G3, 2024, 14(4), jkae031
https://doi.org/10.1093/g3journal/jkae031
Advance Access Publication Date: 14 February 2024
Investigation

An efficient and robust ABC approach to infer the rate and strength of adaptation

Jesús Murga-Moreno [1], Sònia Casillas [2,3], Antonio Barbadilla [2,3], Lawrence Uricchio [4,‡*], David Enard [1,‡*]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85719, USA
[2]Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain
[3]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain
[4]Department of Biology, Tufts University, Medford, MA 02155, USA

*Corresponding author: Department of Biology, Tufts University, Medford, MA02155, USA. Email: Lawrence.Uricchio@tufts.edu; *Corresponding author: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ85719, USA. Email:denard@arizona.edu
‡These authors contributed equally to this work.

Inferring the effects of positive selection on genomes remains a critical step in characterizing the ultimate and proximate causes of adaptation across species, and quantifying positive selection remains a challenge due to the confounding effects of many other evolutionary processes. Robust and efficient approaches for adaptation inference could help characterize the rate and strength of adaptation in nonmodel species for which demographic history, mutational processes, and recombination patterns are not currently well-described. Here, we introduce an efficient and user-friendly extension of the McDonald–Kreitman test (ABC-MK) for quantifying long-term protein adaptation in specific lineages of interest. We characterize the performance of our approach with forward simulations and find that it is robust to many demographic perturbations and positive selection configurations, demonstrating its suitability for applications to nonmodel genomes. We apply ABC-MK to the human proteome and a set of known virus interacting proteins (VIPs) to test the long-term adaptation in genes interacting with viruses. We find substantially stronger signatures of positive selection on RNA-VIPs than DNA-VIPs, suggesting that RNA viruses may be an important driver of human adaptation over deep evolutionary time scales.

Keywords: Mcdonald and Kreitman test; natural selection; viral interacting proteins

Introduction

Genomes contain a record of the evolutionary processes that shape diversity within and across species, and software tools that use genomic sequences to infer aspects of the evolutionary past are now an integral part of population genetics research. Of particular interest to evolutionary biologists are methods that can disentangle various processes that may contribute to diversification between species, such as adaptation and genetic drift. Such methods have the potential to resolve fundamental questions about the evolutionary (e.g. Corbett-Detig *et al.* 2015; Galtier 2016; Galtier and Rousselle 2020) and biological (e.g. Enard *et al.* 2016; James *et al.* 2016) drivers of diversification at the genomic level. Though numerous methods have been proposed to this end, it remains challenging to generate accurate and unbiased estimates. Studies addressing the potential biases of the available approaches unaccounted for evolutionary processes and assessing evidence for genome adaptation is still a highly active area of research in molecular population genetics (McDonald and Kreitman 1991; Gillespie 1994; Smith and Eyre-Walker 2002; Hahn 2008; Fay 2011; Tataru *et al.* 2017; Kern and Hahn 2018; Jensen *et al.* 2019; Johri *et al.* 2020; Johri, Aquadro, *et al.* 2022; Johri, Eyre-Walker, *et al.* 2022).

The development of methods that are both computationally efficient and reasonably robust to model misspecification remains a major challenge. Most computational approaches that infer the rate of long-term adaptation at the DNA level derive from the McDonald and Kreitman (MK-test) framework (McDonald and Kreitman 1991) or the related Poisson random field (PRF) framework (Sawyer and Hartl 1992). Both methods use red observed diverged sites (D) and polymorphic sites (P) to estimate the proportion of nonsynonymous substitutions fixed by positive selection in coding sequences, comparing alleles that are likely to have fitness effects (putatively selected) to those less likely to be under selection (putatively neutral). In the MK framework specifically, the ratio of nonsynonymous (D_N) and synonymous (D_S) fixed differences ($\frac{D_N}{D_S}$) is compared to levels of polymorphism in the same classes ($\frac{P_N}{P_S}$). Nonsynonymous sites are assumed to be putatively functional and possibly subject to selection because of their effect on the primary protein sequence, while synonymous sites are assumed to be selectively neutral. Additionally, all observed polymorphic sites are assumed to evolve neutrally—in other words, selection is strong enough on the subset of alleles that are under selection that fixation or loss occur very rapidly and selected polymorphisms are rarely observed in sequence data. Under these assumptions, $\frac{P_N}{P_S}$ is taken as a null expectation for the relative proportion of fixed differences in the functional class under a neutral evolutionary model. When $\frac{D_N}{D_S}/\frac{P_N}{P_S}$ significantly exceeds one, this is taken as a signal of positive selection on the putatively functional alleles. The rate of adaptation is often summarized by the quantity α, which is defined as the proportion of nonsynonymous (or putatively functional) fixed differences that

Received on 10 October 2023; accepted on 29 January 2024

were under positive selection along a particular evolutionary branch. Smith and Eyre-Walker (2002) applied a simple theoretical model of directional selection relating polymorphism and divergence with adaptation rate and showed that the rate of adaptation α could be estimated with the quantity

$$\alpha = 1 - \left(\frac{D_S}{D_N}\frac{P_N}{P_S}\right). \quad (1)$$

When α is close to 1, then positive selection is the predominant determinant of molecular divergence. If α is close to 0, then drift dominates sequence divergence. This convenient formula has been widely applied to estimate molecular adaptation, in part because of its simplicity. Indeed, the quantities on the right-hand side of equation (1) are commonly inferred by comparing a population sample of sequenced individuals (sequencing sample) to a closely related outgroup species. Although widely used, it should be noted that MK-test and PRF-based approaches have multiple drawbacks that could bias the estimation of α. For instance, several modeling and empirical studies have argued that weakly selected alleles can attain both low and high frequencies and may cause substantial biases in inferences that use equation (1) (Balloux and Lehmann 2012; Lanfear *et al.* 2014; Booker and Keightley 2018; Galtier and Rousselle 2020; Rousselle *et al.* 2020).

Though weakly deleterious alleles are less likely to reach fixation, they can reach relatively high frequencies and contribute to the class P_N. This causes overestimation of the neutral mutation rate in the putatively functional category, which makes the estimation of α downwardly biased (Charlesworth and Eyre-Walker 2008). Fixation of weakly deleterious alleles could also cause overestimation of α after a long population bottleneck since the reduced population size would make the selection less efficient to purge deleterious alleles (Eyre-Walker 2002; Eyre-Walker and Keightley 2009). Conversely, weakly beneficial alleles can be found at intermediate or high frequencies (Tataru *et al.* 2017; Uricchio *et al.* 2019), especially when the rate of strongly beneficial mutations is high (as might be expected in a large population) or weakly beneficial alleles contribute substantially to polymorphism (as might be expected under some polygenic selection models). The presence of weakly selected alleles has been addressed over the last decade by MK-test and PRF-based methods mainly by explicitly modeling the distribution of fitness effects (DFE) for negatively selected variants (Boyko *et al.* 2008; Eyre-Walker and Keightley 2009; Messer and Petrov 2013; Racimo and Schraiber 2014; Galtier 2016), along with the recent incorporation of positively selected variants in such DFE models (Galtier 2016; Tataru *et al.* 2017; Sendrowski and Bataillon 2024). Despite the development of several methods that account for weakly selected polymorphism, some empirical observations remain challenging to explain under existing models, such as the apparent low rate of adaptation in primates, the constrained range of genetic diversity across species, and differences in the rate of adaptation among taxa (Galtier 2016; Castellano *et al.* 2018, 2019a). Generating a deeper biological and evolutionary understanding of the drivers of differentiation across species may require new methods and models that can efficiently estimate the DFE while simultaneously accounting for many (potentially confounding) evolutionary processes.

Demographic processes (such as population contractions, expansions, and migrations) are another primary potential source of bias in the inference of selection (Jensen *et al.* 2019; Johri *et al.* 2020; Johri, Aquadro, *et al.* 2022; Johri, Eyre-Walker, *et al.* 2022), just as the selection is an essential potential confounder in the inference of demography (Schrider *et al.* 2016; Torres *et al.* 2018). The developers of robust inference methods have typically sought to account for both selection and demographic processes simultaneously. The cost of incorporating both demography and selection is accrued in terms of model complexity and loss of efficiency, as it is much more challenging to compute likelihoods or summary statistics under joint demography/selection models. There is some hope, however, that methods based on the asymptotic MK-test (aMK-test) (Messer and Petrov 2013) may have some inherent robustness since these approaches rely on summary statistics that involve ratios of functional and (putatively) nonfunctional alleles. Hence, some of the effects of demography should be absorbed into the ratio, as both categories of alleles will be affected.

Despite the intrinsic robustness, MK-based methods such as aMK-test have some potential limitations, some of which were explored by Uricchio *et al.* (2019). First, aMK-test and most other existing methods do not account for the possibility of weakly beneficial polymorphism. Weakly beneficial polymorphisms can result in downward biases in adaptation rates, and unlike strongly beneficial mutations, may be susceptible to being lost due to background selection (BGS) across the genome. Second, MK-based approaches are not able to estimate adaptation strength. Based on these considerations, Uricchio *et al.* (2019) suggested that a method interrogating the aMK-test summary statistic $\alpha_{(x)}$ patterns as a function of BGS might be able jointly to infer the rate of adaptation and strength of adaptation in the presence of weakly and strong beneficial alleles—when strongly beneficial mutations make up most of the adaptation signal, α should not decrease with BGS strength, whereas α should decrease with BGS if weakly beneficial mutations drive adaptation. Because it is not straightforward to calculate the site frequency spectrum (SFS) through analytical theory when both selection and demographic processes affect genomic variation, the authors leveraged this potential covariation between α and BGS following a generic Approximate Bayesian Computation (ABC) algorithm in which (1) they ran forward simulations with a fixed Distribution of Fitness Effects (DFE) over nonsynonymous mutations accounting for BGS and known demography to simulate the model; (2) they estimated aMK-test patterns from forward simulations following a resampling strategy that avoids simulating the entire model for different parameter combinations; (3) they supplied a set of summary statistics into a ABC framework to infer parameters (Uricchio *et al.* 2019).

Here, we develop an extension of the Approximate Bayesian Computation MK-test (ABC-MK) presented in Uricchio *et al.* (2019) that significantly improves the efficiency of inferring the strength and rate of selection from genomic data. The previously published approach relied on resampling of genetic variants from the output of computationally expensive forward simulations. Any change to the demographic model would require rerunning these simulations, reducing the practical utility of the approach. We have now developed a simpler and much more computationally efficient ABC-based inference procedure, based on a simple random sampling scheme which accounts for sampling and process variance. Therefore the method links the analytical approximation of α with the empirical estimation of adaptation, accounting for the DFE of deleterious and beneficial alleles, BGS, and partial recombination between selected genomic elements while avoiding forward-in-time simulations. We describe the inference procedure, assess its performance and robustness to nonequilibrium demographic scenarios and different intensities of adaptation, and apply it to human genomic data. We show that the new ABC-MK extension is reasonably robust to

nonequilibrium events or different fitness values of adaptation, use it to provide additional evidence for a substantial effect of different RNA-viruses on human adaptation rates, and discuss caveats and potential extensions of our work.

Methods

Preliminaries

Our first goal is to calculate the expected rate of fixation and the expected site frequency spectrum of neutral and selected polymorphism under a model of directional selection with incomplete recombination. To do so, we follow the results of Uricchio *et al.* (2019), which in turn extended the results of several earlier studies (e.g. Eyre-Walker and Keightley 2009; Messer and Petrov 2013). In subsequent sections, we extend these calculations by developing a random sampling scheme that accounts for the Poisson variance in mutation and fixation rates, and allows us to develop a simple inference pipeline. We briefly review the core aspects of the theoretical framework.

Our ultimate goal is to estimate α, the proportion of nonsynonymous substitutions fixed by positive selection, as well as the DFE over de novo nonsynonymous mutations and substitutions. We suppose that selection is directional, with both positively selected and negatively selected mutations. We first consider the case where each selected locus evolves independently, and in subsequent sections we consider cases with BGS and selective interference.

The DFE over beneficial alleles consists of two point masses, one representing strongly beneficial alleles and the other representing weakly beneficial alleles. The rate of adaptation can therefore be decomposed into weakly and strongly beneficial components, $\alpha = \alpha_W + \alpha_S$. The substitution rate for nonsynonymous alleles is denoted as d_N, with d_{N_+}, d_{N_-}, and d_{N_0} representing the rates for positively selected, negatively selected, and neutral alleles, respectively (note that $d_N = d_{N_+} + d_{N_-} + d_{N_0}$ by definition—since there is a DFE for beneficial and deleterious alleles we could further decompose d_{N_-} and d_{N_+} into more categories, but we collapse the DFE into two compartments here and consider the full DFE in later sections). In the same way, we denote as d_S the substitution rate of synonymous mutations, which are assumed to be neutral. We can write α as

$$\mathbf{E}[\alpha] = \frac{d_{N_+}}{d_N} = \frac{d_N - (d_{N_-} + d_{N_0})}{d_N} = 1 - \frac{(d_{N_-} + d_{N_0})d_S}{d_S d_N}. \quad (2)$$

Note that we define α as the realized proportion of positively selected substitutions along the branch, and hence equation (2) is an expression for the expectation of α. As noted by Messer and Petrov (2013), $\frac{d_S}{d_N}$ can be estimated from sequence alignments with the ratio $\frac{D_S}{D_N}$ under the assumption that the observed number of substitutions along a branch should be proportional to the rate. The ratio $(d_{N_-} + d_{N_0})/d_S$ is more complex to estimate, because it relies on partitioning substitutions by their fitness effects. Under the assumption that polymorphic alleles are rarely selected (because deleterious sites are removed from the population quickly and beneficial sites go to fixation rapidly), previous work (Smith and Eyre-Walker 2002) showed that this ratio can be approximated by substituting $\frac{P_N}{P_S}$ into equation (2), providing a point estimate of α,

$$\alpha \approx 1 - \left(\frac{P_N}{P_S}\frac{D_S}{D_N}\right). \quad (3)$$

However, if selected polymorphisms segregate in the sample, then P_N in equation (3) will be inflated relative to the true rate of mutation for neutral nonsynonymous alleles, which results in underestimation of α. A potential solution is to exclude alleles with derived allele frequencies lower than some threshold from the quantities P_N and P_S, since most (negatively) selected alleles should be constrained to lower frequency (Fay *et al.* 2001). While this solution works well for some DFEs (for example, when all deleterious alleles are strongly selected), weakly deleterious alleles can reach appreciable frequencies and bias inference regardless of the selected frequency threshold. Messer and Petrov (2013) extended this idea by developing a very simple estimator of α (aMK-test) that uses all frequencies simultaneously by rewriting the estimator of Smith and Eyre-Walker (2002) as

$$\alpha_{(x)} \approx 1 - \left(\frac{P_{N(x)}}{P_{S(x)}}\frac{D_S}{D_N}\right), \quad (4)$$

where $P_{N(x)}$ and $P_{S(x)}$ are the number of non/synonymous alleles at frequency x in a sequencing sample. This method improves the quality of α estimates by using all of the frequency data simultaneously and providing confidence intervals for α. The estimate of α results in an exponential function of the form: $\alpha = a + b \cdot e^{-cx}$, where the best fit of the exponential at $x = 1$, eliminates the effect of a weakly deleterious allele. In the following section, including empirical and analytical results, we defined $P_{N(x)}$ and $P_{S(x)}$ as all nonsynonymous and synonymous polymorphisms above frequency x which we called the cumulative SFS. As noted in Uricchio *et al.* (2019) this approach scales better as sample size increases since most common allele frequencies x have very few polymorphic sites in large samples and both $P_{N(x)}$ and $P_{S(x)}$ quantities trivially have the same asymptote. Despite the improvement, the method does not estimate the DFE and assumes that beneficial alleles do not contribute to P_N, which turns into α underestimation in cases where selection is predominantly weak.

Expected fixation rates and frequency spectra

A complementary approach to that of Messer and Petrov (2013) is to directly model the effects of the DFE for beneficial and deleterious alleles on the shape of the $\alpha_{(x)}$ curve, and to infer the best fitting model parameters. Nonetheless, note that while $\mathbf{E}[\alpha_{(x)}] = 1 - \mathbf{E}[\frac{D_S P_{N(x)}}{D_N P_{S(x)}}]$ is not straightforward to calculate because it depends on the ratio of several random variables, the expectation of each component in equation 4 ($P_{S(x)}$, $P_{N(x)}$, D_S, D_N) is easily calculated in a directional selection model from first principles using diffusion theory (Evans *et al.* 2007). Therefore, we make a first-order approximation

$$\mathbf{E}[\alpha_{(x)}] \approx 1 - \frac{\mathbf{E}[D_S]\mathbf{E}[P_{N(x)}]}{\mathbf{E}[D_N]\mathbf{E}[P_{S(x)}]}. \quad (5)$$

Our model will assume that positively selected mutations have fitness effects drawn from a point mass distribution with two values defining either strong or weak adaptation (although such an assumption is relaxed in the ABC), and negatively selected mutations were Gamma-distributed following previous studies (Eyre-Walker *et al.* 2006; Boyko *et al.* 2008; Eyre-Walker 2010), replacing $\theta_s = 4N\mu s$ with Gamma distribution $\Gamma[\alpha, \beta]$ over selection coefficients in equations (6) and (7).

In general, our approach can be applied to any distribution for which we can analytically solve the fixation rates and expected frequency spectra. Descriptions of these calculations follow this section and can be found in the online documentation for our software at web address jmurga.github.io/MKtest.jl.

Expected frequency spectrum

The expected number of alleles at frequency x is estimated from the standard diffusion theory for the site frequency spectrum in an equilibrium population (e.g. see equation 31 of Evans et al. 2007).

$$\Psi(x) = \int_s \theta_s \frac{1}{x(1-x)} \frac{e^{4Ns}(1 - e^{-4Ns(1-x)})}{e^{4Ns} - 1} \, ds, \quad (6)$$

where $\Psi(x)$ is the expected alleles at frequency x in a population of size N and $\theta_s = 4N\mu s$ is the population-scaled mutation rate for mutations with selection coefficient s. To obtain the down-sampled frequency spectrum in a finite sample of $2n$ chromosomes, we convoluted equation (6) with the binomial distribution.

Expected number of fixations

Considering the distribution of selection coefficients over new mutations μ_s (selection coefficient underlying the mutation rate) and the fixation probability π_s, we calculate the expected fixation rate as

$$\mathbf{E}[d] = \int_s 2N\mu_s\pi_s \, ds. \quad (7)$$

Following Uricchio et al. (2019), the expected fixation rate can be decomposed by its fitness effect replacing $\theta_s = 4N\mu s$ in equations (6) and (7) with a Gamma distribution $\Gamma[\alpha, \beta]$ over negative selection coefficients and point mass distribution over positive selection coefficients,

$$\begin{aligned}\mathbf{E}[d] = \mathbf{E}[d_+] + \mathbf{E}[d_-] + \mathbf{E}[d_0] = p_+(1 - e^{-2s}) \\ + p_-\left(2^{-\alpha}\beta^{\alpha}\left(-\zeta\left[\alpha, \frac{2+\beta}{2}\right] + \zeta\left[\alpha, 1/2(2 - \frac{1}{N} + \beta)\right]\right)\right) \\ + (1 - p_- - p_+)\frac{1}{2N},\end{aligned}$$

where p_+ and p_- are the probability that an allele is beneficial or deleterious, respectively, and ζ is the Riemann Zeta function (Eyre-Walker 2010).

Background selection and adaptive divergence

BGS (Charlesworth et al. 1993; Hudson and Kaplan 1995; Nordborg et al. 1996) and selective interference (e.g. Hill–Robertson interference, Hill and Robertson 1966) could affect the rate of fixation of weakly deleterious or beneficial alleles. Up to this point, we have considered only selected loci that evolve independently of all other selected loci. In this section, we will relax this assumption by exploring the effects of selective interference on fixation rates and the frequency spectrum. We will follow approximations that will apply in some circumstances (in particular, when BGS is the predominant driver of selective interference) but may fail when strongly beneficial alleles interfere (Good et al. 2014; Cvijović et al. 2018; Garcia and Lohmueller 2021; Ragsdale 2022).

To explore $\alpha_{(x)}$ accounting for recombination and BGS impact, we focused on a model in which the coding locus is flanked on each side by loci of length L, which contain deleterious alleles. We modeled deleterious alleles with a population-scaled selection coefficient $-2Nt$ undergoing persistent deleterious mutation at rate $4N\mu_-$ and the whole flanking loci recombined at a rate r per-base, per-generation. The effects of BGS on fixations and frequency spectra have been subject of much theoretical work (Charlesworth et al. 1993; Charlesworth 1994; Barton 1995; Hudson and Kaplan 1995; Nordborg et al. 1996). Previous work has shown that diversity at the coding locus (π) is decreased relative to its neutral expectation (π_0), and closed form expressions for the expected reduction in diversity are available (Hudson and Kaplan 1995; Nordborg et al. 1996). While patterns of sequence variation induced by BGS can be quite complex (Nicolaisen and Desai 2013; Good et al. 2014; Torres et al. 2018, 2020), to a first approximation the effect of BGS can be thought of as a reduction in the effective population size N_e, with $N_e = N\frac{\pi}{\pi_0}$ (McVicker et al. 2009). To account for the role of BGS on the fixation rate of deleterious alleles, we replace N in the prior equations with N_e after accounting for BGS. We also replace N with N_e in formulae for the frequency spectra.

For beneficial alleles, the effects of selective interference are slightly more complex. Strongly beneficial alleles are essentially unaffected by BGS, in that their fixation probabilities almost do not depend on the reduction in neutral diversity (Uricchio et al. 2019). Weakly beneficial alleles can have their fixation probabilities substantially reduced by BGS. We followed Barton (1995) to derive formulae for the reduction in fixation rate of weakly and strongly beneficial alleles after accounting for BGS, as described in the Supplementary material of Uricchio et al. (2019) (see *Background selection and adaptive divergence* section). The reduction in fixation probability for a weakly beneficial allele with selection coefficient s under interference with deleterious alleles with selection coefficient t is given by

$$\phi(t, s) = e^{\left[\frac{-2\mu}{t\left(1 + \frac{rL}{t} + \frac{2s}{t}\right)}\right]}, \quad (8)$$

where l is the distance in base pairs from the region of interest, $1 \le l \le L$ (see equation 17d of Barton 1995). Multiplying across all deleterious linked sites and factoring in flanking sequences to both the left and right of the focal site (which requires us to square the product below), we find that the reduction in the probability of fixation relative to the case with no linkage (Φ) is

$$\Phi = \prod_1^L \phi(t, s) = e^{\frac{-2\mu\left(\psi\left[1, \frac{r+2s+L}{r}\right] - \psi\left[1, \frac{r(L+1)+2s+t}{r}\right]\right)}{r^2}} \quad (9)$$

where ψ is the polygamma function. We note that these expressions are not expected to hold under very high rates of mutation for beneficial alleles and will be less accurate for strongly beneficial alleles than weakly beneficial alleles. Given these expressions, we can replace the fixation rates for beneficial alleles in our prior formulae with the fixation rates after accounting for selective interference.

Poisson-sampling process

The previous sections described the expectation of fixation rates and frequency spectra under a model of directional selection and selective interference. We now develop a simple random sampling scheme around these expectations that accounts for sampling and process variance, linking analytical calculations to an ABC procedure which avoids computationally expensive forward simulations. We note that the model we explore is quite similar to the BGS model in DeGiorgio et al. (2016), though while we are interested in the long-term accumulation of fixations, DeGiorgio et al. (2016) is primarily interested in the nonequilibrium signature of a recent or ongoing selective sweep.

Following the PRF model (Sawyer and Hartl 1992), we supposed that the number of fixed differences and polymorphic sites were Poisson random distributed variables with mean values given by the expectations in the previous sections.

To avoid performing branch length estimations in our computation, we assumed that the empirically observed number of fixations should be proportional to the length of the evolutionary branch of interest, T, the locus length L and mutation rate μ. We take the observed total number of fixations (including both nonsynonymous and synonymous sites) as a proxy for the expected number, and then sample weakly deleterious, neutral, and beneficial substitutions proportional to their relative expected rates for a fixed set of model parameters. The expected number of substitutions for positively selected substitutions is then

$$\lambda_{D_{N_+}} = D \cdot \left(\frac{\mathbf{E}[d_+]}{\mathbf{E}[d_+] + \mathbf{E}[d_-] + \mathbf{E}[d_0]} \right), \quad (10)$$

where D is the observed number of substitutions in a dataset of interest.

It should be noted that both sampling variance and process variance affect the number of variable alleles at any particular allele frequency in a sequencing sample. The process variance arises from the random mutation-fixation process along the branch, while the sampling variance arises from the random subset of chromosomes that are included in the sequencing data. We sampled a Poisson distributed number of polymorphic alleles at frequency x relative to their rate given the expected frequency spectra. The expected frequency spectra were downsampled using a binomial (with probability of success given by the frequency $\binom{x}{2n}$ in a sample of $2n$ chromosomes) to account for the sampling variance. In a manner exactly analogous to fixed variants as described above, to account for the process variance

$$\lambda[P_N] = \sum_{x=0}^{1} P_{(x)} \cdot \left(\frac{\mathbf{E}[\Psi_{+(x)}] + \mathbf{E}[\Psi_{-(x)}]}{\mathbf{E}[\Psi_{+(x)}] + \mathbf{E}[\Psi_{-(x)}] + \mathbf{E}[\Psi_{0(x)}]} \right). \quad (11)$$

To account for background selection, we solved equation (9) using any expected B values from McVicker *et al.* (2009) at each polymorphic or fixed site. Then, we adjust the fixation probability of deleterious alleles by the expected value of B and the fixation probability for weakly beneficial alleles given the predicted reduction by equation (9). In practice, we bin sites into B value ranges, such that all sites within (for example) a 2.5% B value range of 0.675 to 0.7 experience the same N_e and the same strength of selective interference (for example, $N_e = 0.675N$, which is the midpoint of this B value window). Because we note that inferred background selection strength is unavailable for most species, our software can run without this information by inputting any expected B value range into the software. Nonetheless, the analysis can be limited to an empirical B value or B value range as an approximation of the expected reduction in diversity if downstream empirical datasets account for the inferred background selection strength estimated from McVicker *et al.* (2009) or Murphy *et al.* (2022).

Computational workflow

Our goal is to infer α, α_W, and α_S given a set of observed $\alpha_{(x)}$ values from a sequencing dataset. Since α in our framework is a model output and not a parameter per se (i.e. α depends on the random number of fixations along an evolutionary branch, which in turn depend on the parameters of the evolutionary model), we cannot immediately obtain the corresponding fixation rates and frequency spectra values for a given set of expected α values without first solving for the mutation rates and fixation probabilities considering the input model (see Table 1). Given the α and α_W (which together uniquely determine α_S), a DFE over negatively selected alleles, a known B value, a selection coefficient for beneficial alleles, a selection coefficient for flanking deleterious alleles, a recombination (ρ) and mutation rate on the coding locus (θ), we numerically solve for the probability of fixation of beneficial alleles and the mutation rate on the flaking locus that correspond to the desired α values and BGS strength. This allows us to rapidly calculate the expected frequency spectra and fixation rates that will correspond to the desired α values and generate a sample of $\alpha_{(x)}$ values following the Poisson-sampling process under the corresponding evolutionary model.

Table 1. Model parameter definitions.

γ	Population-scaled selected coefficient of deleterious alleles
s_w	Population-scaled selected coefficient of weakly beneficial alleles
s_S	Population-scaled selection coefficient of strong beneficial alleles
α_W	Proportion of weakly adaptive substitutions
α	Proportion of adaptive substitutions
$\theta_{flaking}$	Population-scaled mutation rate at flanking noncoding locus
θ_{coding}	Population-scaled mutation rate at coding locus
θ	Scale parameter
β	Shape parameter
B	B value from McVicker *et al.* (2009)
l_w	Fixation probability of weakly beneficial alleles
l_S	Fixation probability of strong beneficial alleles
N	Effective population size
n	Sample size
L	Flanking locus length
t	Population-scaled selected coefficient of flanking deleterious alleles

Inferring parameters with ABC

We used a generic ABC algorithm to infer the rate and strength of adaptation. ABC first samples the parameter values from prior distributions; second, simulate the model using these random parameter values and calculates informative summary statistics from the simulations; and third, compares the simulated summary statistics to observed empirical data. The summary statistics that best match the observed empirical data form an approximate parameter posterior distribution. Our approach used empirical data to both perform computational workflow and the Poisson-sampling scheme described above to sample $\alpha_{(x)}$ generating summary statistics corresponding to different evolutionary scenarios and to finally compare such summary statistics to empirical $\alpha_{(x)}$ estimations. Since we do not know a priori the values of any model parameter, to estimate summary statistics, we sample 10^5 sets of parameters randomly from a prior uniform distribution (see Tables 1 and 2), which allows for flexibility in the DFE of deleterious and beneficial alleles. We supplied the summary statistics and empirical $\alpha_{(x)}$ into ABCreg (Thornton 2009) to estimate the empirical values of α_W, α_S, and α while accounting for BGS in bins of 2.5% from $\frac{\pi}{\pi_0} = 0.1$ to $\frac{\pi}{\pi_0} = 1$.

We conducted ABC inferences using two different kinds of data, as explained in the following sections. First, for simulated data, we bootstrapped the simulated replicas to generate 100 datasets to test the method. For empirical data, we followed the bootstrap procedure extensively described at Enard and Petrov (2020) and

Table 2. ABC-MK prior values.

Scenarios	N	n	α	s_W	s_S	γ	β	t	B
Demographic equilibrium	500	500	[0.0,0.9]	[1,10]	[200,2,000]	[−1,000,−200]	$0.184 \cdot 2^{[-2,2]}$	[−1,000,−500]	[0.1,0.999]
Tennensen	1,000	661	[0.0,0.9]	[1,10]	[200,2,000]	[−1,000,−200]	$0.184 \cdot 2^{[-2,2]}$	[−1,000,−500]	[0.1,0.999]
Twoepoch B = 0.4	1,000	50	[0.0,0.9]	[1,10]	[200,2,000]	[−1,000,−200]	$0.184 \cdot 2^{[-2,2]}$	[−1,000,−500]	[0.1,0.999]
Twoepoch B = 0.999	1,000	50	[0.0,0.9]	[1,10]	[200,2,000]	[−1,000,−200]	$0.184 \cdot 2^{[-2,2]}$	[−1,000,−500]	[0.0,0.999]
Thousand Genomes Project data	1,000	661	[0.0,0.9]	[1,10]	[200,2,000]	[−1,000,−200]	$0.184 \cdot 2^{[-2,2]}$	[−1,000,−500]	[0.1,0.999]

Di *et al.* (2021) to generate the control sets required to create null distributions and compare adaptation levels between non-VIPs and VIPs genes.

As summary statistics, we used the value of $\alpha_{(x)}$ for $x \in (2, 5, 20, 50, 200, 661, 925)$, $x \in (2, 5, 20, 50, 200, 500, 700)$, $x \in (2, 4, 5, 10, 20, 30, 50, 70)$ regarding the input data and sample size. These values are proportionally similar to the ones used in Uricchio *et al.* (2019). However, we excluded singletons because very low-frequency alleles are particularly sensitive to sequencing errors and distortions due to demographic processes or other model misspecifications. We inferred posterior distributions for each bootstrapped dataset, each one using the same 10^5 sets of parameters randomly as prior to estimating summary statistics. We set the tolerance threshold in ABCreg to 0.025 such that $2.5 \cdot 10^3$ values were accepted from posterior distributions.

Forward-in-time simulations

We used SLiM 3 (Haller and Messer 2019) to generate simulated sequence variation data under our model and test the performances of our approach. We performed three different sets of simulations with and without the inclusion of nonequilibrium demographics, in which we modeled diverse rates of adaptation and BGS. For each set of simulations, we considered a 5.5 Myears branch length mimicking the human split from chimpanzee. In simulations accounting for demography, we added demographic events following a two-epoch model considering diverse expansion and bottleneck. We set the population size change to happen 1,000 generations ago so testing $\alpha_{(x)}$ patterns depending on the SFS distortion due to recent demographic changes. We vary population size by a factor of 5 and 50 for a two-epoch expansion model and by 5 and 20 for a two-epoch bottleneck model. Note that unlike previous studies, we were interested in the $\alpha_{(x)}$ distortion under strong recent demographic changes (Eyre-Walker and Keightley 2009; Galtier 2016). In addition, we simulate a more *realistic* human demography following Tennessen *et al.* (2012) to model the variation in the 661 African individuals whose genomes are included in the 1,000 Genome Project (Auton *et al.* 2015).

Each simulation represents a coding locus of $2 \cdot 10^3$ bp flanked on each side by a 10^5 bp loci, which contain deleterious alleles driving background selection. A total of $5 \cdot 10^4$ genes were simulated, accounting for 10^8 bp of coding sequence. We performed the simulations following previously estimated values of negative selection of human proteins (Boyko *et al.* 2008), where the distribution of deleterious alleles follows a gamma-distribution with scale and shape parameters of 0.184 and 0.000402, respectively, which implies a mean fitness of $2\,Ns = -457$ for negatively selected nonsynonymous alleles. Strongly and weakly beneficial alleles followed a point-mass distribution given the population-scaled selection coefficients of $2\,Ns = 10$ and $2\,Ns = 500$, respectively. Our simulations supposed that 25% of mutations in each coding locus are synonymous while and 75% are nonsynonymous. We used a mutation within each coding locus of $\theta = 4\,N\mu = 0.001$ and a mean human recombination rate in the flanking sequence of $\rho = 4\,Nr = 0.001$. Given these parameters, we set an α value of 0.4 for equilibrium, and Tennessen simulations, and we decreased α to 0.2 for the two-epoch to vary the rate of adaptation tested. In the same way, because $\alpha_{(x)}$ pattern can also be affected by the number of individuals sampled (n), we tested different sample sizes for equilibrium ($n = 500$), Tennesen ($n = 661$), and two-epoch ($n = 50$) simulations. See Tables 3 and 4 for parameter values of forward simulations. We rescaled ancestral population sizes ($N_e = 10,000$) for equilibrium, and nonequilibrium model simulations according to our computational resources.

Human divergence and polymorphism

We also tested our model with empirical data. We retrieved polymorphic and divergence data in coding sequences from human hg38 assembly. Human assembly, coding sequences, and annotations were retrieved from Ensembl release 109 (Cunningham *et al.* 2022). We estimated fixed substitutions in the human branch by maximum likelihood and ancestral sequence reconstruction based on human, chimpanzee, gorilla, and orangutan alignments. First, we used pblat (Wang and Kong 2019) to align human CDS sequences on panTro6, gorGor6 and ponAbe3 assemblies downloaded from UCSC database (Nassar *et al.* 2023). To ensure that short exons at the edge of CDS were included, we used the pblat-fine option and a minimum identity threshold (-minIdentity) of 60%, keeping only the best scoring hit. The hit with the highest percentage of identity was retrieved if two or more equal best-score hits. Next, chimpanzee, gorilla, and orangutan sequences were pblatted back to hg38 human assembly in order to make sure that the overall middle genomic position for each isoform was not greater than 1 kb. By doing so, we finally retrieved 17,538 protein-coding genes accounting for a total of 94,030 CDS isoforms, which allowed us to identify the best reciprocal orthologous hits. Then, human, chimpanzee, gorilla, and orangutan sequences were aligned using MACSE v2 to explicitly account for codon structure on protein-coding alignments (Ranwez *et al.* 2018). The total number of nonsynonymous and synonymous fixations on the human branch were quantified from substitution maps using Hyphy v2.5 (Kosakovsky Pond *et al.* 2020). The process was divided into two steps: first, we fit the MG94xREV codon substitution model, a modification of the Muse and Gaut (1994) model, including a generalized time-reversible mutation rate matrix. The fit was done using MACSE alignments, and the great apes' phylogenetic tree and phylogenies were re-estimated during the fitting allowing d_N/d_S estimation by branch. Finally, each fitted model was used to reconstruct ancestral sequences and substitution maps. Fixations were summarized by gene taking into account all possible isoforms. Sites in multiple isoforms were counted only once, considering the most constraint annotation. Therefore, synonymous substitutions were considered nonsynonymous if the same position were nonsynonymous in at least one of the gene isoforms.

Polymorphic sites and derived allele frequencies were estimated from 1,000 GP phase 3 high-coverage phased data

Table 3. SLiM simulated values to demographic equilibrium and Tennessen *et al.* (2012) simulations.

Scenarios	N_{anc}	$\mu_{flanking}$	r	μ_{coding}	s_W[a]	s_S[a]	l_W[b]	l_S[c]	α_W	α	B
Demographic equilibrium	500	1.32E−06	5.00E−07	5.00E−07	10	500	3.80E−03	3.90E−04	0.1	0.4	0.2
	500	1.32E−06	5.00E−07	5.00E−07	10	500	7.70E−03	2.60E−04	0.2	0.4	0.2
	500	1.32E−06	5.00E−07	5.00E−07	10	500	1.16E−02	1.30E−04	0.3	0.4	0.2
	500	7.52E−07	5.00E−07	5.00E−07	10	500	3.80E−03	3.90E−04	0.1	0.4	0.4
	500	7.52E−07	5.00E−07	5.00E−07	10	500	7.70E−03	2.60E−04	0.2	0.4	0.4
	500	7.52E−07	5.00E−07	5.00E−07	10	500	1.16E−02	1.30E−04	0.3	0.4	0.4
	500	1.83E−07	5.00E−07	5.00E−07	10	500	3.80E−03	3.90E−04	0.1	0.4	0.8
	500	1.83E−07	5.00E−07	5.00E−07	10	500	7.70E−03	2.60E−04	0.2	0.4	0.8
	500	1.83E−07	5.00E−07	5.00E−07	10	500	1.16E−02	1.30E−04	0.3	0.4	0.8
	500	8.22E−10	5.00E−07	5.00E−07	10	500	3.80E−03	3.90E−04	0.1	0.4	0.999
	500	8.22E−10	5.00E−07	5.00E−07	10	500	7.70E−03	2.60E−04	0.2	0.4	0.999
	500	8.22E−10	5.00E−07	5.00E−07	10	500	1.16E−02	1.30E−04	0.3	0.4	0.999
Tennesen model	7,310	1.32E−06	3.42E−08	3.42E−08	10	500	3.80E−03	3.90E−04	0.1	0.4	0.2
	7,310	1.32E−06	3.42E−08	3.42E−08	10	500	7.70E−03	2.60E−04	0.2	0.4	0.2
	7,310	1.32E−06	3.42E−08	3.42E−08	10	500	1.16E−02	1.30E−04	0.3	0.4	0.2
	7,310	7.52E−07	3.42E−08	3.42E−08	10	500	3.80E−03	3.90E−04	0.1	0.4	0.4
	7,310	7.52E−07	3.42E−08	3.42E−08	10	500	7.70E−03	2.60E−04	0.2	0.4	0.4
	7310	7.52E−07	3.42E−08	3.42E−08	10	500	1.16E−02	1.30E−04	0.3	0.4	0.4
	7,310	1.83E−07	3.42E−08	3.42E−08	10	500	3.80E−03	3.90E−04	0.1	0.4	0.8
	7,310	1.83E−07	3.42E−08	3.42E−08	10	500	7.70E−03	2.60E−04	0.2	0.4	0.8
	7,310	1.83E−07	3.42E−08	3.42E−08	10	500	1.16E−02	1.30E−04	0.3	0.4	0.8
	7,310	8.22E−10	3.42E−08	3.42E−08	10	500	3.80E−03	3.90E−04	0.1	0.4	0.999
	7,310	8.22E−10	3.42E−08	3.42E−08	10	500	7.70E−03	2.60E−04	0.2	0.4	0.999
	7,310	8.22E−10	3.42E−08	3.42E−08	10	500	1.16E−02	1.30E−04	0.3	0.4	0.999

[a]Population-scaled selection coefficient ($2N_e s$).
[b]Fixation probability of weakly beneficial alleles.
[c]Fixation probability of strong beneficial alleles.

Table 4. SLiM simulated values to the two-epoch model simulations.

Scenarios	N_{anc}	$\mu_{flanking}$	r	μ_{coding}	s_W[a]	s_S[a]	l_W[b]	l_S[c]	α_W	α	B		
Two –epoch model	1,000	50	1,000	6.38E−08	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.4
	1,000	50	1,000	6.38E−08	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.999
	1,000	200	1,000	8.22E−10	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.4
	1,000	200	1,000	6.38E−08	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.999
	1,000	5,000	1,000	8.22E−10	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.4
	1,000	5,000	1,000	8.22E−10	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.999
	1,000	50,000	1,000	6.38E−08	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.4
	1,000	50,000	1,000	6.38E−08	2.50E−07	2.50E−07	10	500	2.91E−03	7.79E−05	0.1	0.2	0.999

[a]Population-scaled selection coefficient ($2N_e s$).
[b]Fixation probability of weakly beneficial alleles.
[c]Fixation probability of strong beneficial alleles.

(Byrska-Bishop *et al.* 2022) across seven African ancestry populations. Nonsynonymous and synonymous polymorphism were annotated using VEP (McLaren *et al.* 2016). Ancestral and derived allele frequencies were estimated using Ensembl v109 human ancestral allele information from EPO multialignments. Polymorphic sites only were counted if they overlapped those parts of human coding sequences that were previously aligned. In total, 17,538 orthologues were included in the analysis.

Bootstrap test

We followed the bootstrap procedure extensively described at Enard and Petrov (2020) and Di *et al.* (2021) to compare adaptation levels between VIPs and non-VIP genes. The bootstrap procedure allows us to quantify adaptation specific to VIPs, comparing them with a control set that matches with VIPs on the number and average values of several confounding factors that also can determine the rate of adaptive evolution compared to the rest of the genome (Castellano *et al.* 2019b; Huang 2021). The bootstrap test relies on a straightforward control set-building algorithm that adds control genes to a progressively growing control set (Enard and Petrov 2020); in such a way, the growing control set has the same range of values of confounding factors as the tested set of genes of interest. Following this procedure, we avoid matching VIPs genes individually with non-VIPs control but match the overall VIPs set. Otherwise, we would drastically reduce the pool of potential control genes. When increasing the non-VIPs control set, we set a 5% matching limit over the VIPs set confounding average and limit each non-VIP to represent up to 1,2% of the control set. In total, we match nine potential confounding factors between VIPs and non-VIP control datasets (other confounding factors previously included in Di *et al.* (2021) were found to not impact the present comparison of VIPs and matched non-VIPs):

- Average overall expression in 53 GTEx v7 tissues (THE GTEX CONSORTIUM 2020). We used the log (in base 2) of Transcripts Per Million (TPM).
- Expression (log base 2 of TPM) in GTEx lymphocytes. Expression in immune tissues may impact the rate of sweeps.
- Expression (log base 2 of TPM) in GTEx testis. Expression in testis might also impact the rate of sweeps.
- deCode recombination rates in 50 kb and 500 kb gene-centered windows. The average recombination rates in the gene-centered windows are calculated using the most recent

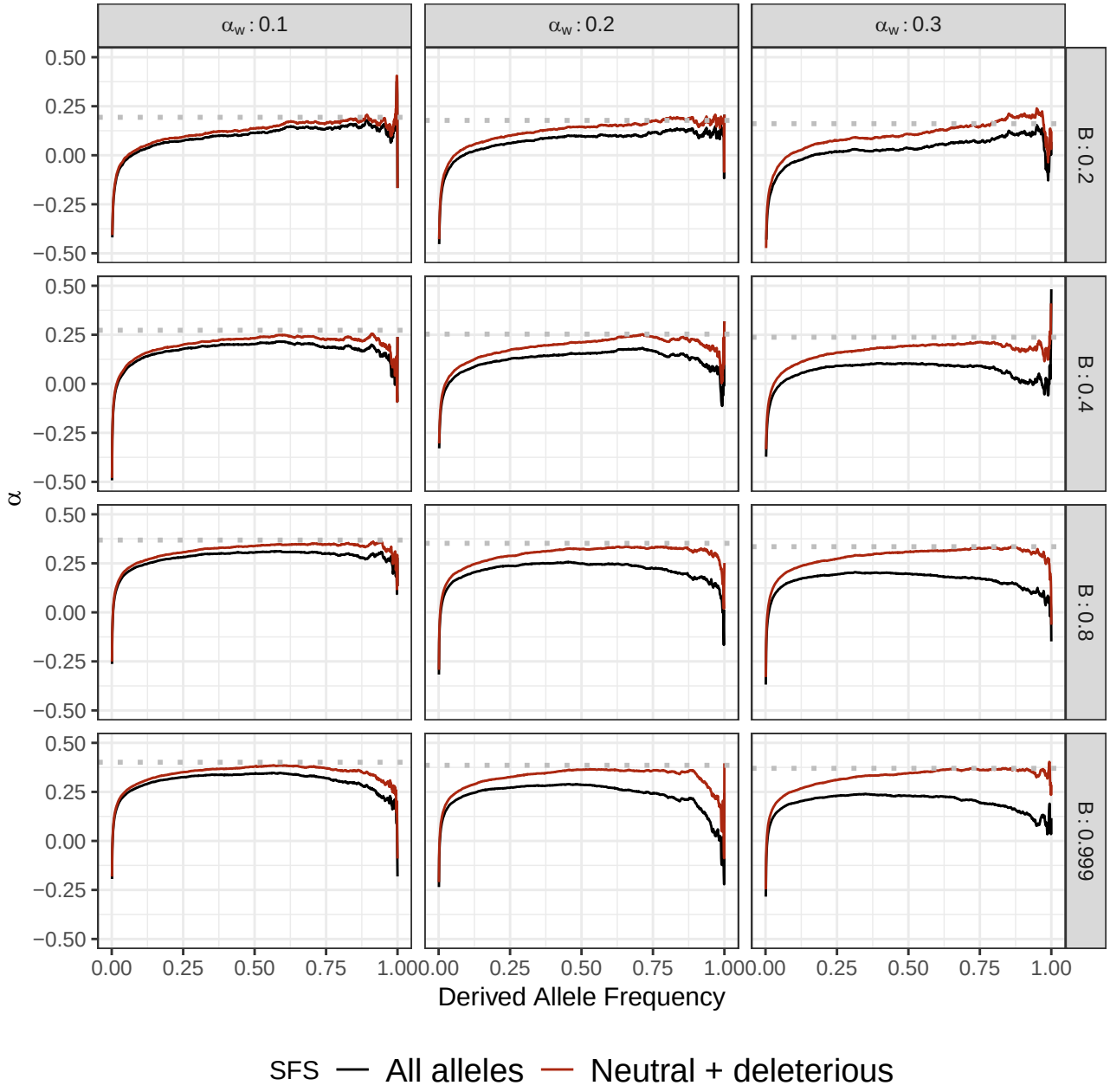

Fig. 1. Effect of weakly advantageous mutation in the presence of BGS in equilibrium demography. We simulated the effect of weakly advantageous allele and BGS effect on $\alpha_{(x)}$. Each row represents a B value. Each column represents a proportion of α_W assuming a total proportion of adaptive substitutions of $\alpha = 0.4$ in the absence of BGS. Dotted line is the true simulated α. Greater proportion of α_W largely biases $\alpha_{(x)}$ patterns which would affect aMK estimations. Stronger BGS removes beneficial mutations from $\alpha_{(x)}$.

deCode recombination map (Halldorsson *et al.* 2019). We use both 50 kb and 500 kb window estimates to account for the effect of varying window sizes on the estimation of this confounding factor (same logic for other factors where we also use both 50 kb and 500 kb windows).

- GC content at the third position of codons in the aligned Ensembl coding sequences.
- The number of protein-protein interactions (PPIs) in the human protein interaction network (Luisi *et al.* 2015). The number of PPIs has been shown to influence the rate of sweeps (Luisi *et al.* 2015). We use the log (base 2) of the number of PPIs.
- The aligned coding sequence length.
- The proportion of immune genes. The matched control sets have the same proportion of immune genes as disease genes, immune genes being genes annotated with the Gene Ontology terms GO:0002376 (immune system process), GO:0006952 (defense response), and/or GO:0006955 (immune response) as of May 2020 (The Gene Ontology Consortium 2021).

Results

Evolutionary processes affecting adaptation inference

We used our forward simulations to retrieve polymorphism and divergence data with a priori known adaptation parameters. To input the data in our model, we pooled the SFS and number of fixations of $5 \cdot 10^4$ genes. Note that this is about 2.5 times as many genes as appear in the human genome -we use this larger number

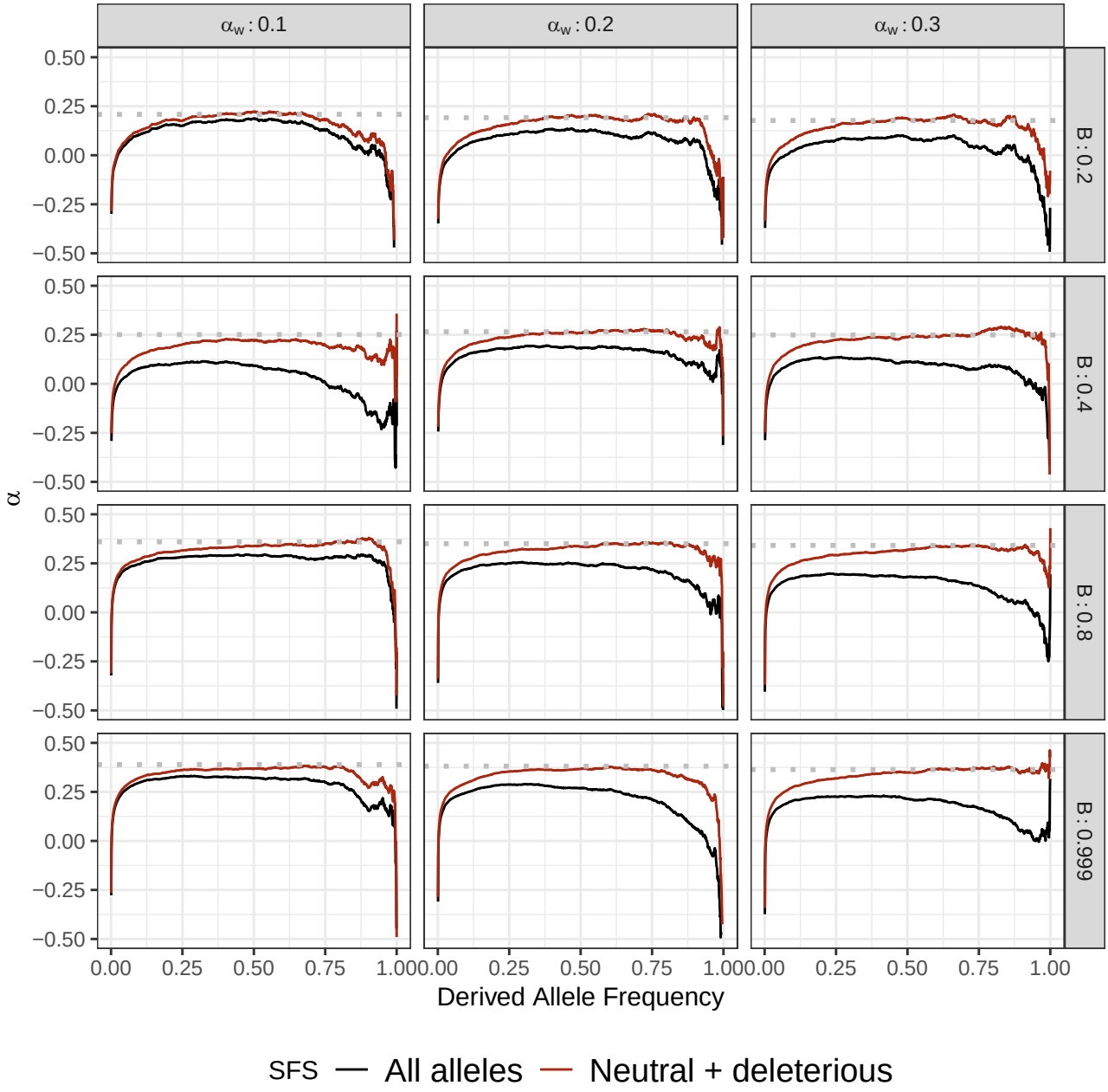

Fig. 2. Effect of weakly advantageous mutation in the presence of BGS and under a nonequilibrium demography. We simulated the effect of weakly advantageous allele and BGS effect on $\alpha_{(x)}$ under the model of Tennessen *et al.* (2012). Each row represents a B value. Each column represents a proportion of αW assuming a total proportion of adaptive substitutions of $\alpha = 0.4$ in the absence of BGS. Dotted line is the true simulated α. Greater proportion of α_W largely biases $\alpha_{(x)}$ patterns which would affect aMK estimations. Stronger BGS removes beneficial mutations from $\alpha_{(x)}$.

of genes such that the trends in the simulated data will be clear and not dominated by noise, while noting that noise may play a larger role in real datasets for some species with limited proteome coverage.

As demonstrated previously, the frequency spectrum (Tataru *et al.* 2017) and $\alpha_{(x)}$ (Uricchio *et al.* 2019) are substantially affected by the presence of weakly beneficial alleles when the rate of such mutations is relatively high (see Figs 1, 2). The asymptote of the $\alpha_{(x)}$ curve bends below the true value of $\alpha_{(x)}$, which is caused by an excess of high frequency nonsynonymous variants under weak positive selection. This results in downward bias of α estimates when aMK-test (Messer and Petrov 2013) is applied to the data (see Table 5). In real sequencing datasets we cannot a priori separate beneficial and deleterious alleles, but in our simulated datasets we can remove the weakly beneficial alleles and test whether this will fix the downwards bias in aMK. When removing weakly beneficial alleles we observed an increase in α estimates from aMK which tend towards the true value of α (see Fig. 1 and Table 5). In addition, $\alpha_{(x)}$ can be substantially affected by BGS, especially when weakly beneficial alleles contribute to the frequency spectrum (see Fig. 1). In cases where α is dominated by strong adaptation, both asymptotic values (accounting for all alleles, or just neutral and deleterious) tend to be similar, because strongly beneficial alleles are not substantially impeded by selective interference with linked deleterious variation.

We also tested the effect of recent demographic events with a simulation of the Tennessen *et al.* (2012) demographic model, specifically for the African continental group. We used demographic

Table 5. α and aMK estimations.

	True α_w	B	True α	$aMK_{(neutral+deleterious)}$	$aMK_{(all\ alleles)}$
Demographic equilibrium	0.1	0.2	0.194	0.163	0.132
	0.2	0.2	0.178	0.161	0.102
	0.3	0.2	0.161	0.151	0.058
	0.1	0.4	0.273	0.222	0.185
	0.2	0.4	0.253	0.214	0.141
	0.3	0.4	0.238	0.191	0.078
	0.1	0.8	0.369	0.337	0.294
	0.2	0.8	0.352	0.317	0.222
	0.3	0.8	0.335	0.308	0.172
	0.1	0.999	0.401	0.358	0.312
	0.2	0.999	0.386	0.338	0.241
	0.3	0.999	0.370	0.354	0.203
Tennesen model	0.1	0.2	0.209	0.16	0.102
	0.2	0.2	0.192	0.184	0.069
	0.3	0.2	0.177	0.165	0.030
	0.1	0.4	0.251	0.202	0.047
	0.2	0.4	0.265	0.245	0.140
	0.3	0.4	0.249	0.248	0.094
	0.1	0.8	0.36	0.32	0.254
	0.2	0.8	0.35	0.322	0.205
	0.3	0.8	0.341	0.323	0.152
	0.1	0.999	0.389	0.357	0.293
	0.2	0.999	0.381	0.354	0.229
	0.3	0.999	0.364	0.358	0.179

parameters following Adrion *et al.* (2020). To improve performance we simulated the African population in isolation, rather than including the full multipopulation model—consequently our simulations do not include the effects of migration on the frequency spectrum. We observed similar patterns to equilibrium simulation regarding the overall shape and asymptotic values of $\alpha_{(x)}$ (Fig. 2). Since this model includes a recent and rapid population expansion, we observe distortions to the frequency spectrum at extremely low and high frequencies due to the excess of rare alleles relative to an equilibrium demographic model.

The same effects were observed in the two-epoch expansion model. Overall, we found the same expected differences on aMK-test and $\alpha_{(x)}$ curves in the two-epoch expansion considering α were reduced from 0.4 to 0.2 compared to equilibrium and Tennessen *et al.* (2012) simulations (see Fig. 3 and Table 6). Hence, BGS and the proportion of weakly advantageous polymorphism would drive the $\alpha_{(x)}$ curve differences despite the demographic model and the overall α. Nonetheless, as in Tennessen *et al.* (2012) simulations, $\alpha_{(x)}$ are biased in extremely low and high frequencies due to the excess of rare alleles, depending on the event strength. However, like in the Tennessen *et al.* (2012) simulations, aMK-test estimations are almost unaffected compared to the equilibrium simulation when removing weak advantageous mutations. That is not the case in the two-epoch bottleneck model. Although weakly advantageous polymorphism affects $\alpha_{(x)}$ curves similarly depending on the BGS strength, aMK-test estimations are biased compared to the equilibrium simulation. Since population contraction makes selection less efficient in removing deleterious mutations, such scenarios lead to an increment in P_N and underestimation of α, depending on the strength of the demographic event (see Fig. 3 and Table 6). Because our method is based on the $\alpha_{(x)}$, such scenarios must be treated particularly carefully, and we discuss them in the following sections.

ABC evaluation using simulation

To compare true parameter values to inferred values, we estimated the mode and 95% CI for each posterior distribution that we obtained from ABC-MK as an approximation of α. Our method can distinguish both weak and strong adaptive contributions to α while performing reasonably accurate estimations. Table 7 shows inferred values and the associated error for each parameter and simulation. In all cases, the posterior distribution overlaps the distribution of the true values from bootstrapped datasets. Figure 4 presents the simulated values and mode estimates from posterior distributions on equilibrium simulations. We also compare ABC-MK to several MK-test extensions, including alternatives which exclude all variants below a frequency threshold for the derived frequency allele (Fay *et al.* 2001; Murga-Moreno *et al.* 2022) and Grapes (Galtier 2016), a Maximum Likelihood software fitting the DFE to infer α. We run Grapes using GammaZero and GammaExpo models since both models best fit the simulated DFE (see Materials and methods section). Figure 5 shows only α estimation since other methods cannot distinguish between α_W and α_S. We note that other more straightforward implementations based on frequency threshold as Fay, Wycoff, and Wu (FWW) (Fay *et al.* 2001) or imputed MK-test (Murga-Moreno *et al.* 2022) extensions achieved similar accuracy to ABC-MK in cases where the proportion of weak adaptation is low. Nonetheless, such results are not only highly dependent on the proportion of strongly positively selected alleles but also on the DFE shape as noted in Murga-Moreno *et al.* (2022), which makes them a poor option for genomes with a highly leptokurtic deleterious DFE. Because Grapes became computationally expensive for large sample sizes and numerically unstable (see Supplementary Fig. 6 and Discussion section), we projected the SFS to a sample size of 20 individuals to perform α inferences. As noted by Al-Saffar and Hahn (2022) Grapes achieved more accurate results for both GammaZero and GammaExpo models than other MK-test implementations. Although α estimates of GammaZero and ABC-MK are comparable in many cases, we note highly variable α inference when fitting the GammaExpo model (see Fig. 5).

We also tested our method using simulations performed under a two-epoch model, including moderate and strong ancient expansion and bottleneck demographic, and the demographic model of Tennessen *et al.* (2012) (see Tables 7 and 8). Although the demographic events affect the number of segregating sites and

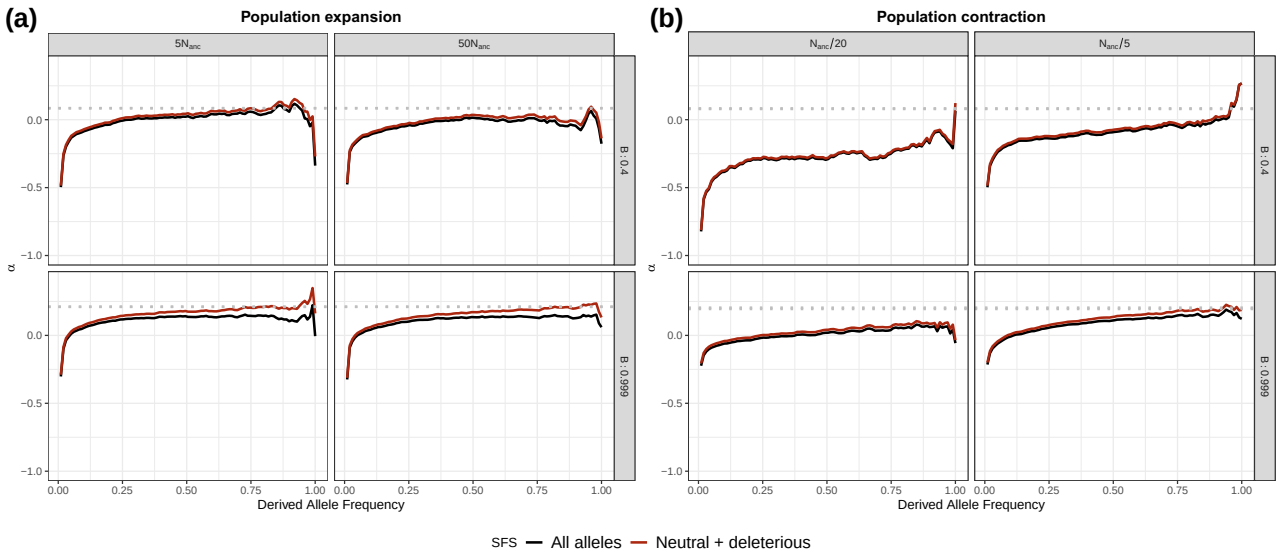

Fig. 3. Effect of weakly advantageous mutation in the presence of BGS and population size change strength. We simulated the effect of weakly advantageous allele and BGS effect on $\alpha_{(x)}$ under a two-epoch model. Each row represents a B value. Each column represents the strength of the contraction or expansion of a proportion assuming a total proportion of adaptive substitutions of $\alpha = 0.2$ in the absence of BGS. Dotted line is the true simulated α. α_W were set to contribute a half of the total α. Stronger BGS removes beneficial mutations from $\alpha_{(x)}$.

Table 6. α and aMK estimations on the twoepoch model.

	N_{anc}	N_B	**True** α_W	**B**	**True** α	$aMK_{(neutral+deleterious)}$	$aMK_{(all\ alleles)}$
Two-epoch model	1,000	50.0	0.1	0.4	0.079	−0.202	−0.220
	1,000	50.0	0.1	0.999	0.195	0.077	0.050
	1,000	200.0	0.1	0.4	0.083	0.075	0.057
	1,000	200	0.1	0.999	0.204	0.188	0.147
	1,000	5,000	0.1	0.4	0.084	0.059	0.030
	1,000	5,000	0.1	0.999	0.210	0.194	0.130
	1,000	50,000	0.1	0.4	0.085	0.013	−0.014
	1,000	50,000	0.1	0.999	0.212	0.189	0.129

the shape of the SFS (which is not modelled in our calculations), our method is reasonably robust to these distortions, including an ancient expansion in the ancestral population and a period of recent rapid growth when excluding low-frequency variants (DAC <5 i.e. DAF <0.0038; Fig. 6, Figs. 7, 9). The inference is biased towards weak selection when we include variants at which the SFS is most distorted, such as singletons and very high-frequency variants. However, the overall value of α is not strongly affected (Fig. 6). Such a pattern likely reflects the qualitative similarity of recent growth events and weakly beneficial alleles in terms of their effects on the SFS, as both will disproportionately increase the number of nonsynonymous variants at a low frequency relative to an equilibrium model. In Fig. 6, we explore a more comprehensive range of parameters using the best-performing set of summary statistics (i.e. excluding derived allele counts under 5). Tables 7 and 8 show the estimated parameters and their corresponding errors for both the Tennessen *et al.* (2012) and two-epoch model simulations. Figure 8 shows the same patterns as the comparisons we made in equilibrium scenarios, where FWW and imputed MK-test just show slight bias because of the simulated DFE shape when the proportion of weak adaptation is low and GammaZero and GammaExpo are the most accurate results along with ABC-MK. However, GammaExpo results are highly variable compared to the other tested MK-tests.

We noted that in cases where the strength of the bottleneck in the ancestral population is larger, our software performed worse α estimation than in the expansion scenarios (see Fig. 9, Table 8). As shown in Figs 8 and 10, Grapes and ABC-MK perform similarly in all cases, but the inference also fails when the ancient bottleneck and the BGS are strong. In addition, although the simulation included a higher proportion of weakly adaptive alleles (with parameters $\alpha = 0.2$ and $\alpha_W = 0.1$), we observed that the GammaExpo model provided less accurate estimates compared to the GammaZero model in all simulated scenarios and GammaExpo results vary more than any other method (see Figs. 8 and 10).

Interestingly, although Grapes assumes independent segregating sites, we noted that in some cases it achieves similar α estimates to ABC-MK in the simulations, including strong to moderate BGS scenarios. Note that Grapes implements the r nuisance parameter correction proposed in Eyre-Walker *et al.* (2006). The r parameter was initially designed to capture the effect of demographic changes on the fate of synonymous and nonsynonymous mutations (Eyre-Walker *et al.* 2006). In practice, the r parameter allows each SFS category to account for each own mutation rate (see equations 6 and 7 from Galtier 2016). Considering the frequency category i at the SFS, r_i is estimated as the relative effective mutation rate of the i frequency compared to singletons at neutral sites (Eyre-Walker *et al.* 2006; Galtier 2016; Tataru *et al.* 2017). Multiple studies have shown that correcting the expected SFS by r can at least partially absorb other SFS distortions like the one produced by ascertainment bias, nonrandom

Table 7. ABC inference in equilibrium and Tennessen *et al.* (2012) simulated datasets.

Scenario	True α	True α_W	B	α_W	α_s	α	$\Delta\alpha_W$	$\Delta\alpha_S$	$\Delta\alpha$
Equilibrium	0.4	0.1	0.2	0.122 [0.074–0.171]	0.072 [0.047–0.096]	0.185 [0.158–0.212]	0.088	0.089	0.009
	0.4	0.1	0.4	0.089 [0.042–0.141]	0.178 [0.15–0.205]	0.248 [0.228–0.271]	0.04	0.046	0.025
	0.4	0.1	0.8	0.092 [0.05–0.126]	0.277 [0.252–0.306]	0.357 [0.341–0.373]	0.021	0.022	0.012
	0.4	0.1	0.999	0.08 [0.046–0.114]	0.326 [0.3–0.353]	0.394 [0.381–0.408]	0.001	0.003	0.007
	0.4	0.2	0.2	0.128 [0.089–0.164]	0.067 [0.046–0.088]	0.184 [0.152–0.214]	0.06	0.042	0.006
	0.4	0.2	0.4	0.116 [0.077–0.154]	0.121 [0.1–0.142]	0.222 [0.2–0.245]	0.015	0.031	0.031
	0.4	0.2	0.8	0.163 [0.127–0.194]	0.199 [0.18–0.222]	0.345 [0.33–0.361]	0.016	0.006	0.007
	0.4	0.2	0.999	0.153 [0.114–0.189]	0.232 [0.209–0.258]	0.372 [0.352–0.389]	0.01	0.009	0.013
	0.4	0.3	0.2	0.09 [0.056–0.121]	0.039 [0.025–0.052]	0.123 [0.094–0.149]	0.015	0.017	0.038
	0.4	0.3	0.4	0.163 [0.12–0.203]	0.071 [0.052–0.092]	0.219 [0.19–0.25]	0.004	0.008	0.019
	0.4	0.3	0.8	0.245 [0.207–0.281]	0.122 [0.101–0.14]	0.35 [0.331–0.371]	0.017	0.015	0.015
	0.4	0.3	0.999	0.214 [0.18–0.248]	0.161 [0.139–0.182]	0.36 [0.342–0.381]	0.041	0.046	0.01
Tennesen model	0.4	0.1	0.2	0.015 [−0.029–0.071]	0.209 [0.177–0.235]	0.206 [0.187–0.23]	0.022	0.037	0.003
	0.4	0.1	0.4	0.21 [0.169–0.254]	0.096 [0.08–0.114]	0.286 [0.254–0.316]	0.038	0.017	0.035
	0.4	0.1	0.8	0.114 [0.074–0.153]	0.293 [0.263–0.32]	0.391 [0.374–0.407]	0.037	0.01	0.031
	0.4	0.1	0.999	0.178 [0.14–0.211]	0.28 [0.253–0.31]	0.442 [0.428–0.457]	0.094	0.025	0.053
	0.4	0.2	0.2	0.067 [0.025–0.106]	0.147 [0.121–0.173]	0.195 [0.174–0.217]	0.009	0.03	0.004
	0.4	0.2	0.4	0.125 [0.087–0.167]	0.187 [0.164–0.212]	0.292 [0.272–0.312]	0.015	0.033	0.026
	0.4	0.2	0.8	0.213 [0.167–0.256]	0.19 [0.164–0.217]	0.385 [0.369–0.402]	0.056	0.003	0.035
	0.4	0.2	0.999	0.194 [0.152–0.23]	0.231 [0.205–0.259]	0.408 [0.392–0.427]	0.02	0.024	0.027
	0.4	0.3	0.2	0.106 [0.062–0.146]	0.093 [0.069–0.123]	0.185 [0.158–0.211]	0.011	0.033	0.008
	0.4	0.3	0.4	0.254 [0.201–0.303]	0.095 [0.079–0.115]	0.329 [0.297–0.362]	0.083	0.018	0.08
	0.4	0.3	0.8	0.272 [0.235–0.313]	0.139 [0.118–0.161]	0.386 [0.362–0.409]	0.03	0.04	0.044
	0.4	0.3	0.999	0.28 [0.244–0.322]	0.153 [0.136–0.174]	0.409 [0.391–0.429]	0.023	0.045	0.046

Error were measured using the mean difference in 100 simulations replicas.

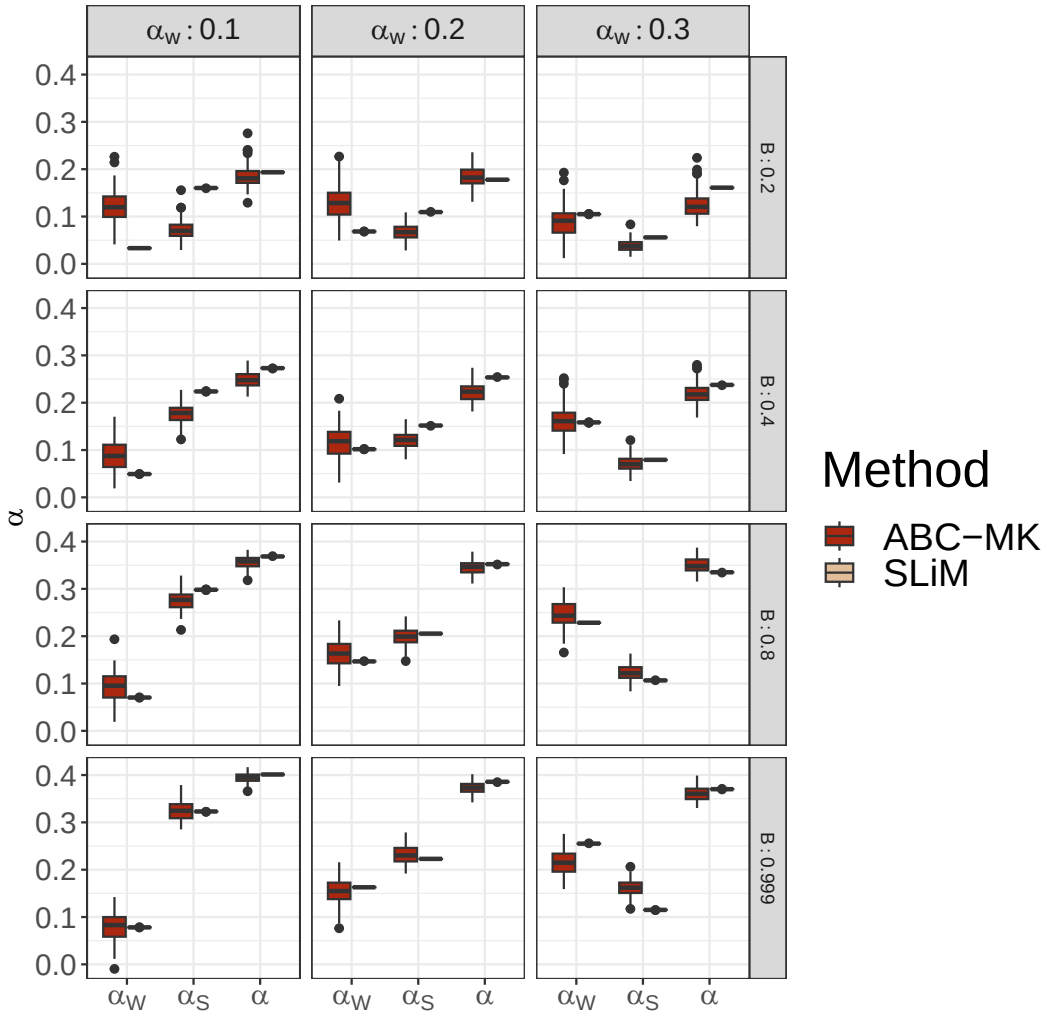

Fig. 4. Mode distribution of 100 sets of summary statistics per parameter value. SLiM values correspond to the true observed α_W, α_S and α estimated from simulations. ABC inferences were performed using ABCreg (Thornton 2009).

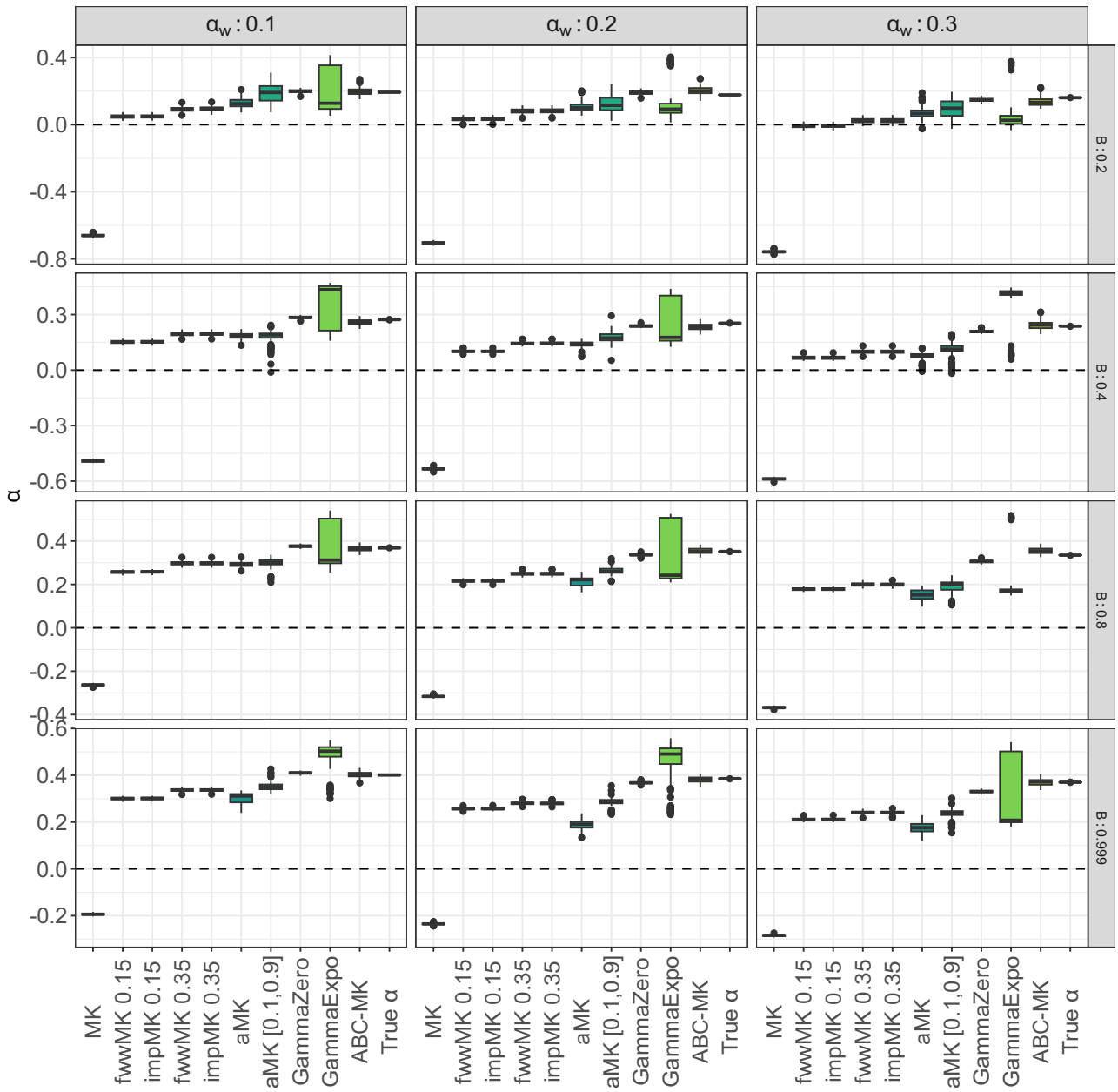

Fig. 5. MK-test comparison at equilibrium. ABC-MK were measured through the mode distribution of 100 sets of summary statistics per parameter value. SLiM values correspond to the true observed α_W, α_S, and α estimated from simulations. GammaZero and GammaExpo model were run using the projecting the SFS in a sample of 20 individuals. ABC inferences were performed using ABCreg (Thornton 2009).

sampling, population substructure or linked selection (Galtier 2016; Tataru *et al.* 2017; Galtier and Rousselle 2020). Similarly, we found that the *r* parameter can absorb unexpected polymorphic patterns in α estimates due to linked selection (Galtier 2016; Tataru *et al.* 2017; Galtier and Rousselle 2020). Still, as noted in Galtier and Rousselle (2020), we observed that absorbing linked selection distortions leads to a significant underestimation of the inferred parameters of the DFE for deleterious alleles (γ and shape parameters) by Grapes not only in the strong bottleneck scenario but also in the equilibrium and the Tennessen *et al.* (2012) scenarios. Despite the relative accuracy of its α estimates, Grapes's DFE model inferences are downward biased, especially for the γ parameter (Table 9). We observe higher precision in estimates of the DFE-related parameters derived from ABC-MK, especially when BGS is strong (see Supplementary Figs. 1, 2, 3, 4, 5, and 6).

Application of ABC-MK: human virus-interacting proteins

As an example application, we estimated adaptation in human genes that interact with viruses. Several studies have found that viral infections have driven adaptation in the human genome (e.g. Nédélec *et al.* 2016; Castellano *et al.* 2019b). Genomic analysis of patterns of variation within experimentally determined viral interacting proteins (VIPs) has repeatedly uncovered signals of both frequent and strong adaptation (Enard *et al.* 2016; Uricchio *et al.* 2019). Multiple lines of evidence suggest that the selective pressure imposed by viruses on hosts appears to be an important driver of strong adaptation during both long term (Enard *et al.* 2016; Castellano *et al.* 2019b; Uricchio *et al.* 2019) and more recent human evolution (Deschamps *et al.* 2016; Racimo *et al.* 2017; Enard and Petrov 2018).

Table 8. ABC inference in two-epoch model simulated datasets.

	N_{anc}	N_B	True α	True α_W	B	α_W	α_S	α	$\Delta\alpha_W$	$\Delta\alpha_S$	$\Delta\alpha$
Two-epoch model	1,000	50	0.2	0.1	0.4	0.04 [0–0.093]	0.005 [−0.002–0.015]	0.054 [0.01–0.12]	0.007	0.045	0.03
	1,000	50	0.2	0.1	0.999	−0.055 [−0.088–0.02]	0.169 [0.137–0.205]	0.107 [0.09–0.122]	0.144	0.054	0.097
	1,000	200	0.2	0.1	0.4	0.106 [0.074–0.133]	0.01 [0.006–0.013]	0.123 [0.091–0.152]	0.075	0.039	0.044
	1,000	200	0.2	0.1	0.999	0.005 [−0.009–0.021]	0.029 [0.009–0.056]	0.035 [0.013–0.058]	0.082	0.08	0.16
	1,000	5,000	0.2	0.1	0.4	0.05 [0.025–0.079]	0.049 [0.026–0.077]	0.1 [0.073–0.129]	0.019	0.004	0.015
	1,000	5,000	0.2	0.1	0.999	0.146 [0.094–0.197]	0.117 [0.084–0.149]	0.25 [0.232–0.272]	0.055	0.002	0.04
	1,000	50,000	0.2	0.1	0.4	0.005 [−0.011–0.025]	0.028 [0.008–0.052]	0.036 [0.014–0.066]	0.026	0.026	0.049
	1,000	50,000	0.2	0.1	0.999	0.058 [0.014–0.103]	0.146 [0.108–0.182]	0.195 [0.179–0.216]	0.032	0.025	0.017

Errors were measured using the mean difference in 100 simulations replicas.

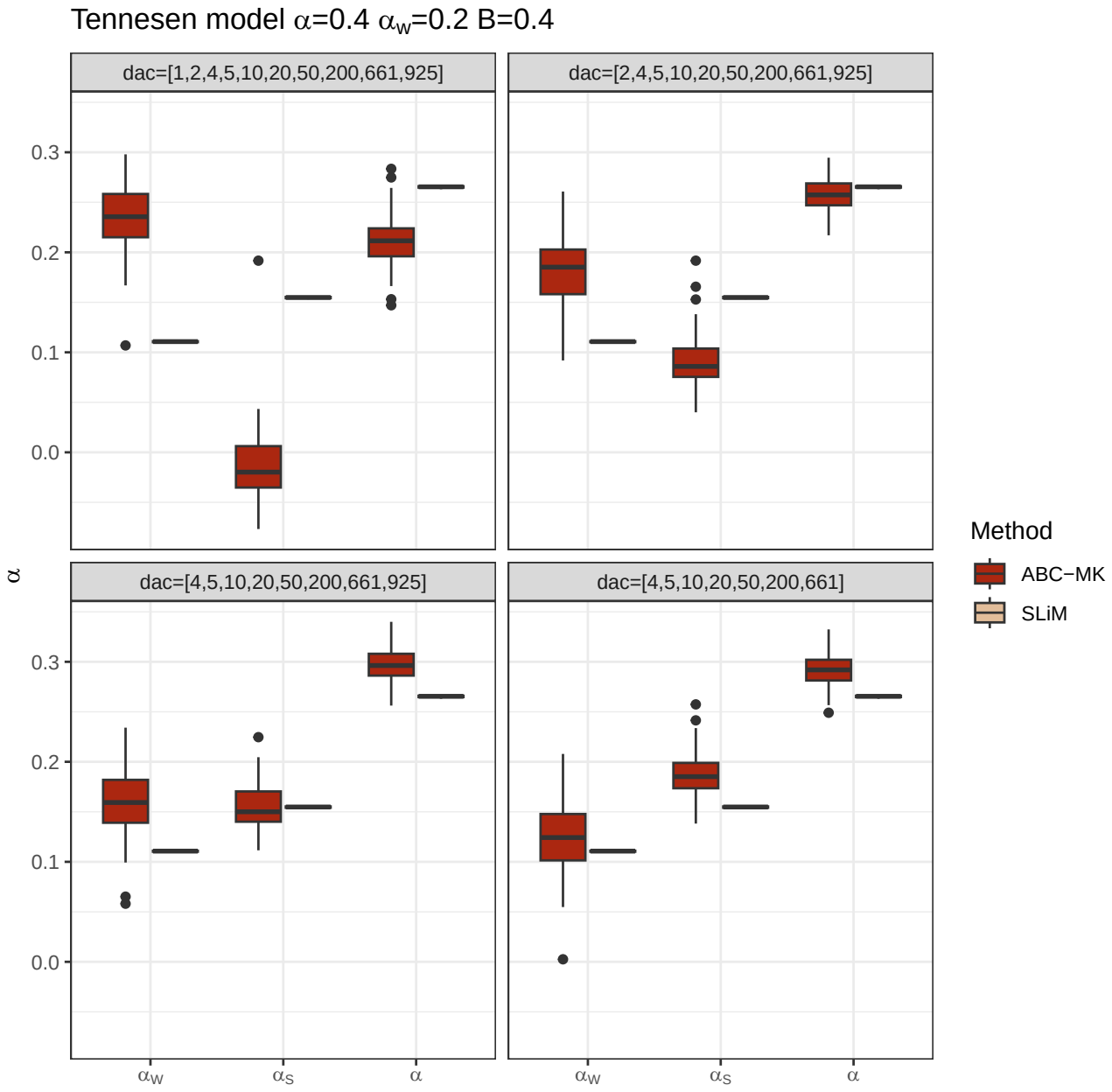

Fig. 6. Summary statistics selection. ABC inference excluding low-frequency variants in Tennessen *et al.* (2012) demographic simulations. SLiM values correspond to the true observed α_W, α_S, and α estimated from simulations.

We recently found that RNA viruses have driven more recent human evolution (in the past 50,000 years) at RNA virus-interacting VIPs (RNA VIPs) than DNA viruses at DNA virus-interacting VIPs (DNA VIPs) (Enard and Petrov 2018, 2020). Here we used ABC-MK to test whether the same was true during the millions of years of human evolution since divergence with

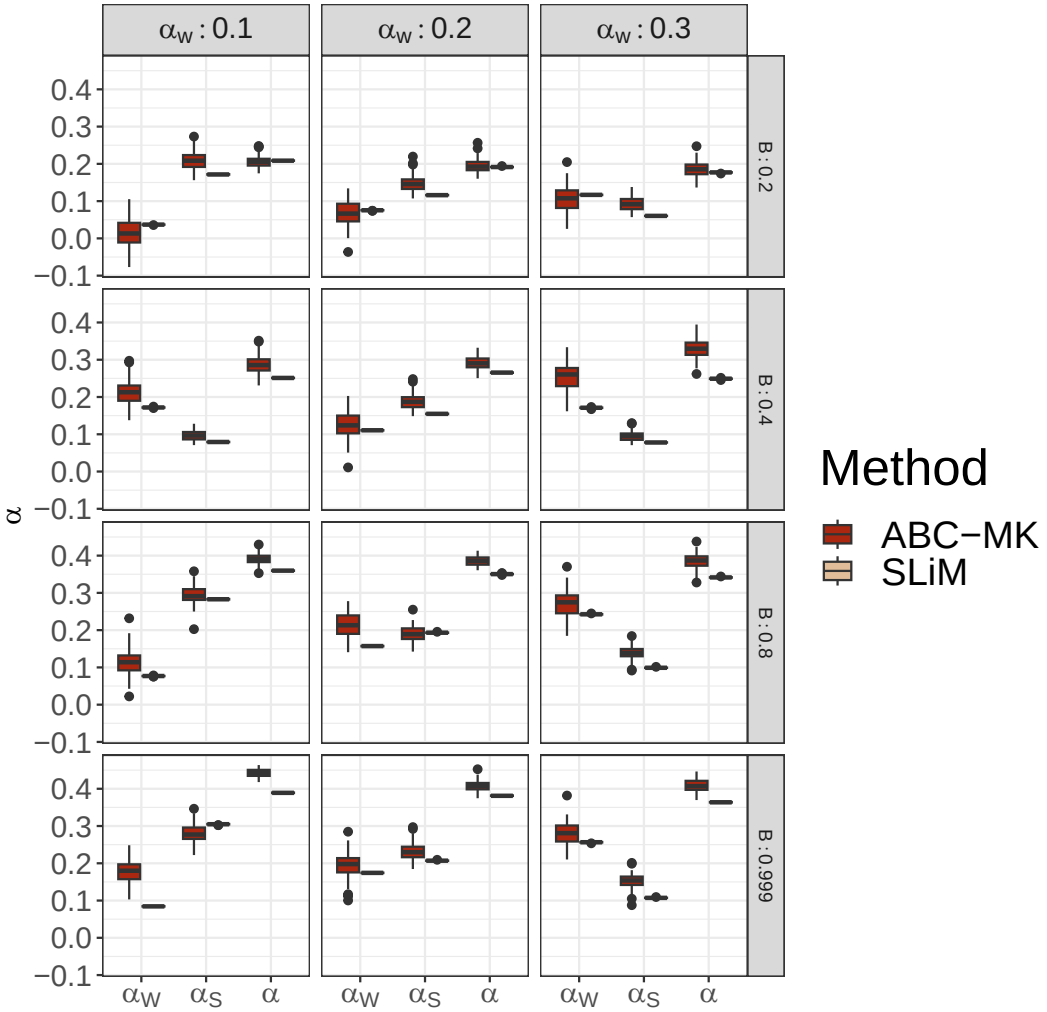

Fig. 7. ABC-MK α_W, α_S, and α inferences of Tennessen *et al.* (2012) demographic model simulations. Mode distribution of 100 datasets. ABC inference was performed using ABCreg (Thornton 2009). SLiM values correspond to the true observed α_W, α_S, and α estimated from simulations.

chimpanzees. To this end, we limit our analyses to a subset of VIP and non-VIP mammals orthologs because gene age can significantly impact the rate of adaptation since they can evolve faster and experience mutations with stronger fitness effects to achieve their optimum fitness (Moutinho *et al.* 2022). Orthologs were identified through human coding sequence alignment with 261 mammal genomes following (Bowman *et al.* 2023) to obtain human, chimpanzee, gorilla, and orangutan orthologs. First, the pipeline uses blat (Kent 2002) to detect ortholog sequences with at least 251 out of the 261 genome assemblies, followed by searching the top hit to the human query back to the human genome. Second, it uses a combination of MACSE v2 and HMMcleaner to align orthologs that can explicitly account for codon structure, identify frameshifts, and readjust reading frames accordingly while cleaning visibly ambiguous or erroneous segments, finding a total of 13,495 orthologs. Thus, to compare the impact of RNA and DNA viruses, we measured adaptation rates separately for the specific VIPs of nine different RNA viruses and the specific VIPs of six different DNA viruses each with more than 180 VIPs (Figs. 11,12).

We subset genomic data corresponding to the 661 individuals of African descent sampled in the Thousand Genomes Project, and fixed-substitution on human-branch were identified by maximum likelihood based on alignments of human (hg38 assembly), chimpanzee (panTro6 assembly) and orangutan (ponAbe3 assembly). Adaptation rates on VIPs were estimated following the previous bootstrap procedure described at Enard and Petrov (2020) and Di *et al.* (2021) (see Materials and Methods). The bootstrap procedure allows us to build random non-VIPs control sets that match the same average values of several confounding factors that might explain the adaptation level rather than the virus interactions themselves. We estimated non-VIPs null distributions by sampling 1,000 non-VIPs sets of the same size as the analyzed VIPs category. Because we defined we defined $P_{N(x)}$ and $P_{S(x)}$ as any polymorphisms above frequency x, we excluded any variants above 70% frequency to avoid mispolarization errors (see Fig. 11). α values for VIPs and non-VIPs datasets were inferred through the mode and 95% CI estimates from posterior distributions following the previously described ABC scheme for each dataset.

We inferred values of α, α_W, and α_S that were very similar to those inferred in Uricchio *et al.* (2019). Looking at the VIPs for specific RNA or DNA viruses, we find that a subset of RNA virus families drove the vast majority of adaptation in VIPs (Fig. 12). More specifically, RNA viruses drove strong adaptation but not weak adaptation (Fig. 12), with five RNA viruses (RHINOV, HIV, ZIKAV, IAV, and COV) with VIPs with significantly elevated strong adaptation, versus only one DNA virus (HPV). We further find this is true using either α or ω_a (Fig. 12, Table 10, Supplementary Fig. 7,

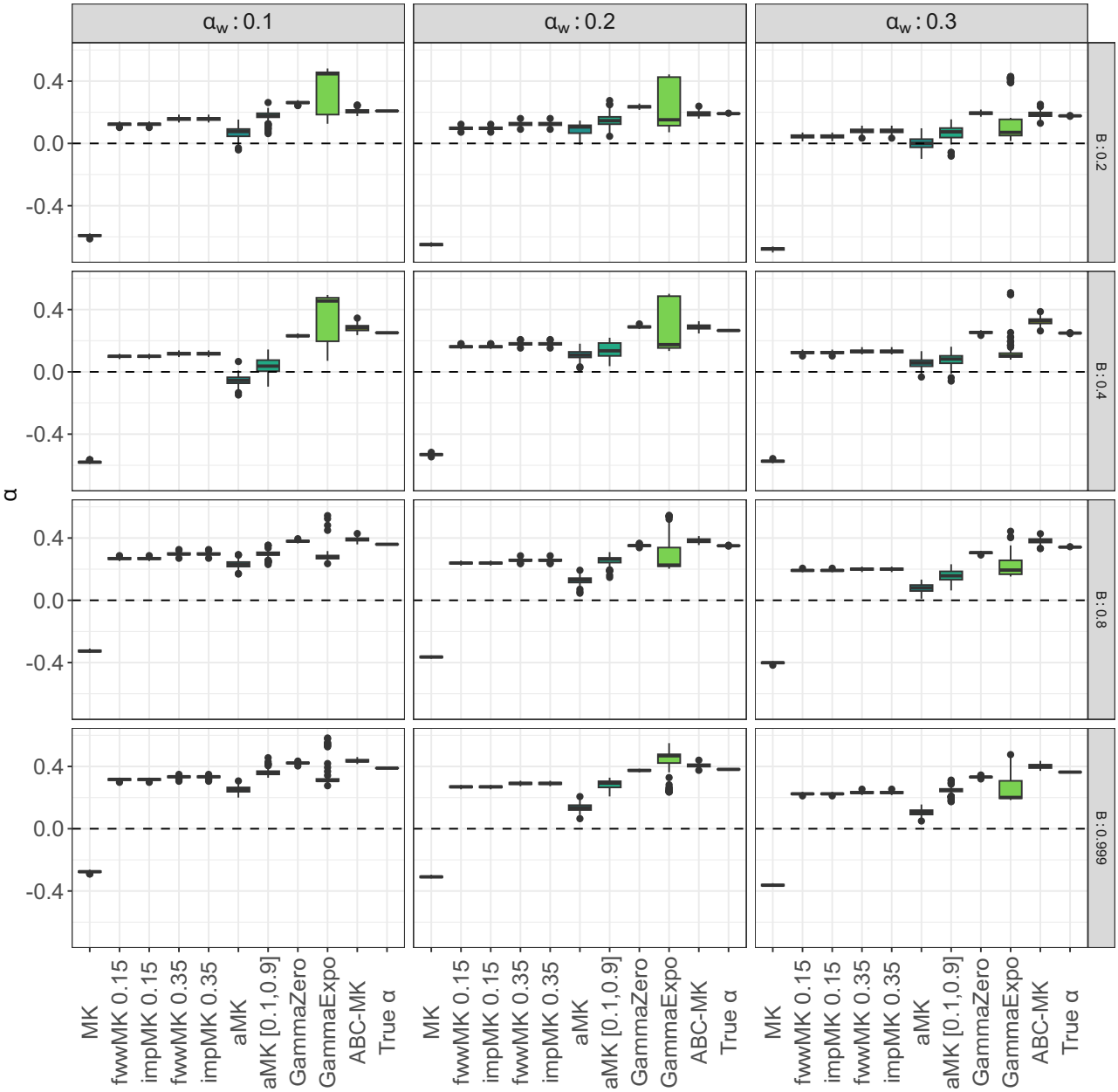

Fig. 8. MK-test comparison at Tennessen *et al.* (2012) simulation. ABC-MK were measured through the mode distribution of 100 sets of summary statistics per parameter value. SLiM values correspond to the true observed α_W, α_S and α estimated from simulations. GammaZero and GammaExpo model were run using the projecting the SFS in a sample of 20 individuals. ABC inferences were performed using ABCreg (Thornton 2009).

Supplementary Table 1) (Galtier 2016) to compare VIPs and controls. These results are coherent with growing evidence that RNA viruses drove strong recent adaptation more frequently than DNA viruses (Enard and Petrov 2018, 2020).

Discussion

Though several MK-test extensions were developed to bypass limitations of the classic MK framework, generating accurate and unbiased estimates remains challenging. Here, we proposed a new version of the method presented in Uricchio *et al.* (2019), ABC-MK, which can efficiently deal with the presence of deleterious, beneficial alleles, partial recombination, and linked selection. any aspects of the input data must be considered before performing α estimations from the MK-test. Thus, our analyses tested ABC-MK and other MK-tests extensions, including several intrinsic data features that can potentially bias α estimation, not only including the presence of slightly beneficial allele and linked selection but also the effect of the sample size or the recommended minimum amount of polymorphic and divergent data. We note that similar studies have been carried out to evaluate MK-test strength and robustness under different conditions, including the frequency threshold to remove weakly selected segregating alleles, the effect of the folded and the unfolded Site Frequency Spectrum, the gene-by-gene analyses, the underlying DFE or the role of selected rare, strong mutation on DFE inferences (Charlesworth and Eyre-Walker 2008; Booker 2020; Al-Saffar and Hahn 2022; Murga-Moreno *et al.* 2022). We aim to expand such studies by comparing the most up-to-date and newly proposed ABC-MK method. Hence, we analyzed a battery set of simulations

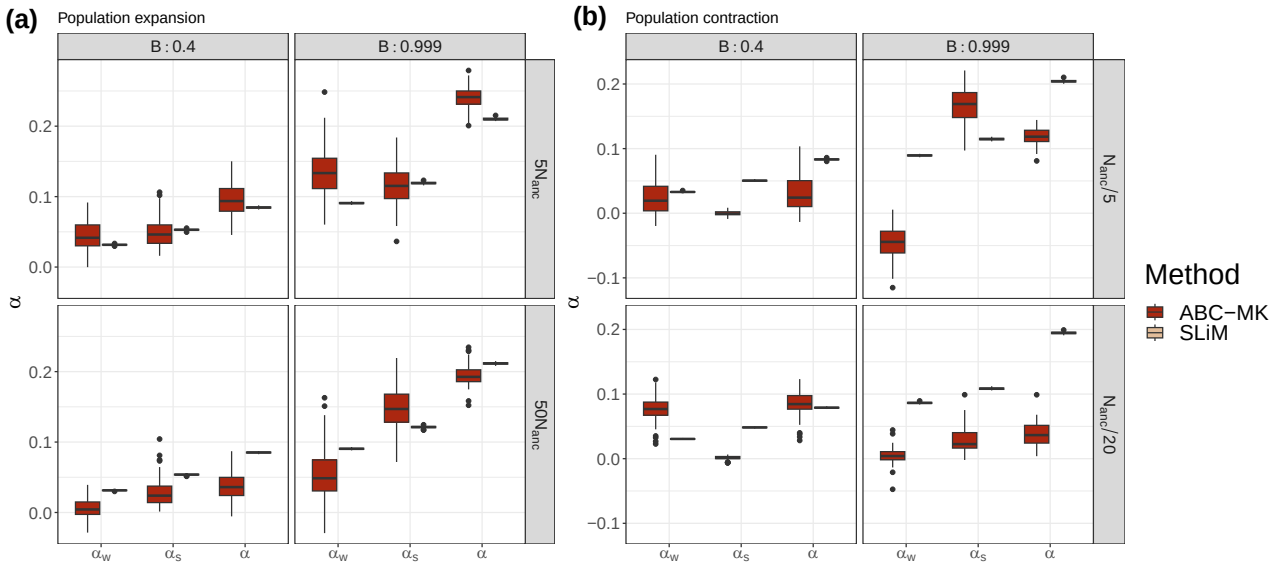

Fig. 9. ABC-MK α_W, α_S, and α inferences of the two-epoch demographic model simulations. Mode distribution of 100 datasets. ABC inference was performed using ABCreg (Thornton 2009). SLiM values correspond to the true observed α_W, α_S, and α estimated from simulations. a) Two-epoch expansion simulations. b) Two-epoch bottleneck simulations.

Table 9. DFE inference in two-epoch model simulated datasets using ABC-MK and grapes.

Parameter	N_{anc}	N_B	True α	True α_W	B	True value	ABC-MK	GammaExpo	GammaZero
Shape (β)	1,000	50	0.2	0.1	0.4	0.184	0.827	2.048	1.242
	1,000	50	0.2	0.1	0.999	0.184	0.707	1.866	1.273
	1,000	200	0.2	0.1	0.4	0.184	1.126	2.556	1.509
	1,000	200	0.2	0.1	0.999	0.184	0.502	2.192	1.305
	1,000	5,000	0.2	0.1	0.4	0.184	0.906	1.386	1.034
	1,000	5,000	0.2	0.1	0.999	0.184	0.971	1.713	1.141
	1,000	50,000	0.2	0.1	0.4	0.184	0.706	0.993	0.764
	1,000	50,000	0.2	0.1	0.999	0.184	0.813	1.523	1.079
Negative selection coefficient (γ)	1,000	50	0.2	0.1	0.4	457	1.253	0.089	0.060
	1,000	50	0.2	0.1	0.999	457	0.931	0.175	0.159
	1,000	200	0.2	0.1	0.4	457	1.103	0.041	0.024
	1,000	200	0.2	0.1	0.999	457	1.227	0.096	0.096
	1,000	5,000	0.2	0.1	0.4	457	1.164	0.297	0.172
	1,000	5,000	0.2	0.1	0.999	457	0.898	0.354	0.294
	1,000	50,000	0.2	0.1	0.4	457	1.203	0.442	0.199
	1,000	50,000	0.2	0.1	0.999	457	0.939	0.419	0.313

Each method shows the mean ratio between the inference and the true value of 100 simulations replicas.

including multiple BGS levels to test MK-tests under the explicit assumption that the SFS can account for the presence of deleterious and beneficial alleles while linked selection can heavily affect the rate of adaptation as well as distort the SFS. We discuss the limits and advantages of the most up-to-date MK-test methods to provide valuable guidelines on the MK-test analyses design and test ABC-MK limitations.

First, the new ABC-MK is a user-friendly version that deals with genomic data without the cost of forward-in-time simulations. Despite the community effort to provide valuable and fast simulation approaches (Haller and Messer 2019; Adrion *et al.* 2020), including BGS and large population sizes at simulations has a substantial computational cost. Hence, α inferences in ABC approaches based on forward-in-time simulation, like the previous ABC-MK, became unrealistic in terms of computational time and resources, and previous analysis took days on High-Performance Computing cluster (Castellano *et al.* 2019b; Uricchio *et al.* 2019). In addition, the first ABC-MK implementation employed a resampling-based approach that allows querying parameter values but uses the same set of forward simulations to improve efficiency. Such an approach avoids simulating the entire model for different parameter combinations. However, it relies on the same set of simulations, limiting the analyses to a particular level of polymorphic and divergent sites, known demographic parameters, as well as a set of priors values of BGS, mutation rate, recombination rate (Uricchio *et al.* 2019).

The new ABC-MK is based on a simple random sampling scheme around analytical expectations, avoiding expensive forward-in-time simulations. Such a strategy accounts for sampling and process variance, and proved to be a reasonably robust to nonequilibrium demography. The computational workflow can be completed in a few hours at first run because it requires

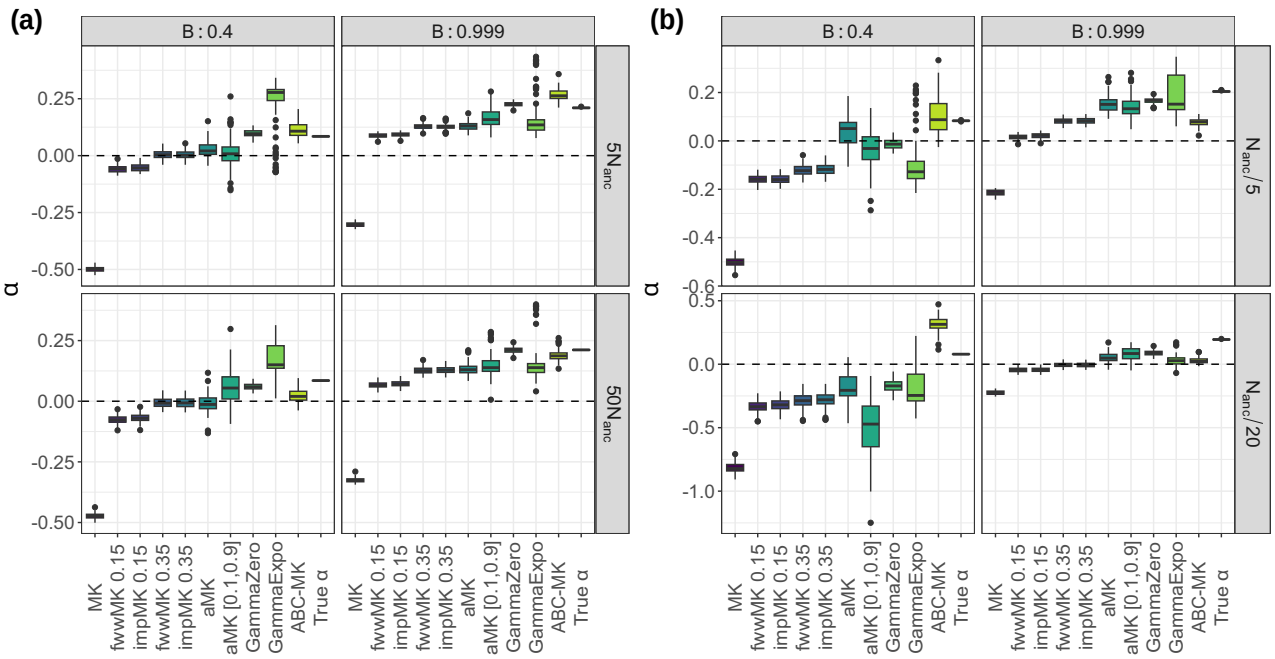

Fig. 10. ABC-MK α_W, α_S and α inferences of the two-epoch demographic model simulations. Mode distribution of 100 datasets. ABC inference was performed using ABCreg (Thornton 2009). a) Two-epoch expansion simulations. b) Two-epoch bottleneck simulations.

estimating enough analytical parameter combinations to perform ABC estimations. As an example, the first run step takes about an hour to estimate the 3.7 million parameter combinations for each selected sample size used in this study on an i7-7700HQ laptop CPU with eight cores. Although such timing is relative to computational resources, the selected population size, BGS levels, and the total number of parameter combinations, it should be noted that we bypassed the previous requirement of utilizing a High Performance Computing cluster. Most importantly, once the analytical expectations are estimated, subsequent estimations can be done in a few seconds, including several inferences at a time.

Second, in this study, we assume the user has no a priori knowledge about BGS level, adaptation rate or selection strength. Hence, in addition to circumventing forward-in-time simulations and the resampling strategy, we showed that the new ABC-MK can achieve accurate α and DFE parameters estimations without prior knowledge about the expected adaptation rate, the DFE or the BGS level. Table 2 show prior values where a considerable uniform distribution range is considered for each parameter. Such a feature became especially interesting since whole genome sequencing (WGS) became cheaper and more nonmodel species genomic data became available. Although rigorous adaptation studies require good-quality WGS data, a priori knowledge about the underlying DFE, recombination, mutation rate, or BGS levels can be unknown and require extra effort. Especially background selection data can pose a challenging analysis, requiring good quality annotations and recombination maps (McVicker *et al.* 2009; Murphy *et al.* 2022). Nonetheless, once WGS data is available, good quality annotation, demography, recombination maps, and BGS estimation are convenient steps that can be achieved even for a few samples and increase the accuracy of the adaptation rate estimation. For the specific case of the ABC-MK, such information can help limit prior parameters, resulting in more accurate ABC inferences. In other approaches, such as exome sequencing, it became impossible to establish some previous knowledge, like BGS estimation. We encourage the user to use WGS instead of exome sequencing whenever possible to avoid any associated errors such as variant calling, homology errors, mapping paralog or bad-quality gene annotation. Still, we configure ABC-MK to be as flexible as possible; hence, the user can fix parameters, or limited parameter ranges to achieve more accurate estimations rather than rely on a remarkably large uniform prior distribution. Such an option can lead to more accurate estimations whenever parameter information is available.

Third, as we thoroughly describe in the previous sections, ABC-MK exploits $\alpha_{(x)}$ while being able to discern the difference between weakly, strongly advantageous variants and BGS. Unlike strongly advantageous mutations, weakly advantageous mutations do not reach fixation so quickly that they become negligible to nonsynonymous polymorphism. Hence, weakly advantageous variants tend to segregate before eventually reaching fixation, which results in a downward trend of the $\alpha_{(x)}$, especially at higher frequencies. Although we can discern such a trend, it is essential to note that the $\alpha_{(x)}$ patterns may also trend downward due to mispolarization error biasing α inferences due to an unrealistic inference of α_W. Therefore, based on the findings of Hernandez *et al.* (2007) about mispolarization when considering positive selection, we recommend running ABC-MK using nonsynonymous and synonymous variants with a derived allele frequency less than 0.7 or at least 0.9 as in Haller and Messer (2017) in cases where the user requires to maximize the input data. In addition, unlike Grapes, FWW or imputed MK-test, ABC-MK cannot deal with the folded

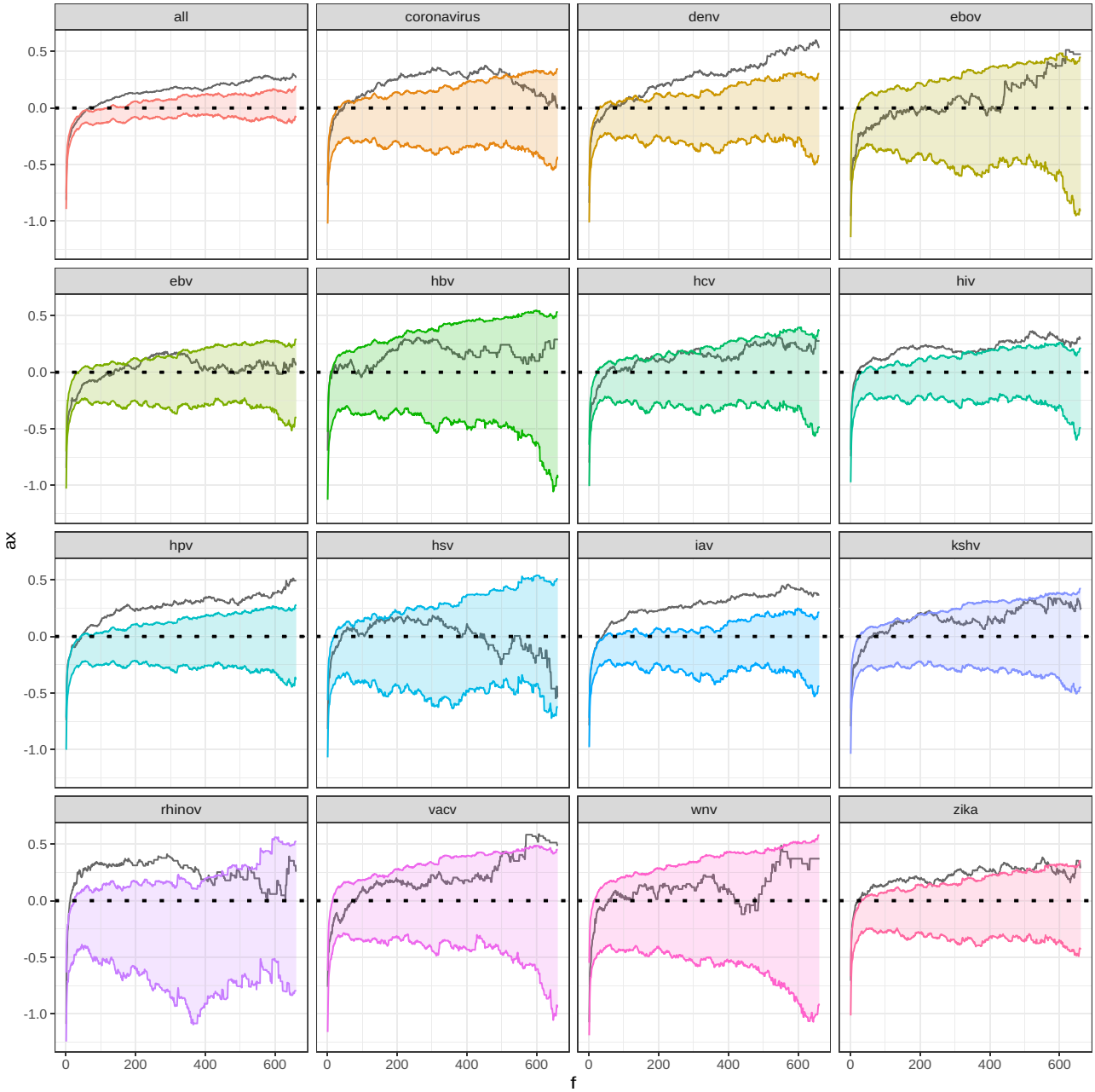

Fig. 11. α_x patterns on VIPs and non-VIPs control. Colored range represent 95% $\alpha_{(x)}$ confidence interval.

SFS (see Murga-Moreno *et al.* 2022 to further discussion) because the $\alpha_{(x)}$ pattern here exploded would be modified and low-frequency variants will be indistinguishable among deleterious and weakly advantageous alleles and the effect of mispolarization significantly increased.

Finally, similar to other methods like Grapes (Galtier 2016), polyDFE (Tataru *et al.* 2017), or DFE-alpha (Eyre-Walker and Keightley 2009), ABC-MK is not intended for applications to individual genes or very small or subsets of genes. For such applications, other methods may be preferred (see Murga-Moreno *et al.* 2022 for further discussion) and we strongly discourage the use of the method in these cases. Nonetheless, we resample a random equilibrium scenario to establish the ABC-MK's accuracy, robustness and scalability regarding the amount of data and the sampling size. Supplementary Fig. 8 shows how ABC-MK and the other tested MK-tests are affected by such variables. We projected the SFS to a smaller size while resampling and pooling 100 simulations, reducing the total number of simulations by 10%, 1%, and 0.1%. ABC-MK is surprisingly unaffected by sample size but shows similar decreases in performance when reducing the total number of sites to other more straightforward MK-test extensions. One might expect that any estimate reliant on the $\alpha_{(x)}$ becomes

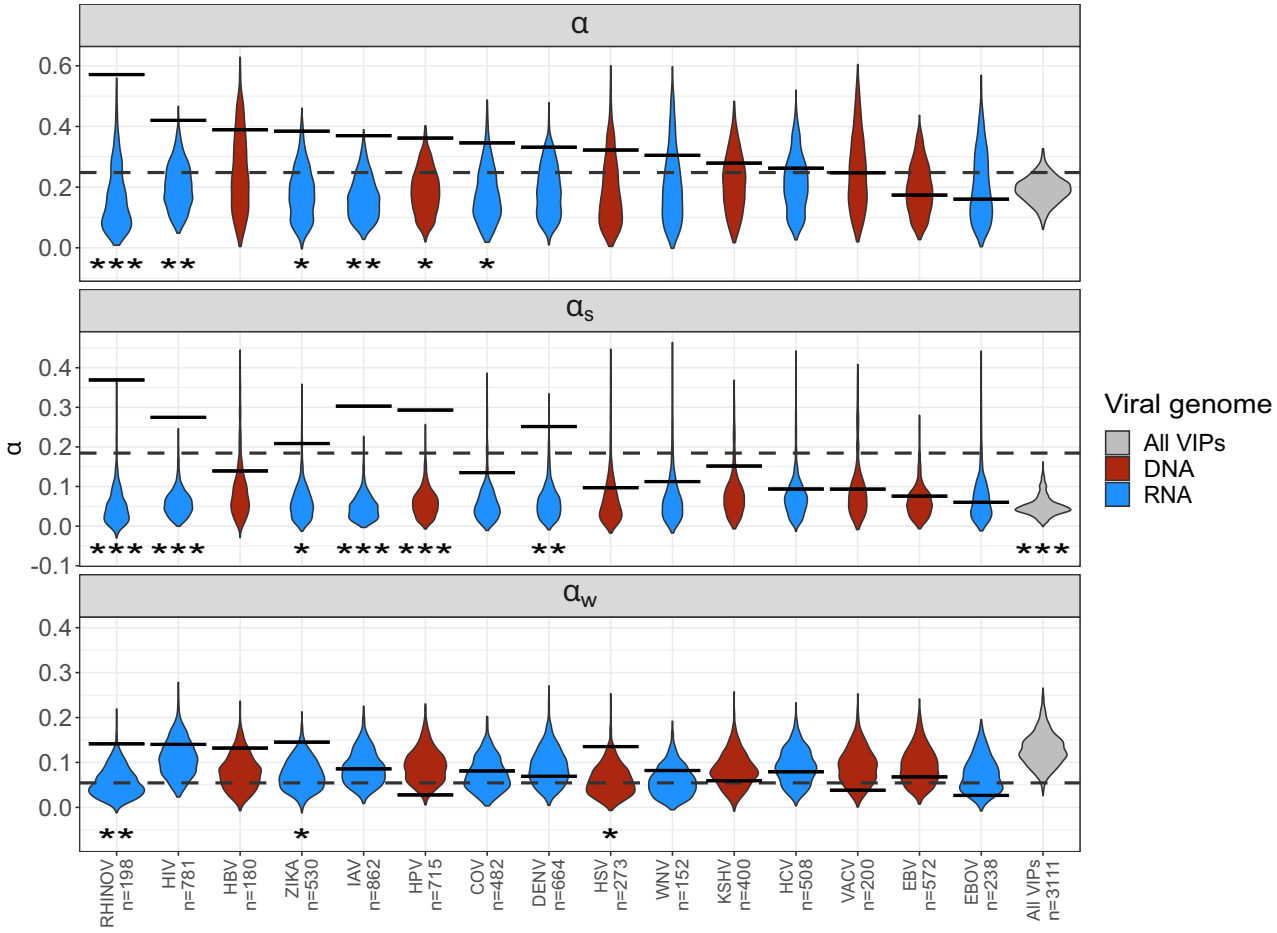

Fig. 12. ABC-MK inference on different RNA and DNA-VIPs categories. Violin plots and solid lines represent inferences on non-VIP bootstrapped datasets and VIP categories according to the virus interaction. The dashed line is the overall VIPs inference.

less reliable than other methods when reducing the SFS information in terms of polymorphic sites. The same might be expected when reducing the sample size since both could lead to a nonasymptotic behaviour on $\alpha_{(x)}$. We suggest that ABC-MK accuracy and robustness to polymorphism and sample size are dependent on the use of the cumulative SFS in the inference step, which seems to scale better with sample size as noted by Uricchio *et al.* (2019), and trivially has the same asymptote as the noncumulative SFS (Uricchio *et al.* 2019). Supplementary Fig. 8 and Supplementary Table 2 show that nor ABC-MK or Grapes are the best option when the total number of polymorphic sites is equal or less than 200–400 since we observe biased estimation in such cases and we note that Grapes GammaExpo became numerically unstable depending on the sample size, leading to biased α inference for the same simulated model. In addition, the computational cost significantly increased when analyzing more than 20 individuals for GammaZero and GammaExpo. ABC-MK takes the same amount of time independent of the sample size, and a simple analysis of our TGP-like simulation can improve GammaZero and GammaExpo from seconds to hours.

In sum, our implementation of ABC-MK provides robust model inferences of adaptation rate and strength under a range of evolutionary scenarios that are likely to be applicable to nonmodel genomes. Our user-friendly implementation runs on a desktop computer, can handle genome-scale data, and does not require a priori knowledge of demographic history or background selection strength. Since it is feasible and efficient to use for both large and small samples, it could accelerate research into the environmental, biological, and evolutionary drivers of adaptation across the tree of life.

Data availability

We developed user-friendly software to run ABC-MK and the bootstrap procedure described at Enard and Petrov (2020) and Di *et al.* (2021). The software includes and automates all the different MK-tests tested in the study, including Grapes, which is automatically downloaded and installed as a Conda package. ABC inference can be performed using ABCreg (Thornton 2009) or a custom implementation of the ABC local linear regression function from R package (Csilléry *et al.* 2012). A custom implementation of DADI (Gutenkunst *et al.* 2009) and moments (Jouganous *et al.* 2017) projection to downsample the SFS is also available to facilitate Grapes use. The software is based on Julia language and supports multithreading, interactive environments as well as Command Line Interface usage. Docker and Singularity

Table 10. Mode and 95% CI of α_W, α_S, α, negative selection coefficient ($\gamma = 2N_e s_-$), shape parameter (β) and p-values in human datasets.

Dataset	Viral genome	α_W	α_S	α	γ	β	P-value α_W	P-value α_S	P-value α
RHINOV	RNA	0.142 [0.033–0.664]	0.369 [0.044–0.621]	0.571 [0.36–0.841]	471.161 [210.552–968.289]	0.394 [0.299–0.504]	0.001	0.001	0.009
HIV	RNA	0.14 [0.007–0.523]	0.275 [−0.002–0.411]	0.42 [0.285–0.623]	567.58 [230.068–990.292]	0.265 [0.213–0.358]	0.006	0.001	0.237
HBV	DNA	0.132 [0.019–0.562]	0.139 [0.006–0.475]	0.389 [0.182–0.709]	632.614 [231.317–993.048]	0.244 [0.161–0.36]	0.167	0.169	0.097
ZIKA	RNA	0.145 [0.011–0.517]	0.209 [−0.012–0.405]	0.384 [0.243–0.628]	309.872 [223.871–982.288]	0.252 [0.196–0.347]	0.017	0.023	0.030
IAV	RNA	0.086 [−0.072–0.457]	0.303 [0.001–0.44]	0.37 [0.246–0.552]	528.845 [215.823–977.793]	0.263 [0.214–0.348]	0.003	0.001	0.492
HPV	DNA	0.028 [−0.065–0.46]	0.293 [−0.031–0.427]	0.362 [0.217–0.527]	404.733 [229.822–980.814]	0.242 [0.182–0.322]	0.015	0.001	0.981
COV	RNA	0.081 [−0.049–0.455]	0.135 [0.002–0.411]	0.346 [0.162–0.571]	528.202 [219.173–988.197]	0.242 [0.178–0.34]	0.044	0.056	0.389
DENV	RNA	0.069 [−0.059–0.471]	0.252 [−0.05–0.41]	0.331 [0.188–0.523]	724.572 [237.58–997.376]	0.247 [0.185–0.333]	0.054	0.007	0.670
HSV	DNA	0.135 [−0.003–0.507]	0.097 [−0.011–0.379]	0.322 [0.12–0.634]	655.98 [223.942–976.381]	0.236 [0.165–0.351]	0.144	0.225	0.028
WNV	RNA	0.082 [−0.012–0.468]	0.113 [−0.012–0.435]	0.305 [0.089–0.613]	322.908 [219.774–980.638]	0.286 [0.209–0.404]	0.195	0.173	0.240
KSHV	DNA	0.059 [−0.057–0.425]	0.152 [−0.047–0.382]	0.279 [0.103–0.518]	484.308 [232.981–981.482]	0.236 [0.172–0.335]	0.266	0.089	0.727
HCV	RNA	0.079 [−0.029–0.42]	0.094 [−0.037–0.332]	0.262 [0.101–0.497]	454.763 [213.725–974.311]	0.219 [0.167–0.319]	0.289	0.257	0.598
VACV	DNA	0.038 [−0.069–0.428]	0.093 [−0.045–0.389]	0.247 [0.036–0.519]	536.851 [211.195–964.98]	0.222 [0.147–0.331]	0.476	0.364	0.917
EBV	DNA	0.068 [−0.032–0.368]	0.076 [−0.027–0.266]	0.174 [0.057–0.445]	886.085 [226.868–980.951]	0.21 [0.164–0.296]	0.568	0.315	0.693
EBOV	RNA	0.027 [−0.052–0.383]	0.06 [−0.043–0.331]	0.16 [0.005–0.465]	577.142 [227.886–994.858]	0.241 [0.159–0.339]	0.571	0.492	0.904
All VIPs	DNA/RNA	0.054 [−0.087–0.331]	0.184 [−0.008–0.316]	0.248 [0.165–0.414]	339.151 [210.262–962.819]	0.212 [0.182–0.281]	0.093	0.001	0.988

containers are also available. Software, tutorials, and documentation are freely available at https://github.com/jmurga/MKtest.jl. The code and Jupyter notebooks used to perform the analyses are available at https://github.com/jmurga/abcmk_simulations. Supplemental material available at G3 online.

Funding

David Enard is funded by NIH NIGMS MIRA grant 5R35GM142677.

Conflicts of interest

The author declares no conflict of interest.

Literature cited

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, *et al.* 2020. A community-maintained standard library of population genetic models. eLife. 9:e54967.

Al-Saffar SI, Hahn MW. 2022. Evaluating methods for estimating the proportion of adaptive amino acid substitutions. Pages: 2022.08.15.504017 Section: New Results.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, *et al.* 2015. A global reference for human genetic variation. Nature. 526:68–74.

Balloux F, Lehmann L. 2012. Substitution rates at neutral genes depend on population size under fluctuating demography and overlapping generations. Evolution. 66:605–611.

Barton NH. 1995. Linkage and the limits to natural selection. Genetics. 140:821–841.

Booker TR. 2020. Inferring parameters of the distribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare. G3: Genes, Genomes, Genetics. 10:2317–2326.

Booker TR, Keightley PD. 2018. Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. Mol Biol Evol. 35:2971–2988.

Bowman JD, Silva N, Schüftan E, Almeida JM, Brattig-Correia R, Oliveira RA, Tüttelmann F, Enard D, Navarro-Costa P, Lynch VJ. 2023. Pervasive relaxed selection on spermatogenesis genes coincident with the evolution of polygyny in gorillas. Pages: 2023.10.27.564379 Section: New Results

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, *et al.* 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, *et al.* 2022. High-coverage whole-genome sequencing of the expanded 1,000 genomes project cohort including 602 trios. Cell. 185:3426–3440.e19.

Castellano D, James J, Eyre-Walker A. 2018. Nearly neutral evolution across the Drosophila melanogaster genome. Mol Biol Evol. 35:2685–2694.

Castellano D, Macià MC, Tataru P, Bataillon T, Munch K. 2019a. Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics. 213:953–966.

Castellano D, Uricchio LH, Munch K, Enard D. 2019b. Viruses rule over adaptation in conserved human proteins. bioRxiv. p. 555060.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res (Camb). 63:213–227.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303.
Charlesworth J, Eyre-Walker A. 2008. The McDonald–Kreitman test and slightly deleterious mutations. Mol Biol Evol. 25:1007–1015.
Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13:e1002112.
Csilléry K, François O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 3: 475–479.
Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, *et al.* 2022. Ensembl 2022. Nucleic Acids Res. 50:D988–D995.
Cvijović I, Good BH, Desai MM. 2018. The effect of strong purifying selection on genetic diversity. Genetics. 209:1235–1278.
DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. S weep F inder 2: increased sensitivity, robustness and flexibility. Bioinformatics. 32:1895–1897.
Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L. 2016. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am J Hum Genet. 98:5–21.
Di C, Moreno JM, Salazar-Tortosa DF, Lauterbur ME, Enard D. 2021. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. eLife. 10:e69026.
Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. eLife. 5:e12469.
Enard D, Petrov DA. 2018. Evidence that RNA viruses drove adaptive introgression between neanderthals and modern humans. Cell. 175:360–371.e13.
Enard D, Petrov DA. 2020. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. Philos Trans R Soc B Biol Sci. 375:20190575.
Evans SN, Shvets Y, Slatkin M. 2007. Non-equilibrium theory of the allele frequency spectrum. Theor Popul Biol. 71:109–119.
Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics. 162:2017–2024.
Eyre-Walker A. 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci USA. 107:1752–1756.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26:2097–2108.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 173:891–900.
Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. Trends Genet. 27:343–349.
Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics. 158:1227–1234.
Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet. 12:e1005774.
Galtier N, Rousselle M. 2020. How much does ne vary among species? Genetics. 216:559–572.
Garcia JA, Lohmueller KE. 2021. Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. PLoS Genet. 17:e1009676.
Gillespie JH. 1994. The Causes of Molecular Evolution. Oxford (England): Oxford University Press.
Good BH, Walczak AM, Neher RA, Desai MM. 2014. Genetic diversity in the interference selection limit. PLoS Genet. 10:e1004222.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.
Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution. 62:255–265.
Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, *et al.* 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 363:eaau1043.
Haller BC, Messer PW. 2017. asymptoticMK: a web-based tool for the asymptotic McDonald–Kreitman test. G3 Genes, Genomes, Genetics. 7:1569–1575.
Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol. 36:632–637.
Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol. 24:1792–1800.
Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res (Camb). 8:269–294.
Huang YF. 2021. Dissecting genomic determinants of positive selection with an evolution-guided regression model. Mol Biol Evol. 39. https://doi.org/10.1093/molbev/msab291.
Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. Genetics. 141:1605–1617.
James JE, Piganeau G, Eyre-Walker A. 2016. The rate of adaptive evolution in animal mitochondria. Mol Ecol. 25:67–78.
Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, Charlesworth B. 2019. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 73:111–114.
Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, Keightley PD, Lynch M, McVean G, Payseur BA, *et al.* 2022. Recommendations for improving statistical inference in population genomics. PLoS Biol. 20:e3001669.
Johri P, Charlesworth B, Jensen JD. 2020. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics. 215:173–192.
Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. 2022. On the prospect of achieving accurate joint estimation of selection with population history. Genome Biol Evol. 14:evac088.
Jouganous J, Long W, Ragsdale AP, Gravel S. 2017. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. Genetics. 206:1549–1567.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.
Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. Mol Biol Evol. 35:1366–1371.
Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, *et al.* 2020. HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Mol Biol Evol. 37:295–299.
Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. Trends Ecol Evol. 29:33–41.
Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H. 2015. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol. 7:1141–1154.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351:652–654.
McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. Genome Biol. 17:122.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:1–16.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. PNAS. 110:8615–8620.

Moutinho AF, Eyre-Walker A, Dutheil JY. 2022. Strong evidence for the adaptive walk model of gene evolution in Drosophila and Arabidopsis. PLoS Biol. 20:e3001775.

Murga-Moreno J, Coronado-Zamora M, Casillas S, Barbadilla A. 2022. impMKT: the imputed McDonald and Kreitman test, a straightforward correction that significantly increases the evidence of positive selection of the McDonald and Kreitman test at the gene level. G3 Genes, Genomes, Genetics. 12:jkac206.

Murphy DA, Elyashiv E, Amster G, Sella G. 2022. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. eLife. 12:e76065.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11: 715–724.

Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT, *et al.* 2023. The UCSC genome browser database: 2023 update. Nucleic Acids Res. 51:D1188–D1195.

Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ, Hebert S, *et al.* 2016. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell. 167:657–669.e21.

Nicolaisen LE, Desai MM. 2013. Distortions in genealogies due to purifying selection and recombination. Genetics. 195:221–230.

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection*. Genet Res (Camb). 67: 159–174.

Racimo F, Marnetto D, Huerta-Sánchez E. 2017. Signatures of archaic adaptive introgression in present-day human populations. Mol Biol Evol. 34:296–317.

Racimo F, Schraiber JG. 2014. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. PLoS Genet. 10:e1004697.

Ragsdale AP. 2022. Local fitness and epistatic effects lead to distinct patterns of linkage disequilibrium in protein-coding genes. Technical Report. bioRxiv. Section: New Results Type: article

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol Biol Evol. 35: 2582–2584.

Rousselle M, Simion P, Tilak MK, Figuet E, Nabholz B, Galtier N. 2020. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. PLoS Genet. 16:e1008668.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics. 132:1161–1176.

Schrider DR, Shanku AG, Kern AD. 2016. Effects of linked selective sweeps on demographic inference and model selection. Genetics. 204:1207–1223.

Sendrowski J, Bataillon T. 2024. fastDFE: fast and flexible inference of the distribution of fitness effects. bioRxiv. Pages: 2023.12.04.569837 Section: New Results

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics. 207:1103–1119.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, *et al.* 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 337:64–69.

The Gene Ontology Consortium. 2021. The gene ontology resource: enriching a GOld mine. Nucleic Acids Res. 49:D325–D334.

THE GTEX CONSORTIUM. 2020. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science. 369: 1318–1330.

Thornton KR. 2009. Automating approximate Bayesian computation by local linear regression. BMC Genet. 10:35.

Torres R, Stetter MG, Hernandez RD, Ross-Ibarra J. 2020. The temporal dynamics of background selection in nonequilibrium populations. Genetics. 214:1019–1030.

Torres R, Szpiech ZA, Hernandez RD. 2018. Human demographic history has amplified the effects of background selection across the genome. PLoS Genet. 14:e1007387.

Uricchio LH, Petrov DA, Enard D. 2019. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol. 3: 977–984.

Wang M, Kong L. 2019. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics. 20:28.

Editor: J. Comeron