequence are the sites, where a beneficial mutation can occur, $1 \rightarrow 0$. Thus, the rate of beneficial mutation at any fixed position of the genome must be proportional to the frequency of less-fit allele f at this position. If the system is fully adapted, we have $f = 0$, and no beneficial mutations are possible.

Experiment description. The experiments, shown in Fig 1, consider naturally occurring evolution and count beneficial and deleterious mutations emerging in an adapting population. The authors evolve a population of bacteria or virus for a short time in culture. Newly emerging beneficial mutations result in spontaneous increase in the best-fit allele frequency in time (selection sweeps). Although exact protocols differ, the count occurs for naturally occurring mutations, not for random mutagenesis. In one experiment (Acevedo, et al. 2014), an experimentalist uses a deep sequencing technique CirSeq to monitor the arising frequency of minority alleles at each genomic site as a function of time and fits it with a simple one-site evolution model expression to estimate s for each site. In another, the experimentalist is focused on beneficial mutations in *E.Coli.* (Imhof and Schlotterer 2001), he measures selection coefficient s for each selection sweep from time series, and then count the number of sweeps at sites belonging to an interval of the selection coefficient (X-axis in Fig. 1). Therefore, all these experiments measure the naturally occurring mutation density, DFE, and not intrinsic DFE.

In the last experiment, a beneficial mutation event occurs spontaneously, with a small probability, at a rare less-fit site. If it survives random drift, it gets fixed in the population. We can present the results of these experiments on beneficial mutations (Y axis, Fig 1) as the product $DFE(s)\pi(s)$, where the observed DFE is given by

$$DFE(s) = f(s)g(s), \tag{1}$$

$f(s)$ is the frequency of target sites available for beneficial mutations, $\pi(s)$ is fixation probability of beneficial mutation s, and $g(s)ds$ is the number of sites with the selection coefficient in interval $[s, s+ds]$. Therefore, we conclude that the raw distribution of selection coefficient across different sites, intrinsic DFE $g(s)$, is not the same as the observable distribution DFE(s), given by Eq. 1. The first distribution is a property of the virus and the cell type and is fixed. The DFE(s) observed in the experiments (Fig. 1) depends on the state of the population and evolves in time, since $f(s)$ evolves in time, which explains the aforementioned observation in (Eyre-Walker and Keightley 2007). DFE(s) given by Eq. 1 serves as the input density parameter for the models of evolution (Good, et al. 2012) .

Below mutant frequency $f(s)$ is assumed to have pre-evolved before the experiment for a long time, reflecting pre-history of the population under similar conditions, but is not in mutation-selection drift equilibrium yet, i.e., it is not best adapted yet to the conditions of the experiment. We will describe this pre-evolution of $f(s)$ by simulations and analytically. After predicting the form of $f(s)$, we will use it to estimate intrinsic distribution of $g(s)$ from data. We will show that $f(s)$ depends sharply (exponentially) on s, while both experimental dependence $g(s)$ and fixation probability $\pi(s)$ depend on s relatively weakly. Therefore, the exponential dependence in mutant frequency $f(s)$ dominates experimentally measured DFE(s) (see Eq. 1), which explains the results shown in Fig. 1.

Model. We consider an asexual organism, which evolved for some time but is still far from the mutation-selection equilibrium before the experiment. A haploid population has N binary sequences, where each genome site (nucleotide position) numbered by i =1, 2, …, L carries one of two possible genetic variants (alleles), denoted K_i =0 or K_i =1. Each site (nucleotide position) has one of two alleles: the better-fit (for example, A), or the less-fit (for example, G). We note that beneficial mutations are rare, which is why it unlikely that two occur at the same nucleotide. We focus here on the short-term adaptation to a new constant environment, where the bi-allelic model is a fair approximation.

The genome is assumed to be very long, $L >> 1$. Time is discrete and measured in units of population generations. The evolution of the population is described by a standard Wright-Fisher model, which includes the factors of random mutation with genomic rate μL, natural selection, and

random genetic drift. Recombination is assumed to be absent. Once per generation, each individual genome is replaced by a random number of its progeny which obeys multinomial distribution. The total population stays constant with the use of the broken-stick algorithm. To include natural selection, the average progeny number (Darwinian fitness) of sequence K_i is set to e^W. We consider the simplest case when the fitness effects of mutations, s_i, are additive over sites:

$$W = \sum_{i=1}^{L} s_i K_i$$

The reference genome, $\{K_i=0\}$, can be chosen in arbitrary way. For our aim, it is convenient to set it to be the same as the best fit sequence, so that all selection coefficients s_i are negative. Each site i with deleterious allele, $K_i=1$, is a target site for a possible beneficial mutation. Vice versa, a site with the favorable allele, $K_i=0$, can have a deleterious mutation. A more general version of fitness model that accounts for pairwise epistatic interactions is considered in (Pedruzzi, et al. 2018) and, for macroscopic epistasis, in (Good and Desai 2015). Here we focus on additive contributions of single sites to the fitness landscape. We note that most sites usually do not have epistatic partners, so the approximation is fair.

The fitness cost of a deleterious allele s is distributed in a complex way among genomic sites. In general, the inherent distribution $g(s)$ is unknown and depends on a virus, host cell type, and a protein. Its measurement requires an experiment with site-directed mutagenesis along the entire genome. The genome has to be mutated artificially, site by site, and then the value of s is measured for each mutation. Below we make no assumptions regarding $g(s)$ and demonstrate that the exponential shape in the less-fit allele frequency $f(s)$ arises automatically and independently on the form of $g(s)$. Later on, we will show how $g(s)$ can be calculated from sequence data for the influenza virus and demonstrate that it is an unremarkably slow function within an interval of s.

Our work applies only far from mutation selection equilibrium when system is still adapting. It is well known that, in equilibrium, the dependence $f(s)$ is not exponential, but close to

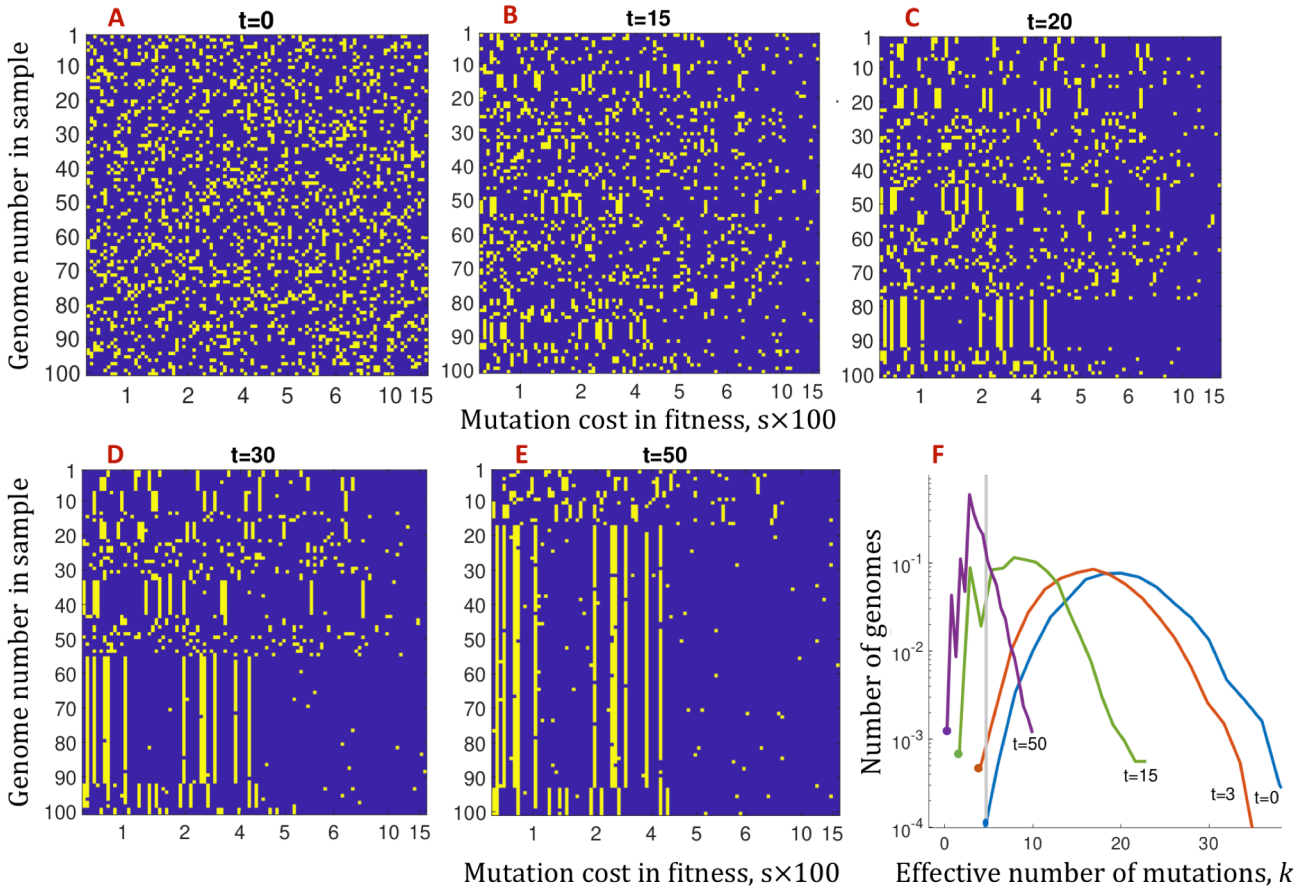

Fig. 2: Deleterious alleles with higher values of fitness cost, *s*, are the first to be depleted during the process of adaptation. In other words, beneficial mutations with higher s are fixed first. (a-e) Evolution of a sample of 10^2 sequences. Violet dots: better-fit alleles, yellow dots: less-fit alleles. X-axis: the cost in fitness, s, multiplied by 100. The values of s are randomly distributed with the half-Gaussian distribution, $s > 0$, with the average $s_{av} = 0.05$. Genomic sites are ordered by the value of s. Y-axis: genome number in the sample. The initial population is randomized with the average frequency of deleterious alleles $f_{in} = 0.2$. Time points in generations are shown. (f) Evolution of the genome distribution in fitness. X-axis: the effective number of deleterious alleles, defined as $k = -W/s_{av}$. where W is fitness. Different colors show discrete time intervals from 0 to 5. Vertical grey line shows the best fitness class of genomes at t=0. The emergence of clonal structure in (a-e) coincides with the transition from the selection of pre-existing sequences to the traveling wave regime. Parameters: $f_{in} = 0.2$, $N = 10^4$, $L = 100$, $s_{av} = 0.05$, $\mu L = 0.05$.

$f = \mu/s$. For example, computer simulations in [Keightly NRG] show that the DFE evolves away from an exponential distribution when approaching equilibrium.

Monte-Carlo simulation. We start from an initial population of N genomes that has a fraction of deleterious alleles randomly distributed among genomic sites (Fig 2a). Evolution of a sample of hundred sequences in a representative Monte-Carlo run is shown in Fig. 2. For the sake of visual convenience, we have re-ordered genomic sites in the ascending order of the value of selection coefficient s_i.

In the process of evolution, we observe increasing redistribution of deleterious alleles among genomic sites as follows (Fig. 2). The sites with a relatively high mutation cost loose deleterious alleles due to natural selection. The asymmetry becomes evident from t = 20. Finally, at t = 50 (Fig 2e), mutations on the right side are almost absent. Thus, deleterious alleles with higher values of mutation cost vanish earlier, which represents a qualitative explanation of the observed exponential dependence of DFE on s (Fig 1)**.**

We note that in our example, we set a rather large value of initial f, which is convenient for numerical computations. In real life, mutant frequency f may be much smaller than the value we choose. However, our results do not depend on this initial condition assumption. Later, we provide our analytic derivation which is general and applies to very low f, as long as they are not in mutation selection equilibrium.

In addition to the observed re-distribution of less fit alleles, we also observe the emergence of group of identical sequences, which we explain by evolutionary process as follows. In Fig 2, two intervals of adaptation can be discerned. Early on, new mutations can be neglected, and the critical evolutionary factor is the natural selection of pre-existing genomes (Fig 2a, b). It was previously revealed by a combination of modeling and experimental evolution of vesicular stomatitis virus (Dutta, et al. 2008). In time interval, $t << 1/s_{av}$, where s_{av} is the average of $g(s)$, the distribution of alleles over genomes remains random.

In contrast, in the second time interval, which starts around $t \sim 1/s_{av}$, new beneficial mutations become crucial for further evolution, because they give birth to new highest-fit genomes (Fig 2b-e). To explain the formation and subsequent growth of groups of identical sequences (Fig 2b-e), we address to traveling wave theory of evolution (Fig 2f).

Formation of these clones occurs at the edge of the traveling wave of fitness distribution (Rouzine, et al. 2003; Desai and Fisher 2007; Hallatschek 2011; Good, et al. 2012) (Fig 2f). The fitness distribution moves in time towards higher values of fitness, i.e., smaller numbers of

deleterious alleles. At early times, the distribution is broad and symmetric. In this regime, as was mentioned earlier, the main force is the selection of preexisting genomes. After a while ($t \sim 1/s_{\text{av}}$), the profile becomes asymmetric, and the high-fitness edge starts to move to the left together with the peak due to new beneficial mutations (Fig 2f). The genomes, appearing on the left side from the initial high-fitness edge (grey line in Fig 2f) share the initial genetic background. Hence, they produce observed groups of sequences identical at most sites (yellow vertical lines, Fig 2b). As the wave progresses, the clonal structure grows, and eventually, most genomes in the population become an offspring of the same ancestor (Fig 2f).

Analytic derivation of universal DFE.

Short times. As the above simulation shows, the evolution of genomes occurring at short times $t << 1/s_{\text{av}}$ is mainly due to the selection of preexisting variation and new mutations are not important (*Methods*). The probability of having a deleterious allele at a site with mutation cost s at time t has the form

$$f(s,t) = \frac{f_{in}}{(1-f_{in})\, e^{ts} + f_{in}} \tag{2}$$

where f_{in} is the initial mutant frequency. The slope of the distribution of deleterious alleles is defined as

$$\beta = -\frac{\partial \log(f)}{\partial s} \tag{3}$$

We observe that the formula in Eq 2 does not depend on the initial distribution of selection coefficients among sites, $g(s)$. At a small initial mutant frequency f_{in}, the formula can be

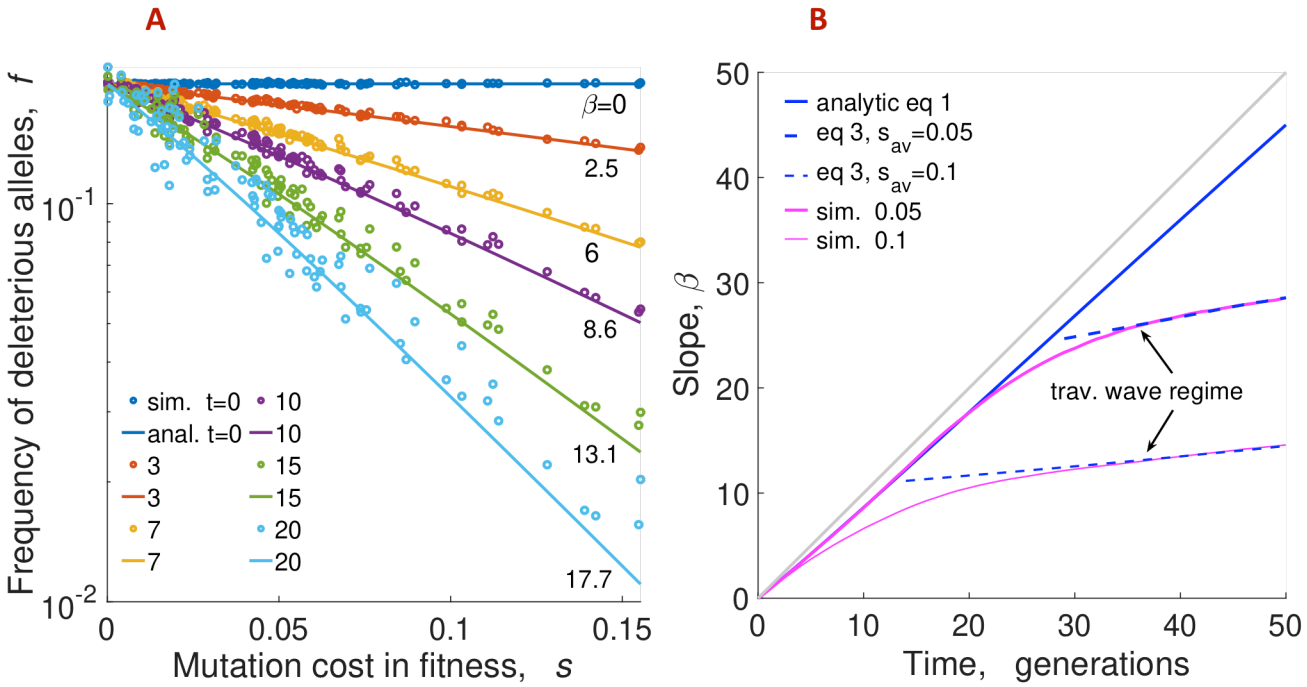

Fig. 3: The frequency of deleterious alleles decays exponentially with their fitness effect, with the slope increasing in time. (A) Analytic prediction (Eqs. 1) for the frequency of deleterious alleles agrees with Monte-Carlo simulation. X-axis: Mutation cost of deleterious allele at a genomic site, *s*. Y-axis: Frequency of deleterious alleles at such a site, $f(s)$. The mutant frequency f is averaged over 20 random simulation runs. Different colors show different times, symbols are simulation, and lines are analytic prediction (Eq. 1). The numbers on the curves are the values of the slope. Parameters as in Fig 2. (B) The slope of the distribution of deleterious alleles β, analytic (blue lines) and simulation (purple lines), as a function of time, t. Parameters f_{in} and s_{av} different from those in Fig 2 are shown on the legend. The log-slope for the simulated curves of mutant frequency in (A) is obtained by an exponential fit. We observe that the deviation of the simulated slope from the analytic prediction Eq. 1 at long times coincides with the establishment of the traveling regime, which occurs later for smaller s_{av} (Fig. 2f). At long times, the traveling wave prediction Eq. 3 applies (dashed blue lines). Grey diagonal shows $\beta(t) = t$.

approximated with an exponential, $f(s,t) \approx \exp(-ts)$. The exponential slope is approximately equal to time, $\beta = t$ (see Fig. 3). This is an early regime where the evolution of different sites is effectively independent.

Long term. At longer times $t > 1/s_{\text{av}}$, beneficial mutations become essential, and the above approximation does not apply anymore. We need to use the results of the traveling wave theory (Rouzine, et al. 2003; Desai and Fisher 2007; Hallatschek 2011; Good, et al. 2012). In the stationary regime of traveling wave (Fig 2f), fixation of beneficial alleles is the process that

dominates the loss of deleterious alleles. Let t_0 be the characteristic time when the traveling wave regime starts. In *Supplementary Methods*, we solve a dynamic equation for allelic frequency f and obtain

$$f(s,t) = f_{in}\, e^{-t_0 s - \mu N \int_{t_0}^{t} \pi(s)dt} \quad (4)$$

where $\pi(s)$ is the probability of fixation of a beneficial mutation with fitness gain s derived previously (Good, et al. 2012). By expanding the argument of the exponential in Eq. 4 in s, the slope takes the form

$$\beta(t) = t_0 + \mu N \pi'(0)(t - t_0) \quad (5)$$

Please note that Eq. 4 for adaptation regime neglects deleterious mutation events and is valid far from equilibrium.

Previously, a more general argument was used to predict the exponential shape of the DFE (Pedruzzi, et al. 2018). We assumed that mutant frequency $f(s)$ has evolved for some time before the experiment measuring DFE, but that the population was not in equilibrium yet, so that deleterious mutation events (reverse mutations) are negligible. Under these condition, the system is in quasi equilibrium, where all the variables change slowly adjusting to the slow change of the average fitness in time. Hence, given the fitness distribution of genomes, the distribution of alleles over sites and genomes is given by the condition that the entropy is in the maximum. However, the full equilibrium does not occur until much later on scaled much larger than 1/<s>.

Further, the fitness distribution is narrow, as follows from traveling wave theory, $\Delta W << W$. Therefore, the system entropy is at the conditional maximum S restricted by the average value of fitness -W. From these assumptions, we obtained the probability to have a deleterious allele at a given site

$$f = (1 - f)e^{-\beta s} \quad (6)$$

where $\beta = \frac{\partial S}{\partial W}$. Eq. 6 is general, while Eqs. 2 and 4 provide explicit expressions for β.

Thus, Eqs. 2 to 4 demonstrate that the exponential dependence $f(s)$ arises in the course of evolution at any initial conditions after the evolution time $t \sim 1/s_{\text{av}}$, and that the resulting exponential slope is robust to the initial conditions. The pre-factor at the exponential depends on the time of pre-evolution and f_{in}. We assume that the system evolved for time longer than the inverse average $<s_i>$ but is not in equilibrium yet.

Later, we propose the procedure of estimation of this pre-factor directly from data.

Monte-Carlo simulation confirms theory. To test our analytic theory, we compare the frequency of deleterious alleles obtained by analytic prediction (Eq. 2) *f (s)* with the results of Monte-Carlo simulation averaged over 20 random runs at several time points (Fig. 3a). At $t = 0$, simulated and predicted mutant frequencies are constant, since all sites have the same probability of deleterious allele, f_{in}. Thus, the slope β is equal to 0 (blue line). At later times, we observe that the slope increases gradually in time and the frequency of deleterious alleles *f(s,t)* depends exponentially on selection coefficient. Apart from some residual fluctuations, our analytical formula (Eq. 1) demonstrates excellent agreement with simulation. Since the sites with possible beneficial mutations with given mutation gain s are the sites with deleterious alleles with fitness cost s, we confirm that distribution of beneficial fitness effects acquires and maintains an exponential shape in a broad interval of time.

Then, we compared the analytic prediction for the log-slope of DFE β (Eq. 2) with simulation, for different values of the average selection coefficient s_{av} and initial allelic frequency f_{in} (Fig. 3b). We observe a good match with the analytic formula that predicts the linear increase of the slope in time at early times. At longer times, $t > 1/s_{av}$, our analysis and simulation deviate because the time dependence of the simulated slope becomes slower than linear in time. The results are not very sensitive to initial frequency f_{in} or variation of other model parameters (Fig. 3b). Note that, in this regime, although the slope increases more slowly than predicted by Eq. 2, the exponential dependence on mutation cost is conserved. For longer times, the fluctuations increase with time, which is related to strong stochastic effects in the traveling wave regime.

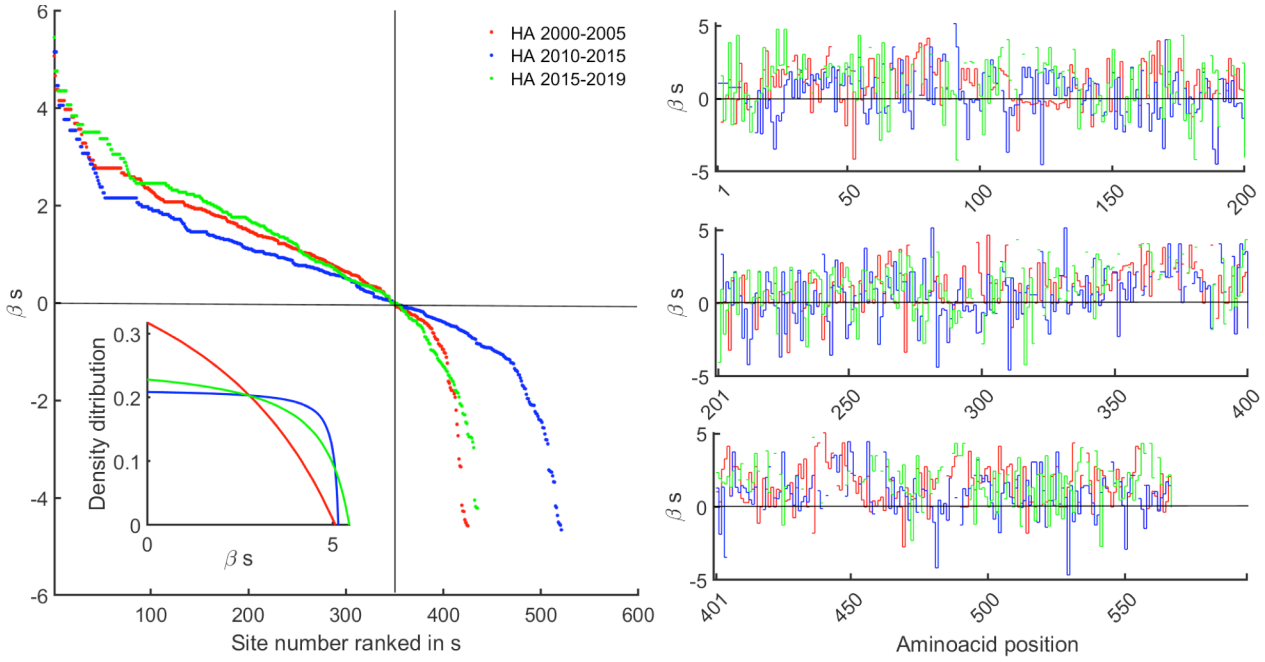

Fig. 4: The method determines selection coefficients for all genetically diverse sites of influenza H1N1 Hemagglutinin (HA). (A) The ranked relative values of selection coefficient βs calculated from Eq. 6 for three 5-year windows (shown). Only sufficiently variable sites (f > 5%) are shown. For balanced sampling, sequences with overall mutation frequency < 5% were down-weighted by 25% through resampling 25 times (Pedruzzi and Rouzine 2019). (Insert) The normalized distributions of selection coefficients across amino acid positions for $s > 0$ in the three time windows. (B-D) The values of βs at their actual positions in HA.

The differences between the prediction of Eq. 2 and simulation at long times are caused by entering the traveling wave regime. In this regime, the wave moves beyond the best-fit sequence present in the initial population due to beneficial mutations (Fig 2f). To predict the slope analytically, we need to account for the effect of beneficial mutations (Good, et al. 2012). Using the analytic result in Eq. 4 derived in *S1 Appendix*, we obtain a good agreement with long-term simulation results (Fig. 3). These results were averaged over several independent simulation runs. Thus, our model of evolution provides a simple explanation for the long-standing puzzle of the exponential DFE (Fig. 1).

Calculating selection coefficients from a virus protein sequence set. Our results have an important practical application. They enable us to estimate the relative value of the selection coefficient, *s,* for each genetically diverse site using a *one-time* sequence set, as long as the system is far from steady-state. We focus on one of two surface proteins of Influenza A H1N1, Hemagglutinin (HA), which contains targets of neutralizing antibody response and a subject of

intense scrutiny. HA sequences were downloaded from a public database (https://www.fludb.org). The sequences were collected worldwide between years 2000 and 2019. We classified them into three 5-year time windows before and after pandemic influenza (2005-2010), during which replication inhibitors were administered broadly. After binarization of all sequences into consensus 0 and non-consensus 1, we determined allelic frequency f_i for each site i. Based on our analytic result in Eq. 3, the relative value of the selection coefficients at aminoacid position i can be estimated from

$$\beta(t)s_i = -\log\left[\frac{f_i(t)}{f_{norm}}\right] \quad (7)$$

The presence of an additional factor f_{norm} in Eq. 7 is due to the fact that, in Eq. 7, f_i represents the frequency of less-fit alleles at site i, hence, $s_i > 0$ for all sites, by definition. In real life, in experiment, the best-fit sequence is not known, or may not even exist, see the discussion of influenza below. Hence, f_i has to be redefined as the frequency of minority alleles with respect to the consensus sequence determines at a fixed (usually initial) time point. Therefore, some sites will have negative s_i, and we need to introduce factor $f_{\rm norm}$ to account for such sites: they correspond to $f_i > f_{\rm norm}$.

Note that the left-hand side in Eq. 7 factorizes into a product of two term: one depends only on time, and another only on site. We can use this fact to determine the normalization factor f_{norm}, as follows. In each time window, we rank genomic sites in the descending order in $-\log f$. We observe the intersection between the ranked curves obtained at different times (Fig. 4a). Then, we add a constant to the ranked $\log f$ to obtain $s = 0$ at the intersection point between the curves. The resulting estimate of $\beta(t)s_i$ from Eq. 7 represents the selection coefficient at site i in relative units. Further, taking the inverse derivative from each ranked s curve, we obtain the distribution density of selection coefficient over non-conserved sites $g(s)$, which is broad and becomes almost uniform after the pandemic of 2005-2010 (inset in Fig. 4a). Finally, we can re-order the ranked sites back and plot the relative values of selection coefficient, βs, against their aminoacid positions (Fig 4 b-d).

Discussion

In summary, we proposed an evolutionary explanation for the exponential DFE of beneficial mutations in terms of the mutation gain in fitness. Using an asexual population model, we predicted a gradual depletion of deleterious alleles with higher fitness costs accompanied by the emergence of a clonal structure after $t \approx 1/s_{av}$. First, neglecting new mutations, we obtained an exponential dependence of allelic frequency on fitness. The logarithmic slope is equal to time, which corresponds to the virtual absence of linkage effects at early times. The formula is in agreement with Monte-Carlo simulations for early times until $t \approx 1/s_{av}$. At longer times, when beneficial mutations become crucial for the generation of new highly fit genomes, we obtained another expression based on the traveling wave theory. Our results confirm the previous work (Pedruzzi, et al. 2018) where an exponential dependence for deleterious allele frequency was predicted using a rather general argument based on the maximum of entropy. This work confirms this result and, moreover, calculates a specific logarithmic slope for the exponential.

Based on the experiments cited in *Introduction*, many models assume an exponential distribution of fitness effects as a starting assumption (Gerrish and Lenski 1998; Good, et al. 2012; Walczak, et al. 2012). Our findings provide an evolutionary justification for this assumption and update these theories by predicting that the distribution is not constant but shrinks in time. However, when mutation selection balance is approached, reverse mutations demolish selection as well as exponential dependence in DFE.

Other groups attempted to explain the universality of the exponential DFE using formal statistical arguments, such as the extreme-value theory (Gillespie 1982; Orr 2003; Joyce, et al. 2008). There are essential differences between this pioneering work and our findings. In the cited work, the aim was to prove an exponential distribution for the raw distribution of selection coefficient among all possible genomic sites, *g(s)*, in the limit of large *s*.

In contrast, we take into account our conclusion that the exponential dependence of DFE on selection coefficient is mostly determined by the evolved occupation numbers of sites $f(s, t)$ in the broad range of s and that $g(s)$ is a relative slow function of s. Also, the cited approach (Gillespie 1982; Orr 2003; Joyce, et al. 2008) predicts a constant slope of the exponential, while our analysis, simulation, and experimental data prove that it changes in time.

Applying our theory to influenza A virus sequence data, we estimated selection coefficients for each diverse site of hemagglutinin in a broad time range. Note that predicted s_i depend on a time interval (Fig 4b-d). The time dependence is expected, because the evolution of influenza occurs under time-dependent selection conditions created by accumulating memory cells in recovered individuals (Rouzine and Rozhnova 2018). The sharp change in the shape of distribution of $g(s)$ (Fig. 4a inset), as well as the drop in the exponential slope before and after pandemic (Fig 4a), may be related to epistatic effects caused by rapid development of virus resistance to antiviral treatment used during pandemics of 2005-2010 (Pedruzzi and Rouzine 2019). Analogously, it has been argued that the high level of HIV diversity in chronically infected patients is caused by adaptation to the individual immune response, including escape mutations in 20-30 epitopes and numerous epistatic sites per each escape mutation (Rouzine and Coffin 1999).

To conclude, we demonstrated that the exponential DFE observed in viruses and bacteria is a natural consequence of the process of adaptation. We derived analytical expressions for the log-slope of exponential distribution in a broad range of times and parameter values. We showed how these theoretical findings can be used to measure the Darwinian fitness effect of mutations, in relative units, from a single protein sequence sample.

Materials and Methods

See online Supplement

Funding: This work has been supported by Agence Nationale de la Recherche grant J16R389 to I.M.R.

Author contributions:

A.B.: Formal analysis, Investigation, Software, Visualization, Writing-original draft

G.P.: Data curation, Software, Visualization

I.M.R.: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Writing-review & editing

Competing interests: Authors declare no competing interests.

Data and materials availability: All data are from public database https://www.fludb.org .

Supplementary Materials: Materials and Methods

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686-690.

Brunet E, Rouzine IM, Wilke CO. 2008. The stochastic edge in adaptive evolution. Genetics 179:603-620.

Desai MM, Fisher DS. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176:1759-1798.

Dutta RN, Rouzine IM, Smith SD, Wilke CO, Novella IS. 2008. Rapid adaptive amplification of preexisting variation in an RNA virus. J Virol 82:4354-4362.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet 8:610-618.

Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. Genetica 102-103:127-144.

Gillespie JH. 1982. A Randomized Sas Cff Model of Natural-Selection in a Random Environment. Theoretical Population Biology 21:219-237.

Good BH, Desai MM. 2015. The impact of macroscopic epistasis on long-term evolutionary dynamics. Genetics 199:177-190.

Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM. 2012. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci U S A 109:4950-4955.

Hallatschek O. 2011. The noisy edge of traveling waves. Proc Natl Acad Sci U S A 108:1783-1787.

Illingworth CJ, Mustonen V. 2012. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. PLoS Pathog 8:e1003091.

Imhof M, Schlotterer C. 2001. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A 98:1113-1117.

Joyce P, Rokyta DR, Beisel CJ, Orr HA. 2008. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. Genetics 180:1627-1643.

Kassen R, Bataillon T. 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38:484-488.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251-2261.
Neher RA, Hallatschek O. 2013. Genealogies of rapidly adapting populations. Proc Natl Acad Sci U S A 110:437-442.
Neher RA, Shraiman BI, Fisher DS. 2010. Rate of adaptation in large sexual populations. Genetics 184:467-481.
Orr HA. 2003. The distribution of fitness effects among beneficial mutations. Genetics 163:1519-1526.
Pedruzzi G, Barlukova A, Rouzine IM. 2018. Evolutionary footprint of epistasis. PLoS Comput Biol 14:e1006426.
Pedruzzi G, Rouzine IM. 2019. High-fidelity analysis of epistasis predicts primary and secondary drug resistant mutations in influenza. , submitted for publication.
Rouzine IM, Brunet E, Wilke CO. 2008. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theoretical Population Biology 73:24-46.
Rouzine IM, Coffin JM. 2005. Evolution of human immunodeficiency virus under selection and weak recombination. Genetics 170:7-18.
Rouzine IM, Coffin JM. 2007. Highly fit ancestors of a partly sexual haploid population. Theoretical Population Biology 71:239-250.
Rouzine IM, Coffin JM. 2010. Multi-site adaptation in the presence of infrequent recombination. Theoretical Population Biology 77:189-204.
Rouzine IM, Coffin JM. 1999. Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. J Virol 73:8167-8178.
Rouzine IM, Rozhnova G. 2018. Antigenic evolution of viruses in host populations. PLoS Pathog 14:e1007291.
Rouzine IM, Wakeley J, Coffin JM. 2003. The solitary wave of asexual evolution. Proc Natl Acad Sci U S A 100:587-592.
Stern A, Bianco S, Yeh MT, Wright C, Butcher K, Tang C, Nielsen R, Andino R. 2014. Costs and benefits of mutational robustness in RNA viruses. Cell Rep 8:1026-1036.
Tsimring LS, Levine H, Kessler D. 1996. RNA virus evolution via a fitness-space model. Phys. Rev. Lett. 76:4440-4443.
Walczak AM, Nicolaisen LE, Plotkin JB, Desai MM. 2012. The structure of genealogies in the presence of purifying selection: a fitness-class coalescent. Genetics 190:753-779.
Wrenbeck EE, Azouz LR, Whitehead TA. 2017. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nat Commun 8:15695.

Supplementary Methods

The evolutionary origin of the universal mutation spectrum

A. Barlukova , G. Pedruzzi, and I. M. Rouzine*

Sorbonne Université, Institute de Biologie Paris-Seine, Laboratoire de Biologie Computationnelle et Quantitative, LCQB, F-75004 Paris, France
*To whom correspondence should be addressed:
igor.rouzine@sorbonne-universite.fr

Sample balancing. After alignment and binarization of sequences by setting consensus residues to 0 and non-consensus to 1, we balanced sequence sampling for the reasons described in (*1*). Specifically, we selected the sequences with frequency of 1s per genome less than a preset value dv (5%). These were randomly sampled 25 times and down-weighted by a coefficient D_w, set to 25%. Then, we followed the procedure described in the main text and in the legend to Fig. 5 to predict the distribution of selection coefficients.

Model. We consider a haploid population of N binary sequences, where each genome site (nucleotide position) numbered by i =1, 2, ..., L carries one of two possible genetic variants (alleles), denoted K_i =0 or K_i =1. The genome is assumed to be very long, $L >> 1$. Time is discrete and measured in units of population generations. The evolution of the population is simulated using a standard Wright-Fisher model, which includes the factors of random mutation with genomic rate μL, natural selection, and random genetic drift. Recombination is assumed to be absent. Once per generation, each individual genome is replaced by a random number of its progeny which obeys multinomial distribution. The total population stays constant with the use of the broken-stick algorithm. To include natural selection, the average progeny number (Darwinian fitness) of sequence K_i is set to e^W. We consider the simplest case when the fitness effects of mutations, s_i , are additive over sites:

$$W = \sum_{i=1}^{L} s_i K_i \qquad (8)$$

The reference genome, $\{K_i=0\}$, can be chosen in arbitrary way. For our aim, it is convenient to set it to be the same as the best fit sequence, so that all selection coefficients s_i are negative. Each site i with deleterious allele, $K_i=1$, is a target site for a possible beneficial mutation. Vice versa, a site with the favorable allele, $K_i=0$, can have a deleterious mutation. A more general version of fitness model that accounts for pairwise epistatic interactions is considered in (*2*) and, for macroscopic epistasis, in (*3*). Here we focus on additive contributions of single sites to the fitness landscape.

In our simulations, selection coefficients are chosen randomly at each site from the half-normal distribution

$$g(s) = \frac{2}{s_{av}\pi} \exp\left(-\frac{s^2}{\pi {s_{av}}^2}\right) \tag{9}$$

Here, s_{av} is the average mutation cost in fitness for the initial state.

Early evolution. We focus on one genomic site of L sites, whose selection coefficient is assumed to be known and equal to -s, where $s > 0$. The other selection coefficients are assumed to vary following some random distribution with density $g(s)$. We make no assumption regarding the shape of $g(s)$, but assume that distribution of s at different sites is independent. We assume also that initial population has random distribution of alleles among sites with average frequency of less fit alleles denoted by f_{in}. Biologically this condition corresponds to a population that has changed suddenly its conditions. First, we will also neglect the effect of new mutations, which simplification is shown to be accurate at early times, when evolution is dominated by decay of standing variation. In the next subsection, we will include the effect of new mutations. Denote by I_0 the proportion of all possible sequences having 0 (wild-type) at given site. The proportion of possible sequences having 1 (mutation) at the site denote by I_1. Then, the corresponding mutant frequency can be found as ratio

$$f = \frac{I_1}{I_0 + I_1} \tag{10}$$

To obtain frequency in time we account for the action of selection. Selection causes decay of the number of each sequence $\{K_i\}$, by time dependent factor $e^{-t\sum_{i=1}^{L-1} s_i K_i}$. Frequencies I_0 and I_1 represent the average over all values of s and K_i for all sites

$$I_0 = (1 - f_{in}) \prod_{i=1}^{L-1} \left(\int_0^{+\infty} ds_i \, e^{-t\sum_{i=1}^{L-1} s_i K_i} \sum_{K_i=0}^{1} p(K_i) g(s_i) \right)$$

$$I_1 = f_{in} \; e^{-ts} \prod_{i=1}^{L-1} \left(\int_0^{+\infty} ds_i e^{-t\sum_{i=1}^{L-1} s_i K_i} \sum_{K_i=0}^{1} p(K_i) g(s_i) \right)$$

probabilities of having less-fit and better-fit alleles is $p(1) = f_{in}$ and $p(0) = 1 - f_{in}$, respectively. Then, I_0 and I_1 can be rewritten by taking exponential term inside of parentheses and separating variables at different sites i. The result takes the form

$$f = \frac{e^{-ts} f_{in}}{1 - f_{in} + e^{-ts} f_{in}} \tag{11}$$

Traveling wave regime. In the traveling wave regime, which starts around $t > 1/s_{\text{av}}$, beneficial mutations have to be included into consideration because they create new best-fit genomes of the traveling wave (*4-7*). Let t_0 be the characteristic time of the beginning of traveling wave regime and $\pi(s)$ be the fixation probability

of beneficial mutations. In this regime, most of deleterious alleles are located at uniformly deleterious sites (Fig. 2, yellow columns). Hence, their loss occurs mostly due to fixation of new beneficial alleles at these sites. Then, the dynamic equation for the frequency of deleterious alleles for $t > t_0$, $t_0 \sim 1/s_{\text{av}}$, can be written as follows

$$\frac{\partial f(s,t)}{\partial t} = -\mu N \pi(s) f(s,t) \tag{12}$$

The initial condition for Eq. 8 can be obtained from the estimate of f in the initial time interval where selection of pre-existing genomes is the dominant process. From Eq. 7

$$f(s,t_0) \approx f_{in} e^{-t_0 s}$$

The solution of Eq. 8 with these initial conditions is given by Eq. 3 in the Main text. Thus, the problem is reduced to the result of the previous work (*8*) for the fixation probability of mutations $\pi(s)$

$$\pi(s) \sim e^{-\frac{s^2}{2v}} \left(\frac{e^{-s\frac{x_c}{v}} - 1}{s} \right) + \frac{e^{\frac{{x_c}^2}{2v}}}{v x_c} \int_{x_c}^{\infty} dx\, x e^{-\frac{(x-s)^2}{2v}}, \quad \pi(0) \approx \frac{1}{N} \tag{13}$$

Here, v is the rate of adaptation, x_c is the characteristic value of fitness. Parameters v and x_c can be found numerically from two transcendental equations

$$2 = \frac{U_b}{\sigma} \sqrt{\frac{2\pi\sigma^2}{v}} \left[1 + \frac{\sigma}{x_c} + \frac{v}{\sigma x_c - v} \right] e^{\frac{(x_c - \frac{v}{\sigma})^2}{2v}} \tag{14}$$

$$1 = N U_b \left[\frac{x_c^2}{v} - 1 + \frac{2x_c\sigma}{v} + \frac{2\sigma^2}{v} \right] e^{-\frac{1}{\sigma}\left(x_c - \frac{v}{2\sigma}\right)} \tag{15}$$

where U_b is the probability of beneficial mutation per genome per generation. Thus, the time evolution of the frequency of deleterious alleles can be found numerically using Eqs. 3, 9, 10, 11. We note that parameter t_0 in Eq. 3 is unknown, since Eq. 3 is the asymptotic expression at large times, $t \gg 1/s_{av}$ To connect two regimes of evolution we find optimum value of t_0, such that simulated and analytical curves for the slopes are at the best match. The value of σ is the average effect of beneficial mutation for the sites that can have such a mutation and, hence, is equal to $1/\beta$, where β found from self-consistent condition given by Eqs. 2 and 3. Note that Eq. 3 predicts an exponential decay at small to moderate s; at large $s \sim 1$, it predicts a faster decay with s.

1. G. Pedruzzi, I. M. Rouzine, High-fidelity analysis of epistasis predicts primary and secondary drug resistant mutations in influenza. *, submitted for publication*, (2019).

2. G. Pedruzzi, A. Barlukova, I. M. Rouzine, Evolutionary footprint of epistasis. *PLoS Comput Biol* **14**, e1006426 (2018).
3. B. H. Good, M. M. Desai, The impact of macroscopic epistasis on long-term evolutionary dynamics. *Genetics* **199**, 177-190 (2015).
4. I. M. Rouzine, J. Wakeley, J. M. Coffin, The solitary wave of asexual evolution. *Proc Natl Acad Sci U S A* **100**, 587-592 (2003).
5. M. M. Desai, D. S. Fisher, Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759-1798 (2007).
6. I. M. Rouzine, J. M. Coffin, Highly fit ancestors of a partly sexual haploid population. *Theor Popul Biol* **71**, 239-250 (2007).
7. R. A. Neher, B. I. Shraiman, D. S. Fisher, Rate of adaptation in large sexual populations. *Genetics* **184**, 467-481 (2010).
8. B. H. Good, I. M. Rouzine, D. J. Balick, O. Hallatschek, M. M. Desai, Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc Natl Acad Sci U S A* **109**, 4950-4955 (2012).