Senescence and entrenchment in evolution of amino acid sites

A. V. Stolyarova1✉, E. Nabieva1,2, V. V. Ptushenko3,4, A. V. Favorov5,6, A. V. Popova7, A. D. Neverov7 & G. A. Bazykin1,2

Amino acid propensities at a site change in the course of protein evolution. This may happen for two reasons. Changes may be triggered by substitutions at epistatically interacting sites elsewhere in the genome. Alternatively, they may arise due to environmental changes that are external to the genome. Here, we design a framework for distinguishing between these alternatives. Using analytical modelling and simulations, we show that they cause opposite dynamics of the fitness of the allele currently occupying the site: it tends to increase with the time since its origin due to epistasis (“entrenchment”), but to decrease due to random environmental fluctuations (“senescence”). By analysing the genomes of vertebrates and insects, we show that the amino acids originating at negatively selected sites experience strong entrenchment. By contrast, the amino acids originating at positively selected sites experience senescence. We propose that senescence of the current allele is a cause of adaptive evolution.
Fitness landscape is the key concept of evolutionary biology, and its description is necessary to fully understand adaptive evolution and speciation\(^1^–^4\). Unfortunately, the large dimensionality of even the landscapes of individual proteins makes them impossible to measure comprehensively in a direct experiment\(^5^–^6\). Still, methods of comparative genomics can be used to assess the integral features of fitness landscapes. The simplest informative unit of landscape structure is the single-position fitness landscape (SPFL)\(^7\), i.e., a vector of fitness values of all possible alleles at an individual genomic position. SPFLs change with time\(^8^–^15\); this may affect the optimality of the allele that is currently prevalent at this site, influencing subsequent evolution.

One factor entailing changes of SPFL is substitutions at other sites of the genome. For this to be the case, these substitutions need to affect the relative fitness of different variants at the considered site, i.e., these sites have to be involved in epistatic interactions. Epistasis has been postulated to be a prevalent factor of protein evolution and divergence across species\(^16^–^18\). One is expected manifestation of genomewide epistasis is entrenchment, or the evolutionary Stokes shift\(^12^,^25^,^26\) — a phenomenon whereby the relative fitness of the allele currently prevalent at the site increases as substitutions at interacting sites accumulate. The reason for this increase is the constraint imposed by the site in consideration onto epistatically interacting sites. The evolution of the remaining sequence is constrained to preserve the high fitness of the resident allele, and may even increase it; at the same time, this sequence is free to evolve to become less compatible with other variants not currently present at the site. Over time, this leads to an increase in the fitness of the current allele relative to other alleles, including those that resided at this site earlier.

Entrenchment was demonstrated both in simulated protein evolution\(^12^,^25^,^26\) and in evolution of real-life proteins. For example, it was shown that reversals of past substitutions are suggestive of entrenchment: their rate declines with time, indicating that they become more deleterious, i.e., that the current allele becomes more preferable compared to the previous one\(^27^–^30\). The decline in the rate of reversals is caused both by the increase in the fitness of the current allele and the decrease in the fitness of the replaced allele\(^28\).

However, the SPFL may change due to environmental fluctuations even in the absence of epistasis. If such changes are recurrent, the fitness landscape becomes a time-dependent “seascape”\(^16^,^31^–^36\). This leads to recurrent positive selection (fluctuating selection) in favor of the newly beneficial alleles and to adaptive evolution\(^35^,^37^–^40\). Nowadays, the way fluctuating selection shapes the dynamics of the relative fitness of the current allele remains poorly studied.

Here, we characterize the effects of epistasis and of fluctuating selection on SPFL changes and estimate the contribution of these forces in past evolution.

Results

Environmental fluctuations make the current allele less fit. First, we ask how fluctuating selection affects the relative fitness of different alleles at a site. If changes of the SPFL are random with regard to the identity of the allele currently residing at the site, we expect that they, on average, will reduce its relative fitness (see formal proof in Supplementary Note 1). This is because, due to natural selection, the relative fitness conferred by the current variant is, on average, higher than that of a random variant at this site. An episode of positive selection triggered by an SPFL change may then cause the spread of a novel variant, which would confer high fitness till the next SPFL change.

To illustrate this, we simulate amino acid evolution on a randomly changing fitness landscape. In this simulation, the fitness values for each of the 20 possible amino acids are drawn from a predefined distribution, and the amino acid substitutions occur with probabilities determined by the corresponding selection coefficients. At random moments of time, fitness values are redrawn from the same distribution (Fig. 1a–c).

As a result of selection, the fitness of the current allele is on average higher than that of other alleles (Fig. 1b, c); in particular, if selection is strong, the site is typically occupied by the best-possible allele (Fig. 1b). However, as the landscape changes randomly, the fitness of this original allele, on average, decreases with time, gradually approaching the mean fitness across all possible variants (Fig. 1e, f). We call this process senescence of the current allele\(^41\). This effect is more pronounced for the rugged landscape, when one allele is highly more beneficial than others (Fig. 1e), and less pronounced when selection is weaker (Fig. 1f).

The decline in fitness of the current allele due to fluctuating selection leads to an increase in the rate at which it is lost (Fig. 1h, i), in line with the quenched theory of fluctuating selection\(^34^,^37\) (Supplementary Note 2).
after the ancestral substitution, but a reduced rate later on. By contrast, under senescence ($k < 0$), the rate of substitutions increases with time since the allele gain (Fig. 2b). An alternative mode of simulation of senescence, whereby random changes in SPFL and the molecular evolution caused by them are modeled explicitly, gives the same results (Supplementary Fig. 1).

Besides the phylogenetic distribution of substitutions, the mode and rate of SPFL change also affect the overall rate of molecular evolution (Fig. 2c). Compared to a static landscape of the same shape, entrenchment reduces the substitution rate, as the time-averaged fitness of the current allele is higher, and therefore its substitutions rarer. Conversely, under senescence, many of the substitutions of the current allele are advantageous, increasing the overall rate of evolution. Importantly, senescence doesn’t necessarily result in overall evolution rate exceeding the neutral rate, which is a hallmark of positive selection. Indeed, if an allele is strongly preferred, a drop in its fitness over the course of senescence may still leave it the optimal one, so that negative selection will still maintain it (as observed for the rugged SPFL, Fig. 2c right).

**Senescence and entrenchment at single-allele resolution.** Large phylogenies allow detecting changes in substitution frequencies for individual alleles. Each originating allele, e.g., an amino acid arising at a specific site from an ancestral amino acid substitution, can be inherited by multiple descendant lineages leading to different extant species. Ancestral state reconstruction can then be used to infer the lineages at which this allele has been lost due to a reversion or substitution to a different amino acid. If enough such lineages are available, this allows to trace the decline or increase in the rate of allele substitution since its origin, i.e., entrenchment or senescence.

We applied binomial logistic regression to detect changes in substitution frequencies with the age of the current allele along the phylogeny for five mitochondrial genes of Metazoa. The regression was performed separately for each allele $B$ with a known time of origin (corresponding to allele gain $A \rightarrow B$) at each site. Among the 42,637 such alleles, we identified 28 alleles for which the frequency of replacement significantly increased with time since their origin (i.e., senescing alleles), and 21 alleles...
where it decreased (i.e., entrenched alleles) at 5% false discovery rate (Fig. 3a, Supplementary Table 1). The examples of phylogenies indicating allele replacements at senescing and entrenched alleles are shown in Fig. 3b, c. The overall number of substitutions occurring along the phylogeny was similar between sites containing entrenched and senescing alleles (Supplementary Fig. 13, sign test p value = 0.84), indicating that the substitution rate was insufficient for distinguishing between these scenarios.

**Heterogeneity leads to an artifactual signal of entrenchment.** While phylogenies spanning hundreds and thousands of species, like those available for mitochondrial proteins, allow to measure the
changes in the substitution rate for individual alleles, in smaller phylogenies, the number of substitutions experienced by an allele can be insufficient for such an analysis. Still, it may be possible to identify the prevailing patterns of substitutions by pooling alleles together. However, such pooling can be problematic: even in the absence of SPFL changes, the rate of substitution can appear to change with time since allele origin if the pooled alleles have different time-invariant substitution rates, confounding inference of SPFL changes.

Indeed, consider a set of alleles, each characterized by its own substitution rate that is stationary (constant in time) but differs between alleles. While the replacement rate may be constant for each allele, so that the time to replacement is characterized by an exponential distribution, it will not, in general, be exponentially distributed in the resulting heterogeneous dataset. Instead, the frequency of substitution will appear to decline with time (Fig. 4a), making it non-stationary and mimicking entrenchment of the current allele. The problem of data heterogeneity leading to decreasing hazard function is well known in demographic studies, where the problem of data heterogeneity leading to decreasing hazard function is well known in demographic studies, where the number of substitutions experienced by an allele can be insufficient for such an analysis. Still, it may be possible to identify the prevailing patterns of substitutions by pooling alleles together. However, such pooling can be problematic: even in the absence of SPFL changes, the rate of substitution can appear to change with time since allele origin if the pooled alleles have different time-invariant substitution rates, confounding inference of SPFL changes.

To show this, we simulate molecular evolution on static SPFLs of different shapes. If all alleles have the same fitness, i.e., if all substitutions are neutral (“flat” landscape), the substitution rate is independent of time since allele origin (Fig. 4b). By contrast, if the fitness values of alleles are drawn from a gamma distribution, so that different alleles have different fitness, the substitution frequency decreases with the age of the current allele, even though the SPFL doesn’t change (Fig. 4c). On a more rugged SPFL, when one allele is much more fit than all others, this decline is even sharper (Fig. 4d).

Inferring senescence and entrenchment from phylogenies.

To address the problem of heterogeneity of alleles, we reconstructed the SPFL dynamics while accounting for differences between alleles in mutation rates and “baseline” selection. In the absence of prior information about the distribution of these characteristics, we fitted the approximate Bayesian computation (ABC)47 approach to obtain the posterior distribution of the rate of substitution change with time since allele origin. The ABC depends on a set of summary statistics to evaluate the difference between the simulation results and the data. We used two summary statistics, each aggregating overall individual alleles, which reflect the age-dependent dynamics of substitution rates (see Methods, Supplementary Fig. 2).

We used two models for parameter inference. Under the two-parameter model, we assumed that log fitness values for individual alleles were drawn from a gamma distribution with rate and shape parameters denoted by $\alpha$ and $\beta$, and the log fitness of the current allele at all sites changed linearly with rate $k$. Under the three-parameter model, the fitness changed only for a fraction of alleles, while the fitness of the remaining alleles was invariant (see Methods for details).

Simulations show that both models perform well in identifying senescence and entrenchment under a broad range of parameters, and are robust to overall substitution rate, phylogeny shape, and the value of $k$ for them depended on the mode of selection acting at them, on the basis of the ratio of non-synonymous and synonymous substitutions per site ($\omega$). Under the two-parameter model, the fitness changed only for a fraction of alleles, while the fitness of the remaining alleles was invariant (see Methods for details).

Simulations show that both models perform well in identifying senescence and entrenchment under a broad range of parameters, and are robust to overall substitution rate, phylogeny shape, and the value of $k$ for them depended on the mode of selection acting at them, on the basis of the ratio of non-synonymous and synonymous substitutions per site ($\omega$). Under the two-parameter model, the fitness changed only for a fraction of alleles, while the fitness of the remaining alleles was invariant (see Methods for details).
senescence (Fig. 5b, middle panel). A similar pattern is observed in phylogenies of insects (Fig. 5d).

While senescence is observed at sites that undergo rapid substitution ($\omega > 1$), it is distinct from an increase in the overall substitution rate. Similarly, while entrenchment is observed at constrained sites ($\omega < 1$), it is distinct from a reduction in substitution rate. To illustrate this, we simulated evolution under different substitution rates but constant SPFL on the phylogenies of vertebrates and insects, using the same distribution of $\omega$ values as in the real data, and estimated $k$ using the ABC pipeline (Fig. 5a, c). No senescence or entrenchment was detected for the datasets simulated using the stationary model under $\omega = 1$ or $\omega > 1$. A weak spurious signal of entrenchment was detected for the simulated datasets with $\omega < 1$, resulting from the high
heterogeneity in evolution rates (see above); however, it was much weaker than that observed in the data for this category of sites.

Discussion

While the direction of changes in fitness in the course of evolution is unpredictable for an individual allele, there are certain statistical regularities. Previous work has shown that, at a site involved in epistatic interactions with other sites, the relative fitness conferred by the incumbent allele is expected to increase with time since its origin\textsuperscript{9,18,22,23}. Acting alone, i.e., if the overall fitness landscape is static, this process of entrenchment should make the propensities existing at individual sites more pronounced. This, in turn, should limit the level of divergence between highly divergent sequences, although reaching this level may take a very long time\textsuperscript{9,10,48–51}.

Here, we consider the dynamics of allele fitness due to changes in the overall fitness landscape itself. We show that, if the direction of these changes is independent of the current position of the population in the genotype space, the expected mean dynamics — senescence — is opposite to that of entrenchment. We design a method to distinguish the two patterns from the phylogenetic distribution of substitutions and find that entrenchment is ubiquitous at negatively selected sites, while senescence is prevalent at sites undergoing adaptive evolution under positive selection. While senescence is a distinct phenomenon from positive selection (Supplementary Note 3), the observed concordance between the direction of non-stationarity

Fig. 3 Senescence and entrenchment of individual alleles in the mitochondrial genes of Metazoa. a Manhattan plot of senescing and entrenched alleles. Only the alleles with a known phylogenetic position of origin, i.e., those that were not yet present in the tree root, were analyzed; a single genomic site can contain zero, one or several alleles. P values are calculated using binomial logistic regression. The alleles demonstrating significant senescence under 5% FDR are shown in red; the alleles demonstrating entrenchment are shown in green. No amino acid sites contained more than one significantly senescing or entrenched alleles. The list of significantly senescing or entrenched alleles is shown in Supplementary Table 1. b Examples of senescing (COX2 position 56, red) and entrenched (ATP6 position 71, green) alleles. The contiguous segment of the phylogeny carrying the derived allele is shown in color. c Distribution of substitutions along the lifetime of alleles shown in (b). For the senescing allele, the phylogenetic branches corresponding to allele replacements (red) originate later than the branches without replacements (gray). Conversely, the entrenched allele is more frequently replaced soon after its origin (green).
and selection mode imply that these phenomena are closely related, and that senescence causes many of the adaptive substitutions.

What causes allele senescence? Firstly, it can sometimes result from negative epistasis with an allele arising at an interacting site. While this could result in senescence occasionally, on average epistasis results in entrenchment\textsuperscript{25,26}. Secondly, senescence could result from changes in selection pressure external to the genome. We have previously described the acceleration of substitutions over the course of allele lifetime in the evolution of influenza A virus\textsuperscript{41}. This pattern has been mainly observed at sites associated with avoidance of the host immune system pressure, and we interpreted it as evidence for negative frequency-dependent selection actively disfavoring the current allele\textsuperscript{41}. Here, we show that this type of selection is not a prerequisite for senescence. Instead, senescence is expected whenever selection alters without regard to the identity of the current allele. How much of senescence, and positive selection resulting from it, is due to random changes of the fitness landscape, and how much is due to systematic selection against the current allele, can be a subject of further research.

Table 1 Senescence and entrenchment in evolution of genomes of vertebrates and insects.

|        | # of sites   | # of alleles  | $k$       | Fraction of senescing or entrenched alleles |
|--------|--------------|---------------|-----------|--------------------------------------------|
|        |              |               |           |                                            |
| Vertebrates |              |               |           |                                            |
| $\omega < 1$ | 2,598,701 (87.8%) | 574,137 (36.4%) | 62.3 (25.1, 100.0) | 0.99 (0.95, 1.00) |
| $\omega = 1$ | 348,047 (11.8%)  | 917,340 (58.1%) | 45.3 (−36.8, 98.9) | 0.42 (0.15, 0.90) |
| $\omega > 1$  | 13,189 (0.4%) | 86,852 (5.5%)  | −23.3 (−94.7, −3.5) | 0.18 (0.0, 0.86) |
| Total       | 2,959,937 (100%) | 1,578,329 (100%) |           |                                            |
| Insects    |              |               |           |                                            |
| $\omega < 1$ | 2,699,432 (93.2%) | 429,800 (48.8%)  | 47.8 (−1.1, 87.9)  | 0.81 (0.47, 1.00) |
| $\omega = 1$ | 185,829 (6.4%)  | 413,101 (46.8%) | 9.4 (−80.2, 91.4) | 0.16 (0.01, 0.92) |
| $\omega > 1$  | 9698 (4.4%) | 38,577 (4.4%)  | −22.6 (−88.7, 52.0) | 0.25 (0.03, 0.92) |
| Total       | 2,894,959 (100%) | 881,478 (100%)  |           |                                            |

The analyzed datasets, the corresponding estimates of the rate of entrenchment ($k > 0$) or senescence ($k < 0$) and the fraction of alleles that experience these processes. The values show the median of the ABC posterior distribution of parameter values; numbers in parentheses represent the 95% posterior probability intervals. The total number of sites and alleles in both datasets is shown in bold.

Fig. 5 Senescence and entrenchment in protein sequences of vertebrates and insects. Plots provide ABC estimates of the rate of senescence or entrenchment $k$ and the fraction of alleles with changing fitness for protein sequences of vertebrates and insects. For each dataset, the posterior distribution of parameters under 1% acceptance threshold after local ridge regression adjustment is shown. Positive values of $k$ correspond to entrenchment, while negative values of $k$ correspond to senescence of the current allele. a, b Simulated data under the stationary model using the vertebrate (a) or insect (b) phylogeny and distribution of $\omega$ values. c, d In real genomic data of vertebrates (c) or insects (d), sites under negative selection show strong entrenchment, neutral sites demonstrate the intermediate signal, and positively selected sites are senescing.
Methods

Data. We used multiple alignments of exons of vertebrates and insects from the UCSC Genome Browser database together with the corresponding phylogenies22 (Supplementary Fig. 8). Columns with gaps were excluded. From these alignments, we reconstructed the alleles in the internal nodes of phylogenetic trees with codeml23. We re-estimated the lengths of individual branches as the average frequency of amino acid substitutions per site on this branch. We classified codon sites as positively selected (ω > 1), neutral (ω = 1), or positively selected (ω < 1) using the Bayes empirical Bayes method as implemented in the PAML package34. The size of the datasets and the number of sites and substitution subtrees in each bin are shown in Table 1. The mitochondrial dataset consists of the amino acid alignment of five proteins for several thousand metazoan species35.

Simulations. Simulations of amino acid sequence evolution were performed with the SELVa simulator43. SELVa is a forward-time Markov chain simulator that allows the user to model sequence evolution along a predefined phylogenetic tree on static or dynamic SPFLs. The user can specify both the shape of SPFL (i.e., the vector of allele fitnesses for a single position in the genome) and the rule for its change. In this work, we used three types of SPFLs for amino acid sites with 20 possible alleles: flat SPFL corresponding to neutral sites (no substitution leads to change of fitness, log fitness vector is (0, 0, …, 0)); rugged SPFL (one allele is highly preferable over the other ones, log fitness vector is (10, 0, 0, …, 0)); and gamma-distributed SPFL, where the log fitness values for alleles are randomly chosen from the gamma distribution with user-defined parameters. These features of the SPFL change. In the random change mode, the SPFL changes are a Poisson process with a user-defined rate. In this case, fitness values are either reshuffled between alleles (for the flat or rugged SPFL) or redrawn from the same distribution (for gamma-distributed fitnesses). In the current allele-dependence model, the fitness of the current allele decreases or decreases linearly with time. The user can define the rate of this change (k) and the length of the time interval between changes (Δt). The log fitness of the current allele at time t + Δt is then set to

$$f_k(t + \Delta t) = f_k(t) + k \Delta t.$$  

The fitness values for other alleles remain unchanged. Positive values of k correspond to entrenchment of the current allele, which means its fitness increases with time, and negative ones, to senescence, so that its fitness decreases. When a substitution occurs, i.e., the current allele is replaced with another one, the fitness of the new allele starts changing, and the fitness of the new allele starts to change at rate k. In this work, we used Δt = 0.01 d, where Δt is the length of time required for a single substitution at a neutral site; this is small enough to simulate the gradual change of allele fitness.

Individual sites are simulated independently, and interactions between them are not modeled directly. We also do not explicitly model the heterogeneity in substitution rates between alleles, although such heterogeneity arises from the differences in the fitness values of individual alleles drawn from the underlying distribution. To demonstrate the effects of randomly changing SPFLs, patterns of substitutions emerging due to senescence and entrenchment, and the heterogeneity of allele heterogeneity, we used a simple model phylogenetic tree (Supplementary Fig. 12). For ABC-based inference of the dynamics of the current allele fitness in the evolution of real-life protein sequences, we used the corresponding phylogenies of vertebrates and insects (Supplementary Fig. 8).

Substitution subtrees. For every replacement A → B on any internal branch of the phylogeny, we can define the corresponding substitution subtree, namely, the contiguous segment of the phylogeny where every internal node carries the derived variant B. For this subtree, B is the current allele, and the initial substitution A → B is the allele gain (Supplementary Fig. 2). Within a substitution subtree, the current allele B can be replaced by the ancestral variant A or some other variant C in the course of allele loss(es).

A genomic position can carry no substitution subtrees if it is fully conservative, or carry one or more substitution subtrees; the number of substitution subtrees equals the number of substitutions on the internal phylogenetic branches in this position. This means that rapidly evolving sites carry more substitution subtrees than conservative ones.

We define a statistic $s_{branch}$, the frequency at which the allele B that has occupied the considered genomic position at the origin of a specific branch has been replaced at this branch. If a single site is analyzed (as in the analysis of mitochondrial genes), $s_{branch}$ can take the values of 0 or 1; if multiple sites are pooled, $s_{branch}$ for different substitution subtrees are considered separate and $s_{branch}$ can also take values between 0 and 1. $s_{branch}$ is determined by the SPFL and the overall mutation rate of allele B. If the mutation rate is assumed to be constant, changes in $s_{branch}$ with time since the origin of B within the substitution subtree can be used to detect changes in SPFL. We estimate the rate of replacement of B as a function of its age, i.e., the evolutionary time since it was gained.

Inference of fitness changes for groups of alleles. An individual substitution subtree usually does not provide enough data to infer SPFL changes. To identify such changes with confidence, we have to pool data across subtrees and sites. However, pooling data on different subtrees creates a spurious signal of entrenchment due to heterogeneity of evolution rate and unevenness of SPFLs28,46. One approach to adjust for these confounding factors is to estimate the mean substitution rate of a subtree, which combines the mutation rate of the site and the effect of the fitness variance of the current allele, and to use this value for model fitting, e.g., in the maximum likelihood (ML) framework. However, our phylogenies are not deep enough to perform ML estimates: the number of substitutions per site is typically too low. Instead, to account for confounding effects, we used the approximate Bayesian computation approach (ABC).

Since the age-dependent patterns of substitutions are sensitive to data heterogeneity and the shape of SPFLs, we can’t directly measure the rate at which the fitness of the current allele changes. To estimate the strength and abundance of senescence and entrenchment in the evolution of protein sequences, we used rejection ABC (approximate bayesian computation) with ridge regression adjustment as implemented in the abc package for R35. ABC is a popular method for parameter inference if the likelihood function is not known.

Simulations for ABC were also performed with SELVa. Importantly, rather than simulating the full phylogenetic trees, we simulated individual substitution subtrees. For each dataset of interest, we extracted the list of subtrees generated by substitutions (allele gain events) in this dataset. For each substitution subtree, we then ran SELVa with the given parameters, assuming that the number of sites in the simulation equaled the number of cases when this subtree appeared in the data. The results were pooled across subtrees, and summary statistics were calculated. This approach has two advantages in comparison to simulations based on full phylogenetic trees. First, our summary statistics are based on subtrees only, and using the list of substitution subtrees from the data is more informative than simply the number of sites: this way, we don’t have to wait until the ancestral substitution occurs in a simulation, but can start the simulation at the moment we know it has occurred in the data. Second, since the subtrees are smaller, the simulation runs faster.

We used two model functions for ABC. The first one is based on the assumption that all sites in the dataset are susceptible to senescence or entrenchment of the same strength (two-parameter model). It requires two parameters: $\alpha$ rate parameter for the gamma distribution of alleles fitness values (see above) and the rate of change of the fitness of the current allele $k$.

The second model represents a mixture of two categories of sites: those with a static SPFL ($k = 0$) and those under senescence ($k < 0$) or entrenchment ($k > 0$). It takes three parameters as input in addition to $\alpha$: $s_{branch}$ is the fraction of substitution subtrees (which corresponds to the fraction of alleles) under senescence or entrenchment with rate $k$. The simulated values for $k$ were distributed uniformly from $-1$ to $1$ and the fraction of alleles under senescence or entrenchment was also distributed uniformly from 0 to 100%; and $\alpha$ was distributed log-uniformly from $-1.5$ to 1.

The number of amino acid sites in the datasets with different site-specific $\omega$ values, and the number of substitution subtrees at these sites, vary between 10^3 and 10^6 (Table 1). To account for the variance in summary statistics for smaller datasets, the ABC model function takes a list of subtrees and their counts in the given dataset as input and generates simulations of the same size (but not larger than 100,000 substitution subtrees due to runtime restrictions). For all datasets, we used ridge regression algorithm for parameter estimation as implemented in the abc package.

Summary statistics. After evaluating a range of possible summary statistics for ABC, we ended up using two statistics based on the dynamics of allele replacement. All branches across all subtrees in the simulation are pooled together and used to calculate the following linear regression:

$$s_{branch} = a \cdot \text{length}_{branch} + b \cdot \text{age}_{branch} + c,$$

where $s_{branch}$ is the frequency at which the current allele was lost lost on the given branch, length$_{branch}$ is the average length of this branch across all substitution subtrees, and age$_{branch}$ is the age of the current allele, i.e., the distance from the root of the substitution subtree to the branch. As summary statistics, we used the values of $a$ and $b$.

Supplementary Fig. 11 shows how $a$ and $b$ depend on the simulation parameters: they strongly depend on the direction and the rate of current allele fitness change $k$ and the fraction of alleles susceptible to these changes. Summary statistics were calculated with python scripts using numpy, sclearn, Bio and ete3 packages.

ABC validation. Since our method is based on time-dependent dynamics of substitutions, its efficiency is expected to depend on the depth of the phylogenetic tree: short trees will not provide the timescale sufficient to detect changes of the current allele substitution rate. We examined the accuracy of the two-parameter model using test phylogenies with simple topology and internal edges of varying lengths (Supplementary Fig. 3a). Indeed, both the prediction error and the width of the 95% posterior probability interval are the smallest for the phylogenies with internal branches of intermediate lengths (Supplementary Fig. 3c, d). While the shortest trees don’t provide the time range sufficient to detect SPFL changes, the longest ones lack resolution and also result in bigger prediction error.

We validated the ABC pipeline on parameter estimation with SELVa simulations based on the reconstructed phylogeny of 53 vertebrates and the abc package for R35. To cross-validate ABC performance under different tolerance rates
and to evaluate the accuracy of parameter estimation for both two-parameter and three-parameter models, we calculated prediction error for parameters based on 100 randomly chosen simulations with the cross-validation function of the abc package. The prior size was 10^6 simulations for both models. The cross-validation results are shown in Supplementary Table 3 and Supplementary Figs. 4 and 5. The tolerance level 0.01 was chosen for both models. Cross-validation tests for parameter inference with our ABC pipeline show that we can accurately estimate the parameters of both models using the selected set of summary statistics.

Next, we asked whether our method is sensitive to changes in the overall rate of evolution. For each model, we generated the testing set of 100 simulations with randomly chosen parameters with normal and twofold increased substitution rate and then used ABC to infer the parameters. We demonstrate that, although the magnitude of k was overestimated for simulations with accelerated evolution rate, the estimates were not biased in any direction (t-test p value = 0.53, Supplementary Figs. 6 and 7, top panel; Supplementary Table 2). While the fraction of senescent or entrenched alleles in the three-parameter model was underestimated for simulations with accelerated evolution rate, the magnitude of the bias was not large (on average 0.1, t-test p value = 0.01) (Supplementary Table 2).

We also checked whether our method allows to confidently distinguish between senescence and entrenchment. The confusion matrices based on the same testing set of simulations show that the frequency of misclassification is 0% for the two-parameter model and 1% for the three-parameter model, and cases of misclassification were only observed in simulations with low k (<1) (Supplementary Table 2). Furthermore, in the few erroneously classified cases, the 95% probability interval for k overlapped with zero.

We used SELVa to simulate datasets under different modes of selection to test whether the artificial signal of senescence or entrenchment can occur in the stationary model of evolution (Fig. 5a, c).

**Reporting summary** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability** The genomic datasets of vertebrates and insects were downloaded from UCSC database https://hgdownload.soe.ucsc.edu (Rosenblum et al.35). The alignments of mitochondrial genes used in this work were published in https://doi.org/10.1093/gbe/evx025 (Klink and Bazykin46).

**Code availability** The simulated prior distributions used in the current study, the summary statistics calculated from the genomic datasets of vertebrates and insects and source code for the analysis are available at https://github.com/astolyarova/senesence-ABC.

Received: 10 October 2019; Accepted: 20 August 2020; Published online: 14 September 2020

**References**

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Int. Cong. Genet. 1, 356–366 (1932).
2. Smith, J. M. Natural selection and the concept of a protein space. Nature 225, 563–564 (1970).
3. Pal, C. & Papp, B. Evolution of complex adaptations in molecular systems. Nat. Ecol. Evol. 1, 1084–1092 (2017).
4. Fragata, L., Blandaert, A., Dassourov, M. A., Liberles, D. A. & Bank, C. Evolution in the light of fitness landscape theory. Trends Ecol. Evol. 34, 69–82 (2019).
5. Gavrillets, S. Fitness Landscapes and the Origin of Species. (Princeton Univ. Press, 2004).
6. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet. 15, 480–490 (2014).
7. Bazykin, G. A. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. Biol. Lett. 11, 20150315 (2015).
8. Rogozin, I. B., Thomson, K., Cairós, M., Carmel, L. & Koonin, E. V. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov’s law of homologous series. Biol. Direct 3, 7 (2008).
9. Polovolotetskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. Nature 465, 922–926 (2010).
10. Kondrashov, A. S., Polovolotetskaya, I. S., Ivanov, D. N. & Kondrashov, F. A. Rate of sequence divergence under constant selection. Biol. Direct 5, 10 (2010).
11. Usmanova, D. R., Ferretti, L., Polovolotetskaya, I. S., Vlasov, P. K. & Kondrashov, F. A. A model of substitution trajectories in sequence space and long-term protein evolution. Mol. Biol. Evol. 32, 542–554 (2015).
12. Goldstein, R. A., Pollard, S. T., Shah, S. D. & Pollock, D. D. Nonadaptive amino acid convergence rates decrease over time. Mol. Biol. Evol. 32, 1373–1381 (2015).
13. Zou, Z. & Zhang, J. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol. Biol. Evol. 32, 2085–2096 (2015).
14. Klink, G. V. & Bazykin, G. A. Parallel evolution of metazoan mitochondrial proteins. Genome Biol. Evol. 9, 1341–1350 (2017).
15. Klink, G. V., Golovin, A. V. & Bazykin, G. A. Substitutions into amino acids that are pathogenic in human mitochondrial proteins are more frequent in lineages closely related to human than in distant lineages. PeerJ 5, e1413 (2017).
16. Gillespie, J. H. The Causes of Molecular Evolution (Oxford Univ. Press, 1994).
17. Pollock, D. D., Taylor, W. R. & Goldman, N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. 287, 187–199 (1999).
18. Kondrashov, A. S., Sanyave, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl Acad. Sci. USA 99, 14878–14883 (2002).
19. Dimmic, M. W., Hubisz, M. J., Bustamante, C. D. & Nielsen, R. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21, i126–i133 (2005).
20. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 7, e1001301 (2011).
21. de Visser, J. A. G. M., Cooper, T. F. & Elena, S. F. The causes of epistasis. Proc. Natl Acad. Sci. 108, 3617–3624 (2011).
22. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. Nature 490, 535–538 (2012).
23. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. Nature 497, E1–E2 (2013).
24. Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B. & Bazykin, G. A. Coordinated evolution of influenza A surface proteins. PLoS Genet. 11, e1005404 (2014).
25. Pollock, D. D., Thiltgen, G. & Goldstein, R. A. Amino acid coevolution induces an evolutionary Stokes shift. Proc. Natl Acad. Sci. USA 109, E1352–E1359 (2012).
26. Shah, P., McCandlish, D. M. & Plotkin, J. B. Contingency and entrainment in protein evolution under purifying selection. Proc. Natl Acad. Sci. USA 112, E3226–E3235 (2015).
27. Solyomé, O. & Kondrashov, F. A. Estimating the rate of irreversibility in protein evolution. Genome Biol. Evol. 4, 1213–1222 (2012).
28. Naumenko, S. A., Kondrashov, A. S. & Bazykin, G. A. Fitness conferred by replaced amino acids declines with time. Biol. Lett. 8, 825–828 (2012).
29. Risso, V. A. et al. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. Mol. Biol. Evol. 32, 440–455 (2015).
30. Goldstein, R. A. & Pollock, D. D. Sequence entropy of folding and the absolute rate of amino acid substitutions. Nat. Ecol. Evol. 1, 1923–1930 (2017).
31. Gillespie, J. Polymorphism in random environments. Theor. Popul. Biol. 4, 193–195 (1973).
32. Takahata, N., Ishii, K. & Matsuda, H. Effect of temporal fluctuation of selection coefficient on gene frequency in a population. Proc. Natl Acad. Sci. USA 72, 4541–4545 (1975).
33. Huerta-Sanchez, E., Durett, R. & Bustamante, C. D. Population genetics of polymorphism and divergence under fluctuating selection. Genetics 178, 325–337 (2008).
34. Mustonen, V. & Lässig, M. Molecular evolution under fitness fluctuations. Phys. Rev. Lett. 100, 108101 (2008).
35. Mustonen, V. & Lässig, M. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25, 111–119 (2009).
36. Mustonen, V. & Lässig, M. Fitness flux and ubiquity of adaptive evolution. Proc. Natl Acad. Sci. USA 107, 4248–4253 (2010).
37. Mustonen, V. & Lässig, M. Adaptations to fluctuating selection in Drosophila. Proc. Natl Acad. Sci. USA 104, 2277–2282 (2007).

38. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26, 2097–2108 (2009).

39. Benger, E. & Sella, G. Modeling the effect of changing selective pressures on polymorphism and divergence. Theor. Popul. Biol. 85, 73–85 (2013).

40. Cvijovic, I., Good, B. H., Jerison, E. R. & Desai, M. M. The fate of a mutation in a fluctuating environment. Proc. Natl Acad. Sci. USA 112, E5021–E5028 (2015).

41. Popova, A. V. et al. Allele-specific nonstationarity in evolution of influenza A virus surface proteins. Proc. Natl Acad. Sci. USA 116, 21104–21112 (2019).

42. Kimura, M. The Neutral Theory of Molecular Evolution (Cambridge Univ. Press, 1983).

43. Nabieva, E. & Bazykin, G. A. SELVa: simulator of evolution with landscape variation. Preprint at bioRxiv https://doi.org/10.1101/647834 (2019).

44. Proschan, F. Theoretical explanation of observed decreasing failure rate. Technometrics 5, 375–383 (1963).

45. Vaupel, J. W., Manton, K. G. & Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454 (1979).

46. McCandlish, D. M., Shah, P. & Plotkin, J. B. Epistasis and the dynamics of reversion in molecular evolution. Genetics 203, 1335–1351 (2016).

47. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16, 1791–1798 (1999).

48. Kondrashov, F. A. & Kondrashov, A. S. Multidimensional epistasis and the disadvantage of sex. Proc. Natl Acad. Sci. USA 98, 12089–12092 (2001).

49. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59, 1165–1174 (2005).

50. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312, 111–114 (2006).

51. Ferretti, L., Weinreich, D., Tajima, F. & Achaz, G. Evolutionary constraints in fitness landscapes. Heredity 121, 466–481 (2018).

52. Rosenblloom, K. R. et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 43, D670–D681 (2015).

53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591 (2007).

54. Yang, Z., Wong, W. S. W. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118 (2005).

55. Caillière, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3, 475–479 (2012).

Acknowledgements

We thank Dmitrii Ivanov and Alexey Kondrashov for useful comments on drafts of this paper.

Author contributions

G.A.B. conceived the study; A.V.S., E.N., V.V.P., A.V.F., A.D.N., and G.A.B. designed research; A.V.S. and E.N. performed research; V.V.P. developed the formal proof that random changes in SPFL decrease the fitness of the current allele; A.V.S. and A.V.P. analyzed data; and A.V.S., A.V.F., and G.A.B. wrote the paper with comments from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18366-z.

Correspondence and requests for materials should be addressed to A.V.S.

Peer review information

Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permission information

is available at http://www.nature.com/reprints

Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2020