Inferring parameters of the distribution of fitness effects of new mutations when beneficial mutations are strongly advantageous and rare

Tom R. Booker[1,2]

1. Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada.

2. Biodiversity Research Centre, University of British Columbia, Vancouver, Canada

Correspondence: booker@zoology.ubc.ca

Abstract

Characterising the distribution of fitness effects (DFE) for new mutations is central in evolutionary genetics. Analysis of molecular data under the McDonald-Kreitman test has suggested that adaptive substitutions make a substantial contribution to between-species divergence. Methods have been proposed to estimate the parameters of the distribution of fitness effects for positively selected mutations from the unfolded site frequency spectrum (uSFS). However, when beneficial mutations are strongly selected and rare, they may make little contribution to standing variation and will thus be difficult to detect from the uSFS. In this study, I analyse uSFS data from simulated populations subject to advantageous mutations with effects on fitness ranging from mildly to strongly beneficial. When advantageous mutations are strongly selected and rare, there are very few segregating in populations at any one time. Fitting the uSFS in such cases leads to erroneous parameter estimates which may lead researchers to false conclusions regarding the relative contribution adaptive mutations make to molecular evolution. Fortunately, the parameters for the distribution of fitness effects for harmful mutations are typically estimated with high accuracy and precision. The results from this study suggest that the parameters of positively selected mutations obtained by analysis of the uSFS should be treated with caution and suggest that variability at linked sites should be used in conjunction with standing variability to estimate parameters of the distribution of fitness effects in the future.

Introduction

Characterising the distribution of fitness effects for beneficial mutations is central in evolutionary biology. The rate and fitness effects of advantageous mutations may determine important evolutionary processes such as how variation in quantitative traits is maintained (Hill, 2010), the evolution of sex and recombination (Otto, 2009) and the dynamics of evolutionary rescue in changing environments (Orr & Unckless, 2014). However, despite its central role in evolution, relatively little is known about the distribution of fitness effects (DFE) for advantageous mutations in natural populations. The DFE for advantageous mutations can be estimated from data obtained via targeted mutation or from mutation accumulation experiments (e.g. Bank et al., 2014; Böndel et al., 2019), but such efforts may be limited to laboratory systems. Alternatively, estimates of the DFE can be obtained for natural systems using population genetic methods.

When natural selection is effective, beneficial alleles are promoted to eventual fixation while deleterious variants are maintained at low frequencies. Under recurrent mutation, selection and genetic drift interact to shape the distribution of allele frequencies in a population (Wright, 1937). Parameters of the DFE for both advantageous and deleterious mutations can be estimated by modelling population genomic data, specifically the site frequency spectrum (SFS). The SFS is the distribution of allele frequencies present in a sample of individuals drawn from a population. By contrasting the SFS for a class of sites expected to be subject to selection with that of a linked neutral comparator, one can estimate the parameters of the DFE if selected mutations are segregating in the population of interest (reviewed in Eyre-Walker & Keightley, 2007). Typically, the DFE for nonsynonymous sites in protein coding genes is estimated using synonymous sites as the neutral comparator. Several methods have been proposed that estimate the DFE for deleterious mutations from the SFS under the assumption that beneficial mutations contribute

67 little to standing genetic variation (e.g. Barton & Zeng, 2018; Boyko et al., 2008; Keightley & Eyre-
68 Walker, 2007; Tataru et al.,, 2017).
69
70 The DFE for deleterious mutations can be used when estimating α, the proportion of between-
71 species divergence attributable to adaptive evolution (Eyre-Walker & Keightley, 2009). α can be
72 estimated by rearranging the terms of the McDonald-Kreitman test (MK-test), which assesses the
73 extent of positive selection. Under strong purifying selection, the ratio of divergence at
74 nonsynonymous sites (d_N) to that of synonymous sites (d_S) should be exactly equal to the ratio of
75 nucleotide diversity at nonsynonymous (π_N) and synonymous sites (π_S)(McDonald & Kreitman,
76 1991). Adaptive evolution of protein sequences may contribute to d_N such that $d_N/d_S > \pi_N/\pi_S$.
77 Smith & Eyre-Walker (2002) suggested that the excess d_N due to positive selection (α) can be
78 estimated as
79 $$\alpha = 1 - d_S\pi_N/d_N\pi_S.$$
80 Slightly deleterious alleles may contribute to both standing genetic variation and between-species
81 divergence, estimates of α may therefore be refined by subtracting the contribution that
82 deleterious alleles make to both polymorphism and divergence and this can be calculated using
83 the DFE for harmful mutations (Eyre-Walker & Keightley, 2009). Application of such methods to
84 natural populations suggest that α is of the order of 0.5 in a large variety of animal taxa (Galtier,
85 2016). However, if adaptive evolution is as frequent as MK-test analyses suggest, the assumption
86 that advantageous alleles contribute little to standing variation may be violated and ignoring them
87 could lead to biased estimates of the DFE (Tataru et al., 2017).
88
89 When advantageous alleles contribute to standing variation, parameters of the DFE for both
90 deleterious and beneficial mutations can be estimated from the SFS (Schneider et al., 2011; Tataru
91 et al., 2017). When data from an outgroup species are available, variable sites within a focal

92 species can be polarised as either ancestral or derived and the *unfolded* SFS (uSFS) can be
93 obtained. The uSFS is a vector of length 2*n*, where *n* is the number of haploid genome copies
94 sampled. The ith entry of the uSFS is the count of derived alleles observed at a frequency *i* in the
95 sample. Note that when outgroup data are not available, alleles cannot be polarised and the
96 distribution of minor allele frequencies (known as the *folded* SFS) is analysed. There is limited
97 power to detect positive selection from the uSFS, so the DFE for beneficial mutations is often
98 modelled as a discrete class of mutational effects, with one parameter specifying the fitness
99 effects of beneficial mutations, $\gamma_a = 2N_e s_a$ where N_e is the effective population size and s_a is the
100 positive selection coefficient in homozygotes, and another specifying the proportion of new
101 mutations that are advantageous, p_a. Estimates of γ_a and p_a for nonsynonymous sites have only
102 been obtained a handful of species, and these are summarised in Table 1. The positive selection
103 parameter estimates that have been obtained for mice and *Drosophila* are fairly similar (Table 1).
104 Note that the estimates for humans obtained by Castellano et al, (2019) did not provide a
105 significantly greater fit to the observed data than did a model with no positive selection.
106 Furthermore, Castellano et al, (2019) estimated the parameters for numerous great ape species,
107 the parameters shown for humans are representative of the estimates for all taxa they analysed.
108
109 **Table 1** Estimates of the parameters of positive selection obtained from the uSFS for
110 nonsynonymous sites.

Common name	Scientific name	γ_a	p_a	*Reference*	*Method used*¶
House mouse	*Mus musculus castaneus*	14.5	0.0030	Booker & Keightley, (2018)	*DFE-alpha*
Fruit fly	*Drosophila melanogaster*	23.0	0.0045	Keightley et al, (2016)	*DFE-alpha*
Humans	*Homo sapiens*	0.0064†	0.000025	Castellano et al, (2019)	*polyDFE*

111 ¶ - DFE-alpha implements the analysis methods described by Schneider et al., (2011), *polyDFE* implements the
112 methods described by Tataru et al., (2017)
113 † - Castellano et al., (2019) estimated the mean fitness effect for an exponential distribution of advantageous
114 mutational effects.
115

116 Depending on the rate and fitness effects of beneficial mutations, different aspects of population
117 genomic data may be more or less informative for estimating the parameters of positive selection.
118 As beneficial mutations spread through populations, they may carry linked neutral variants to high
119 frequency, causing selective sweeps (Barton, 2000). On the other hand, if advantageous mutations
120 have mild fitness effects, they may take a long time to reach fixation and make a substantial
121 contribution to standing genetic variation. Campos, et al., (2017) used a model of selective sweeps
122 to analyse the negative correlation observed between d_N and π_S in *Drosophila melanogaster* and
123 estimated $\gamma_a = 250$ and $p_a = 2.2 \times 10^{-4}$. An analysis of the uSFS from the same dataset yielded
124 estimates of $\gamma_a = 23$ and $p_a = 0.0045$ for nonsynonymous sites (Keightley et al., 2016). The sharp
125 contrast between the two studies' estimates of the positive selection parameters may be
126 explained if the DFE for advantageous mutations in *D. melanogaster* is bimodal. If this were so, the
127 different methods (i.e. sweep models versus uSFS analysis) may be capturing distinct aspects of
128 the DFE for advantageous mutations. The handful of studies which have attempted to estimate γ_a
129 and p_a from the uSFS have yielded similar estimates of positive selection (Table 1), which may
130 indicate commonalities in the DFE for beneficial mutations across taxa. On the other hand, it could
131 be that uSFS analyses have only found evidence for mildly beneficial mutations because the
132 approach is only powered to detect weakly beneficial mutations. Indeed, it has been argued that
133 rare strongly selected advantageous mutations, which may contribute little to standing variation,
134 will be undetectable by analysis of the uSFS (Booker & Keightley, 2018; Campos et al., 2017).
135
136 The studies describing the two most recently proposed methods for estimating the DFE for
137 beneficial mutations from the uSFS (Schneider et al., 2011; Tataru et al., 2017) performed
138 extensive simulations, but did not test the case of rare advantageous mutations ($p_a < 0.01$) with
139 strong effects on fitness ($\gamma_a > 24$). Testing this case is important, as studies have indicated that the
140 DFE for advantageous mutations contains strongly beneficial mutations in a variety of taxa (Booker

141 & Keightley, 2018; Campos et al., 2017; Elyashiv et al., 2016; Nam et al., 2017). Note that Tataru et
142 al., (2017) did simulate a population subject to frequent strongly beneficial mutations (γ_a = 800
143 and p_a = 0.02), but the parameter combination they tested may not be biologically relevant as the
144 proportion of adaptive substitutions it yielded was far higher than is typically estimated from real
145 data (α = 0.99). The limited parameter ranges tested in the simulations performed by Schneider et
146 al., (2011) and Tataru et al., (2017) leave a critical gap in our knowledge as to how uSFS based
147 methods perform when advantageous mutations are strongly selected and infrequent.
148
149 In this study, I use simulated datasets to fill this gap and examine how uSFS-based analyses
150 perform when beneficial mutations are strongly selected and rare. I simulate populations subject
151 to a range of positive selection parameters, including cases similar to those modelled by Tataru et
152 al., (2017) and cases where beneficial mutations are strongly selected but infrequent. It has been
153 pointed out that estimating selection parameters by modelling within species polymorphism along
154 with between-species divergence makes the assumption that the DFE has remained invariant since
155 the ingroup and outgroup began to diverge (Tataru et al., 2017). By analysing only the
156 polymorphism data, one can potentially avoid that potentially problematic assumption. Using the
157 state-of-the-art package *polyDFE* v2.0 (Tataru & Bataillon, 2019), I analyse the uSFS data and
158 estimate selection parameters for all simulated datasets with or without divergence. The results
159 from this study suggest that, when beneficial mutations are strongly selected and rare, analysis of
160 the uSFS results in spurious parameter estimates and the proportion of adaptive substitutions may
161 be poorly estimated.

162 Methods

163 Population genomic simulations

164 I tested the hypothesis that the parameters of infrequent, strongly beneficial mutations are
165 difficult to estimate by analysis of the uSFS using simulated datasets. Wright-Fisher populations of
166 N_e = 2,500 diploid individuals were simulated using the forward-in-time package *SLiM* (v3.2; Haller
167 & Messer, 2019). Simulated chromosomes consisted of seven gene models, each separated by
168 8,100bp of neutrally evolving sequence. The gene models consisted of five 300bp exons separated
169 by 100bp neutrally evolving introns. The gene models were based on those used by Campos &
170 Charlesworth, (2019), but unlike that study, I did not model the untranslated regions of the genes.
171 Nonsynonymous sites were modelled by drawing the fitness effects for 2/3rds of mutations in
172 exons from a distribution of fitness effects (DFE), while the remaining 1/3 were strictly neutral and
173 used to model synonymous sites. The fitness effects of nonsynonymous mutations were beneficial
174 with probability p_a or deleterious with probability $1 - p_a$. Beneficial mutations had a fixed selection
175 coefficient of $\gamma_a = 2N_e s_a$. The fitness effects of deleterious mutations were drawn from a gamma
176 distribution with a mean of $\gamma_d = 2N_e s_d$ = -2,000 and a shape parameter of β = 0.3 (s_d being the
177 negative selection coefficient in homozygotes). The gamma distribution of deleterious mutational
178 effects was used for all simulated datasets and was based on results for nonsynonymous sites in
179 *Drosophila melanogaster* (Loewe & Charlesworth, 2006). Uniform rates of mutation (μ) and
180 recombination (r) were set to 1×10^{-6} (giving $4N_e r = 4N_e \mu = 0.01$). Note that μ and r are far higher
181 than is biologically realistic for most eukaryotes, I scaled up these rates to model a population with
182 a large N_e using simulations of 2,500 individuals. Across simulations I varied the γ_a and p_a
183 parameters and performed 2,000 replicates for each combination of parameters. Thus, I simulated
184 a dataset of 21Mbp of coding sequence for each combination of γ_a and p_a tested.
185
186 To model the accumulation of nucleotide substitutions after the split of a focal population with an
187 outgroup, I recorded all substitutions that occurred in the simulations. Campos & Charlesworth,
188 (2019) analysed simulations very similar to those that I performed in this study and showed that

189 populations subject to beneficial mutations with γ_a = 250 and p_a = 0.0002 took 14N_e generations
190 to reach mutation-selection-drift equilibrium. In this study I modelled a range of positive selection
191 parameters, so to ensure that my simulations reached equilibrium I performed 85,000 (34N_e)
192 generations of burn-in before substitutions were scored. The expected number of neutral
193 nucleotide substitutions that accumulate per site in T generations is $d_{Neutral} = T\mu$. The point
194 mutation rate in my simulations was set to $\mu = 1 \times 10^{-6}$ per site per generation, so I ran the
195 simulations for 50,000 generations beyond the end of the burn-in phase to model a neutral
196 divergence of $d_{Neutral}$ = 0.05. All variants present in the population sampled at a frequency of 1.0
197 were also scored as substitutions.
198
199 Using the simulated datasets, I constructed 100 bootstraps by sampling the set of 2,000 replicates
200 with replacement. From each bootstrap sample, I collated variants and constructed the uSFS for
201 synonymous and nonsynonymous sites for 20 diploid individuals.
202

203 Analysis of simulation data

204 I calculated several summary statistics from the simulated datasets. Firstly, I calculated pairwise
205 nucleotide diversity at synonymous sites (π_s) and expressed it relative to the neutral expectation
206 of $\pi_0 = 4N_e\mu$ = 0.01. Secondly, divergence at nonsynonymous sites for both advantageous (dN_a)
207 and deleterious mutations (dN_d) was used to calculate the observed proportion of adaptive
208 substitutions, $\alpha_{Obs} = dN_a/(dN_a + dN_d)$. Finally, I recorded the total number of beneficial mutations
209 segregating in simulated populations, S_{Adv}.
210
211 I estimated DFEs from simulated data by analysis of the uSFS using *polyDFE* (v2.0; Tataru &
212 Bataillon, 2019). *polyDFE* fits an expression for the uSFS expected under a full DFE to data from

213 putatively neutral and selected classes of sites and estimates parameters by maximum likelihood.
214 For each set of positive selection parameters, simulated uSFS data were analysed under "Model B"
215 in *polyDFE* (a gamma distribution of deleterious mutational effects plus a discrete class of
216 advantageous mutations). Initial parameters for the maximisation were calculated from the data
217 using the '-e' option and the uSFS was analysed either with or without divergence using the "-w"
218 option in *polyDFE*. Analysing the uSFS without divergence causes the selection parameters to be
219 inferred from polymorphism data alone. For each replicate, I tested whether the inclusion of
220 beneficial mutations in the DFE improved model fit using likelihood ratio tests between the best-
221 fitting model and a model with p_a set to 0.0. Setting p_a = 0.0 means that positive selection does
222 not influence the likelihood, so two fewer parameters are being estimated. Twice the difference in
223 log-likelihood between the full DFE model and the model with p_a = 0.0 was tested against a χ^2
224 distribution with 2 degrees of freedom. Likelihood surfaces were estimated by running *polyDFE*
225 using fixed values for DFE parameters.
226

227 ### Data Availability

228 All code and *SLiM* configuration files needed to reproduce the results shown in this study are
229 available at https://github.com/TBooker/PositiveSelection_uSFS.
230

231 ## Results

232 ### Population genomic simulations

233 I performed simulations that modelled genes subject to mutation-selection-drift balance with
234 fitness effects drawn from a distribution that incorporated both deleterious and advantageous

235 mutations. The DFE for harmful mutations was constant, but I varied the fraction (p_a) and fitness
236 effects (γ_a) of beneficial mutations across simulated datasets (Table 2). For each set of
237 advantageous mutation parameters, 21Mbp of coding sequences was simulated, of which 14Mbp
238 were nonsynonymous and 7Mbp were synonymous sites. Variants present in the simulated
239 populations were used to construct the uSFS for a sample of 20 diploid individuals, a sample size
240 which is fairly typical of current population genomic datasets (e.g. Castellano et al., 2019; Laenen
241 et al., 2018; Williamson et al., 2014).
242

243 **Table 2** Parameters of positive selection assumed in simulations and the proportion of *polyDFE*
244 runs for which modelling positive selection gave a significantly better fit to the data.

			Proportion Significant	
γ_a	p_a	$\gamma_a p_a$	***With Divergence***	***Without Divergence***
10		0.001	0.66	0.09
50		0.005	1	0.13
100	0.0001	0.01	1	0.83
500		0.05	1	0.58
1,000		0.1	1	0.97
10		0.01	1	0.27
50		0.05	1	1
100	0.001	0.1	1	1
500		0.5	1	1
1,000		1	1	1
10		0.1	1	1
50		0.5	1	1
100	0.01	1	1	1
500		5	1	1
1,000		10	1	1

245

246 Across simulations, the strength of selection acting on advantageous mutations ranged from γ_a =
247 10 to γ_a = 1,000. For a given p_a parameter, increasing the strength of selection increased the
248 observed proportion of adaptive substitutions, α_{Obs} (Figure 1A). This is expected and is due to the
249 monotonic increasing relationship between fixation probability and the strength of positive
250 selection first described by Haldane (1927). Additionally, parameter combinations for which $\gamma_a p_a$

251 were equal had similar proportions of adaptive substitutions, for example compare γ_a = 10 and p_a
252 = 0.01 to γ_a = 1,000 and p_a = 0.0001 (Figure 1A). This was also expected because the rate of
253 adaptive substitutions is proportional to $\gamma_a p_a$. In some datasets, particularly when p_a = 0.01 and
254 advantageous mutations were very strongly selected (i.e. $\gamma_a \geq 500$), α_{Obs} exceeded 0.75, which is
255 higher than is typically estimated from empirical data (Galtier, 2016), so these parameter
256 combinations may not be biologically relevant.
257
258 The effects of selection at linked sites varied across simulated datasets. The DFE for deleterious
259 mutations was kept constant across simulations, so the extent of background selection should be
260 fairly similar across all parameter sets and thus variation in π_S/π_0 reflects the effects of selective
261 sweeps. I found that selection at linked sites reduced synonymous site diversity below the
262 expected value of 0.01 in all simulations (Figure 1B). Increasing the fitness effects or frequency of
263 advantageous mutations had a strong effect on genetic diversity at synonymous sites, as shown by
264 π_S/π_0 in Figure 1B. Unlike α_{Obs}, however, π_S/π_0 responded differently to parameter combinations
265 that had the same $\gamma_a p_a$ (Figure 1). Finally, differences in p_a explained most of the variation in the
266 number of segregating advantageous mutations ($S_{Adv.}$) across simulated datasets, but $S_{Adv.}$ also
267 increased with the strength of positive selection (Figure 1C). On the basis of these results, it is
268 clear that there will be lower power to estimate positive selection on the basis of standing
269 variation when advantageous mutations are rare (i.e. p_a = 0.0001) than when they are
270 comparatively frequent (i.e. p_a = 0.01).
271

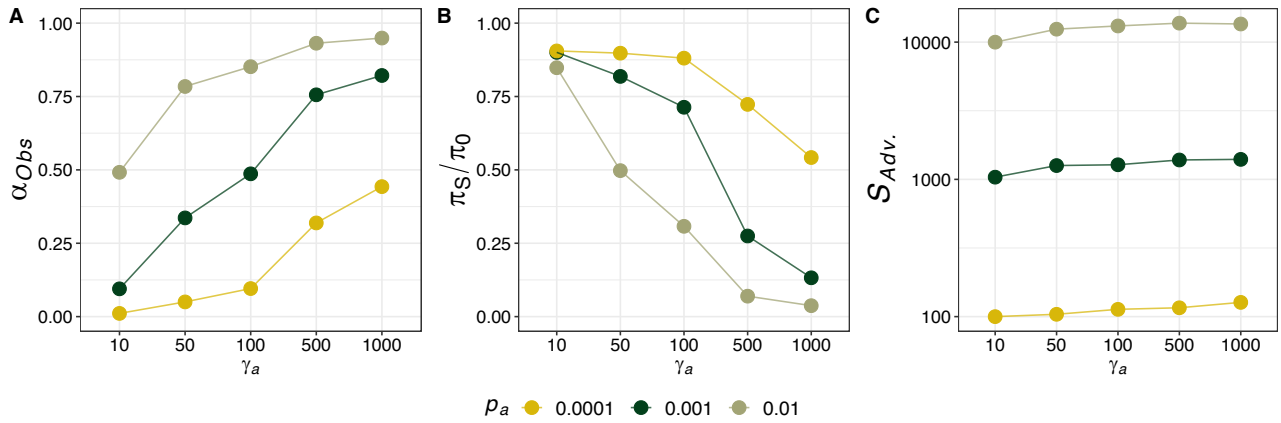


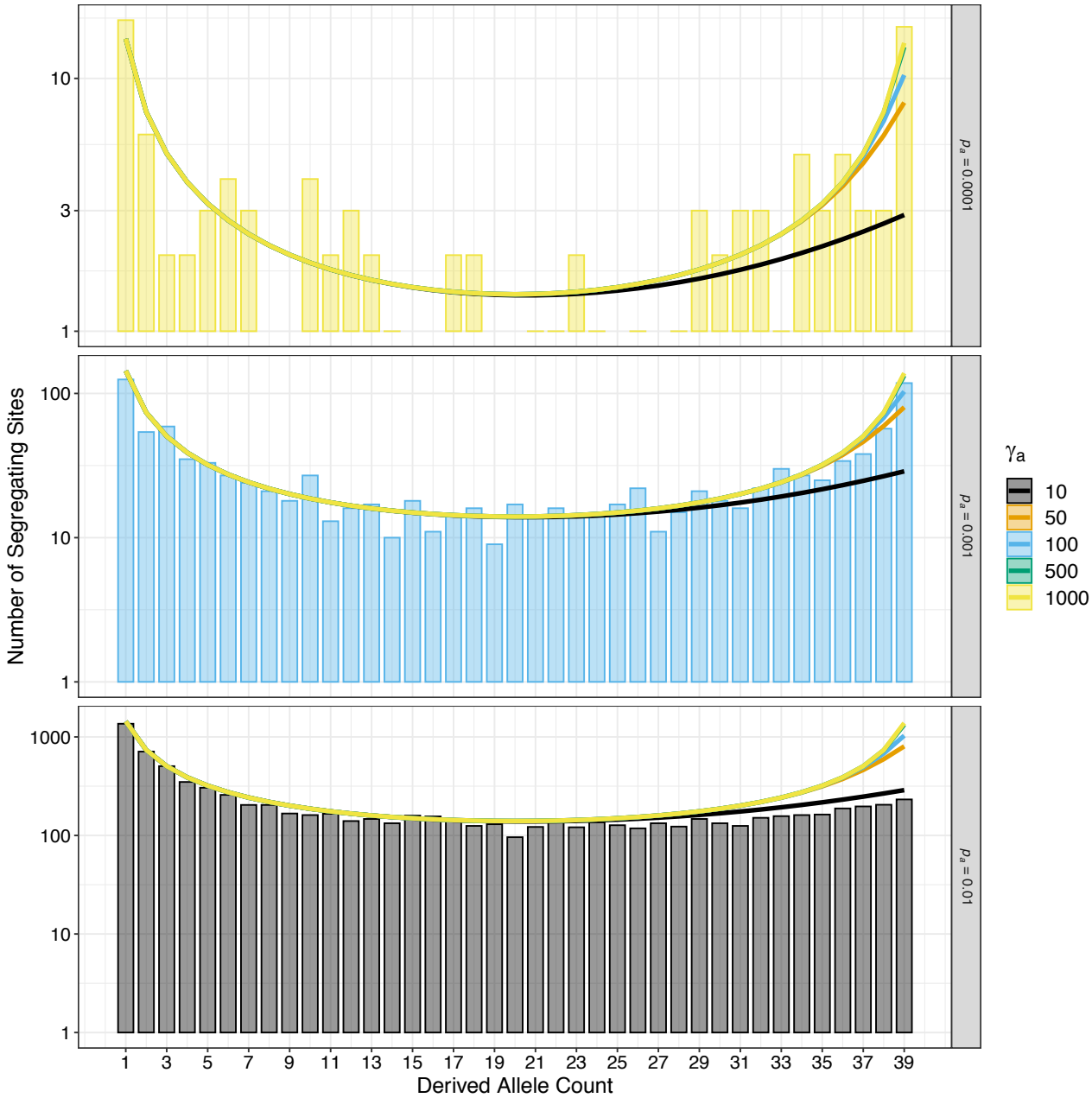
Figure 1 Population genetic summary statistics collated across all simulated genes. α_{Obs} is the observed proportion of substitutions fixed by positive selection. π_s/π_0 is genetic diversity relative to neutral expectation, in the absence of selection, this takes an expected value of 1.0. $S_{Adv.}$ is the number of segregating nonsynonymous sites in the simulated dataset.

Analysis of the unfolded site frequency spectrum

Figure 2 shows the observed (bars) and expected (lines) distribution of derived allele frequencies for beneficial mutations segregating in simulated populations. The three panels of Figure 2 correspond to three parameter combinations (γ_a = 1,000 and p_a = 0.0001, γ_a = 100 and p_a = 0.001 and γ_a = 10 and p_a = 0.01). The three parameter combinations all resulted in similar α_{Obs}, but different π_s/π_0 and $S_{Adv.}$ (Figure 1). The lines in each of the panels of Figure 2 show the analytical expectation for the uSFS of advantageous mutations calculated using Equation 2 from Tataru et al., (2017). The analytical expectation closely matches the observed data for all three parameter combinations (Figure 2). However, for a given value of p_a, the analytical expectation for models with increasing fitness effects were very similar, which likely makes it difficult to distinguish advantageous mutation parameters on the basis of polymorphism alone (Figure 2). For the three parameter sets shown in Figure 2, the overall contribution that advantageous alleles make to the uSFS for nonsynonymous sites is small relative to deleterious ones (Figure S1). Accurate estimation of positive selection parameters from the uSFS requires that the distribution of advantageous

294 alleles can be distinguished from deleterious variants, so when p_a is small it seems likely that uSFS

295 analyses will be unable to easily distinguish competing models.



296

297 **Figure 2** The uSFS for advantageous mutations under different combinations of positive selection
298 parameters. The three bar charts show the observed uSFS under positive selection parameters
299 that yield similar α. The lines in each panel show the expected frequency spectra for different
300 strengths of beneficial mutations and were obtained using Equation 2 from Tataru et al., (2017).
301

302 When analysing a particular uSFS dataset in *polyDFE*, I either modelled the full DFE (i.e. a gamma

303 distribution of deleterious mutations and a discrete class of advantageous mutational effects), or

304 just a gamma DFE for harmful mutations (dDFE). I compared the fit of the two models using

likelihood ratio tests, which tested the null hypothesis that the fit of a full DFE model is similar to that of a model containing only deleterious mutations. For each of the combinations of positive selection parameters shown in Table 2, I ran *polyDFE* on uSFS data from 100 bootstrap replicates. When modelling the full uSFS (i.e. with divergence), *polyDFE* identified models containing positive selection consistently for all but one (p_a = 0.0001 and γ_a = 10) of the parameter combinations tested (Table 2). When the DFE was inferred from polymorphism data alone (i.e. without divergence), models containing positive selection were identified less often, particularly when beneficial mutations were rare (p_a = 0.0001; Table 2).

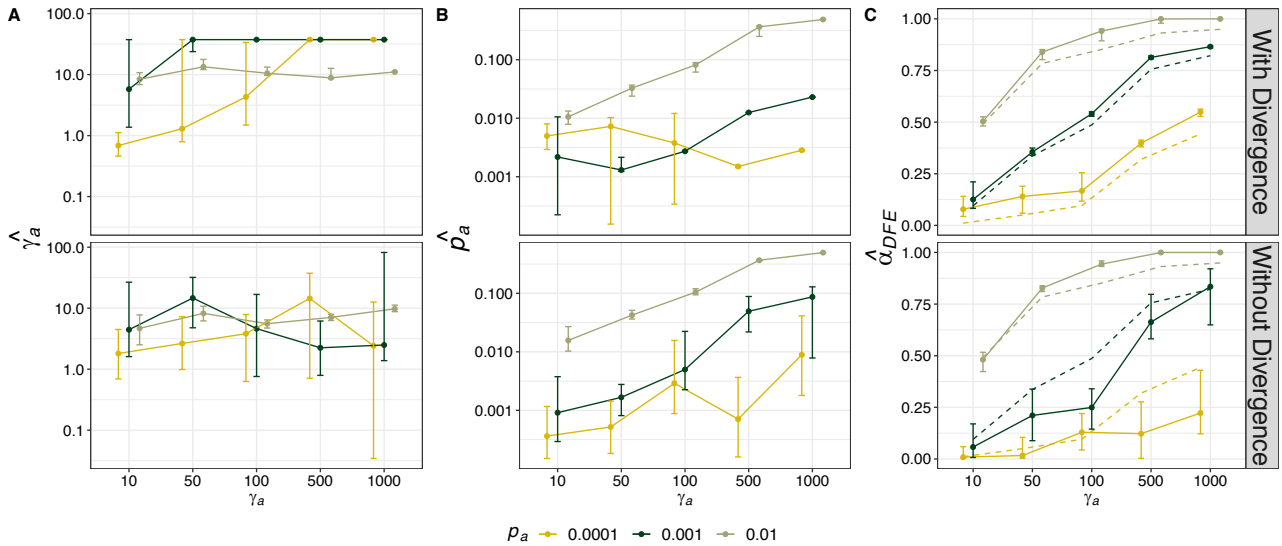


Figure 3 Estimates of the parameters of advantageous mutations and the proportion of adaptive substitutions they imply from simulated datasets. A) γ_a is the inferred selective effect of a new advantageous mutation; B) p_a is the proportion of new mutations that are beneficial; C) α_{DFE} is the proportion of adaptive substitutions expected under the inferred DFE, the dashed lines indicate α_{Obs}, the proportion of adaptive substitutions observed in the simulated datasets.

Figures 3A and 3B show the parameters of positive selection estimated by analysis of uSFS from simulated datasets. I found that when simulated beneficial mutations were mildly advantageous (γ_a = 10) but relatively frequent (p_a = 0.01), both γ_a and p_a were estimated accurately regardless of whether divergence was modelled or not (Figures 3A-B). This finding is consistent with both Schneider et al., (2011) and Tataru et al., (2017). When p_a = 0.01 and γ_a > 10, the analysis of the

326 uSFS with or without divergence yielded very similar parameter estimates, but in both cases, the
327 strength of positive selection seemed to be positively correlated with the estimated p_a (Figure 3).
328 In all cases, when beneficial mutations had $\gamma_a \geq 50$, neither γ_a nor p_a were accurately estimated
329 (Figure 3).
330
331 Tataru et al., (2017) pointed out that, if one had an estimate of the full DFE (i.e. with divergence),
332 the proportion of adaptive substitutions could be obtained by taking the ratio of the fixation
333 probability for a new beneficial mutation over the fixation probability for a random mutation
334 integrating over the full DFE (Equation 10; Tataru et al., 2017). The proportion of adaptive
335 substitutions obtained in this way is denoted α_{DFE}. When modelling the full uSFS, α_{DFE} was
336 estimated with high accuracy, but with a slight upward bias (Figure 3C). When the DFE was
337 inferred without divergence α_{DFE} was underestimated when beneficial mutations were strongly
338 selected and rare (Figure 3).
339
340 In the presence of infrequent, strongly beneficial mutations the parameters of the DFE for
341 deleterious mutations estimated by *polyDFE* were very accurate (Figure S1). Estimates of the DFE
342 for harmful mutations were less accurate when beneficial mutations occurred with $p_a \geq 0.001$ and
343 $\gamma_a \geq 100$. This is presumably because in such cases recurrent selective sweeps eliminate a large
344 amount of neutral diversity and distort the distribution of standing genetic variation at
345 nonsynonymous sites. However, as stated above, the parameter range where the DFE for harmful
346 mutations was poorly estimated in this study may not be biologically relevant.
347
348
349
350

Model Identifiability

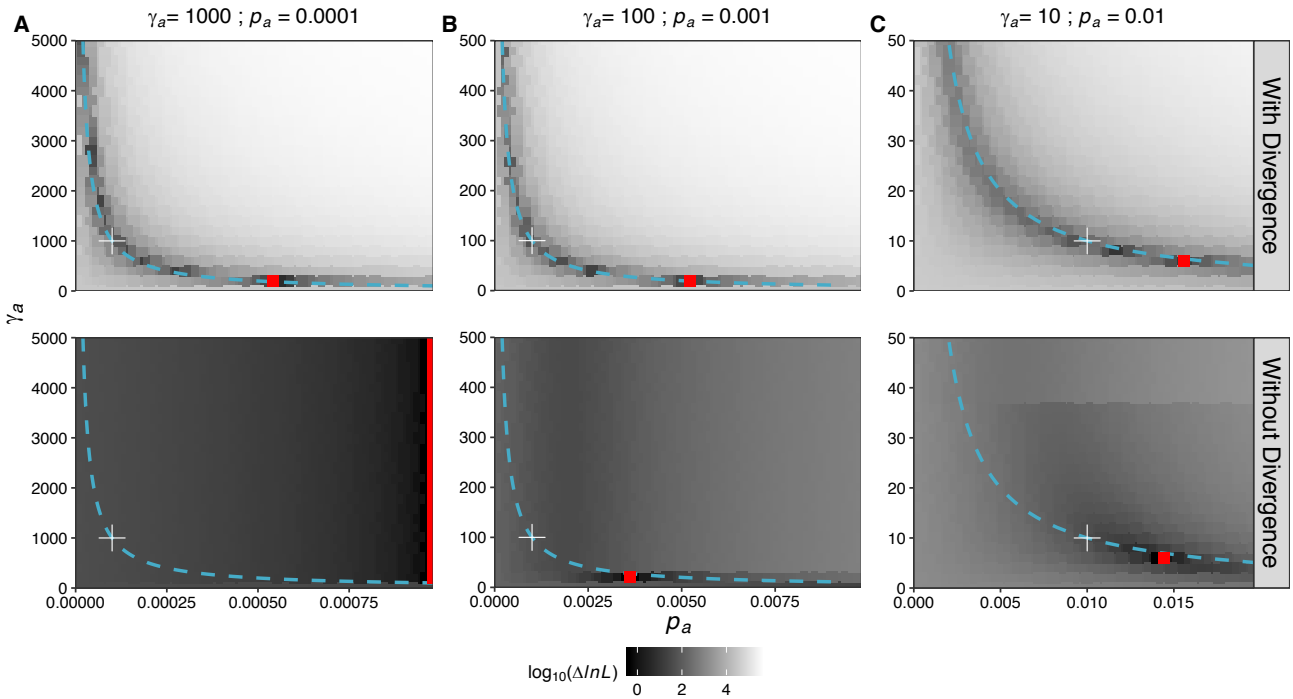


Figure 4 The likelihood surface for the γ_a and p_a parameters for three simulated datasets. Hue indicates differences in log likelihood between a particular parameter combination and the best-fitting model. Best fitting models are indicated by red points and the true parameters are given above the plots and indicated by the white plus signs on the likelihood surface. The relation $\gamma_a p_a$ = 0.1 is shown as a turquoise line and is constant across the three datasets shown.

It is very difficult to tease apart the parameters of positive selection from the uSFS by maximum likelihood. Figure 4 shows the likelihood surface for three sets of positive selection parameters. The three parameter sets all satisfy the condition that $\gamma_a p_a$ = 0.1. The proportion of adaptive substitutions is largely determined by the product $\gamma_a p_a$ (Kimura & Ohta, 1971) and, as expected, the three parameter combinations shown in Figure 4 all exhibit a similar α_{Obs} (Figure 1A). However, the extent by which neutral genetic diversity is reduced and the number of segregating advantageous mutations differ substantially across the three parameter combinations (Figure 1). The top row of panels in Figure 4 shows that when modelling the full uSFS, the likelihood surface closely tracks the relation $\gamma_a p_a$ = 0.1. Focussing on the top panel in Figure 4A, the maximum likelihood estimates (MLEs) of the positive selection parameters (the red dot) are far from the true

370 parameter values (indicated by the plus sign), but the MLEs obtained satisfy $\gamma_a p_a$ = 0.1. The ridge
371 in the likelihood surface was described in both Schneider et al., (2011) and Tataru et al., (2017). It
372 comes about because between-species divergence carries information about α, and α is
373 proportional to $\gamma_a p_a$.
374
375 Inferring the parameters of the DFE from polymorphism alone avoids the assumption of an
376 invariant DFE, but when doing so it may be difficult to distinguish competing models. Indeed,
377 across the three parameter combinations shown, values close to the truth were only obtained
378 from simulated data when γ_a = 10 and p_a = 0.01 (bottom panel Figure 4C). In the case of γ_a = 1000
379 and p_a = 0.0001, the likelihood surface about the true parameters was very flat (Figure 4A).
380 Increasing the p_a parameter increased likelihood for all strengths of selection, so that the MLEs
381 shown in Figure 4A are simply the values with the highest p_a in the range tested (the vertical red
382 line in Figure 4A). When γ_a = 100 and p_a = 0.001, the likelihood surface about the estimates was
383 steep, but the selection parameters identified by maximum likelihood were incorrect (Figure 4B).
384

385 # Discussion

386 In this study, I analysed simulated datasets modelling a range of positive selection parameter
387 combinations. I found that estimates of positive selection parameters obtained by analysis of the
388 uSFS were only accurate when beneficial mutations had $\gamma_a \leq 50$, under stronger selection the
389 individual parameters of positive selection were not accurately estimated (Figure 3). This is not
390 particularly surprising and is consistent with verbal arguments made in published studies (Booker
391 & Keightley, 2018; Campos et al., 2017). However, it is troubling that when beneficial mutations
392 are strongly selected and rare, the uSFS often contained significant signal of positive selection but
393 erroneous parameter estimates were obtained. If one were to analyse an empirical dataset and

394 estimate parameters of positive selection of the order γ_a ~ 10 and p_a ~ 0.01, it would be difficult to
395 know whether those were reflective of the true underlying parameters or an artefact of strong
396 selection. Furthermore, α is accurately estimated when modelling the full uSFS with divergence,
397 but doing so assumes that the DFE has remained invariant over time and that assumption may be
398 problematic. Estimates of α obtained by analysing polymorphism data without divergence avoid
399 this assumption, but may underestimate the true values if beneficial mutations are strongly
400 selected and rare.
401
402 On the basis of this study, it seems that researchers should treat parameters of positive selection
403 obtained by analysis of the uSFS with caution. The expected uSFS for advantageous mutations is
404 very similar when DFE models share the same p_a parameter, and in such cases differing models
405 can only be distinguished by the density of high frequency derived variants (Figure 2). Polarization
406 error when estimating the uSFS can generate an excess in the number of high frequency variants
407 (Keightley & Jackson, 2018), so may generate a spurious signal of strong positive selection.
408 Analysis methods have been proposed which attempt to estimate the rate of polarisation error
409 when modelling the uSFS (Barton & Zeng, 2018; Tataru et al., 2017), but further study is required
410 to determine whether such methods reduce the signal of positive selection in uSFS-based
411 analyses. Parameters of the DFE for harmful mutations were robustly estimated across the
412 simulated datasets.
413
414 Estimates of α based on analysis of the uSFS may be biased when beneficial mutations are strongly
415 selected and infrequent. Calculating α using the rearranged MK-test makes the problematic
416 assumption that the DFE has remained invariant in the time since the focal species began to
417 diverge from the outgroup (Tataru et al., 2017). However, Tataru et al., (2017) pointed out that
418 one can avoid that assumption if α_{DFE} is calculated from a DFE estimated without divergence data.

419 In this study, estimates of α_{DFE} obtained when the full uSFS was analysed were very precise, but
420 with a slight upward bias (Figure 3). When simulated beneficial mutations were strongly selected
421 and rare, the parameters inferred using polymorphism data alone (i.e. without divergence) yielded
422 spurious estimates of α_{DFE} (Figure 3). When analysing datasets from real populations, α_{DFE} may not
423 capture the contribution that strongly beneficial mutations make to molecular evolution. This may
424 make it difficult to contrast α_{DFE} between species with large differences in population size, because
425 the number of segregating advantageous mutations and thus ability to accurately estimate
426 selection parameters will depend on N_e.
427
428 The simulations I performed in this study generated the ideal dataset for estimating parameters of
429 selection from the uSFS. I simulated datasets containing 21Mbp of sites in which genotypes and
430 whether sites were selected or not was unambiguously known. When analysing real data this is
431 not the case and researchers often have to filter a large proportion of sites out of their analyses or
432 choose to analyse a subset of genes that have orthology with outgroups or biological properties of
433 interest. In addition, the populations I simulated were randomly mating and had constant sizes
434 over time. The results I present in this study suggest that even with perfect knowledge of a
435 population that adheres to the assumptions of a Wright-Fisher model, it is inherently difficult to
436 infer the parameters of strongly beneficial mutations from the uSFS, particularly so when
437 beneficial mutation occur infrequently.
438

439 Estimating parameters of positive selection from the uSFS versus 440 estimates from patterns of diversity

441 As discussed above, studies based on analysis of the uSFS and those based on selective sweep
442 models have yielded vastly different estimates of the parameters of positive selection. Patterns of

443 neutral genetic diversity in both humans and wild mice cannot be explained by the effects of
444 background selection alone, and in both species it has been suggested that strongly beneficial
445 mutations are required to explain the observed patterns (Booker & Keightley, 2018; Nam et al.,
446 2017). In the case of wild house mice, positive selection parameters obtained by analysis of the
447 uSFS do not explain dips in nucleotide diversity around functional elements (Booker & Keightley,
448 2018). Recently, Castellano et al. (2019) analysed the uSFS for nonsynonymous sites in great ape
449 species but did not find significant evidence for positive selection. In their dataset, Castellano et al.
450 (2019) had 8 haploid genome sequences for each of great ape species they analysed, and they
451 argued that they were underpowered to detect positive selection on the basis of the uSFS.
452 However, the results given in this study suggest that even with larger samples (recall that I
453 analysed 40 haploid genomes from my simulated populations), parameters of strongly beneficial
454 mutations are very difficult to resolve by analysis of the uSFS.
455
456 In this study, I modelled beneficial mutations using a discrete class of selection coefficients when,
457 in reality, there is likely a continuous distribution of fitness effects. Indeed, studies in both humans
458 and *D. melanogaster* have found evidence for a bimodal distribution containing both strongly and
459 weakly beneficial mutations contributing to adaptive evolution using methods which incorporate
460 linkage information but do not explicitly estimate selection parameters (Elyashiv et al., 2016;
461 Uricchio, Petrov, & Enard, 2019). There are currently no methods that estimate the DFE using an
462 analytical expression for the uSFS expected under the combined effects of BGS and sweeps.
463 Rather, nuisance parameters or demographic models are used to correct for the contribution that
464 selection at linked sites may make to the shape of the SFS (Eyre-Walker, Woolfit, & Phelps, 2006;
465 Galtier, 2016; Tataru et al., 2017). However, as this study shows, the parameters of positive
466 selection may not be reliably detected when analysing the uSFS alone. A way forward may be in
467 using approximate Bayesian computation (ABC) or machine-learning approaches to make use of all

468 of the available data, while not necessitating an expression for the uSFS expected under the
469 combined effects of BGS, sweeps, population size change and direct selection. An advance in this
470 direction has recently been made by Uricchio et al., (2019) who developed an ABC method for
471 estimating α which makes use of the distortions to the uSFS generated by BGS and sweeps. By
472 applying their method to data from humans, Uricchio et al., (2019) found that $\alpha = 0.13$ for
473 nonsynonymous sites, 72% of which was generated by mildly beneficial mutations and 28% by
474 strongly beneficial mutations. Unfortunately, their method does not provide exact estimates for
475 positive selection parameters, but indicates that by using computational approaches, progress
476 could be made in our estimation and understanding of the DFE for both advantageous and
477 deleterious mutations.

478 Acknowledgements

479 I wish to extend gratitude to Peter Keightley, Michael Whitlock and Sam Yeaman for valuable
480 advice and mentorship. Thanks to Sam Yeaman and Peter Keightley for comments on the
481 manuscript. Thanks to Paula Tataru and Thomas Bataillon for help with *polyDFE*.

482 References

483 Bank, C., Hietpas, R. T., Wong, A., Bolon, D. N., & Jensen, J. D. (2014). A bayesian MCMC approach
484 to assess the complete distribution of fitness effects of new mutations: uncovering the
485 potential for adaptive walks in challenging environments. *Genetics*, *196*(3), 841–852.
486 https://doi.org/10.1534/genetics.113.156190
487 Barton, H. J., & Zeng, K. (2018). New methods for inferring the distribution of fitness effects for
488 INDELs and SNPs. *Molecular Biology and Evolution*, *35*(6), 1536–1546.
489 https://doi.org/10.1093/molbev/msy054
490 Barton, N. H. (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, *355*(1403), 1553–1562.

https://doi.org/10.1098/rstb.2000.0716

Böndel, K. B., Kraemer, S. A., Samuels, T., McClean, D., Lachapelle, J., Ness, R. W., … Keightley, P. D. (2019). Inferring the distribution of fitness effects of spontaneous mutations in Chlamydomonas reinhardtii. *PLOS Biology*, *17*(6), e3000192. https://doi.org/10.1371/journal.pbio.3000192

Booker, T. R., & Keightley, P. D. (2018). Understanding the Factors That Shape Patterns of Nucleotide Diversity in the House Mouse Genome. *Mol. Biol. Evol.*, *35*(12), 2971–2988.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., … Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, *4*(5), e1000083. https://doi.org/10.1371/journal.pgen.1000083

Campos, J L, Zhao, L., & Charlesworth, B. (2017). Estimating the parameters of background selection and selective sweeps in Drosophila in the presence of gene conversion. *Proc Natl Acad Sci*, *114*(24), E4762–E4771. https://doi.org/10.1073/pnas.1619434114 10.5061/dryad.vs264)

Campos, José Luis, & Charlesworth, B. (2019). The effects on neutral variability of recurrent selective sweeps and background selection. *Genetics*, *212*(1), 287–303. https://doi.org/10.1534/genetics.119.301951

Castellano, D., Macià, M. C., Tataru, P., Bataillon, T., & Munch, K. (2019). Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes. *Genetics*, genetics.302494.2019. https://doi.org/10.1534/genetics.119.302494

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., … Sella, G. (2016). A genomic map of the effects of linked selection in Drosophila. *PLoS Genet*, *12*(8), e1006130. https://doi.org/10.1371/journal.pgen.1006130

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, *26*(9), 2097–2108. https://doi.org/10.1093/molbev/msp119

Eyre-Walker, A., Woolfit, M., & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, *173*(2), 891–900. https://doi.org/10.1534/genetics.106.057570

Galtier, N. (2016). Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS Genet.*, *12*(1), e1005774.

Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *23*(7), 838–844. https://doi.org/10.1017/S0305004100015644

Haller, B. C., & Messer, P. W. (2019). {SLiM} 3: Forward Genetic Simulations Beyond the {Wright--Fisher} Model. *Mol. Biol. Evol.*, *36*(3), 632–637.

Hill, W. G. (2010, January 12). Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 365, pp. 73–85. https://doi.org/10.1098/rstb.2009.0203

Keightley, P D, Campos, J. L., Booker, T. R., & Charlesworth, B. (2016). Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of Drosophila melanogaster. *Genetics*, *203*(2), 975–984. https://doi.org/10.1534/genetics.116.188102

Keightley, P D, & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, *177*(4), 2251–2261. https://doi.org/10.1534/genetics.107.080663

Keightley, Peter D, & Jackson, B. C. (2018). Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. *Genetics*, *209*(3), 897–906.

Kimura, M., & Ohta, T. (1971). *Theoretical aspects of population genetics*. Princeton Univ. Press.

Laenen, B., Tedder, A., Nowak, M. D., Toräng, P., Wunder, J., Wötzel, S., … Slotte, T. (2018). Demography and mating system shape the genome-wide impact of purifying selection in Arabis alpina. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(4), 816–821. https://doi.org/10.1073/pnas.1707492115

Loewe, L., & Charlesworth, B. (2006). Inferring the distribution of mutational effects on fitness in Drosophila. *Biol Lett*, *2*(3), 426–430. https://doi.org/10.1098/rsbl.2006.0481

McDonald, J. M., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*.

Nam, K., Munch, K., Mailund, T., Nater, A., Greminger, M. P., Krützen, M., … Schierup, M. H. (2017). Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(7), 1613–1618. https://doi.org/10.1073/pnas.1605660114

Orr, H. A., & Unckless, R. L. (2014). The Population Genetics of Evolutionary Rescue. *PLoS Genetics*, *10*(8). https://doi.org/10.1371/journal.pgen.1004551

Otto, S. P. (2009). The evolutionary enigma of sex. *American Naturalist*, *174*(SUPPL. 1). https://doi.org/10.1086/599084

Schneider, A., Charlesworth, B., Eyre-Walker, A., & Keightley, P. D. (2011). A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*, *189*(4), 1427–1437. https://doi.org/10.1534/genetics.111.131730

Smith, N. G. C., & Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature*, *415*(6875), 1022–1024. https://doi.org/10.1038/4151022a

Tataru, P., & Bataillon, T. (2019). polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics*, *35*(16), 2868–2869. https://doi.org/10.1093/bioinformatics/bty1060

566 Tataru, P., Mollion, M., Glemin, S., & Bataillon, T. (2017). Inference of distribution of fitness effects
567 and proportion of adaptive substitutions from polymorphism data. *Genetics*, *207*(3), 1103–
568 1119. https://doi.org/10.1534/genetics.117.300323
569 Uricchio, L. H., Petrov, D. A., & Enard, D. (2019). Exploiting selection at linked sites to infer the rate
570 and strength of adaptation. *Nat Ecol Evol*, *3*(6), 977–984.
571 Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., & Wright,
572 S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved
573 noncoding regions of Capsella grandiflora. *PLoS Genetics*, *10*(9), e1004622.
574 https://doi.org/10.1371/journal.pgen.1004622
575 Wright, S. (1937). The Distribution of Gene Frequencies in Populations. *Proceedings of the*
576 *National Academy of Sciences*, *23*(6), 307–320. https://doi.org/10.1073/pnas.23.6.307
577
578
579
580
581
582
583
584
585
586
587
588

Supplementary Material

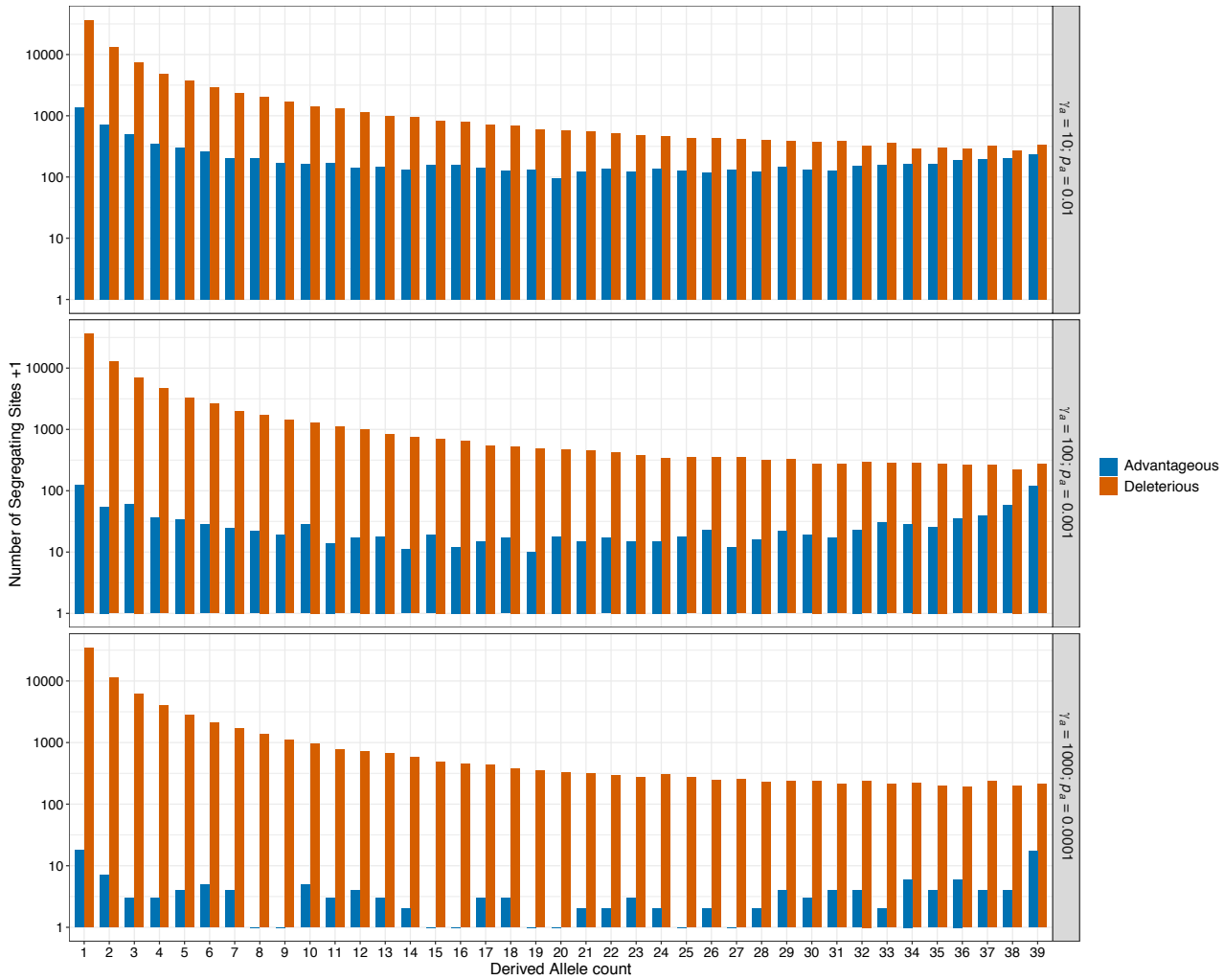


Figure S1 The observed uSFS for nonsynonymous sites for three sets of positive selection parameters. The distribution of deleterious mutations is shown in orange and the distribution of advantageous mutations is shown in blue. For the purposes of visualising the data on a log scale, the number of segregating sites is shown +1.

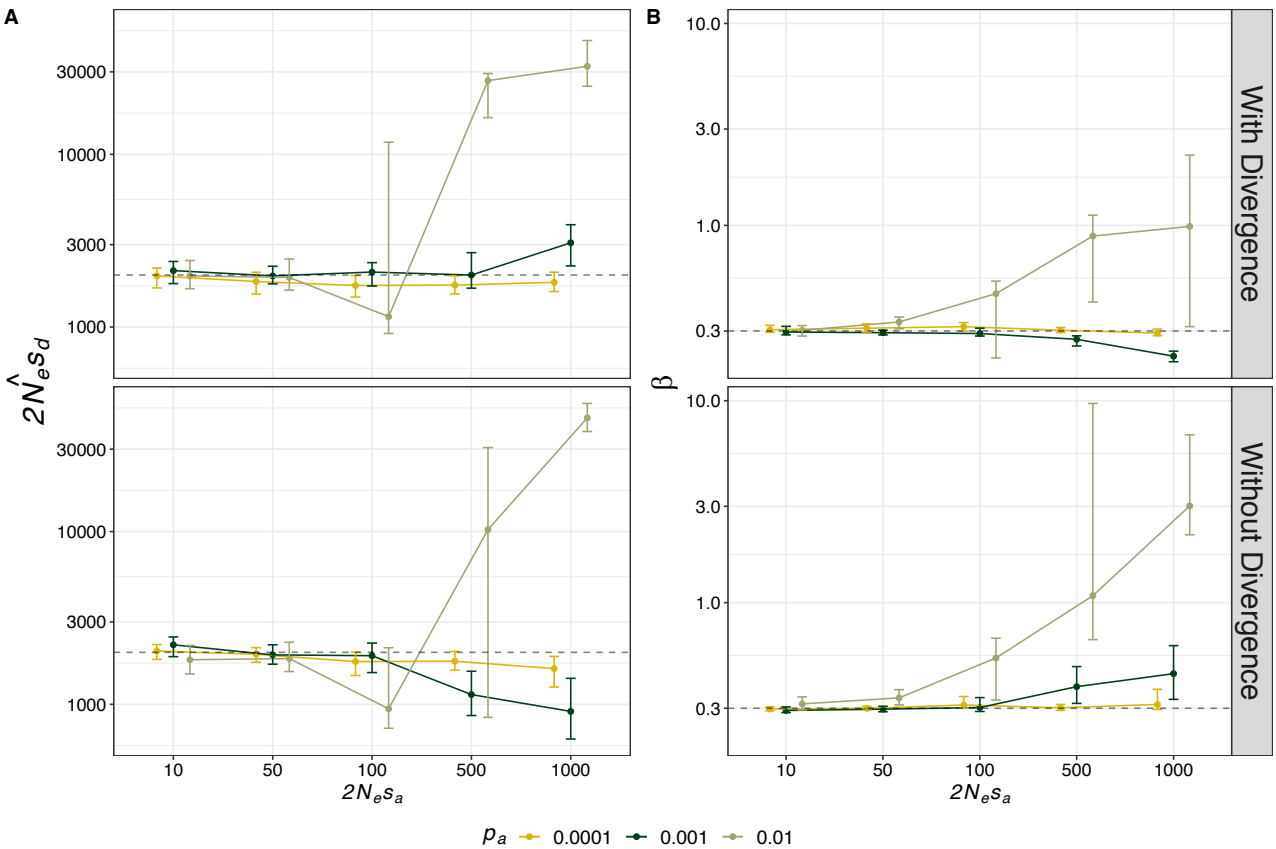


Figure S2 Parameter estimates for the DFE for deleterious mutations obtained from simulated datasets. A) the mean effect of a deleterious mutation and b) the shape parameter of the gamma distribution. Error bars indicate the 95% range of 100 bootstrap replicates.