nature portfolio

https://doi.org/10.1038/s41562-022-01342-6

Supplementary information

Measuring frequency-dependent selection in culture

In the format provided by the authors and unedited

Measuring frequency-dependent selection in culture: Supplementary Information

Mitchell G. Newberry[1,2,*], Joshua B. Plotkin[3,*]

[1] Center for the Study of Complex Systems, University of Michigan 48109
[2] Michigan Society of Fellows, University of Michigan 48109
[3] Departments of Biology and Mathematics, University of Pennsylvania, 19104
[*]Correspondence: mgnew@umich.edu or jplotkin@sas.upenn.edu

Contents

Supplementary Tables

Supplementary Figures

1 Inference method

1.1 Type-exchangeable model of frequency-dependent selection

We describe frequency-dependent selection in a type-exchangeable Wright-Fisher model—that is, a model whose behavior is invariant under relabeling of types. We assume that the fitness of any type at a frequency p follows some function $w(p)$, where the fitness w is related to the corresponding selection coefficient s via $w = e^s$. The Wright-Fisher process allows us to explicitly write the probabilities for types to transition from one set of frequencies to another, with the eventual goal of inferring the fitness function $w(p)$ from the time series of observed transitions by maximizing likelihood.

If we let $X_{i,t}$ be a vector random variable of the count of each type i at time t and let the index $i \in 1, ..., k$ run over existing types (temporarily disregarding mutation), the Wright-Fisher process evolves as the stochastic process

$$\boldsymbol{X}_{t+1}|\boldsymbol{X}_t = x_t \sim \text{Multinom}(n_{t+1}, [\pi_1, \pi_2, ..., \pi_k]), \quad \pi_i = \frac{w(x_{i,t}/n_t)x_{i,t}}{\sum_{j=1}^{k} w(x_{j,t}/n_t)x_{j,t}}. \tag{1}$$

Here $n_t = \sum_{i=1}^{k} x_{i,t}$ denotes the total population of all types $1, ..., k$ at time t. According to this equation, $\boldsymbol{X}_{t+1}$ given $\boldsymbol{X}_t$ is a multinomial sample of individuals, each having probability π_i to be of type i.

We note this conditional probability is invariant under relabeling of types: our concept of type-exchangeability. The π_i here depend on $\boldsymbol{x}_t$ only, and further, the multinomial probability mass function to observe $\boldsymbol{x}_{t+1}$ conditioned on $\boldsymbol{x}_t$,

$$\frac{n_{t+1}!}{x_{1,t+1}! \cdots x_{k,t+1}!} \pi_1^{x_{1,t+1}} \cdots \pi_k^{x_{k,t+1}}, \tag{2}$$

depends on π_i and $x_{i,t+1}$ in such a way that permuting the labels on types merely changes the order of the factors

$$\pi_i^{x_{i,t+1}} = \left(\frac{w(x_{i,t}/n_t)x_{i,t}}{\sum_{j=1}^{k} w(x_{j,t}/n_t)x_{j,t}} \right)^{x_{i,t+1}}.$$

This independence on the label ordering would not hold, for example, if any type k had an idiosyncratic fitness w_k, because the factor $\pi_k = w_k x_{k,t}/(w_k x_{k,t} + \sum_{j=1}^{k-1} w(x_{j,t}/n_t)x_{j,t})$ would generally be changed by label permutation. Likewise, if every type had an idiosyncratic fitness w_i we would need to permute the labels on w_i in the same manner as the x_i to retain invariance of Eq. 2.

The probability density Eq. 2 is also unchanged if we multiply the fitness function $w(p)$ by a constant or, equivalently, add a constant to the selection coefficients $s(p)$. The fitness $w(p)$ is proportional to the expected per-capita reproductive output of a type currently at frequency p in the population prior to enforcing any carrying capacity. That is, $w(p)$ is proportional to the intrinsic growth rate and is hence a relative fitness, relative to some arbitrary standard.

1.2 Likelihood maximization

The Wright-Fisher process above (Supplementary Information section 1.1) specifies the probability of observed counts $\boldsymbol{x}_0, ..., \boldsymbol{x}_T$ over T successive transitions between generations as the product of conditional probabilities

$$\Pr(\boldsymbol{X}_T = \boldsymbol{x}_T|\boldsymbol{X}_{T-1} = \boldsymbol{x}_{T-1}) \Pr(\boldsymbol{X}_{T-1} = \boldsymbol{x}_{T-1}|\boldsymbol{X}_{T-2} = \boldsymbol{x}_{T-2}) \cdots \Pr(\boldsymbol{X}_1 = \boldsymbol{x}_1|\boldsymbol{X}_0 = \boldsymbol{x}_0).$$

Each conditional probability follows the multinomial distribution function, $f(x, \boldsymbol{\pi})$

$$\Pr(\boldsymbol{X}_t = \boldsymbol{x}_t|\boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}) = f(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_{t-1})) = \frac{n_t!}{x_{1,t}! \cdots x_{k,t}!} \pi_1(\boldsymbol{x}_{t-1})^{x_{1,t}} \cdots \pi_k(\boldsymbol{x}_{t-1})^{x_{k,t}}, \tag{3}$$

where $n_t = \sum_{i=1}^{k} x_{i,t}$ is the total population size in generation t. We write $\boldsymbol{\pi}(\boldsymbol{x}_{t-1})$ to emphasize that π_i depends on counts $\boldsymbol{x}$ (Eq. 1). Viewed as a function of the observed counts $\boldsymbol{x}$, f is the distribution function, whereas as a function of parameters, f is the likelihood function. Thus, given a realization of the process $\boldsymbol{x}_t$, the likelihood is $\prod_{t=1}^{T} f(\boldsymbol{x}_t, \boldsymbol{\pi}(\boldsymbol{x}_{t-1}))$. When $w(p) = w(p|\boldsymbol{\theta})$ can be parameterized by a vector of smoothly-varying parameters $\boldsymbol{\theta}$, we can infer the parameters using the method of maximum likelihood, up to an arbitrary constant factor.

Finding the optimal parameter set, in particular when $w(p|\boldsymbol{\theta})$ may have many parameters, can be challenging. When the log-likelihood is concave, good optimization algorithms are available. Taking the log of Eq. 3 gives the log likelihood function

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \ln n_t! - \sum_{i=1}^{k} \ln x_{i,t}! + \sum_{i=1}^{k} x_{i,t} \ln \pi_i(\boldsymbol{x}_{t-1}, \boldsymbol{\theta}), \tag{4}$$

where we now write $\pi_i(\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ to emphasize that by Eq. 1, π_i depends on both $\boldsymbol{x}$ and on the parameters $\boldsymbol{\theta}$ that determine the function $w(p)$. To compute the derivative of the log likelihood, we first write Eq. 4 in terms of w and bundle terms that are not a function of the parameters into a constant c. As the conditional likelihoods always involve a transition from one step to the next, we use a prime ($'$) to denote the succeeding timestep, so that $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ are written $\boldsymbol{x}'$ and $\boldsymbol{x}$ respectively, and likewise $n' = n_t = \sum_{i=1}^{k} x_i'$ and $n = n_{t-1} = \sum_{i=1}^{k} x_i$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}'|\boldsymbol{\theta}) &= \ln n'! - \sum_{i=1}^{k} \ln x_i'! + \sum_{i=1}^{k} x_i' \ln \frac{x_i w(x_i/n|\boldsymbol{\theta})}{\sum_j x_j w(x_j/n|\boldsymbol{\theta})} \\
&= c + \sum_{i=1}^{k} x_i' \ln \frac{x_i w(x_i/n|\boldsymbol{\theta})}{\sum_j x_j w(x_j/n|\boldsymbol{\theta})} \\
&= c + \sum_{i=1}^{k} x_i' \ln(x_i w(x_i/n|\boldsymbol{\theta})) - \sum_{i=1}^{k} x_i' \ln \left(\sum_{j=1}^{k} x_j w(x_j/n|\boldsymbol{\theta}) \right) \\
&= c + \sum_{i=1}^{k} x_i' \ln(x_i w(x_i/n|\boldsymbol{\theta})) - n' \ln \left(\sum_{j} x_j w(x_j/n|\boldsymbol{\theta}) \right).
\end{aligned} \tag{5}$$

Setting the derivative of the log likelihood with respect to $\boldsymbol{\theta}$ to zero gives local extrema of the likelihood. The derivative is

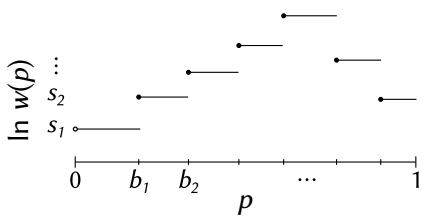
$$\begin{aligned}
&\frac{d}{d\boldsymbol{\theta}} \left(\sum_i x_i' \ln(x_i w(x_i/n, \boldsymbol{\theta})) - n' \ln \left(\sum_i x_i w(x_i/n, \boldsymbol{\theta}) \right) \right) \\
&= \sum_i x_i' \frac{\frac{d}{d\boldsymbol{\theta}} w(x_i/n|\boldsymbol{\theta})}{w(x_i/n|\boldsymbol{\theta})} - n' \frac{\sum_i x_i \frac{d}{d\boldsymbol{\theta}} w(x_i/n|\boldsymbol{\theta})}{\sum_i x_i w(x_i/n|\boldsymbol{\theta})}
\end{aligned}$$

and hence extrema exist when

$$\sum_i x_i' \frac{\frac{d}{d\boldsymbol{\theta}} w(x_i/n|\boldsymbol{\theta})}{w(x_i/n|\boldsymbol{\theta})} = n' \frac{\sum_i x_i \frac{d}{d\boldsymbol{\theta}} w(x_i/n|\boldsymbol{\theta})}{\sum_i x_i w(x_i/n|\boldsymbol{\theta})}. \tag{6}$$

We parameterize frequency-dependent fitness by assuming that $w(p)$ is a piecewise-constant function of p, where the parameters $\boldsymbol{\theta} = \{s_d\}$, $d \in 1, ..., D$ are constant selection coefficients $\{s_d\}$ associated with a set of frequency intervals $\{(l_d, u_d]\}$ that are mutually-exclusive and exhaustive over the domain $(0, 1]$ of $w(p)$ (Supplementary Figure 1). Any given parameterization specifies D frequency bins with $D - 1$ boundaries in $(0, 1)$ labeled $b_1, ..., b_{D-1}$, so that

$$w(p|\boldsymbol{s}) = e^{s_{\mathcal{B}(p)}} = \begin{cases} e^{s_1} & p \in (0, b_1] \\ e^{s_2} & p \in (b_1, b_2] \\ \vdots & \vdots \\ e^{s_D} & p \in (b_{D-1}, 1] \end{cases} \tag{7}$$



Supplementary Figure 1: We let $w(p)$ be a piecewise constant function with discontinuities at b_d, and selection coefficient s_d within each bin d. We let $\mathcal{B}(p)$ indicate the index d such that $p \in (l_d, u_d] = (b_{d-1}, b_d]$.

For notational simplicity we let $b_D = 1$ and $b_0 = 0$ so that $(b_0, b_D]$ is the full domain of $w(p)$, and we use $\mathcal{B}(p)$ to indicate the index d such that $p \in (l_d, u_d] = (b_{d-1}, b_d]$. We thereby infer the most likely fitness of types within each s_d frequency interval.

This $w(p)$ is a discrete frequency-dependent fitness function which is a good approximation to continuous functions given a sufficient number of bins. The bin boundaries themselves are an arbitrary choice and cannot be inferred using maximum likelihood. Using finer frequency intervals gives successively better approximations until too little data is available within each interval to get sufficiently precise parameter estimates. This piecewise constant approximation to the true relationship is analogous to image reconstruction: the resolution limit on frequency bins is analogous to the trade-off between spatial precision and noise in images[84].

What remains is to specialize the maximum-likelihood condition (Eq. 6) to our parameterization of $w(p)$ (Eq. 7), to introduce mutation, and to solve for the maximum-likelihood parameters given an observed time series.

We deal first with mutation. The frequency 0 is not in the domain of $w(p)$. Types that are present in the current generation but were absent— i.e. at frequency 0—in the previous generation must have arrived through some process such as innovation, immigration or recollection. The infinite-alleles Wright-Fisher process provides an accounting for new types, by introducing new mutants stochastically at a constant rate μ per individual per generation. Each mutant is assumed to be of a completely new type and given a new type identity so as to begin at initial frequency $1/N$ where N is the total population size. In data, however, types not seen in the preceding generation may appear at any frequency. We count the total of all such novel individuals appearing in a given generation and call it m_t. We identify these novel individuals with mutants in the infinite-alleles Wright-Fisher process, and we show how to produce a maximum-likelihood estimate $\hat{\mu}$ of μ.

We call these types "mutants" and μ the "mutation rate" in homage to the tradition of population genetics. But this is only a shorthand: we interpret μ to be the rate of individuals with novel type appearing in the data at whatever initial frequency by whatever process (innovation, spelling modification, immigration, proliferation within the data collection interval, etc.), assumed to be independent of the current composition of the population. It is tempting in the context of this study to identify $\hat{\mu}$ with an estimate of some rate of cultural innovation, but our μ conflates many processes generating new types in a population – including for example recalling words from a bygone era.

We will show that the maximum-likelihood estimate for the mutation rate in the infinite alleles Wright-Fisher process makes reference only to the total number of mutant individuals (regardless of how many types at what counts) and is decoupled from selection estimates. This forms a justification for why the selection estimates are irrelevant to the mutational process, allowing us to model selection even when the true mutational process is unknown and differs from the model we use for inference. Furthermore, the mutants (as well as censored types, treated separately in Supplementary Information section 1.10) are part of the population and thus are necessary to determine the frequency of other types as a function of their count. The mutational process also determines the rate at which existing types are diluted by incoming types, which in turn influences the selection coefficient required to stay

at replacement rate in the population, and so estimates of mutation rate will be useful for filling in the missing zero point of relative fitness.

We modify the conditional likelihood expression to accommodate mutation, incorporating m_t observed new mutants and the parameter μ. Again focusing on the transition from generation $t-1$ to t, we use the prime notation ($'$) to denote the later generation, but in so doing we take on a change in semantics: absent mutation, we denoted total population size at time t as $n_t = n' = \sum_{i=1}^{k} x'_i$, where indices $1, ..., k$ enumerate the types present at time $t-1$. When mutations occur, however, n_t cannot simultaneously represent both the total population size at time t and also the sum of pre-existing types, because the latter may exceed the former. We choose to let n_t denote the total population size of all individuals, mutant or otherwise, and retain the definition $n' = \sum_{i=1}^{k} x'_i$, so that n' counts those individuals at time t that arose from types present in the previous generation. We define $m' = m_t$, and hence $n_t = n' + m'$ is the total population size of all individuals at time t, including both pre-existing types and new mutants. The new multinomial conditional log likelihood for this transition between generations, analogous to Eq. 4, is given by

$$\mathcal{L}(\boldsymbol{x}'|\boldsymbol{s}, \mu, \boldsymbol{x}) = \ln(n'+m')! - \ln m'! - \sum_{i=1}^{k} \ln x'_i! + m' \ln \mu + \sum_{i=1}^{k} x'_i \ln \frac{(1-\mu) x_i w(x_i/n)}{\sum_{j=1}^{k} x_j w(x_j/n)} \quad (8)$$

The equivalence to Eq. 4 can be seen by imagining some $x'_0 = m'$ and a corresponding $\pi_0 = \mu$, rewriting the sums from $i = 0$ to k, and multiplying the $\pi_{i\neq 0}$ by $(1-\mu)$ to retain unity in the sum of probabilities. After rearranging terms and simplifying,

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}'|\boldsymbol{s}, \mu, \boldsymbol{x}) &= \ln(n'+m')! - \ln m'! - \sum_{i=1}^{k} \ln x'_i! \\ &\quad + m' \ln \mu + n' \ln(1-\mu) \\ &\quad + \sum_{i=1}^{k} x'_i \ln(x_i w(x_i/n|\boldsymbol{\theta})) - n' \ln \left(\sum_{j=1}^{k} x_j w(x_j/n|\boldsymbol{\theta}) \right). \end{aligned} \quad (9)$$

Considering only terms involving $\boldsymbol{s}$, this log likelihood is equal to Eq. 5: The additional terms involving m' and μ do not affect derivatives with respect to the parameters $\boldsymbol{s}$, and hence Eq. 6 applies with respect to $\boldsymbol{s}$ regardless of the presence or absence of mutation. The full set of parameters $\boldsymbol{\theta}$ however includes the selection coefficients for each frequency bin as well as the mutation rate, which we write $\boldsymbol{\theta} = (\mu, s_1, s_2, \ldots, s_D)$.

We maximize the likelihood by setting the derivatives with respect to all parameters equal to zero. This results in a system of equations. The equation for μ does not involve any other parameters and thus μ can be estimated separately.

$$\frac{\partial}{\partial \mu} \mathcal{L} = \frac{m'}{\mu} - \frac{n'}{1-\mu} = 0, \quad (10)$$

This implies that the maximum-likelihood estimate $\hat{\mu} = m'/(n'+m')$ is simply the observed fraction of mutants. Over the time series then, $\frac{\partial}{\partial \mu} \sum_{t=1}^{T} \mathcal{L}_t = 0$ implies

$$\hat{\mu} = \frac{\sum_{t=1}^{T} m_t}{\sum_{t=1}^{T} n_t}. \quad (11)$$

The best estimate of mutation rate per individual per generation is simply the observed fraction of mutants over the entire time series.

The condition for the maximum-likelihood parameter $\{\hat{s}_d\}$, then, is only the set of equations $\frac{\partial}{\partial s_d} \mathcal{L} = 0$ for all $d \in \{1, ..., D\}$. Thankfully, the derivatives $dw/d\boldsymbol{\theta}$ that appear in Eq. 6 are simply:

$$\frac{\partial}{\partial s_d} w(p) = \begin{cases} e^{s_d} & \mathcal{B}(p) = d \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

Hence we can rewrite Eq. 6 using Eq. 7, $w(x_i/n|\boldsymbol{\theta}) = e^{s_{\mathcal{B}(x_i/n)}}$, and include terms e^{s_d} for $\frac{\partial}{\partial s_d} w(x_i/n|\boldsymbol{\theta})$ whenever x_i/n is in bin d. The factors $(\frac{\partial}{\partial s_d} w(x_i/n))/w(x_i/n)$ in the left side of Eq. 6 are 1 when $\mathcal{B}(x_i/n) = d$ and 0 otherwise: They serves only to indicate for what i to count x_i' in the sum. Writing the maximum likelihood equations then mostly involves notational bookkeeping about which terms to include in which sums. Over a single transition between subsequent generations, this amounts to the system of simultaneous equations

$$\sum_{i:\mathcal{B}(x_i/n)=d} x_i' = n' \sum_{i:\mathcal{B}(x_i/n)=d} \frac{x_i e^{s_d}}{\sum_{j=1}^{k} x_j e^{s_{\mathcal{B}(x_j/n)}}} \quad d \in \{1, ..., D\}. \tag{13}$$

Because the log likelihood of the full time series is the sum of the conditional log likelihoods, and because the derivative is a linear operator, the maximum likelihood equations for the full time series are simply the sum of the Eqs. 13 over all transitions, namely

$$\sum_{t=1}^{T} \left(\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} x_{i,t} \right) = \sum_{t=1}^{T} \left[(n_t - m_t) \left(\frac{\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} e^{s_d} x_{i,t-1}}{\sum_{j=1}^{k} e^{s_{\mathcal{B}(x_{j,t-1}/n_{t-1})}} x_{j,t-1}} \right) \right] \tag{14}$$

for each $d \in 1, ..., D$. This statement for the maximum-likelihood estimator $\hat{\boldsymbol{s}}$, which satisfies these equations, has the interpretation of "observed equals expected": the left hand side is the observed counts emanating from the dth frequency bin and the right side is the expected count given the population configuration at time $t-1$ and number of mutations that occurred between $t-1$ and t. That is, setting the expected data equal to the observed data maximizes the likelihood.

The likelihood expressions already reveal features of the relationships between the parameters μ, $\boldsymbol{s}$, and N. First, the mutation rate is independent of the parameters $\boldsymbol{s}$. This is useful since the mechanisms of mutation are often unknown or ill-defined in cultural contexts. Second, the derivatives of likelihood do not depend on N except where N normalizes counts to frequencies. Thus, the inference depends on the frequencies of types, but not on the total population size or its variation over time.

For a single transition between subsequent generations, the system of equations for the maximum-likelihood parameters (Eq. 13) is linear in $e^{s_1}, \ldots, e^{s_D}$. Nonetheless, the system of equations for the full time series (Eq. 14) is difficult to solve, numerically or otherwise, primarily because all s_j appear in multiple different denominators on the right hand side, which makes the equation for each $\hat{s}_d$ non-linearly dependent on all other $\hat{s}_j$ in the general case.

To find solutions to the system of equations (Eq. 14) we use an MM or "Minorize and Maximize" algorithmic strategy[84]. The strategy is to replace the likelihood $\mathcal{L}(\boldsymbol{s})$ of the parameter vector $\boldsymbol{s}$ with a family of surrogate minorant functions $g(\boldsymbol{s}|\boldsymbol{s}^{(i)})$ where $\boldsymbol{s}^{(i)}$ is the ith iterate designed to approximate the maximum likelihood estimate $\hat{\boldsymbol{s}}$ as i becomes large. The minorant needs two properties: $g(\boldsymbol{s}^{(i)}|\boldsymbol{s}^{(i)}) = \mathcal{L}(\boldsymbol{s}^{(i)})$ and $g(\boldsymbol{s}|\boldsymbol{s}^{(i)}) \leq \mathcal{L}(\boldsymbol{s})$. The minorant is guaranteed to be at most $\mathcal{L}$ everywhere and to match $\mathcal{L}$ and its derivatives with respect to s_d at the point $\boldsymbol{s}^{(i)}$. The consequence is that either $\boldsymbol{s}^{(i)}$ is the optimum of $\mathcal{L}$, in which case the derivatives of g and $\mathcal{L}$ at $\boldsymbol{s}^{(i)}$ are zero, or else, providing that g and $\mathcal{L}$ are smooth, the derivative of g is non-zero and the optimum of g lies somewhere between $g(\boldsymbol{s}^{(i)}|\boldsymbol{s}^{(i)})$ and the optimum of $\mathcal{L}$. If we chose a g that is itself easy to optimize, we can monotonically approach the optimum of $\mathcal{L}$ by finding the optimum $\boldsymbol{s}$ for each minorant $g(\boldsymbol{s}|\boldsymbol{s}^{(i)})$ and choosing each successive optimum to be $\boldsymbol{s}^{(i+1)}$. Each iteration finds a parameter set $\boldsymbol{s}^{(i)}$ that closes the gap between $\mathcal{L}(\boldsymbol{s}^{(i)})$ and the true optimum of $\mathcal{L}(\hat{\boldsymbol{s}})$.

The strategy requires a creative choice of an appropriate minorant. Many minorants are possible, such as a quadratic approximation[86]. Examining the likelihood function Eq. 5, we notice that the very last term,

$$-n' \ln \left(\sum_i x_i w(x_i/n) \right) = -n' \ln \left(\sum_i x_i e^{s_{\mathcal{B}(x_i/n)}} \right)$$

involves a sum inside the log, which gives rise to the troublesome denominator that makes Eq. 14 nonlinear in the e^{s_i}. This term is a convex function of $w(p|\boldsymbol{s})$ and hence it is minorized by its linear-order Taylor expansion in w around $w(p|\boldsymbol{s}^{(i)})$:

$$-n' \ln\left(\sum_j x_j e^{s^{(i)}_{\mathcal{B}(x_j/n)}}\right) - n' \sum_{d=1}^{D}\left(\left(e^{s_d} - e^{s_d^{(i)}}\right)\frac{\partial}{\partial e^{s_d}}\left[\ln\left(\sum_{j=1}^{k} x_j e^{s_{\mathcal{B}(x_j/n)}}\right)\right]_{\boldsymbol{s}=\boldsymbol{s}^{(i)}}\right) \quad (15)$$

Substituting this linear minorant for the convex term in Eq. 5 produces a surrogate likelihood function $g(s|s^{(i)})$. We bundle terms that are not dependent on $\boldsymbol{s}$ into a constant c, which includes those terms that involve only $\boldsymbol{s}^{(i)}$ or μ, n, et cetera:

$$g(\boldsymbol{s}|\boldsymbol{s}^{(i)}) = c + \sum_{d=1}^{D}\left[\sum_{j:\mathcal{B}(x_j/n)=d} x'_j s_d - n' e^{s_d}\left(\frac{\sum_{j:\mathcal{B}(x_j/n)=d} x_j}{\sum_{j=1}^{k} x_j e^{s^{(i)}_{\mathcal{B}(x_j/n)}}}\right)\right] \quad (16)$$

The optimum $\boldsymbol{s}$, obtained by setting derivatives with respect to s_d to zero, is then given by the system of equations analogous to Eq. 13:

$$\sum_{j:\mathcal{B}(x_j/n)=d} x'_j = n' e^{s_d}\left(\frac{\sum_{j:\mathcal{B}(x_j/n)=d} x_j}{\sum_{j=1}^{k} x_j e^{s^{(i)}_{\mathcal{B}(x_j/n)}}}\right), \quad d \in \{1, ..., D\}. \quad (17)$$

Likewise, because a sum of conditional log likelihoods is minorized by a sum of their respective minorants, Eq. 14 has its analog, for $d \in \{1, ..., D\}$,

$$\sum_{t=1}^{T}\left(\sum_{j:\mathcal{B}(x_{j,t-1}/n_{t-1})=d} x_{j,t}\right) = e^{s_d}\sum_{t=1}^{T}\left[(n_t - m_t)\left(\frac{\sum_{j:\mathcal{B}(x_{j,t-1}/n_{t-1})=d} x_{j,t-1}}{\sum_{j=1}^{k} x_{j,t-1} e^{s^{(i)}_{\mathcal{B}(x_{j,t-1}/n_{t-1})}}}\right)\right]. \quad (18)$$

Strikingly, the only difference between the systems Eqs. 14 and 18 is that the s_j in the denominator of the right side have been replaced by $s_j^{(i)}$. This difference is crucial, however, as the solution for $\{s_d\}$ in the system of equations Eq. 18 is now trivial to solve, which produces the next iterate:

$$\begin{aligned} s_d^{(i+1)} = &\ln\left[\sum_{t=1}^{T}\left(\sum_{j:\mathcal{B}(x_{j,t-1}/n_{t-1})=d} x_{j,t}\right)\right] \\ &- \ln\left(\sum_{t=1}^{T}\left[(n_t - m_t)\left(\frac{\sum_{j:\mathcal{B}(x_{j,t-1}/n_{t-1})=d} x_{j,t-1}}{\sum_{j=1}^{k} x_{j,t-1} e^{s^{(i)}_{\mathcal{B}(x_{j,t-1}/n_{t-1})}}}\right)\right]\right). \end{aligned} \quad (19)$$

Here the right side of the equation involves only the previous iterate $\boldsymbol{s}^{(i)}$.

We initialize our MM algorithm by setting $s_d^{(0)} = 0$ for $d \in \{1, ..., D\}$ and we iterate until no further change in $\boldsymbol{s}^{(i)}$ is achieved between iterations. That is, the derivative of the minorant at $\boldsymbol{s}^{(i)}$ has reached approximately zero. At this point, the minorant and $\mathcal{L}$ are both maximized.

As noted previously, the likelihood of the data are unchanged by adding a constant to the selection coefficient s_d in each frequency bin d. We enforce the condition $\sum_{d=1}^{D} s_d = 0$ to achieve a unique maximum-likelihood parameter set. We enforce the constraint by adding the appropriate constant at each iteration. (Absent any enforcement of the constraint, the MM iterates settle on an arbitrary value of $\hat{\boldsymbol{s}}$ that indeed satisfies Eq. 14, but the particular value $\hat{\boldsymbol{s}}$ depends on the initial condition $\boldsymbol{s}^{(0)}$.)

1.3 Binning strategies

The bin boundaries $b_1 ... b_{D-1}$ described in Supplementary Information section 1.2 define the space of possible functions $s(p)$ from which maximizing likelihood selects the optimum.

Choosing a space of functions capable of representing the true form of frequency-dependent selection is an open-ended design problem that depends to some extent on the researcher's intuition and goals, as well as the quality and quantity of data. However, some general principles apply.

The width (frequency range) of a bin represents a trade-off between frequency precision and the precision of the estimate of selection within the bin. Where there is more total data, more precision is available in estimates of selection and narrower ranges of frequency (incorporating fewer data points) are admissible; and so more bins can divide the same overall frequency range. How to allocate the available precision to selection versus frequency depends to some degree on the research question.

Care should be taken so that each bin includes several types. Under a model of pure frequency dependence, types all behave identically and even a single type would be sufficient to estimate the selection on all types. In reality however, types may experience idiosyncratic selection and so it would be faulty generalization to make claims about selection at a given frequency when only one type existed at that frequency in the data. The number of types required to make reliable claims about the average selection in turn depends on the variance in selection on types. Since it takes at least three data points to estimate a variance at all, inference from bins containing fewer than three types should be considered suspect.

Likewise, conclusions should not depend on the precise location of the bin boundaries. If the inclusion or exclusion of a certain type under strong idiosyncratic selection significantly affects the inference for a particular bin, then the conclusion represents that particular type rather than generalities of types at that frequency. Conclusions that only occur when using some particular binning are suspect of "p-hacking"[87], particularly if many binning schemes were attempted and rejected. A simple control for sensitivity to binning is to perturb the bin boundaries or to re-infer under different binning schemes to ensure the conclusions do not depend on the bin boundaries or binning scheme.

Instead of choosing individual bin boundaries by hand, which would increase "researcher degrees of freedom"[88], it is preferable to adopt a binning scheme that specifies bin boundaries as a function of one or a few parameters. For example, Hahn and Bentley[1] bin name frequencies by powers of two times 10^{-4}. This scheme is equally appropriate for measuring frequency-dependent selection, and it resembles Preston's octave classes used in evaluating count abundance data in ecology[23,89].

Evenly dividing the data into D bins also offers a simple prescription that affords the researcher one parameter D to control the trade-off between precision in frequency and precision in the selection estimate. We consider two such schemes, either even division of the frequency range of types in log space—*logarithmic binning*—or else even division of quantiles of type frequencies—*quantile binning*—which allocates roughly the same number of type frequency transitions to each bin. Both of these schemes fully determine the bin boundaries from the frequency distribution (or range) of the data and the number of bins, D. Even division in linear space is also possible, but inappropriate for evolutionary systems: Changes due to selection and drift in typical evolutionary models[90] both occur at magnitudes proportional to $p(1-p)$ for frequency p, and so the same linear range of p might represent a short timescale of evolution for $p \approx 1/2$ and a very long timescale for $p \ll 1/2$. Logarithmic binning on the other hand corresponds to equal proportional changes in p, which approximates $p(1-p)$ for $p \ll 1$.

An ideal binning scheme might evenly allocate the available precision over D bins. Nonetheless quantile binning sometimes performs poorly at this task (cf. Supplementary Figure 3). The reason is that frequency transitions with higher proportional variance are less informative of the underlying mean selection. Proportional variance depends on frequency in both the model and in empirical data, with data points of rare types being less informative. Often, however, ever more types are present at ever lower frequencies, and so logarithmic binning does a good job of evenly allocating precision over bins despite allocating more types to lower-frequency bins. Which binning scheme gives the best results depends on the distribution of type frequencies in the data, but even the simple octave binning scheme performs well on a wide variety of data.

1.4 Regularity conditions

The piecewise-constant parameterization of $s(p)$ contains no *a priori* information about relationships between the parameters s_i. In other words, we do not assume that $s(p)$ approximates a smooth function, or that adjacent values s_i and s_{i+1} should be similar. Making fewer prior assumptions about the form of frequency dependence is a conservative choice. As a result, though, noise in the estimators will result in more jagged inferred $\hat{s}(p)$ curves, especially with finer bins or less data, as reflected in wider confidence intervals on the level of selection within each bin.

Regularity conditions, such as limiting the variability between s_i and s_{i+1}, could enforce smoother inferred $\hat{s}(p)$ curves, limiting the effect of noise on the inference, and may be warranted if much is known about the underlying selection. This topic may be an important area for future research, especially for application to smaller datasets. We do not implement regularity conditions here. If inferred curves of $\hat{s}(p)$ are smooth, despite the lack of regularity conditions, it indicates a property of the data.

1.5 Replacement fitness

The absence of selection ($s \equiv 0$) does not imply that types will tend to maintain the same frequency over time, because the influx of novel types dilutes existing types. And so, we define the concept of replacement fitness, $\bar{w}$, as the fitness that corresponds to zero growth rate in Figure 2 and elsewhere.

The replacement fitness is a property of a time series, and it is the fitness, which, if a type with that (constant, frequency-independent) fitness were introduced at the beginning of the time series, its expected frequency at the end of the time series would would equal its initial frequency.

The replacement fitness (actually, replacement selection coefficient) for generations $t \in \{1, ..., T\}$ and counts $x_{i,t}$ of types $i \in \{1, ..., k_t\}$ at generation t is then

$$\bar{s} = \left(\frac{1}{T}\right) \sum_{t=1}^{T} \log \left(\mu + \sum_{i=1}^{k_t} (x_{i,t}/n_t) e^{s_{\mathcal{B}(x_{i,t}/n_t)}} \right). \tag{20}$$

1.6 Average frequency-dependent selection

We can define the notion of average frequency-dependent selection for a realization of a arbitrary selective process, even if the process is not exchangeable. Suppose that the exact fitness $w_{i,t}$ is known for each type i and time t. That is, in a Wright-Fisher process (cf. Eq. 1), the expected frequencies of type i in the next generation are:

$$\mathrm{E}\left(\frac{X_{i,t+1}}{n_{t+1}}\right) = \pi_{i,t} = \frac{w_{i,t} x_{i,t}}{\sum_j w_{j,t} x_{j,t}}. \tag{21}$$

We define the average fitness $\bar{w}_d$ of the types in bin d as the fitness that would match the expected total reproductive output of types in that bin. This yields an implicit formula for $\bar{w}_d$ by equating the expectations under average fitness over a single generation:

$$\frac{\sum_{i:\mathcal{B}(x_i/n)=d} \bar{w}_d x_{i,t}}{\sum_j w_{j,t} x_{j,t}} = \frac{\sum_{i:\mathcal{B}(x_i/n)=d} w_{i,t} x_{i,t}}{\sum_j w_{j,t} x_{j,t}}. \tag{22}$$

We generalize this definition to multiple generations by viewing each timestep as an independent realization of the reproductive output of whatever types exist in bin d, conducting an arithmetic average over time. We define Z_t to be the population mean fitness,

$$Z_t = \sum_j w_{j,t} \frac{x_{j,t}}{n_t}, \tag{23}$$

and use it to rewrite Eq. 22. The sum over time becomes:

$$\sum_{t=0}^{T-1} \sum_{i:\mathcal{B}(x_i/n)=d} \left(\frac{\bar{w}_d}{Z_t}\right)\left(\frac{x_{i,t}}{n_t}\right) = \sum_{t=0}^{T-1} \sum_{i:\mathcal{B}(x_i/n)=d} \left(\frac{w_{i,t}}{Z_t}\right)\left(\frac{x_{i,t}}{n_t}\right). \tag{24}$$

This formula makes it clear that the definition of $\bar{w}_d$ is independent of the population size history $\{n_t\}$ given the $\{w_{i,t}\}$ and the frequencies in the data, since the frequency-weighted $w_{i,t}/Z_t$ sum to 1, and the counts $x_{i,t}$ enter formulas only as frequencies $x_{i,t}/n_t$.

We can easily solve for $\bar{w}_d$, as it is constant with respect to all the sum indices. We call this $\bar{w}_d$ the realized average fitness of bin d for a given time series $\{x_{i,t}\}$:

$$\bar{w}_d \equiv \frac{\sum_{t=0}^{T-1} \sum_{i:\mathcal{B}(x_i/n)=d} \frac{w_{i,t}x_{i,t}}{Z_t n_t},}{\sum_{t=0}^{T-1} \sum_{i:\mathcal{B}(x_i/n)=d} \frac{x_{i,t}}{Z_t n_t}} \tag{25}$$

The average selection coefficient within bin d is simply $\log \bar{w}_d$. In order to set the zero point of selection relative to the replacement fitness rather than the population average fitness, it may be helpful to subtract the replacement fitness $\bar{s}$. We then take the average frequency-dependent selection coefficient to be

$$\bar{s}_d = \ln \bar{w}_d - \bar{s}. \tag{26}$$

1.7 Wildtype fitness

Sometimes we do not know exact counts of every type in the population or we want to consider frequency-dependent fitness only within a subpopulation such as biblical or non-biblical names. In these cases we make use of a wildtype subpopulation with associated average fitness s_w, and treat all types residing in that subpopulation (if they are known) as residing in their own bin regardless of their frequency. When counts are censored, for example, we do not know exact counts for some types, but we may still know or estimate the total number of censored individuals. Since the sum of types in a given bin in Eqs. 14 is the same whether these types are labeled individually or all belong to an aggregate type, we rewrite Eqs. 14 in terms of aggregate sums of wildtype individuals w_t and w'_t as follows for frequency range d,

$$\sum_{t=1}^{T} \left(\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} x_{i,t} \right) = \sum_{t=1}^{T} \left[(n_t - m_t) \left(\frac{\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} e^{s_d} x_{i,t-1}}{w_{t-1}e^{s_w} + \sum_{j=1}^{k} e^{s_{\mathcal{B}(x_{j,t-1}/n_{t-1})}} x_{j,t-1}} \right) \right] \tag{27}$$

taken simultaneously with the following equation for the wildtype fitness s_w,

$$\sum_{t=1}^{T} \left(w'_{t-1}\right) = \sum_{t=1}^{T} \left[(n_t - m_t) \left(\frac{e^{s_w} w_{t-1}}{w_{t-1}e^{s_w} + \sum_{j=1}^{k} e^{s_{\mathcal{B}(x_{j,t-1}/n_{t-1})}} x_{j,t-1}} \right) \right]. \tag{28}$$

The wildtype aggregation w_{t-1} is the total number of wildtype individuals at generation $t-1$, and w'_{t-1} is the total number of generation-t progeny of wildtype individuals in generation $t-1$. One might expect w'_{t-1} to equal w_t, yet the two accountings may differ. As an example suppose we let the wildtype represent types at frequency below f: if a type transitions from below f at generation $t-1$ to above f at generation t, it contributes to w_{t-1} and w'_{t-1} but not to w_t. The aggregates support an exact analogy between the wildtype counts w_t, wildtype progeny w'_t and wildtype fitness s_w and the counts, $\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} x_{i,t-1}$, progeny $\sum_{i:\mathcal{B}(x_{i,t-1}/n_{t-1})=d} x_{i,t}$, and fitness s_d of frequency-range d. Thus the wildtype aggregate acts as a frequency-independent bin, and we estimate its fitness in the same way as we estimate fitness for types within a frequency bin.

1.8 Subpopulations

Wildtype aggregations (Supplementary Information section 1.7) further allow measurement of frequency-dependent selection among subpopulations. By treating all individuals who are not part of the focal subpopulation as part of the wildtype aggregation, those individuals do not contribute to inferred frequency-dependent selection, but do contribute to replacement fitness. Hence separate estimates for multiple mutually-exclusive subpopulations will be calibrated against the same replacement fitness.

1.9 Confidence intervals

We use two types of confidence intervals (CIs) in this study: a parametric CI derived from the model and a non-parametric CI based on a bootstrap. The model CI is an accurate theoretical prediction of the variance on the estimator subject to the assumption that the data is generated by the model. The bootstrap CI is the 2.5% to 97.5% quantiles of a non-parametric bootstrap[91] that generates an empirical distribution of the estimator by resampling the underlying data. A bootstrap CI makes no assumption that the data are generated from any particular process, but it does assume that a resampling scheme is available that generates equally probable samples, such as multinomially resampling the observed dataset. Either of these assumptions may fail in practice. Corroboration between the two CIs, as we observe in simulations where the underlying model is known (cf. Supplementary Figure 2), supports the accuracy of the model and the underlying assumptions. When the CIs disagree, the wider CI is more conservative, as it may account for failures of the model to describe variability in the data.

We compute model confidence intervals on parameter inferences using the observed Fisher information[92], a measure of the curvature of the likelihood function at its maximum. In general, with an identifiable parameter in the limit of many independent observations, the log-likelihood surface is approximately parabolic near its maximum[74]. The observed Fisher information is the matrix of second partial derivatives of the likelihood function (Hessian matrix), evaluated at the most likely parameters, and its inverse is the variance-covariance matrix of the estimators $\hat{s}_i$. This establishes a variance on the estimator $\sigma^2_{\hat{s}_i}$ as simply the dth diagonal element in the variance-covariance matrix. As noted above, we enforce the relationship $\sum_{d=1}^{D} \hat{s}_i = 0$ among the inferred selection coefficients. And so, in practice, the Hessian has dimensions $(D-1) \times (D-1)$, corresponding to $s_1, \ldots s_{D-1}$. We can compute the corresponding variance in the estimator $\hat{s}_D$ using the relationship $\mathrm{var}(\hat{s}_D) = \sum_{d=1}^{D-1} \mathrm{var}(\hat{s}_d) + 2\sum_{0<e<d<D} \mathrm{covar}(\hat{s}_e, \hat{s}_d)$. This variance on the estimators $\hat{s}_i$ arises from the variance of multinomial samples at the census population sizes n_t. The actual variance in frequency trajectories in the data typically differs from the multinomial sampling variance because, for example, the sampling interval differs from one Wright-Fisher generation (Supplementary Information section 1.12). As a second derivative, the estimator variance is linearly related to the variance in frequency trajectories, so the difference between the multinomial variance at census population size and the actual variance in the data leaves a factor left to be determined. Estimating the effective population size N_e (Supplementary Information section 1.12) establishes the unknown factor $g = N/N_e$ by directly estimating the variance in the data. This establishes the 95% model CIs as

$$\hat{s}_i \pm 1.96\sqrt{g\sigma^2_{\hat{s}_i}}. \tag{29}$$

We compute bootstrap CIs by resampling transitions within the time series. Bootstraps that accommodate the dependence relationships in time series is an open-ended topic[93], because generally there is no resampling scheme that logically guarantees resampled data to be equiprobable under the true generating process, in contrast to the bootstrap argument for independent samples. Nonetheless, in our case, the likelihood calculation sums over observations and expectations of change from one timestep to the next, and it is these transitions that are assumed to be independent across time periods and conditionally independent between types. Dependence between observations enters by either failures of the Markovian

assumption in the true generating process, or by effects of competition between types induced by finite population size. We resample in a way that approximates independence and roughly preserves population size so that resampled transitions have similar frequencies to the originals.

We resample a given timestep by choosing at each frequency transitions of a type $x_i \to x_i'$ from that timestep in the data uniformly with replacement. We determine the number of types to include at each timestep by iteratively including another type sample with probability $\max(1, p - c/e - c)$ where c is population size of the types included thus far, p is the observed population size in the original data, and e is the expected population size increase by including one more type sample. This rule guarantees that the expected resampled population size is equal to the observed population size, but allows statistical variation in population size from generation to generation. This resampling scheme resembles the multinomial resampling of independent observations typical of the Efron bootstrap in that types are sampled uniformly with replacement, but it also preserves the frequency interpretation of the original count data. The frequency associated with a given count is allowed to vary around its expectation in the bootstrap, and thus the bootstrap also represents robustness of the estimator to the inclusion or omission of observations due to sampling of the data, since such sampling omissions would also cause the frequencies in observed data to fluctuate. We therefore consider the bootstrap CIs to be conservative estimates of the true 95% CI.

1.10 Accounting for censorship

We might naively assume that time series are complete and uncensored. In the case of baby names however, most datasets censor annual counts below a certain threshold. In the United States (US), for example, the Social Security Administration censors, each year, any name that occurred fewer than five times. To be conservative with respect to biases due to censorship, we do not calculate or report selection coefficients for frequencies so small that they would correspond to censored counts or counts less than one for any generation in a time series. This frequency is

$$f_{\min} = \max_t f_{\min,t}$$

where $f_{\min,t}$ is the minimum frequency that would still be rounded to an uncensored count at generation t, that is $f_{\min,t} = c - 0.5/n_t$ where c is the minimum uncensored count (e.g. $c = 5$ for US or $c = 1$ if there is no censorship).

We do not ascribe counts at frequency below $f_{\min}$ to any frequency range, ascribing them instead to the wildtype aggregation (see Supplementary Information section 1.7). We count types at frequency below $f_{\min}$ among the wildtype—that is the sum w_{t-1} includes all counts $x_{i,t-1}$ of types i existing in the data at generation $t-1$ such that $x_{i,t-1}/n_{t-1} < f_{\min}$. Likewise w'_{t-1} includes both progeny of those wildtype individuals as well as the total count m_t—the sum of all types that did not appear in the data in generation $t-1$—which may be either mutant types or progeny of unobserved types. Censorship therefore conflates true "mutants" with progeny of types at too low frequency to observe. This entails a new interpretation of novel types: rather than consider types appearing for the first time in the dataset as mutants arising from count zero, we simply consider them as types arising from some frequency too low to be observed in the data. Thus ascribing wildtype fitness s_w to all types originating from below $f_{\min}$ accounts for the net competitive pressure of mutants, types of unknown frequency and types at known frequency below $f_{\min}$. Hence no mutation term enters the replacement fitness calculation (Eq. 20) because the affect of mutation on replacement fitness is subsumed into the contribution of the average fitness s_w of wildtype individuals.

Censorship also biases inferred selection coefficients for frequency ranges above $f_{\min}$, because transitions from a recordable frequency $x > f_{\min}$ to a lower frequency may not be representable in the data or may be censored, whereas transitions from x to a higher frequency are always recorded. The consequence for the associated selection coefficient depends on our interpretation of types that disappear from observation. If we "pessimistically" interpret the disappearance of a type from the data as a transition from frequency x to frequency 0, this negatively biases the associated selection coefficient by substituting zero

for its true frequency, which could be anywhere between zero and the minimum observable frequency. This pessimistic interpretation provides a lower bound on the inferred selection coefficient associated with frequency x. The corresponding "optimistic" interpretation or upper bound imputes the disappearance of a type as a transition to the minimum uncensored count—guaranteed to be greater than the maximum unobservable frequency.

We combine these upper and lower bound interpretations for individual transitions to produce "optimistic" and "pessimistic" interpretations of the full frequency-dependent selection curve by interpreting all partially observed transitions either optimistically or pessimistically. These interpretations are not strict upper or lower bounds on the selection coefficients *per se*, because transitions do not affect selection coefficients superpositively: an upper-bound interpretation of a transition from frequency x may lower the selection coefficient for frequency $y \neq x$. However, because the dependence between inferences at different frequencies is typically weak, we treat the optimistic and pessimistic interpretations as approximate upper and lower bounds on all selection coefficients.

To compute the optimistic and pessimistic interpretations of the censored counts, we adjust Eq. 14 to account for types at frequency below $f_{\min}$ as wildtype and to impute pessimistic and optimistic type counts. We use the following expression, which differs from Eq. 14 in that m_t is subsumed into w'_t, n_t has been replaced with $\hat{n}_t$, and some $x_{i,t}$ have been replaced with $\hat{x}_{i,t}$ as described below:

$$\sum_{t=1}^{T}\left(\sum_{i:\mathcal{B}(x_{i,t-1}/\hat{n}_{t-1})=d}\hat{x}_{i,t}\right)=\sum_{t=1}^{T}\left[\hat{n}_t\left(\frac{\sum_{i:\mathcal{B}(x_{i,t-1}/\hat{n}_{t-1})=d}e^{s_d}x_{i,t-1}}{w_{t-1}e^{s_w}+\sum_{j=1}^{k}e^{s_{\mathcal{B}(x_{j,t-1}/\hat{n}_{t-1})}}x_{j,t-1}}\right)\right]$$

taken simultaneously with the following equation for the wildtype fitness s_w,

$$\sum_{t=1}^{T}\left(w'_{t-1}\right)=\sum_{t=1}^{T}\left[\hat{n}_t\left(\frac{e^{s_w}w_{t-1}}{w_{t-1}e^{s_w}+\sum_{j=1}^{k}e^{s_{\mathcal{B}(x_{j,t-1}/\hat{n}_{t-1})}}x_{j,t-1}}\right)\right].$$

The hatted quantities differ from what they replace as follows. The $\hat{n}_t$ denotes the imputed total population size including individuals that are censored from the data, either by including known aggregate censored counts, independent measurements of total population size, or by imputing the full, uncensored population size. The $\hat{x}_{i,t}$ is simply $x_{i,t}$ if it is present in the dataset, otherwise it is the imputed count that depends on the optimistic or pessimistic interpretation. Under the pessimistic imputation, $\hat{x}_{i,t}=x_{i,t}=0$. Under the optimistic imputation, $\hat{x}_{i,t}=c$ where c is the minimum uncensored count possible in the dataset at that generation. It must be the case for all t that

$$w'_{t-1}+\sum_{d\in 1,\ldots,k_t}\sum_{i:\mathcal{B}(x_{i,t-1}/\hat{n}_{t-1})=d}\hat{x}_{i,t}=\hat{n}_t,$$

where k_t is the number of types present in data at time t. We use this equation to define w'_{t-1} in terms of $\hat{n}_t$ and the $\hat{x}_{i,t}$. It follows that

$$w'_{t-1}>\sum_{i:\hat{x}_{i,t}/\hat{n}_t<f_{\min}}\hat{x}_{i,t}$$

for types i occurring in the data. The range of the sum, the types $i:\hat{x}_{i,t}/\hat{n}_t<f_{\min}$, is the exact complement of types assigned to any frequency bin among types present in the data at generation t. Furthermore,

$$w_{t-1}+\sum_{j=1}^{k_{t-1}}x_{j,t-1}=\hat{n}_{t-1},$$

and hence we use this as a definition of w_{t-1} in terms of $\hat{n}_{t-1}$ and types that appear in the data at frequencies greater than $f_{\min}$. Hence w_{t-1} count both types that appear in the data at frequencies below $f_{\min}$ and types that do not appear in the dataset at all.

The two interpretations of the missing data give high and low estimates for selection coefficients on rare types. However, the fitness ascribed to rare types affects replacement fitness to some extent, so the optimistic estimate for rare types often leads to slightly lower estimates for common types through increasing the replacement fitness. The spread between the estimates across the plotted region is a measurement of potential bias due to censorship and conflation of mutants with offspring of low-frequency types.

1.11 Sampling biases

The real process of naming and the Wright-Fisher process involve similar sampling noise, but a subtle discrepancy in the noise structure can causes bias in estimates, particularly at low frequencies. In the Wright-Fisher process, sampling noise (variance in frequencies from generation to generation) is exactly the multinomial variance due to sampling with replacement from the types present in the previous generation. The real process however samples names not strictly from the previous generation, but from some larger, unobserved population of names in an evolving cultural milieu. The discrepancy sometimes causes measurable bias in estimated selection coefficients, particularly at frequencies corresponding to low absolute counts. For example, the underlying Wright-Fisher process assumes exact lineal continuity from generation to generation, such as would be the case with last names or genes attributable to a particular parent. First names, rather, are copied from many sources, including bygone generations. Hence, some sequences that occur in the data are impossible under the Wright-Fisher process, such as when a name apparently goes extinct then reappears. Aside from being of statistical interest, this very discrepancy has interesting philosophical and scientific implications which we discuss at the end of this section.

We control for this bias by constructing time series of samples in which the only change from generation to generation is sampling noise, then we infer frequency-dependent selection from the time series of samples. First, we construct a fixed, time-independent metapopulation composed of all counts of all types summed over the duration of the dataset. Then we construct an artificial time series based on repeated samples with replacement from the metapopulation, matching population size to the data each generation. The average frequency of every type at every generation in the time series of samples is just its frequency in the metapopulation, and therefore no type changes frequency over time on average. Hence, in principle, there can be no selection and no frequency dependent growth. We then infer frequency-dependent selection in this time series of samples and take the magnitude of departure from neutrality to estimate the magnitude of potential bias due to sampling. We typically find substantial bias at low frequencies where binomial sampling noise predominates.

Inferred frequency dependence in the time series of samples is possible because of apparent growth or decline from generation to generation due to fluctuations in sampled counts: If a type happens to be sampled below its frequency in the metapopulation, it appears to grow on average in the next generation, and the opposite occurs if it is sampled above its true frequency. Typically, these generation-to-generation fluctuations cancel over time and the inference converges to zero selection. However, in certain circumstances cancellation does not occur. For example, when a type appears to go extinct and then reappears, its decrease to extinction registers as negative selection and whereas its subsequent increases in registers as mutation or immigration rather than positive selection, causing drastic bias towards negative selection. A qualitatively similar phenomenon occurs if any type fluctuates below or above a bin boundary from generation to generation. The likelihood of these occurrences diminishes as sampling variance decreases with higher counts.

The net effect of sampling bias depends on the total population size, the bin boundaries, and the distribution of type frequencies. We do not attempt to predict the bias analytically. Instead we simply infer frequency dependence in the resampled time series to determine at what frequencies this bias exists. Experimentally, the bias typically only exists at frequencies corresponding to low counts, with the exact frequencies affected varying by dataset (Supplementary Figure 8).

Finally, differences between cultural and genetic evolution spurred debate[94,95] about how

far a cultural Darwinian metaphor can be carried. Chief among the differences are starkly contrasting mechanisms of inheritance[17]. Extinctions of species and genes are permanent, whereas languages with no speakers have been resurrected, leading to complications in interpreting "extinction" of cultural traits[96]. Whereas family names and lineages may go extinct, a first name, word or language may persist in a cultural repertoire long after its last recorded instance in any dataset. The sampling bias we observe in names is instructive in exactly to what degree these differences are relevant to population-level change. On the one hand, the data manifestly violates the Wright-Fisher process because even a complete birth cohort of names is still a sample from a larger cultural repertoire, and absence of a name from one birth cohort does not imply absence in a subsequent cohort. On the other hand, the differences are negligible above a certain frequency and the dynamics of name frequencies can be meaningfully treated as a Wright-Fisher diffusion[18]. The differences confuse measurements of mutation and extinction, but in measuring frequency-dependent selection, the discrepancies between cultural and genetic models only matter near the boundary at 0 frequency. Discrepancies between models at frequencies near zero is a familiar situation in population genetics, where even alternative models that share the Wright-Fisher diffusion limit often behave differently near the boundaries[97].

1.12 Inferring N_e

Our primary concern in this study is estimating fitness as a function of frequency, which determines the expected change in the frequency of a type over time. The strength of genetic drift, by contrast, determines the variance in type frequencies over time. Here we describe a method to estimate effective population sizes from time series data, which we use to produce simulations that approximately match the stochasticity of the real data. We also use estimates of effective population size to correct confidence intervals for the observed stochasticity in the data.

In an idealized Wright-Fisher population of constant size sampled once every generation, the variance in allele frequency between generations is governed by the population size, N. For example, in the simplest case of two neutral alleles without mutation, for an allele currently at frequency p, the frequency in the next generation is approximately normally distributed with mean p and variance $p(1-p)/N$. However, if the population has nonrandom structure, or if the total population size varies over time, or if the frequencies are measured more or less often than once every generation, then the variance in allele frequency between measurements is governed by the "variance-effective population size", denoted N_e[18]. When N_e is less than the census population size N, then variance in allele frequencies accrues more quickly compared to an idealized Wright-Fisher population.

We infer N_e from annual time series data by directly estimating the variance in allele frequency change from year to year, adapting the approach of Feder et al[98]. First, we infer the selection parameters $\hat{s}_i$ from the data using the maximum-likelihood procedure described above (Supplementary Information section 1.2). The inferred model then produces expectations e_i for the frequency of each type in each year, given the observed frequencies in the preceding year, so the frequencies p_i are approximately distributed according to $\text{Normal}(e_i, e_i(1-e_i)/N_e)$. We then rescale the frequency increments in order to produce homoscedastic updates under the model. The residuals of the data, $r_i = p_i - e_i$ where p_i are the frequencies of types, are distributed normally with mean 0 as

$$r_i \sim \text{Normal}(0, e_i(1-e_i)/N_e),$$

and so if we rescale the residuals according to

$$Y_i = \frac{r_i}{\sqrt{e_i(1-e_i)}}, \tag{30}$$

the rescaled increments Y_i have mean 0 and variance $1/N_e$. We then obtain the estimate $N_e = 1/S$ where S is the sample variance of the Y_i. This effective population size describes the rate at which variance accrues in the empirical frequency trajectories.

In a population of birth-cohort size N, the ratio $g = N/N_e$ quantifies how much faster variance accrues relative to an idealized Wright-Fisher process in a population of size N sampled once per generation[18]. Thus we can simulate a model at population size N in a manner that roughly matches a desired effective population size N_e by introducing $g - 1 = \lfloor N/N_e - 0.5 \rfloor$ rounds of neutral Wright-Fisher sampling for every 1 round of multinomial sampling incorporating selection and mutation.

1.13 Corrections for N_e

Whereas point estimates of selection parameters depend only on the maximum of the likelihood function, the confidence intervals around the estimates depend on the absolute magnitude of the likelihood function (Supplementary Information section 1.9). And so confidence intervals must account for the effective population size N_e when it differs from the census population size N.

The expression for the 95% parametric confidence intervals given in Supplementary Information section 1.9 already accounts for the possibility that $g = N/N_e$ may not equal 1. In particular, when $g > 1$, variance accrues faster in the empirical frequency trajectories than it would an idealized Wright-Fisher process in a population of size N sampled once per generation[18]. And so the confidence intervals on estimated parameters should be wider to account for this extra variability in the data. And, indeed, our expression for the confidence interval (Eq. 29) shows that its width scales with $\sqrt{g}$.

We can understand the effect of N_e on the width of the confidence interval intuitively. If an evolutionary process evolves at effective population size N_e, but is reported at census population size $N = gN_e$ then the counts in the multinomial expression are multiplied by g as $x_{i,t,N} = gx_{i,t,N_e}$. As a result, the terms in the log likelihood of the data that depend on the s_i (Eq. 5) are simply multiplied by constant factor g. This factor in the log likelihood multiplies the observed Fisher information by a factor of g, and therefore multiplies the variance of the estimator $\hat{s}_i$ by a factor $1/g$. Hence, if the Fisher information variance is $\sigma^2_{\hat{s}_i,(N)}$, the true variance on the estimator is $\sigma^2_{\hat{s}_i,(N_e)} = g\sigma^2_{\hat{s}_i,(N)}$, so that the standard deviation is scaled by a factor $\sqrt{g}$.

1.14 Model comparison

The theory of parameter estimation by maximum likelihood also provides a method for model comparison through a likelihood ratio test. This procedure formally tests the goodness-of-fit of one model over another, when the two models are nested—that is, when the more simple model can be obtained from a more general model by specifying the values of some or all of the parameters. In particular, the likelihood ratio test uses a chi-squared statistic for model comparison. After estimating parameters for each of the models by maximizing likelihood, twice the difference in the log likelihood of the two models is asymptotically chi-squared distributed, with the number of degrees of freedom given by the difference in the number of parameters between the two nested models[74].

In the context of our study, the neutral Wright-Fisher model is a special case of frequency-dependent model where $s(p) = 0$ for all frequencies p. Hence, neutrality and frequency dependence are nested models, and the likelihood ratio test provides an established method to compare the two models while accounting for the increased number of parameters of the frequency-dependent model. In models that use a wildtype fitness (Supplementary Information section 1.7) to conflate mutation with selection on types at unobservable frequency, a nested neutral null model is still available which contains one parameter: the value of wildtype selection s_w relative to a constant mean selection s_0 for the non-wildtype population.

We can also use the likelihood ratio test for model comparison in the presence of an effective population size N_e that differs from the census population size N. In this case, as described Supplementary Information section 1.13, the terms involving s_i in the log likelihood of the data (under either the neutral model or the model with selection, Eq 5) are magnified by a factor $g = N/N_e$ relative to what they would be for samples of size N_e, while terms that do not involve the s_i are constant with respect to the parameters and do

not contribute the difference in log likelihoods (the likelihood ratio). Hence, the chi-squared statistic is inflated by a factor of g due to $Ne \neq N$, and we correct this chi-squared statistic by dividing by a factor g.

1.15 Practical guidelines for application to other datasets

The inference procedure described in this section offers strong mathematical and emprical guarantees for the accuracy and consistency of confidence intervals and statistical significance tests, provided that the model conforms to the data-generating process. These guarantees do not hold if there is substantial mismatch between the data-generating process and the model, in particular in the presence of errors or omissions in the data.

We account for the known sources of error that appear in the datasets we analyse by designing controls appropriate to each source of error. However these controls do not cover all possible sources of discrepancy between the model and the data-generating process.

We therefore recommend that researchers applying the method to other datasets design consistency checks similar to the ones we outline in Supplementary Information section 2. A convincing, ideal consistency check is to simulate the data-generating process—including random errors, omissions and discrepancies realistic for the data at hand—in the presence and absence of frequency-dependent selection. After conducting the inference on the output of an ensemble of replicate simulations, if the confidence intervals and false positive rates behave as expected (including the true (known) value 95% of the time when α=0.05), then the method is reliable on real data generated through the same process.

2 Validation

2.1 Theoretical justification

We offer two, principally independent theoretical justifications for the inference method as either the maximum-likelihood estimator of the parameters of a frequency-dependent model, or the average (effective) frequency dependence of a general evolutionary process.

2.1.1 Consistency of maximum likelihood

The first rationale, as a method of inferring frequency-dependent selection parameters, comes from the classical theory of maximum-likelihood parameter estimation, developed by Fisher, Neyman, and Pearson. That theory guarantees that the maximum-likelihood estimators of parameters, under a well-specified model for the data generation process, are asymptotically consistent and normally distributed, provided the likelihood function is smoothly differentiable and the true parameters do not lie at the boundary of a compact parameter space[74].

We therefore know *a priori* that the estimates $\hat{\mu}$ and $\hat{s_i}$ will be asymptotically consistent under the model, because the conditions for Neyman-Pearson theory apply to our likelihood function. The same classical theory further allows us to derive 95% confidence intervals based on the curvature of the likelihood surface at the maximum-likelihood values, called Fisher information[92] (Supplementary Information section 1.9), and provides statistical tests for frequency-dependent selection using likelihood ratio statistics.

2.1.2 Robustness of the estimator

Furthermore, the structure of the estimator and likelihood function (Eqs. 9 and 14) offer additional guarantees. Inferences of selection $\hat{s}_i$ are guaranteed to be independent of the mutational process. This is true because the maximum-likelihood expression for the $\hat{s}_i$ (Eq. 14) does not involve μ and vice versa. A more general intuition is that the counts of selected types conditioned on the number of mutants are still multinomially distributed according to a Wright-Fisher process involving only selection (Eq. 5) as discussed in Supplementary Information section 1.2. Thus even for mutational processes far outside the model, such as a fixed number of mutant types each generation, we may still view the selected types as a

multinomial sample of pre-existing types, and the likelihood maximization Eq. 6, and hence Eq. 14 still applies.

Our inferences of selection $\hat{s}_i$ are also theoretically guaranteed to be unaffected by variation in the total population size. This is because the estimator accounts for the known population size each year – that is, our likelihood expression includes the total population size each year as known quantities, as opposed to treating them as unknown parameters. This situation contrasts with a typical situation in population genetics, in which the demographic history and selection parameters are both unknown and must be jointly inferred, which can often lead to spurious inferences of selection[99]. By extension, the inferred selection coefficients $\hat{s}_i$ are also independent of $g = N/N_e$ as described in Supplementary Information section 1.13. A difference between census and effective population size causes and predictable bias in confidence intervals and in the likelihood ratio test for model comparison. However, a correction factor removes these biases as we describe in Supplementary Information section 1.13 and numerically validation in Supplementary Information section 2.2.

We note in caution, however, that the theoretical justification of maximum likelihood and concomitant guarantees of asymptotic consistency apply when the data are generated by the model, for some choice of parameters. Discrepancy between the likelihood expression and the data-generating process is the problem of model misspecification. When the model is misspecified, the theoretical guarantees do not apply or may apply only approximately[75], and likelihood ratio tests could be biased conservatively or non-conservatively. Outlier data, missing data, coding errors, and other factors in the data-generating process that are not part of the model, constitute forms of model misspecification. We account for two known discrepancies from the pure frequency-dependent model related to incomplete sampling of the population. Censorship of rare names and sampling error in rare counts both could potentially bias parameter estimates $\hat{s}_i$. We control for both of these potential biases as described in Supplementary Information sections 1.10 and 1.11, respectively.

2.1.3 Effective frequency dependence

An alternative line of reasoning provides a distinct theoretical justification for our inference procedure, even when the evolutionary process does not conform to a the frequency-dependent selection model used to derive the estimator (Supplementary Information section 1.1). As we show below, the estimator measures the average total reproductive output of types in each bin, in a manner that is independent of how fitness is assigned to types in the data-generating process. This average, as well as the average in Eq. 25, both estimate the effective frequency-dependent fitness—the mean fitness as a function of frequency in the underlying process.

Here we show that the maximum-likelihood fitness parameters $e^{\hat{s}_d}$ are also unbiased estimators of the same quantity as the average frequency dependent fitness $\bar{w}_d$ defined by Eq. 24 in Supplementary Information section 1.6, which applies to any selection process with arbitrary fitness $w_{i,t}$ assigned to each type i at each generation t. To demonstrate this correspondence, we reconsider the equation Eq. 14 that is satisfied by the $\hat{s}_i$. We make two definitions that will clarify this correspondence. We define population mean inferred fitness as $\hat{Z}_t = \sum_{j=1}^{k} e^{\hat{s}_{\mathcal{B}(x_{j,t}/n_t)}} x_{j,t}/n_t$. We define the realized growth of a type i at generation t to be $r_{i,t} = (x_{i,t+1}/(n_{t+1} - m_{t+1}))/(x_{i,t}/n_t)$ and note that $\sum_i r_{i,t}(x_{i,t}/n_t) = 1$. We substitute these definitions into Eq. 14 and re-index generations to run from $t = 0$ to $t = T - 1$ in order to obtain:

$$\sum_{t=0}^{T-1}(n_{t+1} - m_{t+1}) \left(\sum_{i:\mathcal{B}(x_{i,t}/n_t)=d} \frac{e^{\hat{s}_d}}{\hat{Z}_t} \frac{x_{i,t}}{n_t} \right) = \sum_{t=0}^{T-1}(n_{t+1} - m_{t+1}) \left(\sum_{i:\mathcal{B}(x_{i,t}/n_t)=d} r_{i,t} \frac{x_{i,t}}{n_t} \right) \quad (31)$$

This equation corresponds to Eq. 24 through two identifications: $w_{i,t}/Z_t$ is the expectation of $r_{i,t}$ if the $w_{i,t}$ are known, and $\bar{w}_d/Z_t$ and $e^{\hat{s}_d}/\hat{Z}_t$ are both the average fitness within bin d. The right sides of both Eq. 31 and Eq. 24 compute the average total reproductive output of bin d normalized as described in Supplementary Information section 1.6. Eq. 24 takes a uniform average across generations over the expected reproductive output $w_{i,t}/Z_t$;

whereas Eq. 31 takes a weighted average of the realized reproductive output. The weighting in Eq. 31, the non-mutant population size $n_{t+1} - m_{t+1}$, is proportional to the inverse of variance in a Wright-Fisher process[18]. And so the right side of Eq. 31 is the inverse-variance weighted average of the realized reproductive output: larger samples (populations) are more informative of the underlying fitness.

It is well-known that averages of independent samples converge to their expected values, and this holds regardless of positive re-weighting. Bearing in mind that the fitness assignments $\{w_{i,t}\}$ and trajectories $\{x_{i,t}\}$ may vary randomly between realizations of any given selective process, both of the averages Eq. 24 and Eq. 31 are estimators—given different information—of the same underlying quantity: the mean total reproductive output of types in bin d over realizations of an arbitrary selection process.

We refer to this mean fitness of types as a function of frequency as "effective frequency-dependent fitness". The mean fitness within each frequency bin is a property of any evolutionary process in which the parentage relationship is known from one generation to the next. What we have shown here is that the maximum-likelihood procedure (Supplementary Information section 1.2) estimates effective frequency dependence given any realized growth rates $r_{i,t}$, regardless of whether the underlying evolutionary process is exchangeable or not.

2.2 Numerical validation

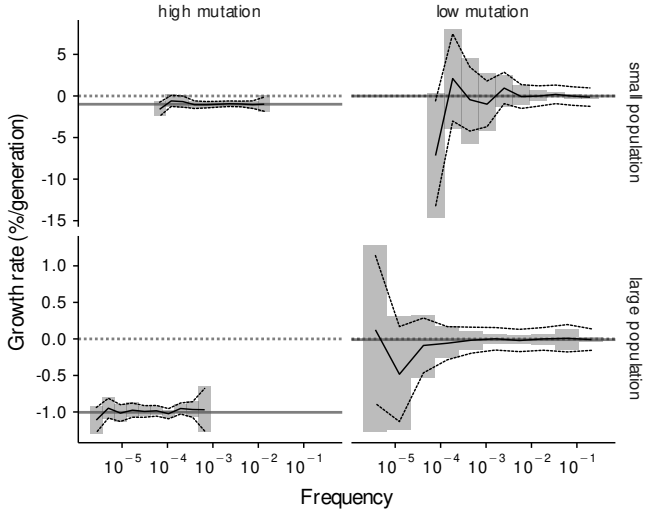
In addition to the theoretical justification above, we numerically validate our inference procedure by simulating known models and applying the inference protocol to the output. We use the inference software we wrote and released[1] under the GNU General Public License. In all cases we find the results in line with statistical expectation for the accuracy of confidence intervals and likelihood ratio tests. This numerical validation controls for mathematical and software implementation errors, and it demonstrates the inference method on straightforward test cases. We simulate neutral populations with various population sizes, mutation rates and demography, and conduct the inference under alternative binning assumptions. We simulate frequency-dependent selection with fluctuating demography, censorship and wild mutation counts and recover the true parameters consistent with the confidence intervals. We simulate subpopulations and recover the contrasting frequency-dependent fitness curves, while demonstrating that the inferred frequency-dependent fitness for the full population is an average of its subpopulations.

2.2.1 Neutral simulations under alternative binning

We validate the false positive rates and confidence intervals of the inference by simulating the neutral null model. That is, we validate that the inference does not detect frequency-dependent selection when selection is in fact absent more often than the significance level α (here assumed to be 0.05).

We first simulated the neutral Wright-Fisher process with two different fixed population sizes (20,000 and 500,000) that represent the range of effective population sizes inferred from any of our datasets (Supplementary Table 1). For each population we simulated a high and low mutation regime, where the high mutation ($\mu = 0.01$) is within an order of magnitude of the mutation seen in baby names data (0.04-0.07) and the low mutation ($\mu = 0.0001$) is two orders of magnitude lower. For each of the four scenarios (population size × mutation regime), we initialized simulations with either a monomorphic population (a single type at frequency 1) or a highly polymorphic population derived from wild name data. We simulated, ignoring the results (burn-in) until the monomorphic and polymorphic initializations had converged in most frequent type to roughly within one order of magnitude, indicating rough approximation to the equilibrium type abundance distribution. This took between 1,000 and 50,000 timesteps of burn-in, depending on the population size and mutation rate. We then conducted inference from time series generated from the next 100 generations of the monomorphically initialized simulation, using 10 log-evenly-spaced bins spanning the frequency range of the data as described in Supplementary Information section 1.2-1.3.

[1] First published at `https://github.com/mnewberry/fdsel`

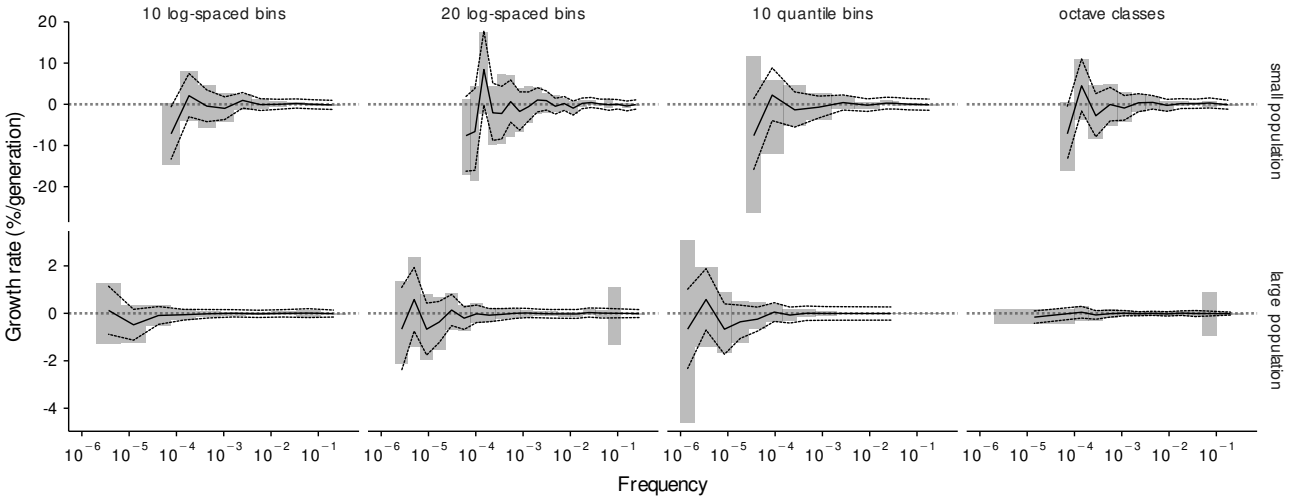


Supplementary Figure 2: **Validation of inference procedure on neutral simulations.** The inference protocol recovers constant fitness, independent of frequency, from simulated neutral populations with different population sizes and mutation rates (small: $N = 20,000$; large: $N = 500,000$; low μ=0.0001, high μ=0.01). Even without selection, extant types are diluted by incoming types each generation, which produces a constant negative growth rate of extant types, independent of frequency. The solid gray horizontal line indicates the predicted average growth rate of extant types, $1 - e^{\mu}$. The 95% confidence intervals include this value approximately 95% of the time, as expected. Grey squares indicate bin width and bootstrap CIs, which here agree well with the CIs derived from Fisher information.

Inferences from these neutral simulations are presented in Supplementary Figure 2. Because of dilution by incoming mutant types, types with no selection have a constant negative growth rate. This growth rate associated with zero selection is determined by the mutation rate, as $(1-e^{\mu})$. The 95% CIs on all bins in all four inferences include this growth rate in 39 out of 40 bins, which is consistent (two-tailed binomial $p = 0.72$) with confidence intervals including the true value 95% of the time. That is, these experiments validate the inference method with respect to its expected false positive rates under controlled conditions. Furthermore, the p-values for rejecting neutrality in these four simulations based on a likelihood ratio test with $D - 1 = 9$ degrees of freedom range from 0.40 to 0.98, none of which pass for significant evidence in rejecting neutrality, as desired.

Each simulated scenario has a different equilibrium number of types and range of type frequencies, and so has different bin widths according to the logarithmic binning scheme. Smaller populations by nature have a higher minimum frequency, while lower mutation rates permit fewer, higher frequency types. Bins with fewer types present and bins at lower frequency have more noise and hence wider confidence intervals. In theory, false positive rates on bins (rate of excluding $1-e^{\mu}$) should not depend on bin width, as the error on the estimate for each bin should be accurately reflected in its confidence intervals. That prediction is borne out in these numerical tests, in that bins with more noise (greater deviation from zero selection) also have wider confidence intervals, so that the propensity for the confidence intervals to exclude the true value remains unchanged.

To further illustrate the relationship between bin size and confidence intervals, we conducted six more inferences on the two low-mutation simulations, using three different binning schemes each (Supplementary Figure 3). In addition to the 10 log-spaced bins as before, we

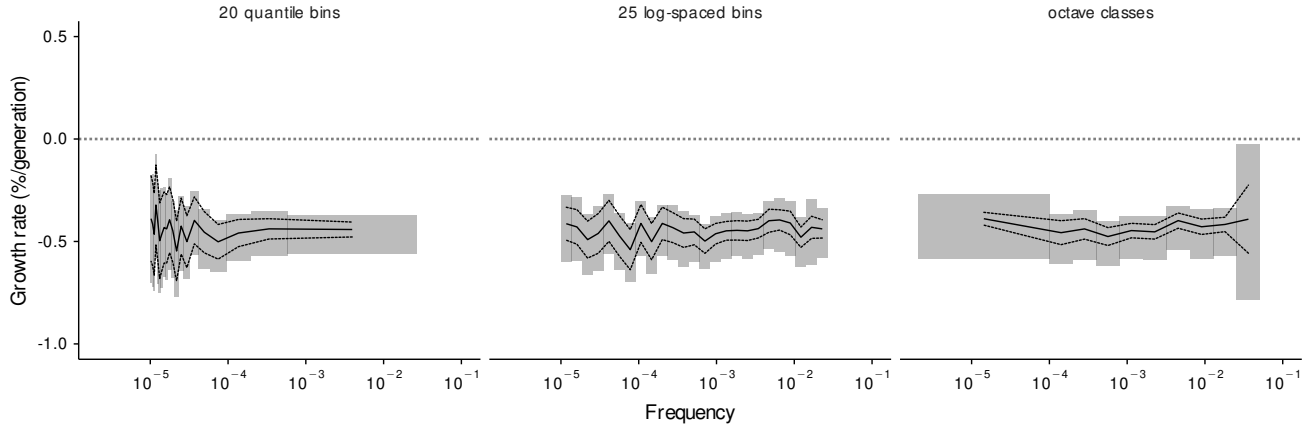


Supplementary Figure 3: **Alternative binning of neutral simulations.** Conducting the inference on the low mutation, large population neutral data produces estimates consistent with constant fitness, independent of bin size. Here we bin the output of the neutral simulations according to several different binning schemes discussed in Supplementary Information section 1.3: logarithmic binning, quantile binning, and the octave class binning actually used on baby names. Narrower frequency ranges and less underlying data lead to wider confidence intervals (Supplementary Information section 1.9). These 95% CIs scale appropriately, so that statistical variability leads to exclusion of the neutral growth rate approximately 5% of the time, independent of bin width or amount of underlying data. The octave class binning performs the best on large populations, in terms of equalizing statistical error across the full frequency range of the population.

inferred using 20 log-spaced bins. Using bins half as wide increases the average $\sigma^2_{\hat{s}_i}$, the variance in the estimator $\hat{s}_i$ (Supplementary Information section 1.9), by factors of 4.5 and 4.4 respectively for the small- and large-population inferences. This increase in variance with narrower bins is expected, due to the trade-off between frequency precision and precision in estimation of the selection coefficients. We also conducted the inference using 10 quantile bins as described in Supplementary Information section 1.3, and finally using the exact binning scheme by powers of 2 (octave classes) used in Figure 2. Here again, the confidence intervals include the true value in 97 out of the 101 bins. These 101 bins count any bin in any scheme that contains at least one transition of at least one type, i.e. any bin i for which $\hat{s}_i$ is defined. This 96.04% success rate resembles the expected 95% rate, however the binomial p-value used above is not meaningful in this case due to pseudoreplication (non-independence) between alternative binning of the same data.

Likelihood ratio tests based on these alternative binning schemes also appropriately fail to reject neutrality. Different binning schemes reapportion noise between different parameter estimates, and so the likelihood ratio test p-value takes on a different value for each binning scheme. However, the likelihood ratio p-value is primarily determined by the probability of the underlying data under the null hypothesis, and so the p-values for alternative binning of the same underlying data tend to cluster around the same value. For the small-population simulation, for example, the likelihood ratio test p-value ranges from 0.40 to 0.56 across the four different binning schemes, whereas for the large-population the p-values range from 0.85 to 0.999.

We note in closing that the octave class binning scheme used to generate Figure 2 per-



Supplementary Figure 4: **Neutral simulations with US demography.** We simulated a neutral population that matched the true US data in the fluctuating total population size and the exact counts of novel types at each timestep, starting from the true initial population state and censoring counts less than 5 in the output. Inference from the simulated time series is consistent with constant fitness, for three different binning schemes. These results indicate that neither binning scheme, fluctuating population size, nor the unspecified mutational process cause confounds that lead us to falsely detect frequency-dependent selection when the underlying model is neutral. Censorship, which can cause deviations from neutrality in theory, did not cause the inference to detect selection at frequencies above 10^{-5}. The negative displacement of the constant growth rate from zero is due to the true influx of novel types derived from the data.

forms best on large populations in terms of equalizing error across bins, equivalent to minimizing the most extreme error.

2.2.2 Simulations matching US demography, mutation, censorship and N/N_e

Changing demography—fluctuating total population size or mutation between generations—might complicate the inference, relative to simulations with fixed population size. Real populations such as the US baby names data include these complications such as changing birth cohort size as well as censorship. In theory, the inference is independent of changing demography and bias due to censorship is accounted for, but we conducted numerical validations to check for mathematical or implementation errors.

We validated the inference procedure with these complications by simulating a neutral population while matching the exact demography of the US baby names dataset in total birth cohort size each year, initial population, and the exact counts of novel types each year. We further reproduce the censorship scheme of the US data by omitting counts less than 5 from the simulation output. Strictly speaking, the simulation cannot both censor a random number of individuals and output population sizes that exactly match the US data. Hence, the simulation uses the total population imputed from the US data each year, but its output population size differs by the randomly fluctuating number of censored individuals.

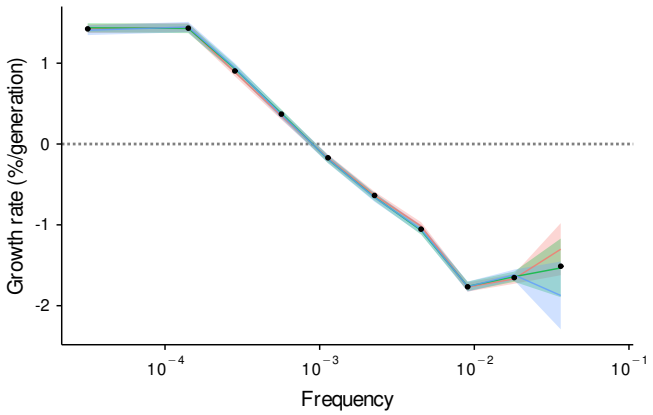
We conduct the inference on the neutral simulated population with US demography exactly as with the true US data: we impute missing counts of censored types and use a wildtype fitness to register the competitive pressure from immigrant types and types at unknown frequency as described in Supplementary Information section 3.1. We compute

upper- and lower-bound inferences based on the imputed counts of missing types as described in Supplementary Information section 1.10. Here, we further conduct the inference under both quantile and logarithmic binning schemes, in addition to the octave class binning used in Figure 2, to test for any binning effects that might arise. In all these cases, we treat types at frequency less than 10^{-5} as part of the wildtype class.

The resulting inference (Supplementary Figure 4) is consistent with constant selection, $s(p) \equiv 0$, and the estimated 95% confidence intervals include the mean inferred selection coefficient in 53 of 55 bins, or 95-98% of the time. The likelihood ratio tests give p-values for rejecting neutrality of 0.98, 0.53 and 0.67 respectively, for quantile binning, logarithmic binning, and the octave class binning scheme used on the real data in Figure 2. Here, as in the neutral simulations with fixed population size and known mutation rate, displacement from the zero growth occurs due to the influx of novel types. In this case however, the true rate of influx of novel types is unknown and cannot even be estimated using the methods of Supplementary Information section 1.2 because of the censorship in the US data. We capture the rate of influx of novel types in the average fitness (net amount of offspring) of the wildtype population, in this case any individual not previously existing or existing at frequencies below 10^{-5}. This wildtype population includes the novel types whose counts are copied from the data in order to match the US demography, and hence the wildtype fitness in this otherwise neutral simulation output is truly positive. We compare the likelihood of the frequency-dependent model (in whatever binning scheme) against the nested neutral model in which the only parameters are the wildtype fitness $\hat{s}_w$ and the constant fitness s_0 of all other types. This likelihood ratio test appropriately fails to reject neutrality, regardless of the binning scheme, as desired.

Here we have limited our attention to frequencies above 10^{-5}. If we instead estimate parameters for lower frequencies, such as the frequency of the maximum censored count in the data, we would again reject neutrality for the correct reason that the upper and lower bound imputations are in fact not neutral (with p=0.02 and $p < 0.001$ respectively). These imputations are not neutral because they are designed to capture the highest and lowest selection values consistent with the data. However these imputations do not differ detectably from neutrality at frequencies above 10^{-5}, confirming that censorship does not affect the inference for US data for the range of frequencies plotted in Figure 2—above 10^{-4}.

Finally, we validated our ability to detect selection when selection is truly present. Again, we simulate a Wright-Fisher process that matched the US data in the initial population state, the total population size each generation, and the exact counts of novel types each timestep, while censoring any counts less than 5. But we otherwise determined type-frequencies through frequency-dependent selection (Eq. 1). In this case the inference model and data generating model match exactly except for the omission of censored types. In both the simulation and inference, we ascribed wildtype fitness to types at frequency below 10^{-5}. Thus types effectively left observation by transitioning to frequency lower than 10^{-5}. The simulation included no random mutational process: types entered observation through novel counts copied from the data. We did not infer μ, instead conflating mutation with wildtype progeny as we did with the real US data. We simulated frequency-dependent selection using a piecewise constant function $s(p)$. We copied its bin boundaries from the US inference depicted in Figure 2, and as parameters, we used the inferred parameters from the US data using $f_{min} = 10^{-5}$ to match the simulation. We conducted 500 independent simulations and inferred parameters from each simulated time series. The first three such inferences are depicted in Supplementary Figure 5. We accounted for the arbitrary zero level of absolute fitness by mean-centering the parameters from both the simulation and inference (the latter variance-weighted according to the Fisher information variance on the estimator) before comparison. For each parameter $s_1 ... s_{10}$ and s_w, the model 95% CIs on the inferred parameters (Supplementary Information section 1.9) included the true parameter value between 462 and 500 times (a range of 92.5% to 100.0%). Over all simulations and all parameters (5,500 cases), the 95% CIs included the true value 98.44% of the time, confirming accuracy of the inferred parameters and a small conservative bias in the estimation of 95% CIs. In addition, we reject neutrality 100% of the time, using a likelihood ratio test, based on the

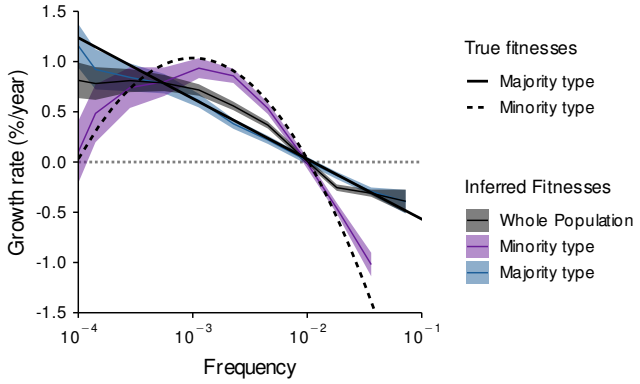


Supplementary Figure 5: **Simulations of frequency-dependent selection with US demography.** Three randomly-selected inferences (colored lines) recover the known, true parameters (black dots) to high precision from simulated time series representing US baby name demography under frequency-dependent selection. Color bands indicate 95% confidence intervals (CIs). The 95% CIs contain the true parameter value used in the simulation 98.44% of the time, over 11 parameters in 500 replicate simulations. The simulations respected the US baby name data in the exact total population size and exact counts of novel types each timestep (year), and they were initialized with the true US initial name counts. Simulated type frequencies evolved according to a Wright-Fisher process with frequency-dependent selection $s(p)$ taken from the inference from US baby names.

output of these simulations that include selection.

We also validated our inference of N_e and confirmed the independence of N_e and the parameter estimates by changing the effective population size in the above simulation, and then estimating the effective population size from the simulation output according to the method described in Supplementary Information section 1.12. To simulate with a different effective population size we introduced $g - 1 = 7$ rounds of neutral Wright-Fisher sampling after each round of the usual multinomial sampling with selection and mutation. Again, we let population size fluctuate according to the actual birth cohort size of the US data, now taking 8 multinomial samples at each birth cohort size in the US data. With $g = 1$, our estimate of N_e was within 3% of the harmonic mean birth cohort size. When $g = 8$ we infer $N/N_{e,g=8} = 8.13 \pm 0.06$ (taking N to be the harmonic mean birth cohort size), which confirms that our inference of the effective population size agrees with our method of simulating different N_e to within an 2% error. The inferred selection parameters $\hat{s}_i$ were also unaffected by simulating with a different effective population size, as expected (Supplementary Information section 2.1).

Confidence intervals and χ^2 statistics for model comparison can be corrected to account for $N_e \neq N$ (Supplementary Information section 1.13, Supplementary Information section 1.14). The confidence intervals in the above simulations with $N/N_e = 8$, corrected by the appropriate factor $\sqrt{g} = \sqrt{8}$, include the true parameters 98.34% of the time over the 10 bins in 500 replicate simulations, confirming once again the accuracy the inferred parameters and a small conservative bias in the estimation of 95% CIs adjusted for N_e. In these 500 selected simulations, the corrected likelihood ratio test rejects neutrality 100% of



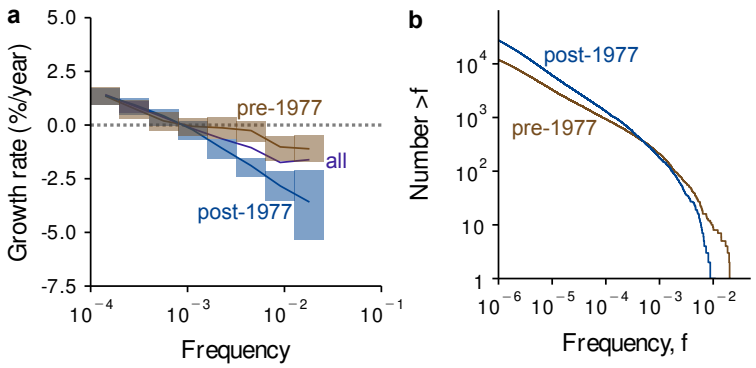
Supplementary Figure 6: **Inferring subpopulation frequency-dependent fitness from simulation.** We recover subpopulation frequency-dependent fitness from a simulated population composed of two subpopulations: a minority type (one third of mutants) with frequency-dependent fitness $w(p) = 1 - 0.01(\log_{10} p + 4)(\log_{10} p + 2)$ (quadratic in log-frequency), and a majority type (two thirds of mutants) with frequency-dependent fitness $w(p) = 1.012 - 0.006(\log_{10} p + 4)$ (linear in log-frequency). Both forms have a zero-growth intercept at 10^{-2}. We simulate a Wright-Fisher process with fixed population size 20,000 and mutation rate 0.0003 for 50,000 generations after 20,000 generations of burn-in from a monomorphic initial population. Confidence bands are analytic 95% confidence intervals derived from Fisher information (Supplementary Information section 1.9).

the time, as desired.

To validate the Type I error rate of the likelihood ratio test, we set $s(p)$ to zero in the above simulation that respects US demographic variation, and simulated with either $g = 1$ or $g = 8$, using 500 replicates each. We computed χ^2 statistics as the twice the log likelihood under the frequency-dependent model minus the log likelihood under the one-parameter neutral model in which s_w is the only free parameter. When $g = 1$, the χ^2 p-value yields a Type I error rate (at $\alpha = 0.05$) of 0.02, indicating a slight conservative bias. When $g = 8$, the corrected χ^2 statistic yields a Type I error rate of 0.03, which is again slightly conservative.

2.2.3 Subpopulations

We validated our ability to recover separate growth laws $s(p)$ between subpopulations when they are in fact different by simulating subpopulations. We simulated a population in which the true fitnesses within two subpopulations were known mathematical functions. In this simulation (with fixed population size 20,000 and fixed μ=0.0003), mutant types were assigned sequentially increasing integer identifiers. Types whose identifier was divisible by 3 were assigned fitness according to a quadratic log-frequency-dependent selection function $s(p) = \log(1 - 0.01(\log_{10} p + 4)(\log_{10} p + 2))$, whereas we assigned fitness to the remaining types according to a linear negative log-frequency-dependent selection function $s(p) = \log(1.012 - 0.0006(\log_{10} p + 4))$. Inference after partitioning the types into the correct subpopulations recovers the input parameters to good approximation (Supplementary Figure 6). Although it might appear that the CIs exclude the continuous functions $s(p)$



Supplementary Figure 7: **Frequency-dependent selection on US first names over time.** Separating the US dataset midway through time (1977), our inference of selection is indistinguishable for rare names, whereas common names display increasing negative selection over time, as observed in other studies[59].

more than 95% of the time from Supplementary Figure 6, this is an artifact of plotting $s(p)$ at the midbin frequency when in fact each inferred $\hat{s}_i$ represents a range of frequencies (and hence a range of values of the true $s(p)$). Making a more accurate comparison would require accounting for the error in representing these continuous functions $s(p)$ with the piecewise-constant parameterization.

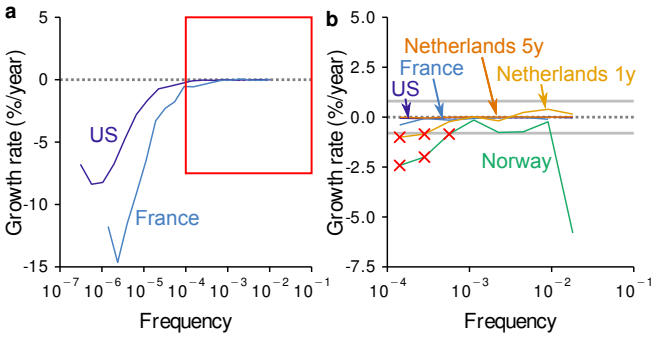
3 Data handling, error estimation, and model comparisons

3.1 Names

Here we describe data cleanup procedures, details of inference and controls appropriate to each of the four baby name datasets. A summary of the relevant properties, inferred parameters, estimates of effective population sizes, rates of bias, and test statistics for rejecting neutrality is given for each dataset in Supplementary Table 1.

For United States baby names, we obtained public data from the United States Social Security Administration database of first names (`https://www.ssa.gov/OACT/babynames/`) 1880-2018. We omitted nine high frequency names that clearly resulted from coding errors around 1985 such as "Christop,M" and "Infant,F". We omitted data up to and including 1935 as non-representative of a complete population of individuals. Births in the database that occur prior to 1935 represent cards issued to children and adults, as opposed to the post-1935 practice of issuing cards at birth. Hence the data prior to 1935 are incomplete and contain unrealistically small birth cohorts. The dataset indicates birth sex as a suffix on the name (either ",M" or ",F"). We consider as distinct names any spelling variants or the identical spelling in persons of different sex. We conducted the inference considering each annual birth cohort to be a generation (timestep) of the Wright-Fisher process, and infer parameters for $\hat{s}(p)$ composed of logarithmically increasing frequency ranges (bins) starting from 0.0001-0.0002 with bin boundaries increasing by powers of two (0.0001, 0.0002, 0.0004, etc.) following Hahn and Bentley[1] as described in (Supplementary Information section 1.3). We additionally infer but do not report fitness parameters for names more rare than 0.0001 whose frequency may or may not be known as described in Supplementary Information section 1.10.

The US data censors counts below 5. Hence, computing an accurate mutation rate is impossible and some fraction of individuals are missing (censored) from the dataset. To account for unobserved individuals, we conduct the inference using a frequency-independent wildtype fitness for censored, mutant and low-frequency types (Supplementary Information sections 1.7 and 1.10). We guess the fraction of missing individuals in the time series by imputing a `CENSORED` type which always counts 5% of the sum of the uncensored population



Supplementary Figure 8: **Sampling bias controls by country.** Inference from synthetic time series composed of generations sampled with replacement from a static metapopulation of name frequencies derived from the data. Inference with 20 bins (a) at all frequencies shows bias at low frequencies in the US and France. The red box indicates the plot range of Figure 2. Panel (b) shows the control inference for the same bins and plot range as Figure 2. Red Xs indicate bins removed from the results for exceeding the error threshold (b, grey lines). We ignore the anomalous control in the highest bin in Norway: It is an artifact that occurs because the highest-frequency name in the metapopulation (Anne) is only slightly less than the bin boundary and occasionally passes into the highest bin only to return, registering low fitness.

each year. We derived the number 5% from the French dataset, in which aggregate counts of censored individuals are listed explicitly and the fraction of counts below 5 is known to be 4.5%. In the US dataset the true count of names at frequency below the maximum censored frequency $f_{\min}$ (censored and uncensored) far exceed the sum of those names that appear uncensored in the data below $f_{\min}$. As described in Supplementary Information section 1.10, we pool the `CENSORED` type together with uncensored counts less than the maximum censored frequency $f_{\min}$=2.1e-06 into the wildtype subpopulation. The inferred wildtype fitness explicitly conflates immigrant types, mutant types, and progeny of types with censored counts into an average growth rate of all unaccounted types. This wildtype fitness is necessary to compute the replacement fitness or zero point of growth but otherwise has no affect on reported results. The error in replacement fitness due to censorship and conflation of mutants with progeny of unobserved types is measured in the spread between optimistic and pessimistic selection coefficients at high frequencies and reported below.

We controlled for sampling bias by constructing time series of samples from the time-independent distribution of name frequencies as described in Supplementary Information section 1.11. The inference from this synthetic time series (Supplementary Figure 8) measures the magnitude of possible bias due to data sampling unaccounted in the Wright-Fisher process.

At frequencies greater than 0.0001 in the US inference, censorship bias affects plotted frequencies by at most 0.06%/year and sampling bias by at most 0.08%/year although frequencies less than 0.0001 were affected by both censorship bias (Supplementary Information section 1.10) and sampling bias (Supplementary Information section 1.11, Supplementary Figure 8). Hence, we do not report inferred fitnesses for names below frequency 0.0001 but otherwise ignore these small biases at plotted frequencies as well within statistical error.

We further acquired baby name data from France 1900-2018 (Insee Fichier des prénoms[2], retrieved July 3, 2020) and Norway 1880-2019 (Statistics Norway Names STATBANK[3], retrieved Aug 18, 2020). Additionally we received complete, uncensored counts from a name

[2]Retrieved from `https://www.insee.fr/fr/statistiques/fichier/2540004/nat2018_csv.zip`

[3]Retrieved from `https://www.ssb.no/statistikkbanken/selectvarval/Define.asp?SubjectCode=al&ProductId=al&MainTable=FornavnFodte&SubTable=1&PLanguage=1&Qid=0&nvl=True&mt=1&pm=&gruppe1=Hele&aggreg1=&VS1=NavnJenter02&CMSSubjectArea=befolkning&KortNavnWeb=navn&StatVariant=&TabStrip=Select&checked=true`

corpus[100] of the Netherlands 1946-2015 anonymized by Gerrit Bloothooft of Meertens Instituut and communicated by email. In France data, we ignored inexhaustive data prior to 1946 as indicated in the file metadata. France censors names that do not occur at least 20 times since 1946 and names that occur less than 3 times in any given year, but lists the totals of censored counts in years `XXXX` and types `_PRENOMS_RARES`. We incorporated the censored counts into a wildtype aggregation with counts below $f_{\min}$=3.4e-06, as with the US inference (Supplementary Information section 1.10), distributing the counts of unknown year uniformly across the dataset. In Norway data, we ignored records prior to 1946 because many contained missing values. Counts less than 4 are censored in Norway, so we again assumed 5% of the data to be censored in each year and incorporated the imputed censored counts into the wildtype aggregate of types below $f_{\min}$=7.8e-05.

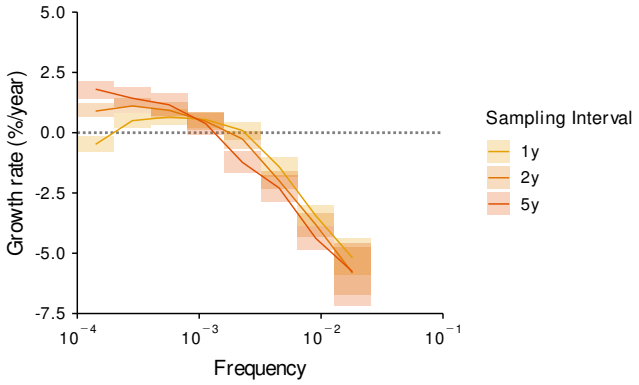
The names and name trajectories in these four datasets are substantially independent of one another. The names themselves are often different: 87% of US names, 59% of French names, and 27% of Norwegian names are unique to their country of origin relative to the other two datasets. (Names are anonymized in the Dutch dataset.) Moreover, there is only a weak correlation between countries in the frequencies trajectories of names over time (Pearson $r^2 = 0.10$ (France vs Norway), 0.13 (US vs France), or 0.22 (US vs Norway)).

We excluded from plots of inferred selection any bins with fewer than 5 types or with censorship or sampling bias exceeding 0.8%. Censorship bias in France and Norway was minimal except in the rarest frequencies in Norway, but sampling bias was substantial in both Norway and the Netherlands due to their small sizes. Hence, we first excluded any frequency bins for which the sampling bias control exceeded 0.8%/year (Supplementary Figure 8) with the exception of the highest-frequency bin in Norway. In the remaining (plotted) bins in all countries, censorship bias (as measured by the difference between optimistic pessimistic interpretations of missing data) was at most 0.08%/year.

With the exception of Norway, the sampling bias control for displayed bins was at most 0.4%/year. Norway was most affected by sampling error due to its small size. We kept the highest bin in Norway despite its control exceeding 0.8%/year because the control is anomalous: The most frequent name in Norway, "Anne", has a frequency 0.0127 over all time, just under the highest frequency bin boundary of 0.0128. This name is the only name to occupy the highest bin in the time series of samples, entering the bin 7 times always to return the following timestep, registering negative growth. In the data by contrast, 17 names populate the highest frequency bin for sustained time periods, so we do not believe this bin is truly substantially affected by sampling bias. The root mean square (RMS) bias in Norway with the exception of the highest bin was 0.5%/year whereas for all other plotted bins in all countries combined, the RMS sampling bias was 0.14%/year. Finally, we observe from Supplementary Figure 8 that sampling bias does not substantially affect the fitness of common types and by extension the replacement fitness.

Lastly, there is error associated with sampling at finite time intervals. We conduct inference using the Wright-Fisher process, but in reality birth is a continuous process. The Wright-Fisher process corresponds to an underlying diffusion process only in the limit that generation times are short. We estimate the error introduced by using finite time intervals by analyzing the uncensored data from the Netherlands (Supplementary Figure 9). When aggregating uncensored Netherlands counts into 1- and 2-year intervals, $\hat{s}(p)$ curves differ by at most 0.6%/year (RMS error: 0.4) due to a combination of bias and pseudoreplication noise. Inference of $s(p)$ from Netherlands data aggregated into 1-year and 5-year bins differ by at most 1.3%/year (RMS: 0.81, Supplementary Figure 9), indicating the bias diminishes with finer sampling intervals. The bias introduced by 1-year sampling is therefore less than 0.6%/year in Netherlands data, which we take to be representative of all name datasets.

We reject neutrality in favor of frequency-dependent selection with $p < 0.001$ in all the datasets we examine (Supplementary Table 1). This is consistent with the simple observation that the graphs of $s(p)$ in Figure 2 do not depict horizontal lines, even considering the variance within each bin. Without accounting for effective population size, the likelihood ratio tests for model comparison yield $\chi^2 = 12312.3$ (9 d.f.) for the US, $\chi^2 = 12085.8$ (10 d.f.) for France, $\chi^2 = 97019.9$ (9 d.f.) for Netherlands, and $\chi^2 = 618.1$ (9 d.f.) for Norway.



Supplementary Figure 9: **Effects of discrete-time sampling.** Uncensored anonymized counts from the Netherlands allow for aggregating data by different time intervals. By creating time series at 1-, 2-, and 5-year intervals, we probe effects arising from temporal discretization, as well as study frequencies lower than the inverse annual birth cohort size. The sampling interval affects sampling error and hence the bias described in Supplementary Information section 1.11, as observed at frequencies below 4E-4 for the smaller sampling intervals. At higher frequencies, a smaller bias remains from discretizing a continuous process into discrete time. The magnitude of this bias increases with sampling interval, and so the 1y estimates are the most accurate.

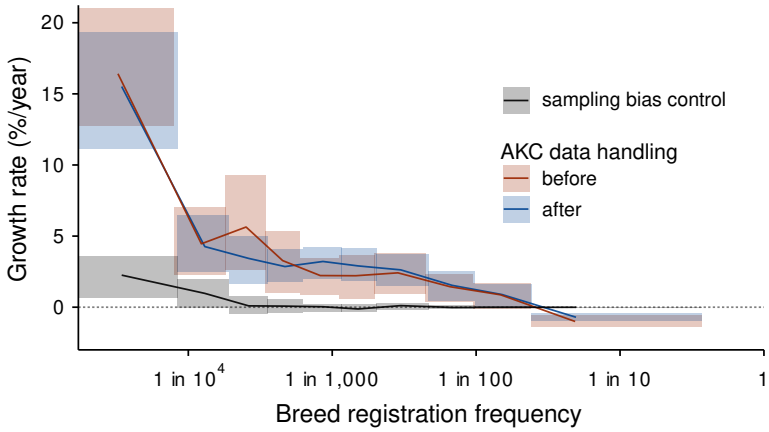
Correcting these likelihood ratio tests to account for inferred values of N/N_e in each of these countries still yields overwhelmingly significant likelihood ratio statistics: $\chi^2 = 1576.8$ for the US, $\chi^2 = 926.0$ for France, $\chi^2 = 50263.2$ for Netherlands, and $\chi^2 = 250.9$ for Norway. The astronomical improbability of neutrality in favor of selection is due to the volume of data and magnitude of deviation of the estimates from $s = 0$.

3.2 American Kennel Club

We obtained from a public depository[2,67] an annual time series of breed registrations from the American Kennel Club (AKC), which includes a total of 53,697,706 dogs of 153 different breeds registered between the years 1926 and 2005. After checking data quality and removing outliers associated with known or imputed errors in breed classification, we inferred frequency-dependent selection using the procedure described in Supplementary Information section 1.2. We also measured sampling bias as described in Supplementary Information section 1.11. In the lowest two bins, sampling bias was appreciable but still less than statistical error, whereas sampling bias was otherwise negligible. The net effect on the inference of sampling bias and data handling including removing outliers, separating time series and ignoring transitions to or from zero is depicted in Supplementary Figure 10.

We used counts of dog breeds as they appear in the original data with the exception of counts 0, which we omitted from the inference procedure because of pervasive non-distinction between 0 and missing data in the original dataset. That is, when $x_i = 0$ or $x'_i = 0$, we simply omit x_i and x'_i from any sums in equations such as Eq. 19. Thus m_t is always zero (we do not infer mutation rate), $n_t = n' = \sum_{i=1}^{k_{t-1}} x_i^t$ where i ranges over the k_{t-1} types which were present (nonzero) at timestep $t-1$, and the sums over types in Eq. 19 involve only the k types which are present at t, ignoring the subpopulation of types for which there is no data. We divided the frequency range into $D = 10$ bins, choosing bin boundaries by quantiles so that each bin incorporated a roughly equal share of transitions $x_i \to x'_i$ (range: 848-894).

We excluded rows in the dataset (nominally, breeds) containing breed subtypes that are summed in other rows, e.g. the rows "Dachshund-XX" and "Dachshund-XX (long-haired)" are summed in "Dachshund", and so we retain only "Dachshund". We retained "Dachshund", "English Toy Spaniel (All)", "Fox Terrier", "Manchester Terrier", "Poodle" and "Spaniel (All

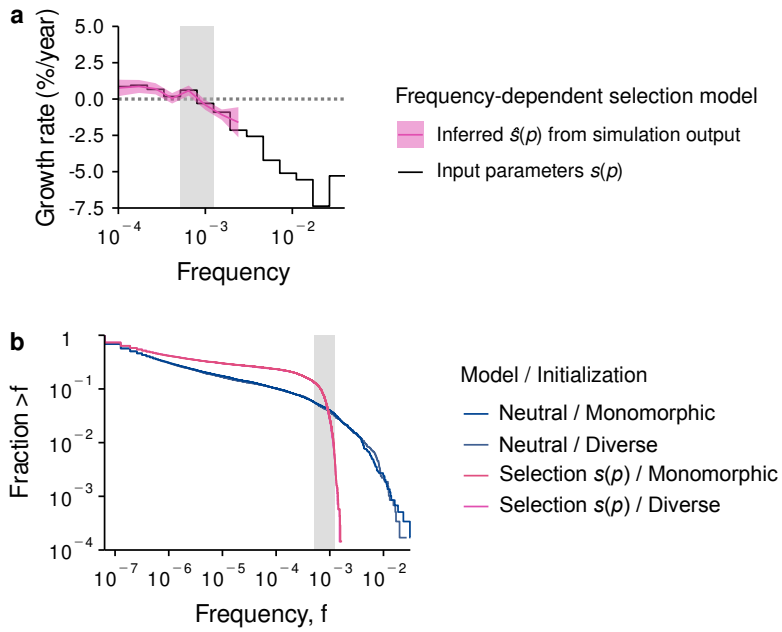


Supplementary Figure 10: **AKC data handling and sampling bias.** In the inference from AKC data depicted in Figure 4 (blue here), we ignored transitions to or from zero, removed outliers, removed data from the year of introduction of dog breeds, and severed time series at apparent discontinuities (Supplementary Information section 3.2). Here we show the inference from the raw data (red) as well as the sampling bias control inference (Supplementary Information section 1.11), which estimates the magnitude of potential bias (black). The sampling bias inference does not exceed 23% proportional error on the magnitude of the inferred selection coefficient (blue curve) or 2.3%/year absolute magnitude, and remains well within the statistical error in each bin.

Cockers)" while excluding all of their subtype rows.

We also removed artifactual discontinuities from the time series. The evolutionary models we fit have approximately continuous frequency trajectories[90], whereas the AKC data contains sharp discontinuities arising from reclassification of breeds or changing interpretation of registration data over time. Hence we omitted certain counts from the data or broke some time series into two separate time series (e.g. Greyhounds-pre1935 and Greyhound-1935onward). We detected anomalous discontinuities by fitting the frequency-dependent model to the raw data and examining residuals. We produced a p-value for each residual transition in frequency as its quantile in the Binomial distribution given its expected frequency and the current population size. The most improbable transition (Chinese Shar-Pei 1992→1993: 90,081→19,465, Bonferroni-corrected $p<10^{-55}$) occurred immediately after the Shar-Pei's year of introduction in the dataset, suggesting that the count during the year of introduction differs in kind from subsequent counts. In fact, 84 breeds were introduced over the course of the time series, and 40 out of these 84 have unlikely transitions ($p<0.1$) immediately following their year of introduction (compared to 654 out of 8810 in the full data, indicating otherwise slightly conservative p-values). We conclude that counts are unreliable in the year a breed is first introduced into the dataset, and so we ignore all 84 such counts. Removing initial points from a time series does not bias the inference under the frequency dependent model.

We detected five other likely artifactual discontinuities by ranking transitions according to the magnitude of proportional change. We removed the count for Poodle in 1939 (effectively removing two transitions) because Poodle is a sum of breed subtypes and the "Poodle (miniature)" subtype was undefined in 1939. We also split the time series for Manchester Terrier at 1945/1946 (where it nominally experienced a transition 880→61), Curly-Coated Retriever at 1932/1933 (82→1), and Greyhound at 1934/1935 (12→1624). Manchester Terrier is a sum over two sub-types, and the end of the "Manchester Terrier (Toy)" time series in 1945 causes the discontinuity. We have no evidence that the abrupt transitions for Greyhound and Curly-Coated Retriever correspond to errors in the data beyond the fact that they are the two highest proportional changes in the entire dataset, and they are much larger than changes known to be artifacts of errors such as breed reclassification. No other abrupt



Supplementary Figure 11: **The effect of negative frequency-dependent selection on the stationary abundances of types.** (a) Introducing frequency-dependent selection creates (b) a cusp in the type abundance distribution, compared to a purely neutral simulation. The location of the cusp coincides with the frequency where selection $s(p)$ transitions from favoring to disfavoring types (grey bars). In the simulation results reported here we ensure that both the neutral- and frequency-dependent simulations have reached equilibrium by verifying that the stationary abundance distributions are the same regardless of the initial configuration of the population (monomorphic or diverse).

transitions in the dataset individually affected the results of inference (Supplementary Figure 10).

As with baby names, we find astronomically low p-values for rejecting neutrality in favor of frequency-dependent selection in dog breeds based on likelihood ratio statistics ($\chi^2 = 3748.9$ with 9 d.f., or $\chi^2 = 125.0$ accounting for $N/N_e = 30$, (Supplementary Information section 5)). We report $p<0.001$ in the main text.

4 Frequency-dependent selection and stationary abundance distributions

We illustrate the effect of frequency-dependent selection on the equilibrium distribution of type abundances by comparing neutral simulations to simulations with frequency-dependent selection (Supplementary Figure 11). Introducing negative frequency-dependent selection creates a cusp in the distribution of type abundances. In particular, the frequency where the cusp occurs coincides with the frequency where $s(p)$ crosses replacement fitness, i.e. the frequency associated with zero growth rate.

To illustrate this point we simulated a 70-generation neutral time series using Wright-Fisher generations with time-varying census population sizes equal to the birth cohort sizes in the Netherlands dataset. We chose the mutation rate $\mu = 0.0003$ by examining which values of μ created populations where the most frequent types achieved frequencies comparable to the most frequent names in the Netherlands dataset (0.02-0.03). These parameters matched the Netherlands data with respect to census population size and the frequency of the most abundant type only. We made no attempt to match any other observed aspects of the Netherlands dataset, including the effective population size (drift rate), mutation or

innovation rate, or type abundance distribution; and so the simulation generally differs from the data in each of these attributes.

We measured the distribution of type abundances (Supplementary Figure 11, also known as the progeny distribution[51]) by recording the frequency of each type that occurred in the simulation, summing all instances of each type over all generations. We ensured that simulations reached equilibrium by simulating 10,000 burn-in generations beginning from two contrasting initial population states: one monomorphic state in which the population consisted of a single type, and one diverse state in which the population was initialized with the first generation of the Netherlands dataset. Convergence of the type abundance distributions independent of initialization indicates that the simulations reached equilibrium (Supplementary Figure 11).

We simulated frequency-dependent selection using a model identical to the neutral simulation except that fitness of each type was assigned according to a curve $s(p)$ inferred from the Netherlands 5-year time series with 25 log-equally-spaced bins (Supplementary Figure 11a).

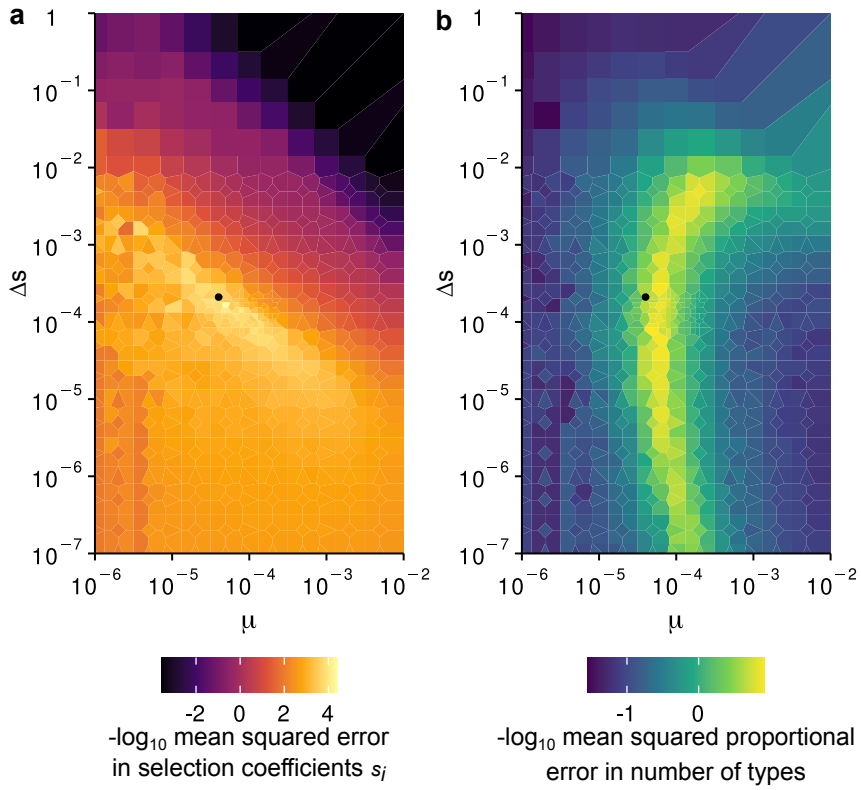
As Supplementary Figure 11 shows, introducing negative frequency-dependent selection imparts a cusp into the abundance distribution of types, at the frequency at which $s(p)$ crosses the replacement fitness.

5 Novelty selection model

We compared the form of frequency dependent selection inferred from the AKC time series to simulations of an evolutionary process that favors novel types—a novelty selection model. The novelty selection model is an infinite-alleles Wright-Fisher process in a population of N individuals with on average $N\mu$ new types introduced per generation. We number each new type sequentially, giving type number 1 fitness 1 (selection coefficient $s = 0$) and type number n selection coefficient $s = n\Delta s$. We use four parameters: the constant census population size N, the constant effective population size N_e, the selection benefit to each novel type Δs, and the per capita per generation mutation rate μ. Thus after long times, both the average selection coefficient in the population (mean fitness) and the selection coefficient of the fittest type increase at rate $N\mu\Delta s$ per generation on average. Hence the relative fitness of the fittest type to the rest of the population is also constant on average after long times. We allow the strength of genetic drift to vary independently of census population size N by simulating g timesteps of evolution for each round of mutation and selection, producing an effective population size $N_e = N/g$. Hence one *generation* of simulation is composed of one Wright-Fisher timestep incorporating mutation and selection and $g - 1$ timesteps of neutral Wright-Fisher evolution.

Notably, types in the novelty selection model are not exchangeable, as in our model of pure frequency dependent selection, but nevertheless novelty selection induces an effective frequency dependent growth rate—namely the average growth rate of types that fall in a given frequency range.

We fitted parameters of the novelty selection model to the AKC data by conducting grid searches across a range of μ (10^{-7} to 0.005) and Δs (10^{-7} to 0.5) on roughly log evenly spaced grids (Supplementary Figure 12). We recorded the average selection coefficient in simulations s_{ave} for types that fall in each frequency bin used in the AKC inference—that is, we measured the equilibrium effective frequency dependence in the novelty selection model. We searched for parameter sets that minimize the squared difference in the effective frequency dependence measured in the simulations versus the frequency dependence inferred from the AKC data, by summing over frequency bins $(s_{\text{ave}} - \hat{s}_{\text{AKC}})^2$. We also looked for agreement in overall diversity of types by tracking the squared difference in the log average number of types in each frequency range in the simulation versus in the AKC data. Parameter sets that agreed in diversity were clustered in a narrow band of mutation rates; whereas parameter sets that agreed in effective frequency dependence were clustered in a narrow diagonal band with $\mu\Delta s \approx$ const.. At all values of N and N_e investigated, these two bands had an identifiable intersection that produced good matches in both frequency dependence



Supplementary Figure 12: **Fitting novelty selection model via** $s(p)$**.** (a) Inferred frequency-dependent selection and (b) type diversity under the novelty selection model match AKC for certain values for the model parameters Δs, μ. The black dot indicates the chosen parameters in Figure 4. Mean squared errors in selection coefficient (a) and proportional number of types ($\log(n_{sim}/n_{AKC})^2$) (b) are computed by averaging across bins and up to 12 replicate simulations.

and diversity, indicating that μ and Δs are identifiable given N and N_e. Near the least-squares minimum, the measured shape of effective frequency dependence from simulations was similar for different values of Δs holding other parameters constant, so we easily matched the range of selection coefficients between the simulation and AKC data (dominated by the growth rate of the rarest frequency bin) by binary search on Δs.

We took N to equal the maximum census population size in the AKC dataset in order to cover the same range of frequencies, namely N=1,435,737. Increasing N beyond this value and refitting μ and Δs as above revealed optimal parameter sets that lie on a line of constant $N\mu\Delta s$. These results concord with the intuition that the average rate of fitness increase due to the appearance of novel types is what primarily determines the strength of the effective frequency dependence. Varying N_e showed that the particular value of $N\mu\Delta s$ that coincided with the best fits was roughly proportional to N_e, indicating that N_e and $N\mu\Delta s$ are not separately identifiable based on diversity and effective frequency dependence alone.

Conversely, determining N_e produced an absolute estimate of $N\mu\Delta s$=1.2% without knowledge of N. Values of N_e close to 10,000 and 50,000 produced stochastic fluctuations in novelty selection simulations that visually match those in the AKC time series, and $N_e \approx 50,000$ gave the best results for intermediate frequencies (Figure 4b vs c). The AKC dataset itself displays different levels of stochasticity over time, as its census population size varies over a factor of 30 from 46,788 in 1931 to 1,435,737 in 1992. To achieve this realistic N_e, we simulated with N=1,435,737 and $g = 30$ giving an $N_e = 47,857.9$. Given these values of N and N_e, and following the fitting procedure described above, we chose $\mu = 0.00004$

(to match the diversity of existing types observed in the AKC time series) and we chose $\Delta s = 0.00021$ (to match the maximum s in any frequency bin in the AKC inference). Since values of N_e ranging from 15,000 to 150,000 are all plausible, and this range of N_e far exceeds the error in finding best fits of μ and Δs given N_e, we estimated $N\mu\Delta s$=0.012 up to roughly a factor of three in either direction.

	United States	France	Netherlands (1y)	Netherlands (5y)	Norway
Country population size	128M – 326M	40M – 67M	9M – 17M	10M – 17M	3M – 5M
Total births	291,183,474	60,314,639	15,408,439	14,351,523	4,317,696
Unique names	104,135	32,341	409,511	400,716	1,870
Censorship	<5 births/year	<20 or <4/year	none	none	<4 births/year
Years used	1936 – 2018	1946 – 2018	1946 – 2014	1950 – 2010	1946 – 2019
Birth cohort size	2,181,299 – 4,410,023	739,217 – 921,077	165,071 – 284,858	859,490 – 1,386,908	44,934 – 72,436
Harmonic mean size	3,592,152	823,123.3	217,961.3	1,079,276	60,732.94
Birth cohort duration	1 year	1 year	1 year	5 years	1 year
% births in top 100	34%	38%	30%	29%	37%
Imputed total births	305,742,609	—	—	—	4,533,582
Inferred μ	—	—	0.067	0.044	—
Inferred $\hat{N}_e$	435,541	57,332	112,919	147,453	20,673
Inferred $\hat{N}_{e,s=0}$	430,075	57,434	137,813	145,038	20,322
Inferred $N/\hat{N}_e$	7.8	13.1	1.9	7.3	2.5
Censorship error	≤0.01 %/year	≤0.02 %/year	—	—	≤0.08 %/year
Sampling error	≤0.08 %/year	≤0.39 %/year	≤0.39 %/year	≤0.04 %/year	≤0.59 %/year*
Neutral null model	$s(p) = -s_w$	$s(p) = -s_w$	$s(p) = 0$	$s(p) = 0$	$s(p) = -s_w$
LR χ^2	1576.8	926.0	50263.2	7466.7	250.9
χ^2 degrees of freedom	9	10	9	9	9
Significance	***	***	***	***	***

*discounting anomalous control in the highest frequency bin.

Supplementary Table 1: **Summary of name datasets.** Here we list general features of each name dataset such as the total number of births, the number of unique names and the type and level of censorship, as well as ancillary imputed and inferred values for each dataset including the number of censored individuals and effective population sizes. Country population size lists the range of reference total number of living persons (in million persons) in each country over the timespan of the inference. The total number of births refers to the number of censored or uncensored births listed in the data. Where the true number of censored individuals is unknown (US and Norway), we list the imputed total number of births. The birth cohort size lists the range of the census population sizes used for inference, including imputed censored counts. The harmonic mean size indicates the harmonic mean of birth cohort sizes across each time series. The percent of births in the top 100 lists the percentage of (imputed) total births over the full time series represented by the top 100 names. The inferred variance-effective population sizes $\hat{N}_e$ and $\hat{N}_{e,s=0}$ assuming $s(p) = 0$ are computed from the frequency-dependent and neutral model residuals, respectively (Supplementary Information section 1.12). Censorship error and sampling error refer to maximum bias (in %/year) within the plotted range. The neutral null model has either a single free parameter, the wildtype fitness, in the case of a country without censorship; or it has no free parameters. The likelihood ratio χ^2-statistic is reported for comparison between the frequency-dependent model versus the nested neutral model, accounting for the effective population size as described in Supplementary Information sections 1.14 and 2.2. Significance refers to the χ^2 p-value for rejecting neutrality in favor of frequency-dependent selection in a likelihood ratio test (Supplementary Information section 2.1). All p-values (***) are <0.001.

Total reigstrations	53,575,962
Unique breeds	153
Years used	1926 – 2005
Registration cohort size	46,788 – 1,435,737
Harmonic mean cohort size	201,048
Sampling error	$\leq$0.95 %/year above 10^{-4}
Neutral null model	$s(p) = 0$
LR χ^2	125.0
χ^2 degrees of freedom	9
Significance	***

Supplementary Table 2: **Summary of AKC dataset.** Here we present summary attributes of the American Kennel Club dataset used for inference after data handling described in Supplementary Information section 3.2. Sampling error varies substantially by bin and is depicted for each bin in Supplementary Figure 10. We compute the LR χ^2 statistic accounting for $N/N_e = 30$ as described in Supplementary Information sections 1.14 and 2.2. Significance *** indicates p<0.001.