PLOS Genetics

METHODS

Isolating selective from non-selective forces using site frequency ratios

Jody Hey*, Vitor A. C. Pavinato

Department of Biology, Temple University, Philadelphia, Pennsylvania, United States of America

* hey@temple.edu

Abstract

A new method is introduced for estimating the distribution of mutation fitness effects using site frequency spectra. Unlike previous methods, which make assumptions about non-selective factors, or that try to incorporate such factors into the underlying model, this new method mostly avoids non-selective effects by working with the ratios of counts of selected sites to neutral sites. An expression for the likelihood of a set of selected/neutral ratios is found by treating the ratio of two Poisson random variables as the ratio of two gaussian random variables. This approach also avoids the need to estimate the relative mutation rates of selected and neutral sites. Simulations over a wide range of demographic models, with linked selection effects show that the new SFRatios method performs well for statistical tests of selection, and it performs well for estimating the distribution of selection effects. Performance was better with weak selection models and for expansion and structured demographic models than for bottleneck models. Applications to two populations of *Drosophila melanogaster* reveal clear but very weak selection on synonymous sites. For nonsynonymous sites, selection was found to be consistent with previous estimates and stronger for an African population than for one from North Carolina.

Author summary

A new statistical method is presented for estimating the distribution of strengths of natural selection acting on mutations in natural populations using the distribution of polymorphic site allele frequencies. In order to isolate the impact of selection, separately from other demographic and genomic factors that can shape allele frequencies, our method uses the ratio of the frequency of candidate selected variants to the ratio of the frequency of neutral variants of the same frequency. An expression for the overall likelihood across the range of frequency ratios is developed using a gaussian approximation. Testing of the method, called SFRatios, finds that it performs reasonably well across a range of strengths of selection and demographic histories. Applications to two Drosophila populations find estimates of the strength of selection on nonsynonymous coding variants consistent with previous estimates and estimates for synonymous variation quite close to selectively neutral.

Check for updates

OPEN ACCESS

Citation: Hey J, Pavinato VAC (2025) Isolating selective from non-selective forces using site frequency ratios. PLoS Genet 21(4): e1011427. https://doi.org/10.1371/journal.pgen.1011427

Editor: Kirk E Lohmueller, University of California Los Angeles, UNITED STATES OF AMERICA

Received: September 12, 2024

Accepted: March 24, 2025

Published: April 21, 2025

Data availability statement: The SFRatios program, along with simulation scripts, as well as script for building the Drosophila data sets, and other scripts used in the analysis of the site frequency spectra from the Drosophila populations, are available at https://github.com/jodyhey/SFRatios.

Funding: This research was supported by National Institutes of Health (NIH.gov) research grant R01GM144468 to JH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Population genomics is often a science of sifting signal from noise as investigators regularly seek to distill the signs of natural selection from the confusing patterns of variation that arises from other factors [1–4]. These other factors are quite diverse with some being especially noise-like (genetic drift, recombination and gene-conversion events, and the effects of linked random mutations) and others that have a directional (i.e., non-random) component, as occurs when the demographic history departs from the assumptions of the investigator's model. Here we describe a new approach for estimating the distribution of selection coefficients acting on mutations, but that does so while largely sidestepping the confounding effects of all these other factors.

For questions about selective effects investigators have often employed a classic body of theory on the distribution of allele frequencies [5,6], also known as the site frequency spectrum, or SFS. An important theoretical advance was the realization that the count of observed sites with an allele at a particular frequency could be modelled as a Poisson random variable, with an expected value that depends on a particular model of directional selection and population demography [7,8]. These Poisson random field (PRF) models provide accessible likelihood formulae, not only for the estimation of single selection coefficients, but also for the estimation of the distribution of selection coefficients [9–11].

The original PRF work was limited to constant size Wright-Fisher (WF) populations. To allow for departures from WF models, these methods have been adapted for joint estimates of selection and demography under models of population size change [12–15] and with gene flow between subpopulations [16]. Nevertheless, real populations can have histories that vary in many ways not accounted for with these methods. Nor do such methods account for the many other non-selective non-demographic factors that can shape the distribution of allele frequencies. For example, if rates of gene conversion and gene conversion bias are high enough, they will alter the site frequency spectrum, as will variation in mutation rates if some sites have a high enough mutation to result in some polymorphic sites being caused by multiple mutations. Variation in recombination rates can also affect allele frequencies by shaping the degree to which some sites, more than others, are affected by selection on linked sites. Finally, the structure of the sample across subpopulations can have a very large effect on the site frequency spectrum, one that may easily not be appreciated if there are unknown subdivisions within the sampled population(s).

One way to improve upon methods that attempt to jointly estimate selection and other factors is to include in the analysis a set of neutral control variants that are thought to be affected only by non-selective factors. In particular, Eyre-Walker and colleagues developed a method that uses the joint likelihood of selected and neutral variants, and represents shared, non-selective factors by a series of nuisance parameters, one for each frequency bin [10,17]. This approach has been adopted in a number of studies and applications [18–21].

However the method of using both selective and neutral sites, while potentially solving one problem, introduces another complication, which is that the respective mutation rates for each of the two classes must be estimated. This can be done

using a previously estimated mutation rate and by assuming a particular demographic history, or by jointly estimating that history. However in most applications it is handled by including in the likelihood function factors L_N and L_S, the number of sampled sites where a neutral or selected mutation could occur, respectively. If these values are known without appreciable error, then the counts of sites that are invariant with respect to a sister species can be obtained, and these can be used to estimate the two mutation rates. One important benefit of this approach is that the divergence measures can be used in turn to estimate the rate of adaptive substitution [9,17,20,21].

For methods that do not include neutral controls or divergence between sister species, and rely only upon the frequency distribution of polymorphisms, the actual mutation rate is not a parameter of much interest. In these cases, it is the changing relative height of the polymorphism count across frequency bins that informs on the effects of selection. However, the methods that use both selected and neutral variants all depend on knowing or successfully estimating the underlying mutation rate, which often depends on knowing the value of the number of sampled neutral positions, L_N. This value will typically include a large number of invariant positions. However if a subset of these is not variable because they are actually under selective constraint, then L_N value will be too large. Another issue that arises when using L values as a means to include divergence in the analyses, is that non-selective factors may have changed over the course of the divergence process [10].

Here we take a new approach to using a neutral control set for isolating the effects of direct selection on a set of variants. But unlike other methods that depend on estimates of the overall mutation rates to selective and neutral variants, our method does not depend on estimates of the mutation rates for each class of variant. The method depends not at all on estimates of species divergence or estimates of the number of sampled selected and neutral positions.

Results

Model

We consider both a neutral set and candidate selected set of biallelic single-nucleotide polymorphisms (SNPs) sampled from n genomes. Assume for the moment that the derived allele at each SNP is known, such that each SNP can be represented simply by the number of derived alleles that occur in the sample, a value that varies between 1 and $n-1$. Under a Poisson random field model, the focus is on the expected numbers of SNPs at each possible sample frequency. For both sets of SNPs the expected counts with i derived allele copies are the product of two terms: the first of these accounts for the rate of incoming mutations, denoted as $\theta/2$ and $\theta_S/2$ for neutral and selected respectively; the second term, denoted as $\mathcal{N}_i$ and $\mathcal{F}_i$ respectively, are both functions of non-selective effects (i.e., population demography, linked selection effects, etc.) and in the case of $\mathcal{F}_i$, the direct effects of selection. Thus, the number of neutral SNPs in bin i follows a Poisson distribution with expectation $\mathcal{N}_i\,\theta/2$, while for selected SNPs the expectation is $\mathcal{F}_i\,\theta_S/2$. For simple diploid Wright-Fisher populations, with free recombination, and lacking any mutational biases, the terms for incoming mutations, $\theta/2$ and $\theta_S/2$, as well as $\mathcal{N}_i$ and $\mathcal{F}_i$, are well understood and can be specified as functions of a small set of parameters. Specifically: $\theta = 4Nu$ where N is the population size and u is the neutral mutation rate; $\theta_S = 4Nu_S$, where u_S is the mutation rate to selected alleles; and $\mathcal{N}_i = 1/i$. The quantity $\mathcal{F}_i$ is a function of the population selection coefficient of the derived allele, $\gamma = 2Ns$, and is found by integration of a term for the distribution of selected allele frequencies in the population over the probability of sampling i alleles in a sample of size n (7):

$$\mathcal{F}(\gamma)_i = \int_0^1 h(\gamma, x)\,\binom{n}{i} x^i (1-x)^{n-i} dx\,. \tag{1}$$

Where $h(\gamma, x)$ is the expected density for derived alleles at frequency x in the population:

$$h(\gamma, x) = \frac{\theta_S\,(1 - e^{-2\,\gamma(1-x)})}{(1 - e^{-2\,\gamma})\,x\,(1-x)}. \tag{2}$$

The approach can readily be extended to accommodate a distribution of fitness effects (DFE). Following Boyko et al., [9] if we have a probability density $\Pr(\gamma = 2Ns) = g(\gamma, \phi)$, where ϕ contains the parameters for the DFE, then an SFS generated under selection will have an expected count for i sampled alleles of $\mathcal{F}_{g,i}\,\theta_S/2$, where

$$\mathcal{F}_{g,i} = \int_{-\infty}^{\infty} g(\gamma, \phi)\mathcal{F}(\gamma)_i d\gamma. \tag{3}$$

We would like to adapt these methods to problems well outside of the constraints of Wright-Fisher assumptions, to include models with complex histories, and do so without additional parameters for demography or any other non-selective factors that might shape the SFS.

For a data set of unknown history, the expected counts can still be envisioned as the product of a term for incoming mutations and a term for the fraction that are sampled in bin i. Now let us suppose that whatever the history, that the ways that non-selective effects shape the expectation for bin i can be mostly captured in a term a_i, and that this term applies to both neutral and selected SNPs. Further suppose that the effects of selection can primarily be captured separately from the effects captured in a_i, by redefining $\mathcal{F}_{g,i}$ as a function not strictly of the probability density of $\gamma = 2Ns$, but rather by shifting the meaning of γ so that it is a function, not of census size, but of effective population size (i.e., $\gamma = 2N_e s$). The change in parameter does not affect the model fitting, but it does serve to highlight the uncertain demographic context. We can use N_e as well in redefining the terms for the incoming mutations, $\theta = 4N_e u$ and $\theta_S = 4N_e u_S$. Then for our population of unknown history, the expected count of neutral mutations in bin $i$will be $a_i\mathcal{N}_i\,\theta/2$ and for selected mutations it will be $a_i\mathcal{F}_{g,i}\,\theta_S/2$.

The motivation for supposing that the bulk of non-selective effects can be separated from the effects of selection is that it opens the door to using the ratio of counts. Of course, the reality is that selection does interact with demography and other factors to shape the distribution of allele frequencies, but a ratio-based approach may still be a useful approximation. Then, for a model of arbitrary non-selective factors, for which the departures from a simple WF model can be captured in a term a_i, and by using effective population size rather than census size, the ratio of expected counts in bin i is:

$$\frac{a_i\mathcal{F}_{g,i}\frac{\theta_S}{2}}{a_i\mathcal{N}_i\frac{\theta}{2}} = \frac{\mathcal{F}_{g,i}u_S}{\mathcal{N}_i u}\,. \tag{4}$$

In other words, by taking the ratio of counts, we may be able to work with the standard WF-PRF theory to estimate selection parameters while ignoring whatever non-selective effects might also have shaped the site frequency distribution.

Let z_i be the observed ratio of selected to neutral counts in bin i. If we knew the probability of z_i as a function of the distribution of the incoming mutation rates and the DFE, $p(z_i, \phi, \theta_S, \theta)$, then we could estimate the several unknowns (i.e., θ_S, θ and ϕ) by maximizing the log-likelihood:

$$\sum_{i=1}^{n-1} Log(p(z_i, \phi, \theta_S, \theta)). \tag{5}$$

The use of a ratio of candidate selected counts to neutral counts is also supported by the fact that the absolute number of SNPs, and thus the components of θ and θ_S corresponding to the rate of incomings mutations, can be quite large when working with whole genome data. With such large data investigators have the option of filtering the data by some criterion or focusing on a particular part of the genome. In such cases of data sets of preset or tailored size, θ and θ_S are, individually, nuisance parameters. In fact, it turns out that it is practical to work without attempting to estimate both mutation rates, but instead to estimate just the ratio of mutation rates.

The probability of an observed ratio, z_i

Let the data be taken as a set of $n-1$ ratios, where z_i is the ratio of the observed count of selected SNPs in bin i to the corresponding count for neutral SNPs. If we use the basic WF-PRF model for both the numerator and denominator, as in [1], then z_i will be the ratio of two Poisson random variables. However, as these are each discrete random variables, their ratio is also a discrete random variable with a complicated probability density [22]. For a more tractable likelihood calculation, we assume that both the numerator and denominator counts follow a gaussian distribution in which the expectations and variances are both equal to the expectations of their respective Poisson distributions. Clearly the gaussian is continuous, but otherwise a gaussian with expectation and variance X provides quite a good fit for a Poisson density with parameter X for values even as low as 10. With genome data it is often possible to have more than 10 SNPs in each bin, even for selected SNPs and even for high values of i.

For the probability density of the ratio of two normal distributions, we use the formulation by Díaz-Francés and Rubio [23]. Because the expectation equals the variance in the case of a Poisson random variable, and thus also for the gaussian distributions used in the approximation, this formulation becomes:

$$p\left(z_i|\phi,\rho,\theta\right)=\left(\frac{e^{\left(\frac{\beta-1}{2\ \delta^2}\right)}\sqrt{\beta}}{\pi(z_i{}^2+\beta)}\right)\left(1+\frac{(1+z_i)\ e^{\left(\frac{(1+z_i)^2\beta}{2(z_i{}^2+\beta)\ \delta^2}\right)}\sqrt{\frac{\pi}{2}}\ Erf\left(\frac{1+z_i}{\delta\sqrt{2\ \frac{z_i{}^2+\beta}{\beta}}}\right)}{\delta\sqrt{\frac{z_i{}^2+\beta}{\beta}}}\right) \tag{6}$$

where $\beta=\frac{\mathcal{F}_{g,i}u_S}{\mathcal{N}_i u}=\rho\frac{\mathcal{F}_{g,i}}{\mathcal{N}_i}$, $\rho=\frac{u_S}{u}$, $\delta=(\mathcal{N}_i\theta)^{-1/2}$, and $Erf()$ is the error function. The only place that a mutation rate term appears outside of a ratio is in δ, which shapes the variance of the density but has modest effect on the expectation. As the individual mutation rate terms are nuisance parameters, we can consider integrating over θ to remove the δ term and generate a simpler version of (6) that is a function of only the DFE and ρ. This we do, numerically, over a uniform log scale density of θ which extends over two orders of magnitude between $10\ \hat{\theta}$ and $\hat{\theta}/10$, where $\hat{\theta}$ is Watterson's estimate of θ.

The result is a density that is a function of the data, i.e., a set of ratios from the selected and neutral SFSs, the ratio of mutation rates, and the parameters of the DFE: $\acute{p}\left(z_i|\phi,\rho\right)$. Then the log-likelihood of a set of ratios for unfolded SFSs is simply

$$\sum_{i=1}^{n-1} Log(\acute{p}\left(z_i|\phi,\rho\right)). \tag{7}$$

For the folded distribution we substitute $\beta=\ \rho\ \frac{(\mathcal{F}_{g,i}+\mathcal{F}_{g,n-i})}{(\mathcal{N}_i+\mathcal{N}_{n-i})}$, and the log-likelihood is summed over $1\leq i\leq\frac{n}{2}$.With large amounts of data, expression (7) should open the door to all the applications for which likelihoods are suitable, including parameter estimation, hypothesis testing and model choice. We have named the use of expression (7) for estimating ρ and ϕ as the "Site Frequency Ratios" method, which we abbreviate as SFRatios.

Working with the ratio of mutation rates, ρ

A key parameter is ρ, the ratio of the total rate of mutation to selected alleles over that portion of the genome screened for selected variants, divided by the corresponding rate for neutral variants. This ratio should not depend on any factors other than these mutation rates (i.e., no effects of selection, demography, linkage, or sampling geography). However, ρ will depend on the relative sampling effort of the two classes of SNPs. For example, if SNPs in the selected pool were sampled from a smaller fraction of the genome than for neutral alleles, then we would expect an estimate, $\hat{\rho}$, to be below one.

Despite being a function of sampling effort, an estimate of ρ can be useful in a at least two different ways. One use of $\hat{\rho}$ is to consider it together with the total counts of selected and neutral SNPs, which we denote by X and Y respectively. If in fact both sets were strictly neutral, then a useful estimate would simply be $\hat{\rho} = X/Y$. But when there is selection on the SNPs in the numerator, the difference between ρ and X/Y is caused by a difference in rates at which mutations that occurred in the population where unsampled when the data were collected. In particular, deleterious mutations will have a higher chance of going unsampled, on average, relative to neutral mutations, because they have been lost from the population or are more likely to be at extreme allele frequencies. Let λ be the probability of not sampling a selected mutation, relative to the probability of not sampling a neutral mutation. Then $\rho = \lambda X/Y$ and we can estimate the relative probability of sampling as

$$\hat{\lambda} = \hat{\rho}\, Y/X. \tag{8}$$

For example, suppose that $X/Y = 0.2$ and $\hat{\rho} = 0.35$, then $\hat{\lambda} = 1.75$, which tells us that a selected mutation is about 75% more likely to go unsampled, relative to a neutral mutation. λ can be considered as a measure of the variation that is missing due to the direct effects of selection on deleterious or beneficial alleles. Missing variants will include those that have been lost or fixed, as well as those that fall in frequency ranges that are less likely to be sampled from.

Another use of $\hat{\rho}$ is to apply it to the estimation of selection parameters for other populations of the same or related species. To see this, suppose that we have a value of $\hat{\rho}$ for one population for which we are relatively confidant that the circumstances are favorable for an accurate estimate. And now allow that there is a second population of the same or closely related species for which we have SNP counts based on the same sampling process as the first population, but for which we are more doubtful that the approach of using the ratios of counts will work for estimating selection parameters. Because the two populations are closely related, and SNPs have been sampled from the same genomic regions, we may be comfortable assuming that the underlying mutation rates are shared by the two populations, and thus that the true value of ρ is the same for both populations. Then we can take $\hat{\rho}$ for the first population and fix ρ for the second population at that value, with the hope that the estimates of selection parameters for the second population will be more accurate than if ρ were also being estimated for that population.

Simulation results

Qualitative assessment of SFRatios. Simulations were conducted to qualitatively assess the underlying rationale of using the ratio of selected to neutral SFS values to isolate the effects of selection. We simulated SFSs, and the selected/neutral SFS ratios, for several demographic models, and compared them to the basic WF model of constant population size. Fig 1A shows mean values of simulated folded SFSs under a constant WF model and three demographic models, for both neutral mutations and mutations with selection coefficients drawn from a lognormal distribution. To aid comparison among models, which vary in their absolute numbers of polymorphic sites, for each model, the values shown are scaled relative to the count for the singleton bin (i.e., $i = 1$, just one observed copy of the rare allele) for the corresponding neutral model, and then plotted on a logarithmic scale. For this demonstration, the ratio of the rate of incoming mutations was set to $\frac{\theta_S}{\theta} = 1.0$. Comparison reveals how the selected sites have an SFS that departs greatly from the neutral case, regardless of the model, and it shows how the models vary considerably in their SFSs, both with and without selection. However, the ratios of these same values, of selected to neutral SFSs, are quite similar for the different demographic models (Fig 1B).

Performance on simple tests of selection. One kind of application of PRF theory is to use a likelihood ratio test to determine whether a WF model that includes selection provides a significantly better fit to a data set than a strictly neutral WF model (8). Such tests can be quite powerful, however, without a way to control for the non-selection-based effects on the SFS, they will be sensitive to virtually any departure from a WF model.

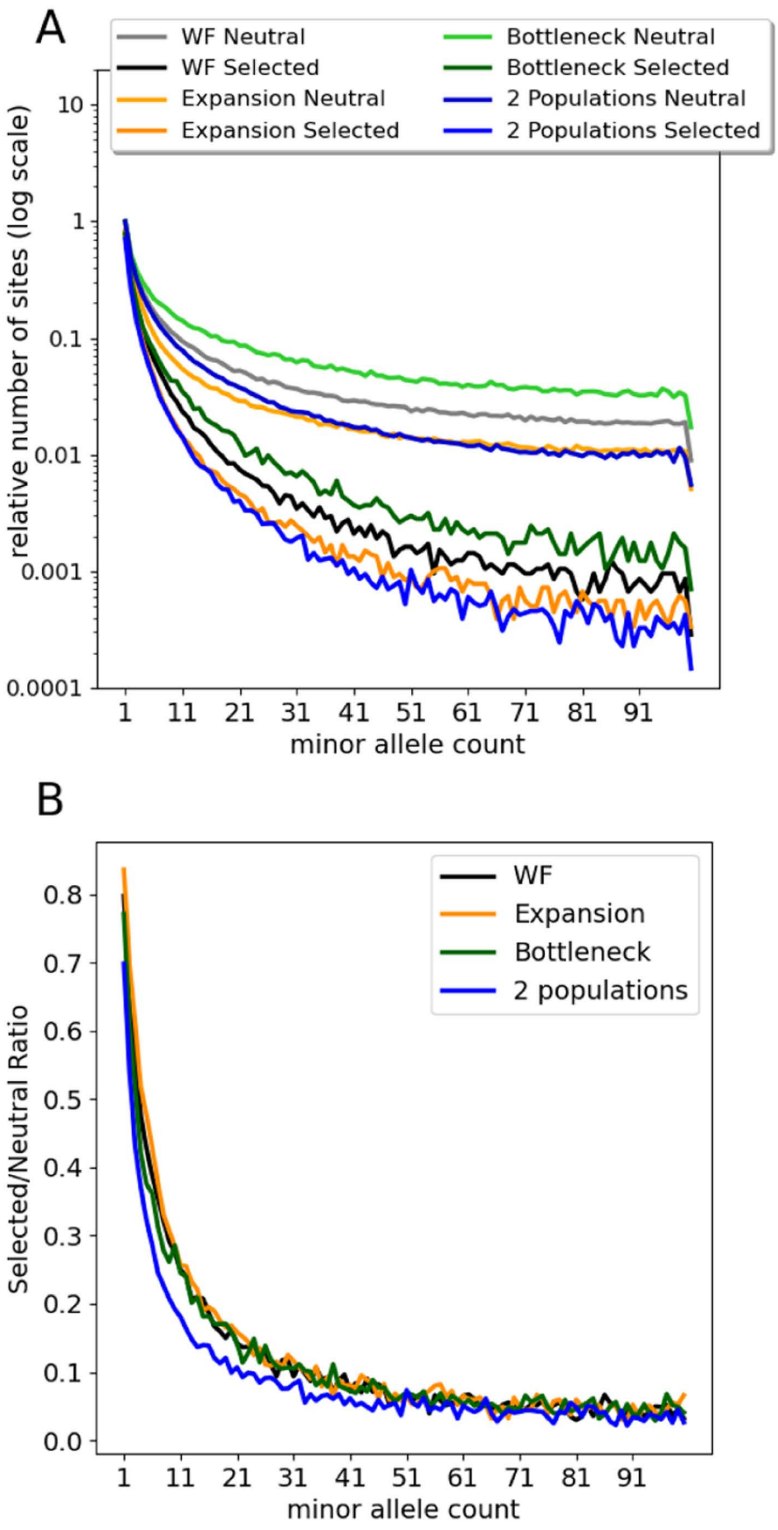

Fig 1. Comparison of SFSs and Selected/Neutral ratio for Wright-Fisher (WF) and other demographic models. Selection model: 2Ns~Lognormal(3.0,1.2) (see methods), with expectation -40.3. For all simulations, $\frac{\theta_S}{\theta} = 1.0$. **A.** Folded SFS values for a sample of 200 chromosomes, scaled to the value for allele count 1. **B.** The ratio of selected to neutral counts for folded SFSs.

https://doi.org/10.1371/journal.pgen.1011427.g001

We considered whether the likelihood in expression (7) can also be used for simple tests of selection by comparing the performance to that based on regular SF-based likelihoods. Computer simulation of SFSs with selection were examined using likelihood ratio tests and subjected to power analyses and receiver operator characteristic (ROC) curve analyses. The same kinds of analyses were then conducted with each simulated data set transformed into a series of ratios using a simulated neutral control and then examined using a likelihood ratio test.

Fig 2 shows results for the statistical performance of basic tests of the hypothesis that the sampled data came from a population with $\gamma = 0$. Each panel compares simple WF-PRF simulations in which the data were sampled from a Poisson distribution for a sample size of 100 genomes, with $\theta_s = 500$, to the SFRatios method in which each data set of selected alleles is paired with a neutral data set of 100 genomes and with $\theta_s = \theta = 500$. In each case, statistical significance was determined by comparing the χ_1^2 value for each of three false positive rates (0.05, 0.01, 0.001) to twice the log of the likelihood ratio. These are

$$2 \times \left(argmax_{\phi,\ \theta} \sum_{i=1}^{n-1} Log(f(k_i, \phi, \theta)) - argmax_{\theta} \sum_{i=1}^{n-1} Log(f(k_i, 0, \theta)) \right), \text{ and}$$
$$2 \times \left(argmax_{\phi,\ \rho} \sum_{i=1}^{n-1} Log(\acute{p}\,(z_i|\phi, \rho)) - argmax_{\rho} \sum_{i=1}^{n-1} Log(\acute{p}\,(z_i|0, \rho)) \right) \quad (9)$$

where $f(k_i, \phi, \theta) = \dfrac{e^{-\mathcal{F}_i\ \theta_S/2}\ (\mathcal{F}_i\ \theta_S/2)^{k_i}}{k_i!}$, for WF-PRF and SFRatios, respectively.

In Fig 2A is shown the statistical power for $-20 \leq\ \phi \leq 20$. Fig 2B shows ROC curves for SFRatios and WF-PRF for the same hypotheses and test distribution (i.e., χ_1^2) as shown in Fig 2A. For this analysis half of the simulations (500) were done when the null hypothesis is true ($\phi = 0$) and half when $-100 \leq\ \phi \leq 1$, with values sampled uniformly at random. Panel C compares the observed distribution of the test statistic to the cumulative χ_1^2 distribution for 1000 simulations when the null hypothesis is true ($\phi = 0$). Results for smaller data sets, including simulations with low variation and small sample size ($\theta = 50, n = 20$) and low variation and large sample size ($\theta = 50, n = 100$), are given in S1, S2 and S3 Figs.

In general, the analyses based on the WF-PRF likelihood showed more statistical power, a closer fit of the likelihood ratio test statistic to the expected χ^2 distribution, and higher area under the curve (AUC) in the ROC analyses, compared to the SFRatios likelihood tests. However, statistical power was high for both methods for all but the smallest selection coefficients; AUC was high for both methods for all 3 sampling schemes, and the χ^2 approximation fitted well except for the smallest sample sizes, even when the fit was accurate in the tail of the distribution associated with the smaller false positive rates. However, even though the SFRatios likelihoods performed well, it is important to recognize that they are all based on twice as much data as the corresponding SF-based likelihoods, in that, each SFRatios data set had both a selected SFS and a neutral SFS.

Estimator bias under Wright-Fisher and other demographies. Forward population genetic simulations were conducted to include linked neutral and selected mutations under several non-WF models. For fixed values of γ, simulations were conducted under a constant population, recent population expansion, recent population bottleneck, and population structure. For simulations under lognormal and gamma DFE, these same demographic models were used, as well as an African-Origin model of human demographic history [24].

Shown in Fig 3 are boxplots of estimates for each of a series of fixed values of γ for each model. In general, weak selection (beneficial or deleterious) was estimated well in all the models, whereas positive $\gamma \geq 10$ was underestimated in all models (i.e., estimate values were closer to zero). Selection at the strongest level in these simulations ($\gamma = -100$) was overestimated (i.e., estimated values were more negative) in the bottleneck and population structure models.

The results of simulations under a lognormal DFE are shown in Figs 4 and 5, each extending from highly negative values to + 1, in order to accommodate deleterious mutations, neutral mutations, and slightly beneficial mutations.. The lognormal distribution is parameterized using the expected value, μ, and the standard deviation, σ, of the random variable's natural logarithm. We considered five sets of parameter pairs, each of which generates a curve with a peak near

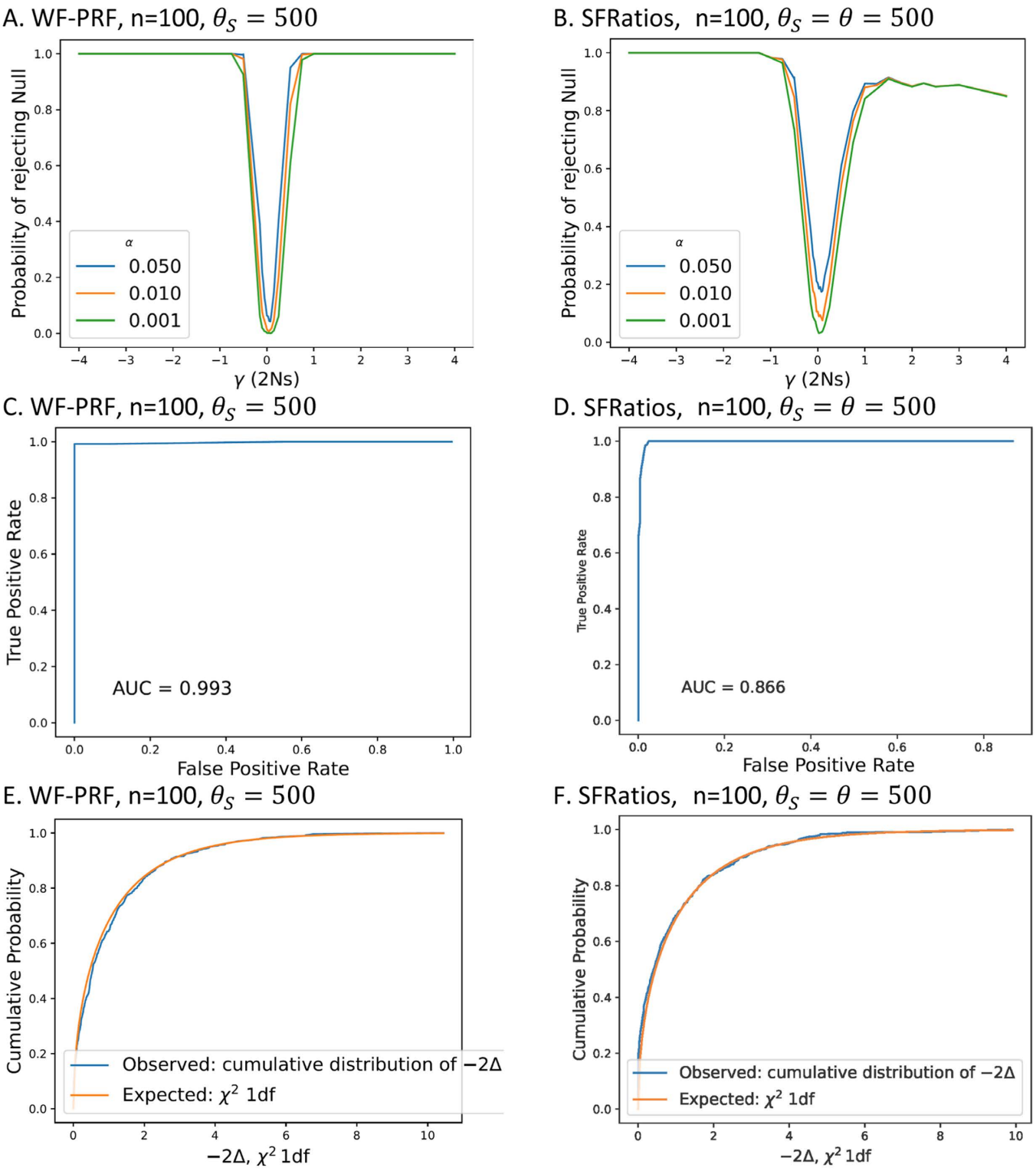

Fig 2. Statistical performance for wright-fisher poisson random field (WF-PRF) and SFRatios. Top row (panels A, **B)**: The probability of rejecting the null (neutral) when the alternative model (selected) is true for different probabilities of false positive (α) and varying strengths of 2Ns. Middle row (panels C, **D)**. Receiver operator characteristic (ROC) curves, with area under the curve (AUC). Bottom row (panels E, **F)**. Cumulative observed distributions of the likelihood ratio test statistic, with χ^2,1df comparison for sets of 500 simulations.

https://doi.org/10.1371/journal.pgen.1011427.g002

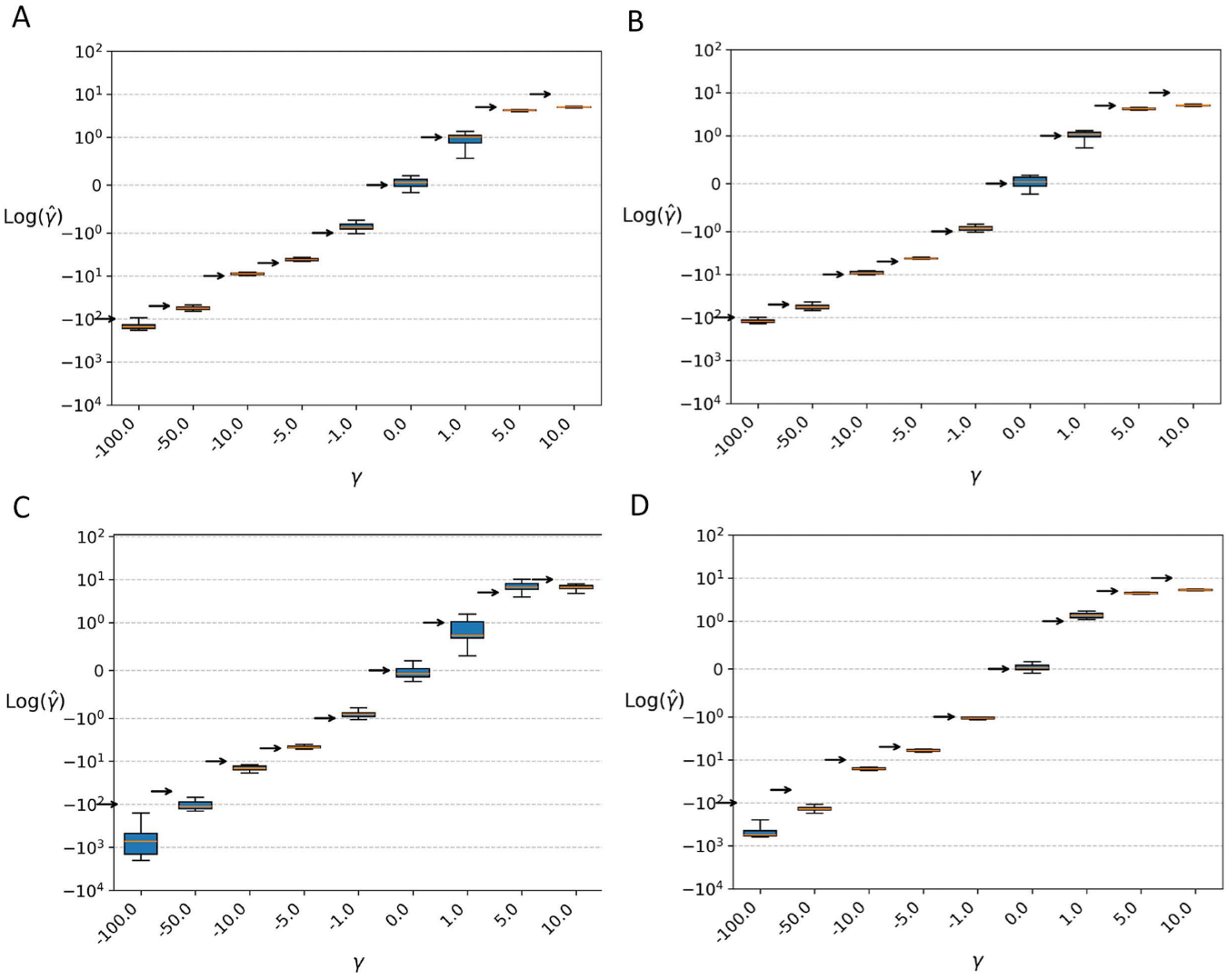

Fig 3. Simulation results for fixed γ values. Boxplots of γ values estimated for simulated data generated under different demographic models. For each box, an arrow indicates the location of the true value. **A.** Population with a constant size. **B.** Population expansion. **C.** Population bottleneck. **D.** Two divergent subpopulations.

https://doi.org/10.1371/journal.pgen.1011427.g003

zero, but with increasingly negative mean values (Fig 4A). Estimator bias for μ and σ is shown in 2D boxplots (Fig 4B-4F) and for ρ in Fig 5.

For the ratio of mutation rates, ρ, bias was near zero or small for all models and parameter sets (Fig 5) with the exception of the lognormal density with the strongest selection (mean ≈ - 1000). For the parameters of the $2Ns$ distribution (Fig 4), statistical bias is low to modest when selection is weak, regardless of the demographic model being used. However for models with strong selection, the statistical bias was higher for some demographies, particularly for the bottleneck and African Origin models. When selection intensity was strongest, estimates had a wider variance (e.g., in the constant and two-population models) or to be underestimated. Overall, estimates of ρ, μ and σ tended to be close to the true values for a wide range of selection models for a wide range of demographies, and this is especially true for weaker selection. With smaller sample sizes (50 chromosomes as opposed to 200 chromosomes for Figs 4 and 5), the results were qualitatively similar but with larger variances (S6 Fig).

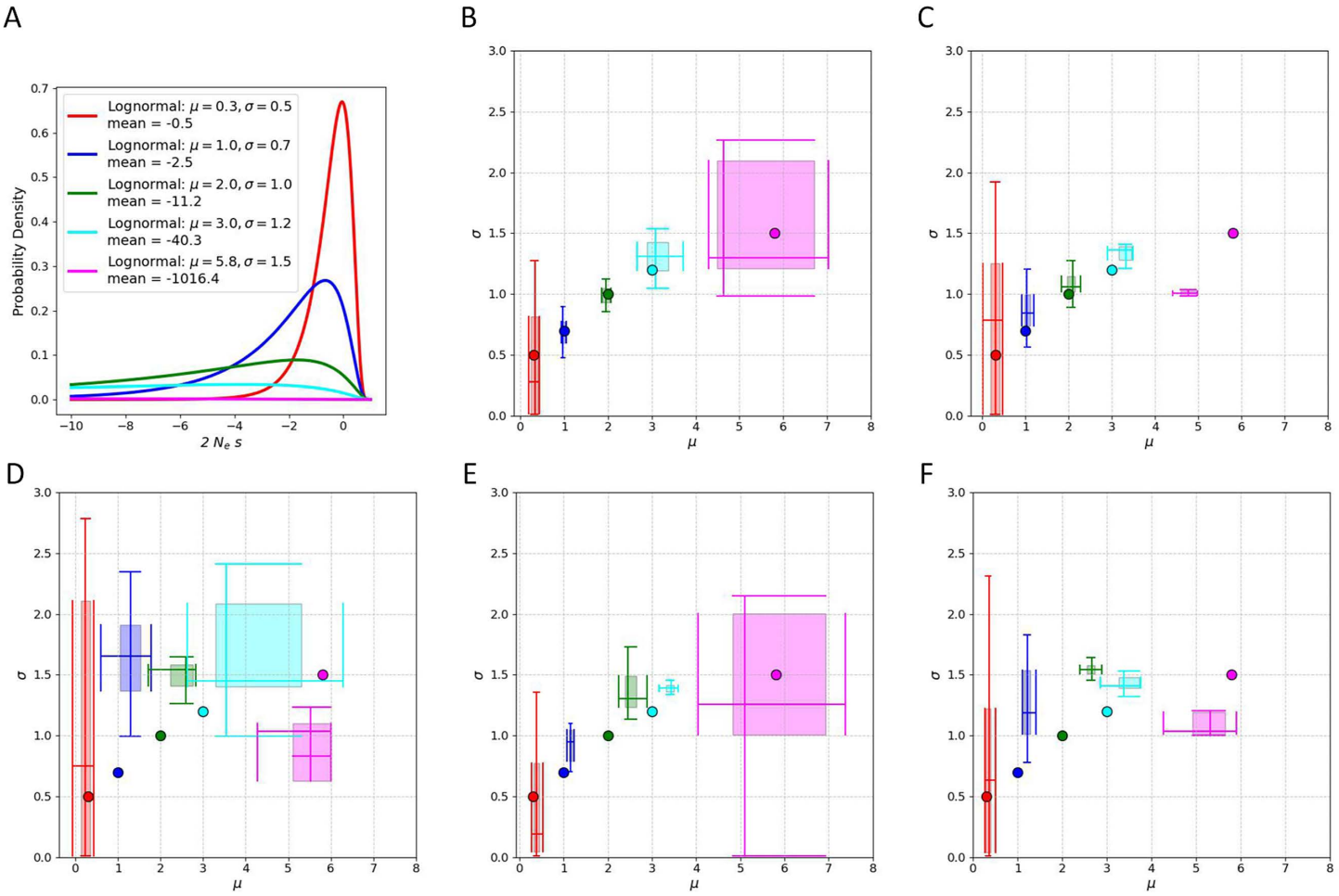

Fig 4. Estimator performance for γ lognormal distribution parameters. For each simulated data set, γ values were drawn from one of 4 lognormal distributions. 20 data sets were simulated for each demographic model and each γ distribution. A. lognormal distributions, for random variable x,$(1 < x < \infty)$,$2Ns = 1\text{-}x$. **B.** Constant population size Wright-Fisher. **C.** Population expansion. **D.** Population bottleneck. **E.** Two populations. **F.** African Origin model.

https://doi.org/10.1371/journal.pgen.1011427.g004

Comparison with fastDFE. fastDFE [19] is an alternative estimator of the distribution of selection intensities that implements a version of the method pioneered by Eyre-Walker and Keightley [17]. Like SFRatios it uses a neutral and a selected SFS, but rather than using the ratio of the two, it estimates the parameters of a selection density by accounting for the covariation in selected and neutral SFSs with a series of nuisance parameters that serve to rescale the neutral SFS to that expected under the standard model. fastDFE, and all methods based on this approach requires counts of invariant sites, which SFRatios does not. However unlike other methods based on the approach of Eyre-Walker and Keightley [17] fastDFE can work with a fully folded SFS, including pooling fixed derived and ancestral sites.

We compared fastDFE and SFRatios for several demographic models under multiple gamma densities (fastDFE does not implement the lognormal density). Results are shown for two models in Table 1 and a larger set of models in S4 and S5 Figs and S1 Table. Results were mixed, with SFRatios performing better under the population expansion model, and with fastDFE performing better in some cases.

Drosophila populations. For two population genomic datasets from *D. melanogaster* we generated folded SFSs for both synonymous and nonsynonymous SNPs. In each case the neutral control set was generated from short introns

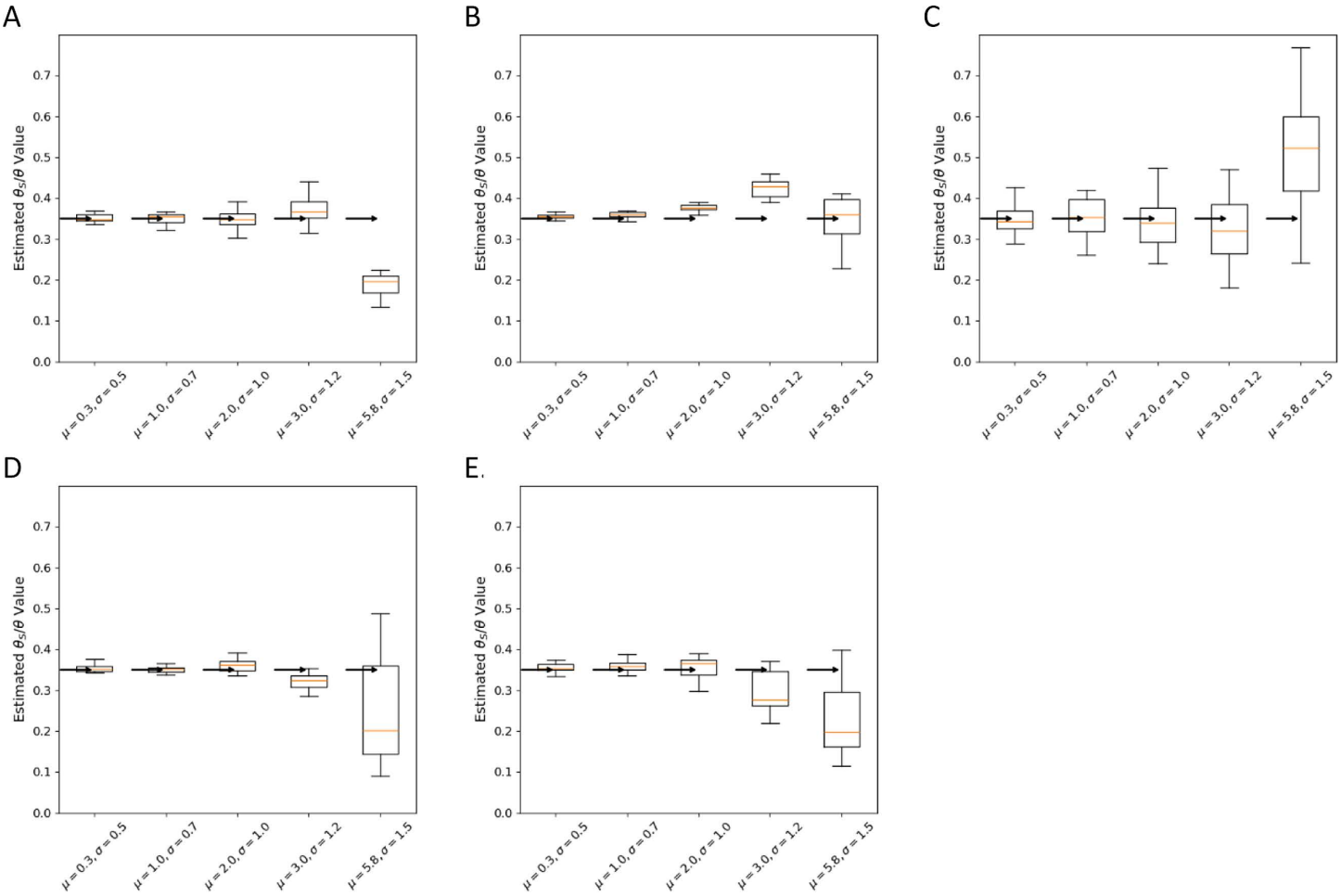

Fig 5. Estimator performance for the ratio of mutation rates, ρ. For each simulated data set, 2Ns values were drawn from one of a 5 lognormal distributions (see Fig 4). 20 data sets were simulated for each demographic model and each 2Ns distribution, each with a true mutation rate ratio of 0.35. **A.** Constant Wright-Fisher **B.** Population expansion. **C.** Population Bottleneck. **D.** Two populations. **E.** African Origin model.

https://doi.org/10.1371/journal.pgen.1011427.g005

[25,26]. To further reduce sources of variance, the neutral control sets were built to reduce variation due to the local sequence context of each SNP. SNPs for the neutral control sets were chosen by identifying for each candidate selected SNP (synonymous or nonsynonymous) the closest short intron SNP having matching flanking bases [18].

The folded SFSs for both populations for nonsynonymous, synonymous, and short-intron SNPs are shown in Fig 6. The relative intensity of selection can be roughly perceived by the flatness of the distributions, with more strongly selected sites showing a greater drop in counts from low frequency bins to higher frequency bins. Given the similarity of the synonymous and short intron SFSs, selection on synonymous sites appears to be nearly neutral.

For both populations we fit three types of continuous distributions (normal, lognormal and gamma) as well as each of these DFEs with the addition of a single point mass at zero of up to 50% of the total density. Model fitting involves estimating either three parameters (i.e., ρ and density parameters) or four parameters (in the case of an added point mass), in which case the same three parameters are estimated as well as the value of density mass at zero. Full results are shown in S2 and S3 Tables with the best fitting models shown in Table 2 and Fig 7.

The SFRatios fitting for nonsynonymous sites suggests stronger selection in the Zambia population (γ mean of -1643,8) than in the North Carolina population (γ mean of -323.8). Given that the Zambia population has had the larger

Table 1. Comparison of SFRatios and fastDFE mean estimates and root-mean-square-error (RMSE) under diverse demographies.

	Demographic model							
	Constant		Expansion		Bottleneck		Two Populations	
Parameter Sets[a]	SFRatios Mean	fastDFE Mean	SFRatios Mean	fastDFE Mean	SFRatios Mean	fastDFE Mean	SFRatios Mean	fastDFE Mean
mean: -1.0	-1.2	-1.8	-1.3	-1.8	-38.2	-6.8	-1.5	-2.1
shape: 1.0	1.8	6.7	4	5.8	1.9	1.8	3.5	6.4
mean: -1000.0	-379.8	-735.4	-278.4	-258.1	-721.8	-209.1	-381.3	-1788.9
shape: 0.50	0.78	0.61	0.83	0.95	0.99	4.1	0.76	0.54
	Constant		**Expansion**		**Bottleneck**		**Two Populations**	
	SFRatios RMSE	fastDFE RMSE	SFRatios RMSE	fastDFE RMSE	SFRatios RMSE	fastDFE RMSE	SFRatios RMSE	fastDFE RMSE
mean: -1.0	0.65	0.98	0.66	1	126.8	7.6	0.6	1.2
shape: 1.0	2.9	7.1	5	6.5	2.6	3.6	4.1	6.9
mean: -1000.0	634.5	312.3	724.1	746.5	415.4	793	629	899.6
shape: 0.50	0.28	0.14	0.34	0.49	0.53	5.3	0.26	0.08

[a]Mean and shape parameters for inverted gamma distribution with a maximum of zero.

https://doi.org/10.1371/journal.pgen.1011427.t001

effective population size, the difference in mean selection strength is in the expected direction. For both populations, the best fitting model included a point mass at zero, and both had the highest overall density at or near zero.

As described above, we can use $\hat{\rho}$ and the counts of selected and neutral SNPs to estimate λ, the relative probability that a selected mutation goes unsampled, relative to a neutral mutation. Because we used pairs of intron and exon SNPs, the ratio of the numbers of polymorphic sites (X/Y in expression (8)) is 1, and so $\hat{\lambda} = \hat{\rho}$ in the present case. For nonsynonymous sites in Zimbabwe and North Carolina the relative number of selected mutations that have gone unsampled, $\hat{\lambda}$, is 2.63 and 3.66, respectively. In other words, nonsynonymous mutations in this population are between 2.6 and 3.7 times more likely than neutral mutations to go unsampled because of selection removing them from the population or driving them to extreme allele frequencies.

The SFRatios fitting for synonymous sites show a consistent pattern across populations and $2Ns$densities. Both populations were fit equally well by the lognormal and gamma distributions (S3 Table), with the values for the lognormal density shown in Table 2 and Fig 7. For both populations $\hat{\rho} \approx 1$ and the mean of the estimated density is close to zero in both cases (-0.148 in the case of Zambia, and -0.311 in the case of North Carolina), and lower than previous estimates [28,29] which were near 1.

Discussion

The promise of the SFRatios method is to enable the estimation of selection intensity without having to consider either divergence between species, or the underlying mutation rates, or the other non-selective factors that will also have shaped polymorphism patterns. Given the simulation results, the method works well across a wide array of demographic scenarios, particularly when selection is not strong. For strong selection (e.g., mean $\gamma = -1000$) most simulation results revealed an estimator bias towards weaker (less negative) selection.

Although not examined here in depth, the approach of using ratios should also be effective for other non-selective factors, in addition to demography, so long as they are shared by the selected and neutral variants. For example, if the two sets of variants are sampled near each other (as in this study), then linked selection effects, including background selection and selective sweeps, will have affected both in similar ways. Similarly, the use of ratios should accommodate factors that affect mutation biases and gene conversion biases if they are shared by selected and neutral variants, as is partly the

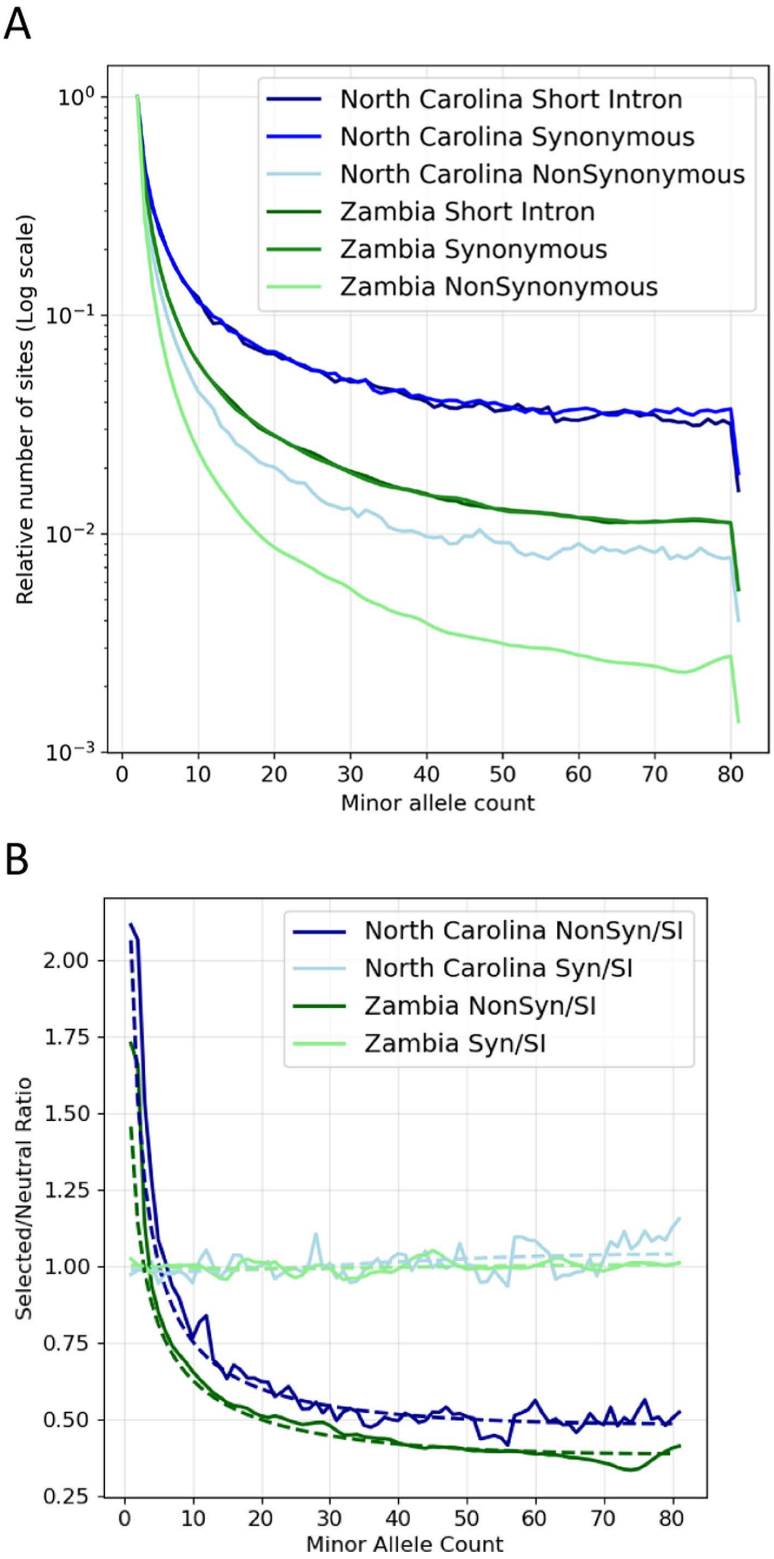

Fig 6. SFSs and ratios for two Drosophila populations. A. SFS counts normalized to that singleton bin count to enable comparisons. **B.** SFS ratios for both populations, for nonsynonymous and synonymous SFSs divided by the short intron SFS. Expected values generated under the best fit models are shown with dashed lines.

https://doi.org/10.1371/journal.pgen.1011427.g006

Table 2. Best fitting models from SFRatios Drosophila analyses.

Pop	DFE[1]	K[2]	AIC[3]	$\hat{\rho}$	λ_1[4]	λ_2[4]	p_+[5]	Mean[6]
Nonsynonymous sites								
Zambia	Lognormal p+	4	-2324.9	2.63 2.62,2.65	4.17 4.16,4.19	2.56 2.54,2.57	0.05 0.04,0.05	-1643.8
North Carolina	Lognormal p+	4	-2128.2	3.66 3.63,3.68	4.34 4.32,4.36	1.77 1.75,1.78	0.11 0.11,0.11	-323.8
Synonymous sites								
Zambia	Lognormal	3	-2328	1.01 1.00,1.01	-0.30 -0.32,-0.28	0.94 0.91,1.00		-0.148
North Carolina	Lognormal	3	-2069	0.99 0.99,1.00	-0.63 -0.67,-0.58	1.34 1.27,1.41		-0.311

[1]The best fitting distribution model for $2Ns$. If the best model included a point mass, $p_+$is the value of that mass and γ_+ is the location (i.e., the $2Ns$ value).

[2]The number of parameters.

[3]Akaike information criterion [27].

[4]Depending on which continuous distribution fit best, λ_1, $\lambda_2$are either the lognormal mean and standard deviation, or the normal mean and standard deviation, or the gamma α, β.

[5]Zero Point mass value.

[6]The expected value of the distribution model for $2Ns$.

https://doi.org/10.1371/journal.pgen.1011427.t002

case in this study. However, if the selected and neutral sets differ because of factors that are not shared, such as differing levels of biased gene conversion due to base composition differences, as occurs in mammals [30], then the ratio-based estimate may suffer.

If investigators have ready access to counts of invariant sites, that otherwise match the criteria for sampling selected and neutral polymorphic sites, they have other tools available, including fastDFE [19]. However counting invariant sites presents a different set of challenges than counting polymorphic sites. For example, it may require assuming some particular fraction of the genome is susceptible to the class of mutations being studied (as when working with synonymous and nonsynonymous variants). A larger difficulty arises when the sampling effort of invariant sites is unknown or uneven, such as when working with VCF files of polymorphic sites, or when working with pooled samples with differing or unknown sampling efforts. These issues are compounded in their difficulty if counts of invariant sites are to be divided into those that are fixed for an ancestral allele and those that are fixed for a derived allele, as required by many methods.

The SFRatios method depends strongly on the quality of the neutral control set of SNPs under three main criteria. The first is selective neutrality, which can be inferred in relative terms using site frequency analyses and divergence patterns [26], but it is difficult to know with certainty. The second criterion is that the neutral set share as many of the non-selective factors as possible with the selected set of SNPs. All parts of the genome share a demographic history, but not all share equally in background selection or mutation and recombination related processes. This second criterion can often be met, at least approximately, by having the neutral SNPs be near to, and interspersed among, the selected SNPs. The third criterion is to avoid increasing the variance of ratios by being sure to record the frequencies of both types of SNPs, selected and control, on the same set of genomes. This criterion arises from the very large effect that unforeseen population structure can have on the SFS of a sample. Consider for example an SFS for 10 genomes sample from a population that, unbeknownst to the investigator, actually includes two divergent subpopulations. In this case the partition of the sample (i.e., the numbers of individuals in each subpopulation) will have a very large effect on the SFS. If the neutral and selected SNPs were drawn from different individuals, then there will be a strong chance that the two sets will have a different partition with respect to the subpopulations, with very large and differing effects on their respective SFSs.

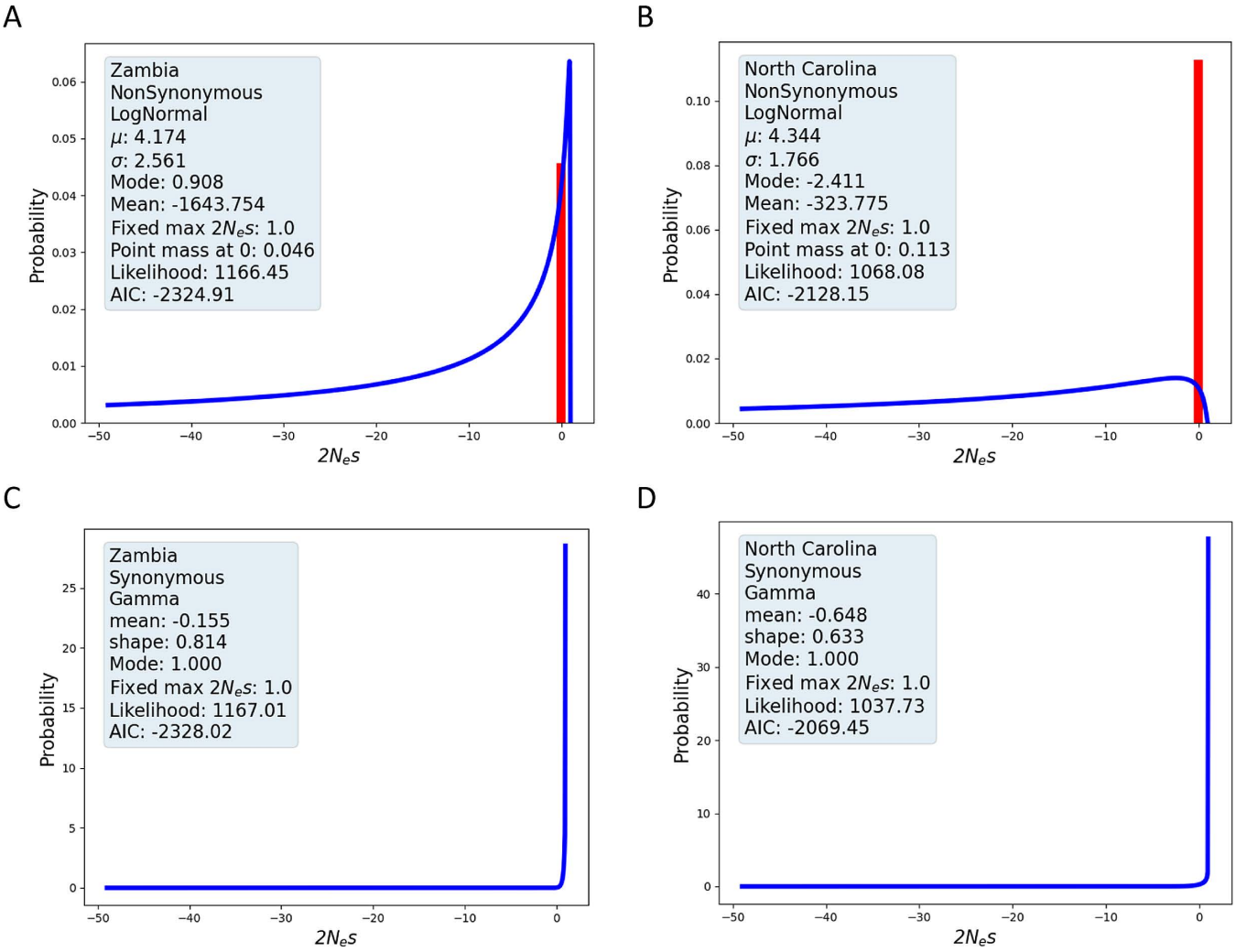

Fig 7. Best-fit estimated 2Ns densities for Drosophila populations. A. Nonsynonymous variation, Zambia. **B.** Nonsynonymous variation, North Carolina. **C.** Synonymous variation, Zambia. **D.** Synonymous variation, North Carolina.

https://doi.org/10.1371/journal.pgen.1011427.g007

In comparison with fastDFE over a wide range of demographies and gamma DFEs, we observed that both SFRatios and fastDFE performed inconsistently, with each method performing better under some of the models. These results highlight the challenge of developing a general estimator that can perform well across a very wide range of DFEs and demographies. It is possible some of the challenge may be computational, particularly if discontinuities in the likelihood surface arise for some parts of the parameter space as a byproduct of approximations used in numerical integration. We did observe that SFRatios performed better for lognormal DFEs (compare Fig 4 with S4 and S5 Figs), which is noteworthy as all of the best-fitting models for the Drosophila data sets were lognormal DFEs (Table 2).

Our study of the Zambia *Drosophila* population can be compared to estimates from other methods that use site frequency data, but that also relied upon mutation rate or divergence measures. Table 3 shows the best fitting model results of three studies, all of which compared lognormal and gamma densities for γ. These estimates bracket that obtained here that had a mean estimate of -1643.8. Huber et al., [31] estimated a mean $2Ns$ of -738.26 in the Zambia sample, while Ragsdale et al., obtained an estimated mean of -2760 on the Zambian sample [32] and another of -7414 on a Rwandan

sample [33]. All these other studies rely upon estimates of the number of invariant sites as well as fitting of a demographic model. In our method, by using the ratio of two SFSs, one being the SFS of putative neutral sites, we avoid both of these complications.

Materials and methods

Simulations

With SLiM3 [34], we simulated a series of diverse demographies with linked selection effects under both a fixed $2Ns$ and a continuous distribution of $2Ns$ values, all without dominance. For the $2Ns$ distribution, we used an inverted lognormal distribution with a maximum set to 1 and a minimum of -100000 and an inverted gamma distribution with a maximum of 0 and a minimum of -100000. For each combination of $2Ns$ distribution and demographic model we ran 20 simulations each with a total genome length of 4 megabase pairs (Mbp) for each of five lognormal distributions that ranged from very weak selection (mean $2Ns = 0.5$) to fairly strong selection (mean $2Ns < -1000$) (see Fig 4). To achieve faster running times, we split each genome into 400 fragments of 10 kilobase pairs (Kbp), allowing us to simulate these smaller fragments faster in independent sub-simulations. Each of these 10 Kbp genomic fragments was designed to model a typical Drosophila gene, including introns and flanking sequences. Each fragment included eight consecutive neutral/selected pairs of size 810 + 324 = 1134 bp, and a neutral segments of 928 bases. Selected fragments had only non-neutral mutations (i.e., $2Ns \neq 0$, while neutral fragments carried only neutral mutations (i.e., $2Ns = 0$). Base diploid population size (N) was set to 1000, with recombination and mutation rates per base pair of 2.5×10^{-7}, giving population level rates ($4Nu$) similar to that seen in human populations (i.e., 10^{-8}).

For each simulation, the site frequency spectra (SFS) for all neutral variants sampled from the sub-simulations were summed (as were those for the selected variants), then ratios calculated as the selected count for a frequency bin, divided by the neutral count for the corresponding frequency bin. All simulations began with a burn-in period of $10\ N$ generations to allow the population to reach mutation-selection-drift equilibrium [35].

For simulations with changing population size, $2Ns$ values were kept constant by rescaling selection coefficients at each generation in the simulation at which population size changed. We chose this approach, rather than the alternative of keeping s constant, to be consistent with our model in which $2Ns$ (or a distribution of $2Ns$) is fixed.

We considered the following demographic models. (1) Constant population size at $N = 1000$. (2). Population expansion, with an initial population of 1000 jumping to 10000, followed by 100 generations before sampling. Population bottleneck, with an initial population of 1000 jumping to 100, followed by 100 generations before sampling. Population structure, with

Table 3. Previous DFE estimates for nonsynonymous mutations.

Study	Population	*n*	DFE	parameterization	Parameter estimates	Mean $2Ns$ [3]
[33][1]	Rwanda	17	Lognormal	μ, σ	12.64,4.9	-7314.71
[32][2]	Zambia	197	Lognormal	μ, σ, *p*-, *p*+, γ+	5.42, 3.36, 0.708, 0.0056, 39.9	-2760.16
[31][1]	Zambia	197	Gamma	α, β	0.35, 2111.2	-738.92

[1]The scale parameter β of the gamma, and the μ of the lognormal distributions were originally reported in the *ss* scale, however here they have been converted to the 2*N*s scale.

[2]The negative part of a DFE as a lognormal or as a gamma distribution, with a point mass for positive 2*N*s values. μ, σ (or α, β) corresponds to the lognormal mean and standard deviation, or gamma shape and scale. The p- and p + are the proportion of negatively and positively selected nonsynonymous mutations and γ+ refers to the point mass value for the strength of positive selection. The expected values in this table were obtained by integrating the negative part over *N*s values giving the corresponding distribution, weighting the negative part by the proportion of negatively selected mutations, and weighting the positive part by the proportion of positively selected mutations times the point mass γ+.

[3]Expected values were obtained by numerical integration over the range −100,000 < 2*N*s < 0.0001.

https://doi.org/10.1371/journal.pgen.1011427.t003

an initial population of 1000, splitting into two populations each of 1000, followed by 1000 generations before sampling equally from both populations.

In addition we simulated the human African-Origin (AO) model as inferred by Gravel et al. [24]. This model has some of the features present in the described models, but also some more complex ones. This model includes multiple populations, a bottleneck at the founding of non-African populations, expansion of European and East Asians subpopulations, bottleneck, and population structure as well gene flow among populations. Because the OA simulations with the inferred population sizes are too computationally expensive, we ran a set of neutral simulations to identify the best scaling factor for the simulation parameters that could recover the site-frequency spectra of the original model without any scaling. We then scaled the simulation parameters: mutation and recombination rates, population sizes, migration rates, and growth rates of the exponential population growth phase with a factor of 10. We did not attempt to simulate whole genomes, but 400 sub-simulations of fragments of size 10 Kbp as was done for the other models. For all models described above, we sampled individuals at the end of the simulation. For the AO model, we sampled at the end by taking an equal proportion of the three populations: Africans, Europeans, and East Asians.

For all demographic models and lognormal $2Ns$ densities, selected and neutral SFSs were generated for samples sizes of both 200 chromosomes and 50 chromosomes to assess the effect of sample sizes. All SFSs were folded before analysis.

2Ns Densities

For distributions of the population selection coefficient we considered a normal distribution, as well as inverted lognormal and gamma distributions, i.e., extending to $-\infty$ rather than $+\infty$. Rather than have the upper limit at zero, which would not allow for the inclusion of strictly neutral mutations, we set the upper limit at 1, to include the possibility of weakly advantageous mutations as well as neutral mutations. These densities are then:

$$Prob(2Ns = \gamma) = \frac{e^{-(Log[1-\gamma]-\mu)^2/(2\sigma^2)}}{(1-\gamma)\sigma\sqrt{2\,\pi}}, \tag{10}$$

for an inverted lognormal density with maximum m, expectation μ and standard deviation σ for the natural logarithm of $m-\gamma$, and

$$Prob(2Ns = \gamma) = \frac{e^{\frac{-\alpha(m-\gamma)}{m-\mu}}(m-\mu)^{\alpha-1}(\frac{m-\mu}{\alpha})^{-\alpha}}{\Gamma(\alpha)}, \tag{11}$$

for an inverted gamma density with maximum m, mean μ and shape parameter α.

We also considered a normal (gaussian) distribution. as well as a mixed distributions, that included a continuous distribution (lognormal, gamma or normal, as described) and a point mass at zero. For any continuous density $f(\gamma)$, the corresponding mixture with a point mass p^+ at zero is

$Prob(2Ns = \gamma) == (1-p^+)\,f(\gamma) + \ p^+\delta(\gamma)$, where $\delta()$ is the Dirac delta function.

fastDFE applications

The fastDFE program [19] was used to analyze SLiM simulated data sets generated under several inverted gamma distributions with a maximum at zero. fastDFE can run on folded SFSs, but also requires the count of invariant sites, which were obtained from the SLiM runs. We used the GammaExpParametrization model in fastDFE with maximum set to zero and divided the estimated mean by 2, as fastDFE is based on a $4Ns$ parameterization of selection strength, in contrast to $2Ns$ as used by SFRatios.

Drosophila applications

We extracted synonymous, nonsynonymous, and short intron site-frequency spectrums from whole-genome sequencing data sets of two *Drosophila melanogaster* populations for the four autosomal chromosome arms. The North Carolina population has 200 inbred lines [36] while the Zambia collection is based on 197 haploid embryos [37]. Because both data sets have missing data in some lines at some positions, allele counts at all SNPs were down sampled to the expected SFS with a uniform sample size of 160. Because of uncertainty as to which allele is truly ancestral for a given SNP, we used folded SFSs [8].

All sequence data were downloaded from the Drosophila Genome Nexus (DGN, https://www.johnpool.net/genomes.html). For each population a VCF file was constructed by first running the 'masking package' (available at DGN) to mask identical-by-descent or admixture tracks when present in one or many genomes, followed by the 'snp-site' program [38] to convert the population multi-alignment FASTA to a VCF. Because snp-site assigns the common allele as the reference, a custom script was used to assign the correct reference base. This script also ensures the genotype data conform to the standard VCF format v.4. An additional filter was applied to only keep bi-allelic SNPs genotyped on 50% or more of individuals in each population. With the DGN data in VCF format, we used GATK LiftoverVcf [39] to shift the SNP coordinate positions from *D. melanogaster* reference genome Dmel 3 to Dmel 6 [40].

We annotated SNPs with SNPEff [41]. For the neutral SFS we used SNPs found in short introns (< 86 bp in length), after removing 8 bp from each side [25,26,42]. To help ensure that the selected and neutral SNPs were as closely matched as possible for local mutational context, we followed Machado et al., [18] by pairing each candidate selective SNP with the nearest short intron (SI) SNP that had the same reference allele and flanking bases. For each candidate selected SNP, a short intron partner was identified by matching a text string that included two nucleotides from the reference genome sequence (1 bp before and after the SNP) and the SNP genotype reference/alternate allele pair (e.g., C/T). As the order of the reference and alternative allele did not matter (all analyses were carried with folded site-frequency spectra) we consider both allele pairs (e.g., C/T and T/C) together when defining the SNP mutational context. For example, if one SNP was C/T, where C was the reference allele, and T was the alternative allele (as in the VCF), and a second SNP was T/C, where T was the reference and C was the alternative, and if they both had an A nucleotide one bp before and after, both SNPs share the same mutational context (e.g., AC/TA or AT/CA). These mutational contexts were defined for every combination of SNP genotypes and flanking bases, giving a list of possible 96 mutational contexts. For each candidate selected SNP, the nearest short intron SNP with matching mutational context, and not previously sampled, was added to the short intron data. Of the three classes of sites nonsynonymous, synonymous, and short intron, the latter class had the fewest SNPs and was the limiting factor for the total number of SNPs included in a data set. We obtained 44,266 and 111,161 short intron and synonymous pairs and 47,208 and 120,777 short intron and nonsynonymous pairs of SNPs with at least 160 genomes genotypes for North Carolina and Zambia, respectively.

Supporting information

S1 Fig. The probability of rejecting the null (neutral) when the alternative model (selected) is true for different probabilities of false positive (α) and varying strengths of 2Ns. Results for Wright Fisher population Poisson Random Field (PRF) likelihood-ratio tests are shown in panels A, C, and E. Results for PRF-Ratio tests are shown in panels B, D and F. Sample sizes: Few genomes (n = 20) low variation (θ=50) in panels A and B; More genomes (n = 100) and low variation (θ=50) in panels C and D; More genomes (n = 100) and high variation (θ=500) in panels E and F.
(TIF)

S2 Fig. Receiver operator characteristic (ROC) curves, with area under the curve (AUC). Results for Wright Fisher population Poisson Random Field (PRF) likelihood-ratio tests are shown in panels A, C, and E. Results for PRF-Ratio tests are shown in panels B, D and F. Sample sizes: Few genomes (n = 20) and low variation (θ=50) in panels A and B;

More genomes (n = 100) and low variation (θ=50) in panels C and D; More genomes (n = 100) and high variation (θ=500) in panels E and F.
(TIF)

S3 Fig. Cumulative observed distributions of the likelihood ratio test statistic, with χ^2,1df comparison for sets of 500 simulations. Results for Wright Fisher population Poisson Random Field (PRF) likelihood-ratio tests are shown in panels A, C, and E. Results for PRF-Ratio tests are shown in panels B, D and F. Sample sizes: Few genomes (n = 20) and low variation (θ=50) in panels A and B; More genomes (n = 100) and low variation (θ=50) in panels C and D; More genomes (n = 100) and high variation (θ=500) in panels E and F.
(TIF)

S4 Fig. SFRatios estimator performance for gamma parameters. For each simulated data set, γ values were drawn from one of 5 gamma distributions. 20 data sets were simulated for each demographic model and each γ distribution. A. Gamma distributions for random variable x,($0 < x < \infty$),$2Ns$ =-x. B. Constant population size Wright-Fisher. C. Population expansion. D. Population bottleneck. E. Two populations. F. African Origin model.
(TIF)

S5 Fig. FastDFE (Sendrowski and Bataillon 2024) estimator performance for gamma parameters. For each simulated data set, γ values were drawn from one of 5 gamma distributions. 20 data sets were simulated for each demographic model and each γ distribution. A. Gamma distributions, for random variable x,($0 < x < \infty$), $2Ns$ =-x. B. Constant population size Wright-Fisher. C. Population expansion. D. Population bottleneck. E. Two populations. F. African Origin model.
(TIF)

S6 Fig. SFRatios estimator performance for sample size of 50 simulated genomes for lognormally distributed γ. For each simulated data set, γ values were drawn from one of 5 lognormal distributions (see Materials and Methods and Fig 4). 20 data sets were simulated for each demographic model and each γ distribution. A. Constant population size Wright-Fisher. B. Population expansion. C. Population bottleneck. D. Two populations. E. African Origin model.
(TIF)

S1 Table. Comparison of SFRatios and fastDFE mean estimates and root-mean-square-error (RMSE) under diverse demographies and diverse gamma DFEs.
(XLSX)

S2 Table. Model fitting to nonsynonymous sites. Estimates and 95% confidence levels are shown for SFRatios. K is the number of model unknowns, and includes $\rho = \theta_S/\theta$ (unless ρ is fixed), as well as the parameters for the density of $2Ns$ (λ_1 and λ_2) and the estimated proportion (p^+) of a point mass at zero if included in the model. For lognormal and normal, $\lambda_1 = \mu$, $\lambda_2 = \sigma$ as described in the text, while for gamma, λ_1 =mean, λ_2 =shape as described in the text.
(XLSX)

S3 Table. Model fitting to synonymous sites. Estimates and 95% confidence levels are shown for SFRatios. K is the number of model unknowns, and includes $\rho = \theta_S/\theta$ (unless ρ is fixed), as well as the parameters for the density of $2Ns$ (λ_1 and λ_2) and the estimated proportion (p^+) of a point mass at zero if included in the model. For lognormal and normal, $\lambda_1 = \mu$, $\lambda_2 = \sigma$ as described in the text, while for gamma, λ_1 =mean, λ_2 =shape as described in the text.
(XLSX)

Acknowledgments

We are grateful to Adam Eyre-Walker for helpful comments, and to Janek Sendrowski for assistance with fastDFE.

Author contributions

Conceptualization: Jody Hey, Vitor A. C. Pavinato.

Formal analysis: Jody Hey, Vitor A. C. Pavinato.

Funding acquisition: Jody Hey.

Investigation: Jody Hey.

Methodology: Vitor A. C. Pavinato.

Software: Jody Hey, Vitor A. C. Pavinato.

Supervision: Jody Hey.

Writing – original draft: Jody Hey, Vitor A. C. Pavinato.

Writing – review & editing: Jody Hey, Vitor A. C. Pavinato.

References

1. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973;74:175–95.
2. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004;13(4):969–80. https://doi.org/10.1111/j.1365-294x.2004.02125.x PMID: 15012769
3. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116:153–9. https://doi.org/10.1093/genetics/116.1.153 PMID: 3110004
4. Miyata T, Yasunaga T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol. 1980;16(1):23–36. https://doi.org/10.1007/BF01732067 PMID: 6449605
5. Wright S. The distribution of gene frequencies in populations. Proc Natl Acad Sci U S A. 1937;23:307–20. https://doi.org/10.1073/pnas.23.6.307 PMID: 16577780
6. Kimura M. Diffusion models in population genetics. J Appl Prob 1964; 1: 177-232.
7. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132:1161–76. https://doi.org/10.1093/genetics/132.4.1161 PMID: 1459433
8. Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. Genetics. 1994;138:227–34. https://doi.org/10.1093/genetics/138.1.227 PMID: 8001789
9. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4(5):e1000083. https://doi.org/10.1371/journal.pgen.1000083 PMID: 18516229
10. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 2006;173(2):891–900. https://doi.org/10.1534/genetics.106.057570 PMID: 16547091
11. Loewe L, Charlesworth B, Bartolomé C, Nöel V. Estimating selection on nonsynonymous mutations. Genetics. 2006;172(2):1079–92. https://doi.org/10.1534/genetics.105.047217 PMID: 16299397
12. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nature Reviews Genetics. 2007;8(8):610–8.
13. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 2005;102(22):7882–7. https://doi.org/10.1073/pnas.0502300102 PMID: 15905331
14. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 2007;3(9):e163. https://doi.org/10.1371/journal.pgen.0030163 PMID: 17907810
15. Bhaskar A, Wang YR, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research. 2015;25(2):268–79.
16. Wakeley J. Polymorphism and divergence for island-model species. Genetics. 2003;163(1):411–20. https://doi.org/10.1093/genetics/163.1.411 PMID: 12586726
17. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 2009;26(9):2097–108. https://doi.org/10.1093/molbev/msp119 PMID: 19535738
18. Machado HE, Lawrie DS, Petrov DA. Pervasive Strong Selection at the Level of Codon Usage Bias in Drosophila melanogaster. Genetics. 2020;214(2):511–28. https://doi.org/10.1534/genetics.119.302542 PMID: 31871131

19. Sendrowski J, Bataillon T. fastDFE: Fast and Flexible Inference of the Distribution of Fitness Effects. Mol Biol Evol. 2024;41(5):msae070. https://doi.org/10.1093/molbev/msae070 PMID: 38577958
20. Tataru P, Mollion M, Glémin S, Bataillon T. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics. 2017;207(3):1103–19. https://doi.org/10.1534/genetics.117.300323 PMID: 28951530
21. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet. 2016;12(1):e1005774. https://doi.org/10.1371/journal.pgen.1005774 PMID: 26752180
22. Griffin TF. Distribution of the ratio of two poisson random variables. Texas: Tech University; 1992.
23. Díaz-Francés E, Rubio FJ. On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. Statistical Papers. 2013;54:309–23.
24. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences. 2011;108(29):11983–8.
25. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol. 2010;27(6):1226–34. https://doi.org/10.1093/molbev/msq046 PMID: 20150340
26. Clemente F, Vogl C. Unconstrained evolution in short introns?–An analysis of genome-wide polymorphism and divergence data from Drosophila. J Evol Biol. 25(10):1975-90. 2012.
27. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19(6):716–23. https://doi.org/10.1109/tac.1974.1100705
28. Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 2007;177(4):2251–61. https://doi.org/10.1534/genetics.107.080663 PMID: 18073430
29. Zeng K, Charlesworth B. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics. 2009;183(2):651–62. https://doi.org/10.1534/genetics.109.101782 PMID: 19620398
30. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311. https://doi.org/10.1146/annurev-genom-082908-150001 PMID: 19630562
31. Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. Proc Natl Acad Sci U S A. 2017;114(17):4465–70. https://doi.org/10.1073/pnas.1619508114 PMID: 28400513
32. Ragsdale AP, Coffman AJ, Hsieh P, Struck TJ, Gutenkunst RN. Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. Genetics. 2016;203(1):513–23. https://doi.org/10.1534/genetics.115.184812 PMID: 27029732
33. Kousathanas A, Keightley PD. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics. 2013;193(4):1197–208.
34. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol Biol Evol. 2019;36(3):632–7. https://doi.org/10.1093/molbev/msy228 PMID: 30517680
35. Wright S. Evolution in Mendelian populations. Genetics. 1931;16:97–159. https://doi.org/10.1093/genetics/16.2.97 PMID: 17246615
36. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482(7384):173-8.
37. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics. 2015;199(4):1229–41. https://doi.org/10.1534/genetics.115.174664 PMID: 25631317
38. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial genomics. 2016;2(4):e000056. https://doi.org/10.1099/mgen.0.000056 PMID: 28348851
39. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020.
40. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, et al. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015;43(D1):D690-D7. https://doi.org/10.1093/nar/gku1099 PMID: 25398896
41. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672
42. Halligan DL, Keightley PD. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Research. 2006;16(7):875–84. https://doi.org/10.1101/gr.5022906 PMID: 16751341