

Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data

Paula Tataru^{*,1}, Maéva Mollion^{*}, Sylvain Glemin^{†,‡} and Thomas Bataillon^{*}

^{*}Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, Aarhus C, 8000, Denmark, [†]Evolutionary Biology Centre, Uppsala University, Norbyvägen 14-18, Uppsala, 752 36, Sweden, [‡]Institut des Sciences de l'Evolution de Montpellier, UMR 5554 CNRS, Place Eugène Bataillon, 34095 Montpellier cedex 5, France

ABSTRACT The distribution of fitness effects (DFE) encompasses deleterious, neutral and beneficial mutations. It conditions the evolutionary trajectory of populations, as well as the rate of adaptive molecular evolution (α). Inference of DFE and α from patterns of polymorphism (SFS) and divergence data has been a longstanding goal of evolutionary genetics. A widespread assumption shared by numerous methods developed so far to infer DFE and α from such data is that beneficial mutations contribute only negligibly to the polymorphism data. Hence, a DFE comprising only deleterious mutations tends to be estimated from SFS data, and α is only predicted by contrasting the SFS with divergence data from an outgroup. Here, we develop a hierarchical probabilistic framework that extends on previous methods and also can infer DFE and α from polymorphism data alone. We use extensive simulations to examine the performance of our method. We show that both a full DFE, comprising both deleterious and beneficial mutations, and α can be inferred without resorting to divergence data. We demonstrate that inference of DFE from polymorphism data alone can in fact provide more reliable estimates, as it does not rely on strong assumptions about a shared DFE between the outgroup and ingroup species used to obtain the SFS and divergence data. We also show that not accounting for the contribution of beneficial mutations to polymorphism data leads to substantially biased estimates of the DFE and α . We illustrate these points using our newly developed framework, while also comparing to one of the most widely used inference methods available.

KEYWORDS distribution of fitness effects, rate of adaptive molecular evolution, beneficial mutations, polymorphism and divergence data, Poisson random field

1 New mutations are the ultimate source of heritable variation.
2
3 The fitness properties of new mutations determine the possible
4 evolutionary trajectories a population can follow (Bataillon and
5 Bailey 2014). For instance, supply rate and fitness effects of ben-
6 efiticial mutations determine the expected rate of adaptation of a
7 population (Lourenço *et al.* 2011), while deleterious mutations
8 condition the expected drift load of a population (Kimura *et al.*

Copyright © 2016 by the Genetics Society of America
doi: 10.1534/genetics.XXX.XXXXXX

Manuscript compiled: Tuesday 5th July, 2016%

¹Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, Aarhus C 8000, Denmark. paula@cs.au.dk

1963). Even a few beneficial mutations with large effects can quickly move a population towards its fitness optimum, while the fitness can be reduced through the accumulation of multiple deleterious mutations with small effects that occasionally escape selection. Genome-wide rates and effects of new mutations influence, among others, the evolutionary advantage of sex (Otto and Lenormand 2002), the expected degree of parallel evolution (Chevin *et al.* 2010b), the maintenance of variation on quantitative traits (Hill 2010), and the evolutionary potential and capacity of populations to respond to novel environments (Chevin *et al.* 2010a; Hoffmann and Sgrò 2011).

Effects of new mutations on fitness are typically modeled as independent draws from an underlying distribution of fitness effects (hereafter DFE) which, in principle, spans deleterious, neutral and beneficial mutations. Lately, there has also been considerable focus on estimation of the DFE of new non-synonymous mutations, and learn more about factors governing the rate of adaptive molecular evolution, commonly defined as the proportion of fixed adaptive mutations among all non-synonymous substitutions, and often denoted α . Therefore, inferring the DFE, both from experimental (Bataillon and Bailey 2014; Bataillon *et al.* 2011; Jacquier *et al.* 2013; Halligan and Keightley 2009; Sousa *et al.* 2011), but also from polymorphism and divergence data (Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Eyre-Walker and Keightley 2009; Keightley and Eyre-Walker 2012; Galtier 2016), has been a longstanding goal of evolutionary genetics.

The McDonald-Kreitman test (McDonald *et al.* 1991) was one of the first attempts to use DNA data to measure the amount of selection experienced by genes. It compares the amount of variation (counts of nucleotide polymorphism) within a species (ingroup) to the variation between species (measured by divergence counts between sequences from the ingroup and an outgroup). The test parses and contrasts the amount of variation found at the synonymous and non-synonymous sites, where the synonymous sites are assumed to be neutrally evolving sites. Smith and Eyre-Walker (2002) further developed this test to also infer the amount of purifying selection, defined as the propor-

tion of strongly deleterious mutations, and α (see also Welch (2006) for a maximum likelihood approach). Building on the Poisson Random Field (PRF) theory (Sawyer and Hartl 1992; Sethupathy and Hannenhalli 2008) and arising as extensions to the classical McDonald-Kreitman test, a series of methods have been developed to not only characterize the amount of selection, but also the DFE (Bustamante *et al.* 2003; Piganeau and Eyre-Walker 2003; Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Keightley and Eyre-Walker 2010; Gronau *et al.* 2013; Kousathanas and Keightley 2013; Racimo and Schraiber 2014), and then used it as a building block to estimate α (Loewe *et al.* 2006; Eyre-Walker and Keightley 2009; Schneider *et al.* 2011; Keightley and Eyre-Walker 2012; Galtier 2016).

Assuming that sites are independent, that new mutations follow a Poisson process and always occur at new sites, these methods then model the observed variation using a Poisson distribution. The variation within the ingroup is given through counts of the site frequency spectrum (SFS), whose mean in each entry of the SFS is calculated as a function of the DFE and other parameters. Selection is assumed to be weak ($s \ll 1$, but note that $4N_e s$ can still be large) and the DFE to be constant in time and the same in both the ingroup and outgroup.

Additionally, in order to disentangle selection from demography and other forces (Nielsen 2005), and in the spirit of the McDonald-Kreitman test, the sequenced sites are divided into two classes of neutrally evolving and selected sites. The DFE is then inferred by contrasting the SFS counts for the neutral and selected sites, by assuming that such forces equally affect the two classes.

Ideally, a full demographic model should be jointly inferred with the DFE parameters from the data. However, this can be computationally very demanding and instead a simplified demography is often assumed, where a single population size change is allowed (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Kousathanas and Keightley 2013), or a somewhat more complex model is inferred (Boyko *et al.* 2008). Alternatively, the explicit inference of demography can be avoided altogether by introducing a series of nuisance parameters that

85 account for the demography and sampling effects. These pa- 123
86 rameters account for distortions of the polymorphism counts 124
87 relative to neutral expectations in an equilibrium Wright-Fisher 125
88 population (Eyre-Walker *et al.* 2006; Galtier 2016). An added ben- 126
89 efit is that controlling for demography effects (either explicitly 127
90 or through nuisance parameters) can also remove bias caused 128
91 by linkage (Kousathanas and Keightley 2013; Messer and Petrov 129
92 2012). The approach of Eyre-Walker *et al.* (2006) can potentially 130
93 be more robust for estimating a DFE than putting a lot of faith 131
94 in a simplified demographic scenario. 132

95 The proportion of adaptive substitutions, α , is typically ob- 133
96 tained as a ratio between an estimate of the number of adap- 134
97 tive substitutions and the observed selected divergence counts 135
98 (Eyre-Walker and Keightley 2009; Loewe *et al.* 2006; Keightley 136
99 and Eyre-Walker 2012; Galtier 2016). The number of adaptive 137
100 substitutions is calculated by subtracting, from the observed 138
101 divergence counts at selected sites, the expected counts accrued 139
102 by fixation of deleterious and neutral mutations. These expected 140
103 counts are calculated from an inferred DFE of deleterious mu- 141
104 tations (henceforth denoted deleterious DFE). The deleterious 142
105 DFE is inferred from the SFS data under the assumption that 143
106 all SNPs at selected sites are only deleterious. Therefore this 144
107 approach for estimating α heavily relies on the assumption that 145
108 the ingroup and outgroup species share the same DFE - or more 146
109 accurately, the same distribution of scaled selection coefficients 147
110 $S = 4N_e s$. Unfortunately, this assumption of invariance might 148
111 not often be met in practice, because the DFE might change, or 149
112 simply because it is unlikely that both ingroup and outgroup 150
113 evolved with the same population size. 151

114 There has been great focus on developing methods inferring a 152
115 deleterious DFE from polymorphism data alone (Keightley and 153
116 Eyre-Walker 2007; Kousathanas and Keightley 2013; Eyre-Walker 154
117 *et al.* 2006; Racimo and Schraiber 2014). These methods rely on a 155
118 crucial assumption: beneficial mutations contribute negligibly 156
119 to polymorphism (SFS counts) and therefore are not modeled 157
120 for this type of data. The reasoning behind this is that strongly 158
121 selected beneficial mutations will fixate very quickly and that “at 159
122 most an advantageous mutation will contribute twice as much

heterozygosity during its lifetime as a neutral variant” (Smith 123
and Eyre-Walker 2002). This assumption is backed by one study 124
(Keightley and Eyre-Walker 2010) discussed in more details in 125
the *Results and Discussion* section. While some DFE methods do 126
model a full DFE (encompassing both deleterious and beneficial 127
mutations) (Bustamante *et al.* 2003; Piganeau and Eyre-Walker 128
2003; Boyko *et al.* 2008; Schneider *et al.* 2011; Gronau *et al.* 2013; 129
Galtier 2016), the majority of them do not estimate α . 130

Here, we develop a hierarchical probabilistic model that com- 131
bines and extends previous methods, and that can infer both 132
the full DFE and α from polymorphism data alone. We use 133
our method and perform extensive simulations to investigate 134
different aspects of the inference quality. We show that the as- 135
sumption that beneficial mutations make negligible contribution 136
to SFS data is unfounded and that a full DFE can also be inferred 137
reliably from polymorphism data alone. Using the estimated 138
full DFE, we show how α can be inferred without relying on 139
divergence data. Performing inference on polymorphism data 140
alone proves more adequate when assumptions regarding the 141
outgroup evolution (for example, that the scaled DFE is shared 142
between the ingroup and outgroup) are not likely to be met. 143
We also demonstrate that when the contribution of beneficial 144
mutations to SFS data is ignored, both the inferred deleterious 145
DFE and α can be heavily biased. We compare our method 146
and illustrate the resulting bias using the most widely used 147
inference method, dfe-alpha (Keightley and Eyre-Walker 2007; 148
Eyre-Walker and Keightley 2009; Schneider *et al.* 2011; Keight- 149
ley and Eyre-Walker 2012). We also investigate the impact on 150
inference of misidentification of ancestral state. 151

152 Hierarchical model for inference of DFE and α

153 In this study, we build on several of the methods using PRF 154
theory to build a hierarchical model to infer, via maximum like- 155
lihood, the DFE from polymorphism (site frequency spectrum, 156
SFS) and divergence counts. Our hierarchical model is combin- 157
ing and extending different features from different approaches. 158
Figure 1 shows a schematic of the data and the model. We offer 159
below a summary of the assumptions and theory underlying

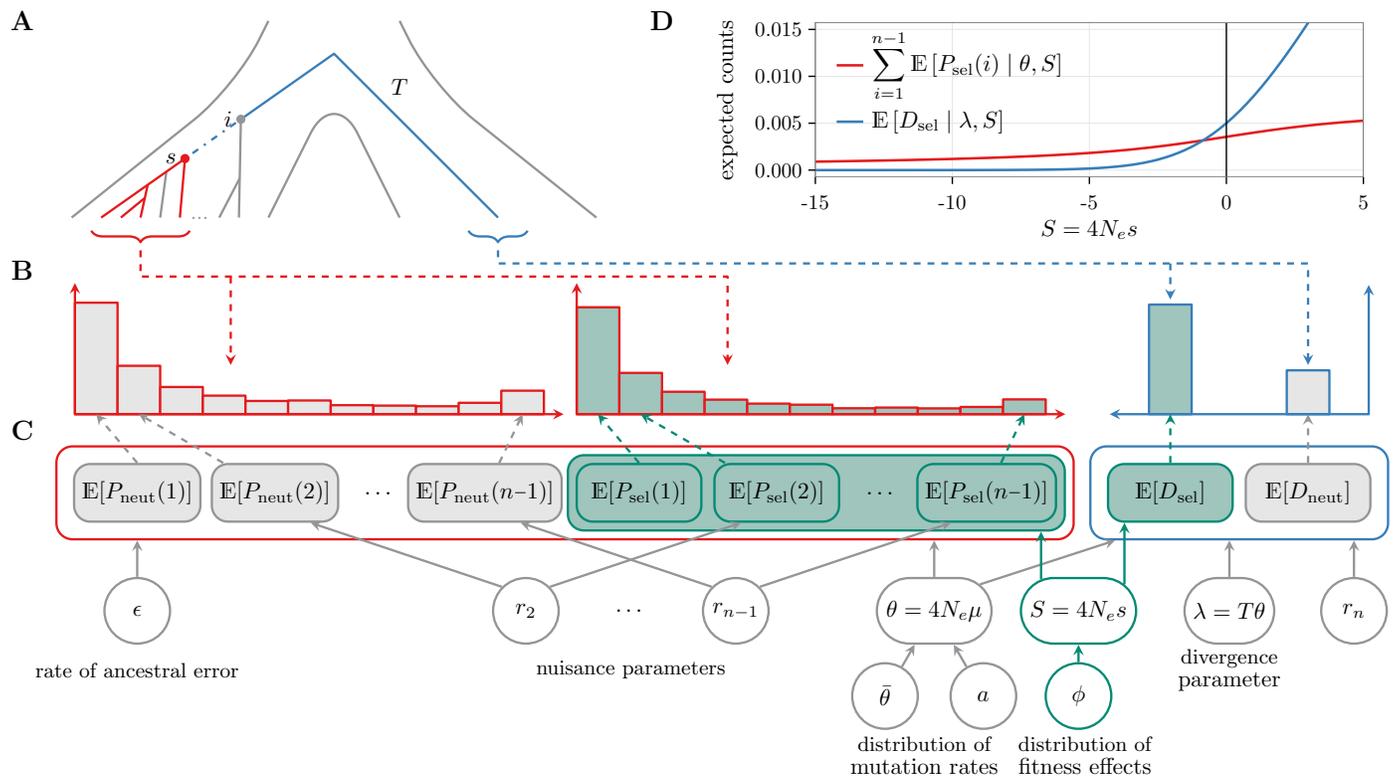


Figure 1 Schematic of data and model. Throughout the figure, gray and green filling indicates sites that are assumed to be evolving neutrally or potentially under selection, respectively, while red and blue outline indicates polymorphism and divergence data (expectations), respectively. (A) The history and coalescent tree of two populations: the ingroup (on the left side), for which polymorphism data is collected, and the outgroup (on the right side), for which divergence counts are obtained. A total of n sequences are sampled from the ingroup (marked in red), with the most recent common ancestor (MRCA) found at s . The MRCA of the whole ingroup population is found at i . From the outgroup we typically have access to one sequence (marked in blue). The total evolutionary time between s and the sampled outgroup sequence can be divided into the time from s to i (blue dot-dash line) and T , the time from i to the sampled outgroup sequence (blue full line). (B) Site frequency spectrum and divergence counts ($p_z(i)$ and d_z , with $z \in \{\text{neut}, \text{sel}\}$ and $1 \leq i < n$). (C) Expected counts ($E[P_z(i)]$ and $E[D_z]$, with $z \in \{\text{neut}, \text{sel}\}$ and $1 \leq i < n$), model parameters and relations between parameters, expectations and data. (D) Expectations as a function of S , for $\theta = 0.001$ and $\lambda = 0.005$.

160 our approach. Further details on the likelihood function, its
 161 implementation and numerical optimization can be found in the
 162 Supplemental Material.

163 **Notations and assumptions**

164 The data is divided into sites that are assumed to be either sites
 165 that evolve neutrally (henceforth marked by the subscript neut),
 166 or sites that bear mutations with fitness consequences and for
 167 which the DFE is estimated (henceforth marked by the subscript
 168 sel). Let the observed SFS be given through $p_z(i)$, where $p_z(i)$
 169 is the count of polymorphic sites that contain the derived allele
 170 i times, $1 \leq i < n$, and l_z the total number of sites surveyed,
 171 where n is the sample size and $z \in \{\text{neut}, \text{sel}\}$. We denote by
 172 $P_z(i)$ the corresponding random variable per site, defined as the

173 random number of sites that contain the derived allele i times,
 174 normalized by l_z . From the PRF theory, $p_z(i)$ follows a Poisson
 175 distribution with mean $l_z E[P_z(i) | \theta, \phi]$, where $\theta = 4N_e \mu$ is the
 176 scaled mutation rate per site per generation, and ϕ is a parametric
 177 DFE (Figure 1B and C) that will be specified later in the
 178 *Results and Discussion* section. Here, we assume additive selection
 179 and we define the selection coefficient s as the difference
 180 in fitness between the heterozygote for the derived allele and
 181 the homozygote for the ancestral allele, leading to fitness of 1,
 182 $1 + s$ and $1 + 2s$ for the ancestral homozygote, heterozygote and
 183 derived homozygote genotypes, respectively.

184 **Expected SFS**

185 From PRF theory (Sawyer and Hartl 1992; Sethupathy and Han-
186 nenhalli 2008),

$$\begin{aligned} \mathbb{E}[P_{\text{neut}}(i) | \theta] &= \frac{\theta}{i}, \\ \mathbb{E}[P_{\text{sel}}(i) | \theta, S] &= \theta \int_0^1 B(i, n, x) H(S, x) dx, \end{aligned} \quad (1)$$

187 where

$$B(i, n, x) = \binom{n}{i} x^i (1-x)^{n-i}$$

188 is the binomial probability of observing i derived alleles in a
189 sample of size n , when the true allele frequency is x , and

$$H(S, x) = \frac{1 - e^{-S(1-x)}}{x(1-x)(1 - e^{-S})}$$

190 Note that due to our scaling of the mutation rate, $H(s, x)$ is pro-
191 portional (with a factor of 1/2) to the mean time a new semidom-
192 inant mutation of scaled selection coefficient $S = 4N_e s$ spends
193 between x and $x + dx$ (Wright 1938). Figure 1D shows the
194 expectations from equation (1) as a function of S .

195 To obtain $\mathbb{E}[P_{\text{sel}}(i) | \theta, \phi]$, we integrate over the DFE,

$$\mathbb{E}[P_{\text{sel}}(i) | \theta, \phi] = \int_{-\infty}^{\infty} \mathbb{E}[P_{\text{sel}}(i) | \theta, S] \phi(S) dS. \quad (2)$$

196 Relative to the expected SFS of independent sites under a Wright-
197 Fisher constant population (equations (1) and (2)), the observed
198 SFS can be distorted due to demography, ascertainment bias,
199 non-random sampling, and linkage. We account for such distor-
200 tions that affect both the neutral and selected sites to a similar
201 extent by using the approach of Eyre-Walker *et al.* (2006) and
202 introduce nuisance parameters r_i , $1 \leq i < n$, that scale the
203 expected SFS, for $z \in \{\text{neut}, \text{sel}\}$,

$$\mathbb{E}[P_z(i) | \theta, r_i, \phi] = r_i \mathbb{E}[P_z(i) | \theta, \phi]. \quad (3)$$

204 To avoid identifiability issues, we set $r_1 = 1$.

205 **Full DFE and divergence counts**

206 Unlike methods that infer only a strictly deleterious DFE, we
207 can incorporate a full DFE that includes both deleterious and
208 beneficial mutations. Additionally, to add flexibility in the infer-
209 ence of the full DFE, we optionally model divergence counts (the
210 number of observed fixed mutations relative to an outgroup)
211 d_z as a Poisson distribution with mean $l_z^d \mathbb{E}[D_z | \lambda, \theta, \phi]$. Here,
212 l_z^d is the number of sites used for divergence counts, and can
213 possibly be different than l_z . We have that

$$\begin{aligned} \mathbb{E}[D_{\text{neut}} | \lambda] &= \lambda, \\ \mathbb{E}[D_{\text{sel}} | \lambda, S] &= \lambda \frac{S}{1 - e^{-S}}, \end{aligned} \quad (4)$$

214 where $\lambda = T\theta$ is a composite divergence parameter that accounts
215 for the number of neutral mutations that go to fixation during the
216 divergence time T from the MRCA of the ingroup population to
217 the outgroup (blue full line in Figure 1A). The term $S/(1 - e^{-S})$
218 accounts for the fixation of a mutation with scaled selection
219 coefficient S , and can be obtained as $\lim_{x \rightarrow 1} H(S, x)$. Figure 1D
220 shows the expectations for the divergence counts at selected
221 sites from equation (4) as a function of S .

222 As divergence counts are calculated by comparing the out-
223 group sequence to the sample of sequences from the ingroup,
224 polymorphism may be misattributed as divergence, i.e. muta-
225 tions that are polymorphic in the ingroup population but fixed
226 in the sample are counted as divergence. This is the case for
227 mutations that occur between the MRCAs of the sample and
228 ingroup (blue dot-dash line in Figure 1A). As noted by Keightley
229 and Eyre-Walker (2012), misattributed polymorphism can lead
230 to biased inference of α . To account for this, we adjust the above
231 means to also incorporate the misattributed polymorphism by
232 increasing the expectations with the contributions coming from
233 mutations present in all n sampled individuals,

$$\begin{aligned} \mathbb{E}[D_{\text{neut}} | \lambda, \theta, r_n] &= \mathbb{E}[D_{\text{neut}} | \lambda] + \theta r_n \frac{1}{n}, \\ \mathbb{E}[D_{\text{sel}} | \lambda, \theta, r_n, S] &= \mathbb{E}[D_{\text{sel}} | \lambda, S] \\ &\quad + \theta r_n \int_0^1 B(n, n, x) H(S, x) dx. \end{aligned} \quad (5)$$

Assuming that the ingroup and outgroup share the same DFE, we integrate over it to obtain

$$\mathbb{E}[D_{\text{sel}} | \lambda, \theta, r_n, \phi] = \int_{-\infty}^{\infty} \mathbb{E}[D_{\text{sel}} | \lambda, \theta, r_n, S] \phi(S) dS. \quad (6)$$

Unfolded SFS and ancestral misidentification

When only a deleterious DFE is inferred, the folded SFS is typically used, where only sums of the form $p_z(i) + p_z(n-i)$ are modeled. This is sufficient for inference of deleterious DFE (Keightley and Eyre-Walker 2007). However, the unfolded SFS contains valuable information for inference of the full DFE, as beneficial mutations are expected to be present in high frequencies (Durrett 2008; Fay and Wu 2000). To obtain an unfolded SFS, the ancestral state needs to be identified, and this is error prone. To account for potential misidentification of the ancestral state, we model the mean of $P_z(i)$, $z \in \{\text{neut}, \text{sel}\}$, as a mixture of sites whose ancestral states were correctly identified (with probability $1 - \epsilon$), or misidentified (with probability ϵ) (Williamson et al. 2005; Boyko et al. 2008; Glémin et al. 2015),

$$\begin{aligned} \mathbb{E}[P_z(i) | \theta, r_i, \epsilon, \phi] &= (1 - \epsilon) \mathbb{E}[P_z(i) | \theta, r_i, \phi] \\ &\quad + \epsilon \mathbb{E}[P_z(n-i) | \theta, r_i, \phi]. \end{aligned} \quad (7)$$

Mutation variability

There is substantial evidence that both substitution and mutation rates vary along the genome (Golding 1983; Yang 1996; Francioli et al. 2015; Hodgkinson and Eyre-Walker 2011; Arndt et al. 2005), with a long tradition of modeling this variability in phylogenetic inferences as a gamma distribution (Golding 1983; Yang 1996). A few DFE inference methods allow for mutation rates to vary in a non-parametric fashion (Bustamante et al. 2003;

Gronau et al. 2013). Here, we model mutation variability by assuming that mutation rates follow a gamma distribution with mean $\bar{\theta}$ and shape a . This is motivated by the phylogenetic approaches, but also by mathematical convenience: if the mean of a Poisson distribution follows a gamma distribution, the resulting distribution is a negative binomial distribution. We assume that the data is divided into m non-overlapping fragments of lengths $l_{\text{neut}}^j + l_{\text{sel}}^j$, $1 \leq j \leq m$, and for each fragment j , we have the SFS $p_z^j(i)$, $1 \leq i < n$. Possibly, we have an additional m^d fragments of lengths $l_{\text{neut}}^{d,j} + l_{\text{sel}}^{d,j}$ for which we have the divergence counts, d_z^j . We assume that each fragment has a constant mutation rate θ , but that mutation rates can vary between fragments. Given the mutation rate θ_j of the fragment j , then $p_z^j(i)$ and d_z^j follow the Poisson distributions with means $l_z^j \mathbb{E}[P_z(i) | \theta_j, r_i, \epsilon, \phi]$ and $l_z^{d,j} \mathbb{E}[D_z | \lambda, \theta_j, r_n, \phi]$, given by equations (1)–(6). Integrating over the mutation rates distribution, we obtain that $p_z^j(i)$ and d_z^j have a negative binomial distribution with shape a and means $l_z^j \mathbb{E}[P_z(i) | \bar{\theta}, r_i, \epsilon, \phi]$ and $l_z^{d,j} \mathbb{E}[D_z | \lambda, \bar{\theta}, r_n, \phi]$, respectively.

Inferring α using divergence or polymorphism data alone

Once the DFE is estimated, α can be calculated from the observed divergence counts as follows (Eyre-Walker and Keightley 2009)

$$\alpha \approx \frac{d_{\text{sel}} - \frac{l_{\text{sel}}^d}{l_{\text{neut}}^d} d_{\text{neut}} \int_{-\infty}^0 \frac{S}{1 - e^{-S}} \phi(S) dS}{d_{\text{sel}}}, \quad (8)$$

where the nominator represents the estimated number of adaptive substitutions, which are obtained by subtracting the expected deleterious and neutral substitutions from the total observed divergence counts at selected sites. Keightley and Eyre-Walker (2012) extended the above estimation of α to account for the misattributed polymorphism. Using our framework, we correct for the misattributed polymorphism by removing from d_{sel} and d_{neut} the expected number of mutations that are in fact polymorphic. These expectations can be readily obtained from equation (5) by setting $\lambda = 0$. Then the new estimate of α is obtained as in equation (8), where d_{sel} and d_{neut} are replaced with the re-adjusted divergence counts d_{sel}^* and d_{neut}^* given by

$$d_{\text{neut}}^* = d_{\text{neut}} - I_{\text{neut}}^d \mathbb{E}[D_{\text{neut}} | \lambda = 0, \theta, r_n], \quad (9)$$

$$d_{\text{sel}}^* = d_{\text{sel}} - I_{\text{sel}}^d \mathbb{E}[D_{\text{sel}} | \lambda = 0, \theta, r_n, \phi].$$

291 This approach to calculate α relies heavily on the assumption
 292 that the ingroup and outgroup share the same scaled DFE.
 293 However, if one has access to an estimated full DFE purely from
 294 polymorphism data, α can still be estimated by replacing the
 295 observed divergence counts with the expected counts from equation
 296 (4). As λ will cancel out in the resulting fraction, α can be
 297 obtained by setting $\lambda = 1$. Then,

$$\alpha \approx \frac{\int_0^{\infty} \mathbb{E}[D_{\text{sel}} | \lambda = 1, S] \phi(S) dS}{\int_{-\infty}^{\infty} \mathbb{E}[D_{\text{sel}} | \lambda = 1, S] \phi(S) dS}. \quad (10)$$

298 In the rest of this paper, we refer to the two above estimates
 299 of α as α_{div} and α_{dfe} , respectively, to distinguish more clearly the
 300 type of information used.

301 Likelihood estimation and comparison of models

302 The hierarchical framework described above allows maximum
 303 likelihood estimation of both evolutionary (mutation rates, DFE
 304 parameters) and nuisance parameters, as well as the error in
 305 the ancestral states, ϵ . Details about the implementation and
 306 optimization of the likelihood function are given in the Supplemental
 307 Material. Note that in our implementation, likelihood
 308 ratio tests (LRTs) can be used to test rigorously whether the polymorphism
 309 data provides evidence for a full DFE, or if a strictly deleterious DFE
 310 is sufficient for accounting for the data. This framework also allows to
 311 decide whether including nuisance parameters and / or ancestral errors
 312 provides a better fit to the data. The p-values for the LRT are obtained
 313 by assuming that likelihood ratio under the null hypothesis (reduced model
 314 is correct) is distributed as χ^2 .

316 Results and Discussion

317 To investigate the statistical performance of our method to infer
 318 the DFE, α and test hypothesis regarding the contribution of ben-

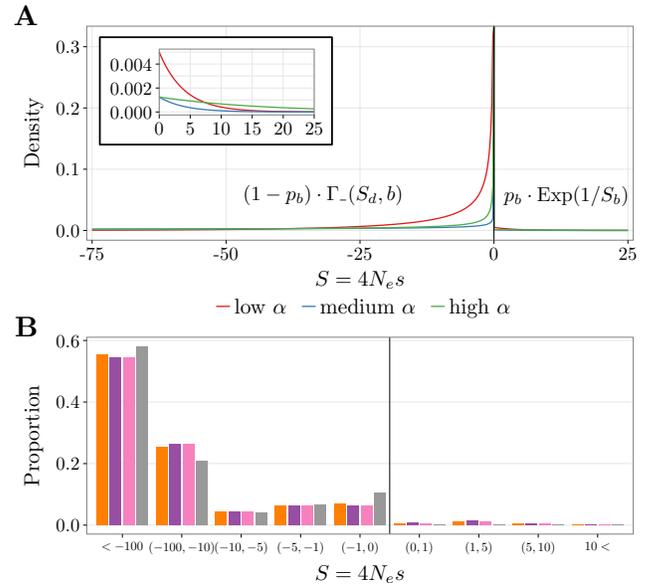


Figure 2 Example of simulated and inferred gamma + exponential DFEs. (A) Three of the simulated DFEs (corresponding to LALSD, MALPB, and HAHSB from Table S1) with different α s (proportion of beneficial substitutions). The DFEs are parameterized by S_d (mean selection coefficient of deleterious mutations), b (shape of distribution of deleterious DFE), p_b (proportion of beneficial mutations), and S_b (mean selection coefficient of beneficial mutations). The inset shows a zoom-in of the beneficial part of the DFE. (B) Simulated discretized DFE (orange, corresponding to MAMSD from Table S1), together with the mean (over the 100 replicates) inferred discretized DFE using both polymorphism and divergence data (purple) and only polymorphism data (pink and gray), where a full DFE (pink) and a deleterious DFE (gray) was inferred.

319 efcial mutations to patterns of polymorphism, we performed
 320 extensive simulations using SFS_CODE (Hernandez 2008). We
 321 explored a wide range of simulated DFEs (12 full DFEs and 5
 322 deleterious DFEs, Table S1), chosen such that the simulated α
 323 had one of four possible values (0%, 20%, 50% and 80%, Figure
 324 2A). Most simulations were performed using a constant
 325 population size and without error in the identification of the
 326 ancestral state. Results are shown for this type of data if not
 327 otherwise specified. These assumptions were later relaxed. The
 328 simulations contained linkage and were performed to resemble
 329 exome data. For each considered simulation scenario (one given
 330 DFE, demographic, linkage, misidentification of the ancestral
 331 state), we simulated 100 replicate data sets. For more details on
 332 the simulated data, see the Supplemental Material.

The general shape of the DFE is not agreed upon (Welch *et al.*

2008; Bataillon and Bailey 2014). The DFE has been modeled using a wide range of functional continuous forms (Boyko *et al.* 2008; Kousathanas and Keightley 2013; Galtier 2016), but also as a discrete distribution (Gronau *et al.* 2013; Kousathanas and Keightley 2013; Keightley and Eyre-Walker 2010). Here, we use a DFE consisting of a mixture between gamma and exponential distributions, that model deleterious and beneficial mutations, respectively. With probability $1 - p_b$, a new mutation is deleterious and its selection coefficient comes from a reflected gamma distribution with mean $S_d < 0$ and shape b , while with probability p_b , a new mutation is beneficial and its selection coefficient comes from an exponential distribution with mean $S_b > 0$ (Figure 2A). We do not explore alternative parametric DFE families. For such studies, we refer the reader to Kousathanas and Keightley (2013); Welch *et al.* (2008).

We inferred the DFE and alpha parameters in our model using three different models: a full DFE was inferred from both polymorphism and divergence data; a full DFE was inferred from polymorphism data alone; an only deleterious DFE was inferred from polymorphism data alone. From the inferred DFEs, we calculated α_{dfe} and α_{div} . For the inference assuming only a deleterious DFE, α_{dfe} is always 0, and for such inference we therefore only calculated α_{div} . The distortion parameters r were always estimated, while the ancestral misidentification error ϵ was fixed to 0, unless otherwise specified. We report the inference performance using $\log_2(\text{estim}/\text{sim})$ on a log-modulus scale. Here, estim is the estimated value, while sim is the simulated value. Unlike the relative error, this log ratio gives equal weight to both overestimation and underestimation of the parameters. For example, the log ratios of 1 and -1 correspond to the estimated value being double or half the simulated value, respectively. When $\text{sim} = \text{estim}$, the ratio is equal to 0. See the Supplemental Material for details.

Inference of deleterious DFE

Using simulations that did not contain any beneficial mutations in the polymorphism data, we first investigated how well we can infer the deleterious DFE and if our method can recover the fact

that all polymorphic mutations are deleterious. We observed that the two parameters determining the deleterious DFE, S_d and b , and α , are inferred accurately when only a deleterious DFE was estimated (Figure 3A and Figure S1). When, instead, a full DFE was inferred from the polymorphism data alone, the parameters showed different amounts of bias (Figure 3A and Figure S1). A crucial question is whether the data allows one to decide correctly which model is most sensible: the full DFE or only deleterious DFE? When using a LRT to compare the relative goodness of fit on simulated data, virtually all data sets tended to reject the full DFE model in favor of the reduced model featuring only deleterious mutations in the DFE (Figure S2). This indicates that while our method can account for the presence of beneficial mutations in the SFS data, it can also accurately detect if there is no empirical evidence for such mutations in the data. So in principle, one can perform estimation under both the full and deleterious DFE models and use the LRT to decide which model is most appropriate for the data.

Inference of full DFE

From the expected contribution of mutations to polymorphism and divergence data, as a function of S (Figure 1D), it is evident that if beneficial mutations occur at any appreciable rate, they should have a non-negligible impact in the polymorphism data. This suggests that it should be possible to infer the full DFE from polymorphism data alone. We investigated this using data generated under a full DFE. As one might expect, the deleterious DFE parameters were inferred equally well regardless of whether the divergence data was used or not (Figure S3). The variance of the estimates seems to be somewhat larger when divergence data is not used, but this is most likely due to the inference using less data. The parameters of the beneficial part of the DFE and α were inferred with different levels of accuracy (Figure 3B and Figure S3). From the simulation scenarios considered here, it is apparent that the value of α predicts the accuracy: the higher α , the better the prediction, for both inference with and without divergence. In short, when beneficial mutations are comparatively rare and of very small effects, estimating their

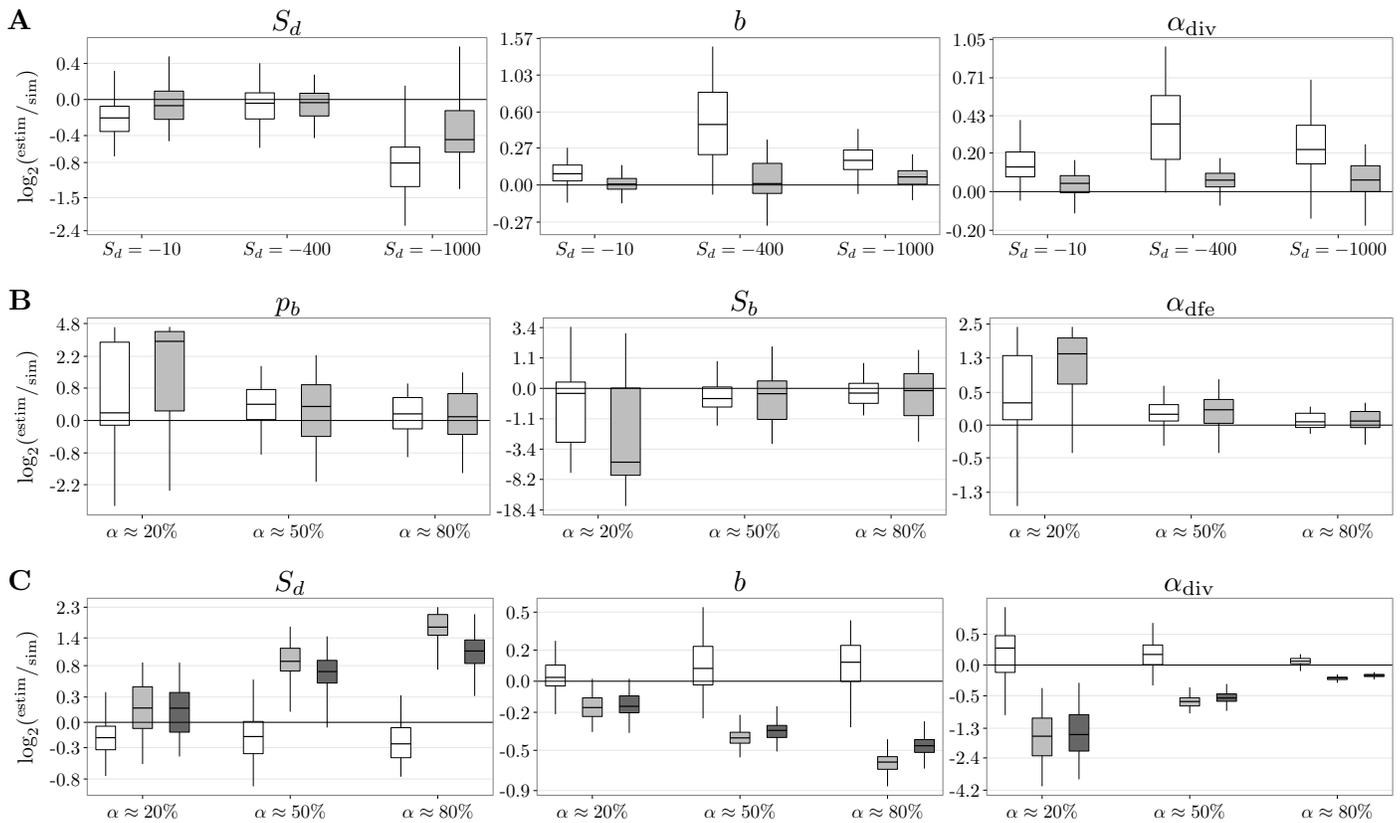


Figure 3 Inference of α (proportion of beneficial substitutions) and DFE parameters: S_d (mean selection coefficient of deleterious mutations), b (shape of distribution of deleterious DFE), p_b (proportion of beneficial mutations), and S_b (mean selection coefficient of beneficial mutations). (A) Quality for inference performed on polymorphism data alone, for three simulated deleterious (corresponding to DelLSD, DelMSD, and DelHSD from Table S1) DFEs with different S_d s. The DFE parameters are inferred using only polymorphism data assuming a full (white boxes) and deleterious (gray boxes) DFE. (B) Quality for inference performed on polymorphism and divergence data, for three simulated DFEs with different α s (corresponding to LALS, MAMS, and HAHSD from Table S1). The DFEs differ only in the simulated value of S_d . The DFE parameters are inferred using both polymorphism and divergence (white boxes) and only polymorphism (gray boxes) data. (C) Quality for inference performed on polymorphism data alone, for three simulated DFEs with different α s (corresponding to LALS, MAMS, and HAHSD from Table S1). The DFEs differ only in the simulated value of S_b . Only polymorphism data is used, and the DFE parameters are inferred assuming a full DFE where ϵ is set to 0 and is not estimated (white boxes) and a deleterious DFE (gray boxes), where ϵ is set to 0 and is not estimated (light gray boxes), or is estimated (dark gray boxes). The data was simulated with $\epsilon = 0$.

408 properties and α is challenging (even with divergence data).
 409 Conversely, when beneficial mutations are relatively common,
 410 they dominate the divergence counts, but also make substantial
 411 contribution to the SFS counts, which alone can allow reliable
 412 estimation of the beneficial fraction of the DFE and α . For the
 413 lower values of α ($\alpha \approx 20\%$), the use of divergence data provides
 414 more accurate estimates than when relying on polymorphism
 415 data alone. This could perhaps be explained by the fact that,
 416 in this case, the polymorphism data is heavily dominated by
 417 deleterious mutations and it is more difficult to tell apart the
 418 amount of beneficial selection from polymorphism data alone.
 419 However, as α increases, the differences in performance between

420 the inference with and without divergence diminishes, strongly
 421 indicating that divergence data is not necessarily needed for
 422 accurate inference.

423 Similar to [Schneider et al. \(2011\)](#), we observe a strong negative
 424 correlation between the proportion of beneficial mutations p_b
 425 and their scaled selection coefficient S_b (Figure S4). This illus-
 426 trates the fact that p_b and S_b are difficult to estimate separately,
 427 but their product, which largely determines α , is more accurately
 428 estimated. This can be seen in Figure 3B and Figure S3, where
 429 even though p_b and S_b might be biased, overall α is inferred
 430 more accurately. [Schneider et al. \(2011\)](#) reported that the estima-
 431 tion of p_b and S_b improves as more sites are available. In our

432 data simulations we used a fixed number of sites, but we do
433 observed that p_b , S_b and α are better estimated as α increases.

434 When inferring a full DFE, we can calculate both α_{div} and
435 α_{dfe} , which should both be good predictors of the true simulated
436 α . We generally found very good correlation between the two
437 estimated values (Figure S5), and perhaps not surprising, the
438 estimates of α_{dfe} obtained when performing inference on both
439 SFS and divergence data were more tightly correlated.

440 We note here that Schneider *et al.* (2011) is the only method
441 that we are aware of that can estimate both a full DFE and α from
442 polymorphism data alone, though the authors did not investi-
443 gate the power to infer α , but rather the product of $p_b S_b$, which
444 is taken as a proxy for α . Additionally, they did not consider
445 in their simulations different deleterious DFEs. Our simulated
446 DFEs were chosen such that they cover cases with the same
447 simulated p_b and S_b , but generate different α s. The differences
448 in α can be driven by the amount of beneficial mutations, but
449 also by the intensity of purifying selection, or, said slightly differ-
450 ently, the properties of the deleterious fraction of the DFE. These
451 simulated data set revealed that the amount and strength of
452 positive selection is not the only determinant in how accurately
453 p_b and S_b are inferred. For example, the results in Figure 3B are
454 given for simulated DFEs that differ only in the value of S_d , i.e.
455 the strength of purifying selection, and we find in this instance a
456 clear difference in the inference performance in terms of relative
457 error.

458 **Bias from not inferring full DFE**

459 Given that divergence data is clearly not necessary for reliable
460 estimates of the full DFE, a question arises: what happens when
461 inference methods ignore the presence of beneficial mutations in
462 the polymorphism data? For this, using the simulated data sets
463 generated using full DFEs, we performed inference only on poly-
464 morphism data where we inferred either a full DFE, like before,
465 or under a reduced model restricted to only a deleterious DFE
466 (Figure 3C and Figure S6). Note that this corresponds to the cur-
467 rent state of the art for empirical studies of DFE from population
468 genomics data, where data tend invariably to be analyzed under

469 the assumption that SFS data is to be fitted exclusively with a
470 deleterious DFE (Racimo and Schraiber 2014; Bataillon *et al.* 2015;
471 Halligan *et al.* 2013; Arunkumar *et al.* 2015; Charlesworth 2015;
472 Harris and Nielsen 2016; Slotte *et al.* 2010; Strasburg *et al.* 2011).
473 When α was $\approx 20\%$, the inferred S_d and b were, at times, more
474 accurate when only a deleterious DFE was used. However, as
475 α increased, the two parameters were increasingly biased. The
476 mean S_d was estimated to be more negative, while the shape b
477 was estimated to be closer to 0: the inferred deleterious DFEs
478 were getting much more leptokurtic than the parametric DFE
479 used to simulate the data. This resulted in inferring DFEs with
480 more probability mass accumulating close to 0. A straightfor-
481 ward interpretation is that the inference method attempted to fit
482 the SFS counts contributed by the weakly beneficial mutations
483 by fitting a DFE that comprised a sizable amount of weakly dele-
484 terious mutations (the best proxy for beneficial mutations). A
485 comparison of the simulated and inferred discretized DFEs (Fig-
486 ure 2B and Figure S7) illustrates this point: the inference with
487 only deleterious DFE overestimated appreciably the amount of
488 mutations with a selection coefficient in the range $(-1, 0)$ (sim-
489 ulated: 0.07, deleterious DFE: 0.11) and $(-5, -1)$ (simulated:
490 0.063, deleterious DFE: 0.067) ranges.

491 DFE methods that do not model beneficial mutations in the
492 polymorphism data use a folded SFS. To mimic this behavior,
493 we allowed for ϵ to be estimated, even though no errors in the
494 identification of ancestral state were simulated. We observed
495 that ϵ reduces partially the bias in the parameters (Figure 3C,
496 Figures S6 and S7).

497 Using a LRT we could test, as before, for the presence of
498 beneficial mutations in the polymorphism data by comparing
499 the inferences with a full or deleterious DFE (Figure S8). We
500 observed that the larger α , the stronger the preference for the full
501 DFE model. We also noticed that, even though α might be rela-
502 tively large, if the mean strength of beneficial selection was very
503 low (Figure S8, MALSB where $S_b = 0.1$), the LRT indicated that
504 there were no beneficial mutations in the polymorphism data.
505 Such mutations can pass as weakly deleterious mutations when
506 fitting the data. The LRT also showed an increasing preference

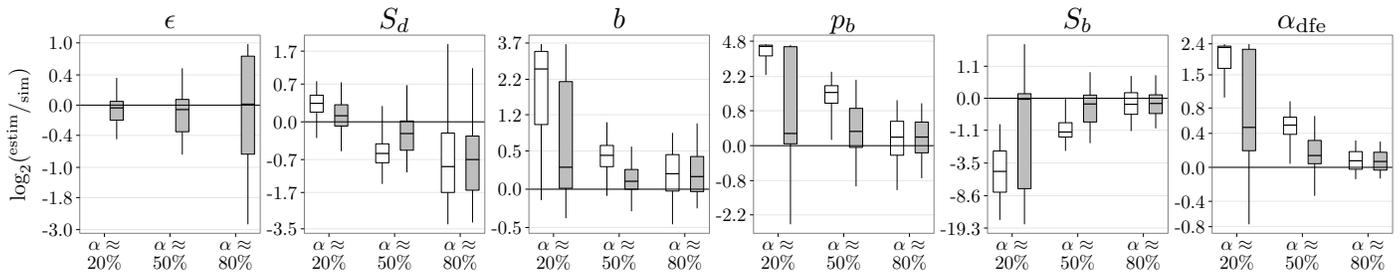


Figure 4 Inference of α (proportion of beneficial substitutions), ϵ (rate of ancestral error), and DFE parameters: S_d (mean selection coefficient of deleterious mutations), b (shape of distribution of deleterious DFE), p_b (proportion of beneficial mutations), and S_b (mean selection coefficient of beneficial mutations). The figure shows the inference quality for three simulated DFEs (corresponding to LALSD, MAMSD, and HAHS from Table S1) with different α s. The DFEs differ only in the simulated value of S_d . A full DFE is inferred from both polymorphism and divergence data, and ϵ is set to 0 and is not estimated (white boxes), or is estimated (gray boxes). The data was simulated with $\epsilon = 0.05$.

507 with α for ϵ when only deleterious DFE is inferred, indicating, 534
 508 as expected, that the model with $\epsilon \neq 0$ could account for some 535
 509 of the weakly beneficial mutations present in the polymorphism 536
 510 data. 537

511 Inferring only a deleterious DFE leads to a consistent bias in 538
 512 α as well. This bias is not that well correlated with the simulated 539
 513 value of α , but it is apparent that a higher α leads to a smaller 540
 514 bias (Figure S6). This is in contrast to the bias observed for S_d 541
 515 and b . To obtain α from a deleterious DFE only, we need to rely 542
 516 on the divergence data. Perhaps, when α is large, the signal 543
 517 of positive selection is so strong in the divergence data that it 544
 518 overrides, to some extent, the bias in S_d and b , leading to a more 545
 519 accurate estimate of α . 546

520 The assumption of negligible contribution of beneficial muta- 547
 521 tions to SFS counts can be traced back to Smith and Eyre-Walker 548
 522 (2002). To support the claim, the authors stated that “if advan- 549
 523 tageous mutations, with an advantage of $N_e s = 25$ occur at 550
 524 one-hundredth the rate of neutral mutations, they will account
 525 for 50% of substitutions, but account for just 2% of heterozygos-
 526 ity”. Our simulated S_b (which is scaled by $4N_e$) was typically 4.
 527 To investigate what happens when selection is much stronger,
 528 we simulated a full DFE with $S_b = 800$ such that only 10% of 551
 529 beneficial mutations (0.2% of all mutations) had a selection coef- 552
 530 ficient of 100 or less. For this, the simulated α was nearly 100% 553
 531 and one would expect that, as selection is so strong, most mu- 554
 532 tations would fix quickly. While the DFE parameters could not 555
 533 be recovered as accurately (Figure S9), the estimated α was very 556

precise, regardless of the model used for inference. This points to
 the fact that, even when positive selection is very strong, there is
 enough information left in the polymorphism data to be able to
 estimate α without relying on divergence data. The bias in S_d , b
 and α (Figure S9) and LRT (Figure S10) from inference with only
 deleterious DFE followed the same trend as before. However,
 even though α was large, p_b and S_b were not that well estimated.
 This is most likely because when S_b is getting very large, the
 expected counts from equation (1) become independent of S ,
 since $H(S, x) \approx \frac{1}{x(1-x)}$ for large S (Figure 1D). This explains
 why the inference method will have trouble narrowing precisely
 the value of S_b .

Keightley and Eyre-Walker (2010) investigated if the presence
 of beneficial mutations in the polymorphism data could poten-
 tially affect the inference when assuming only a deleterious DFE.
 For this, they simulated data using a partially reflected gamma
 distribution, given by

$$\phi(S; S_d, b) = \frac{1}{1 + e^S} \Gamma(|S|; -S_d, b),$$

where $\Gamma(x; m, s)$ is the density of a gamma distribution with
 mean m and shape s . This distribution arises from the assump-
 tion that the absolute strength of selection is gamma distributed
 and that each site can be occupied by either an advantageous
 or a deleterious allele, both having the same absolute selection
 strength $|s|$. Keightley and Eyre-Walker (2010) simulated data

557 with $|S_d| = 400$ and $b = 0.5$. Due to the chosen distribution, the
558 simulated proportion of beneficial selection was $p_b = 0.0214$,
559 while the mean selection coefficient of beneficial mutations was
560 only $S_b = 0.014$ ($p_b S_b = 0.00029$). These values are close to
561 one of our own simulated DFEs with $\alpha \approx 20\%$ (LALSB, Ta-
562 ble S1), with the difference that we simulated an S_b that was
563 approximately 7 times larger. For this simulation scenario we
564 did, indeed, find little bias in S_d and b when inferring only a
565 deleterious DFE (Figures S6 and S7). However, arguably, this
566 strength of beneficial mutations is extremely low. For example,
567 [Schneider et al. \(2011\)](#) inferred the strength of beneficial muta-
568 tions from *Drosophila* and found $p_b S_b$ to be two-three orders of
569 magnitude higher: $p_b = 0.0096$ and $S_b = 18$ ($p_b S_b = 0.1728$).

570 **Impact of ancestral error on inference**

571 The results presented above were based on simulations where
572 the true ancestral state was used. To investigate the conse-
573 quences of misidentification of the ancestral state, we added
574 errors to the simulated data (see Supplemental Material for de-
575 tails). Inferring a full DFE and using divergence data, we found
576 that we can properly account for the rate of misidentification,
577 and the error ϵ is accurately recovered (Figure 4 and Figure S11).
578 As expected, the inference of the DFE and α is biased when
579 the misidentification is not accounted for. A LRT for $\epsilon \neq 0$
580 (Figure S14) supported the use of a model including the joint
581 estimation of ϵ and DFE parameters for the data with errors, but
582 rejected the more complex model for the data without error.

583 [Galtier \(2016\)](#), who also used distortion parameters r_i when
584 inferring the DFE, stated that these parameters are “expected
585 to capture any departure from the expected SFS as soon as it
586 is shared by synonymous and non-synonymous sites”. Our
587 results indicate that the misidentification of the ancestral state
588 cannot be accurately accounted for by the r_i parameters (Figure 4
589 and Figure S11). However, we did find that the resulting bias
590 decreased with α and that the preference (as measured by a LRT)
591 for models inferring $\epsilon \neq 0$ also decreased for data sets with
592 higher α (Figure S14). For data simulations with $\alpha \approx 80\%$, the
593 inference was just as good when ϵ was set to 0. To investigate

594 this in more details, we also ran the inference with $r_i = 1$ (i.e.
595 no distortion correction) and $\epsilon = 0$ on those simulated DFEs.
596 The results showed a large bias in the DFE parameters and α
597 when $r_i = 1$ (Figure S12), and a LRT favored the estimation
598 of r_i s (Figure S15). This illustrated that merely using the r_i
599 parameters without explicitly accounting for misidentification of
600 the ancestral state is not always accurate and can bias inference
601 of DFE and α .

602 Both the presence of beneficial mutations and $\epsilon \neq 0$ create
603 similar patterns in the polymorphism data: the frequency of the
604 common derived alleles increases. We have seen before that ϵ
605 can account for some of the positive selection in the data (Fig-
606 ure 3C, Figures S6 and S7). Similarly, we observed that positive
607 selection can account for some of the misidentification of ances-
608 tral state. On simulations with a deleterious DFE and incorrect
609 ancestral states, we found that when assuming $\epsilon = 0$, the pa-
610 rameters inferred when a full DFE is assumed are, generally,
611 more accurate than when only a deleterious DFE is inferred
612 (Figure S13). A LRT also supported the use of a full DFE (Fig-
613 ure S16). Comparing the inferred p_b and S_b when ϵ is inferred or
614 not (Figure S13) showed that these parameters are higher when
615 $\epsilon = 0$, further indicating that they captured some of the ancestral
616 misidentification errors. Therefore, if the data contains sites that
617 have the ancestral state misidentified, which is virtually always
618 the case in empirical data sets, ancestral misidentification will
619 be wrongly interpreted as positive selection if the misidentifi-
620 cation is not accounted for. If ϵ is inferred jointly with the DFE
621 parameters, a LRT comparing models with full DFE or only dele-
622 terious DFE can correctly detect that the polymorphism data
623 does not contain any beneficial mutations (Figure S16). Our
624 simulation results illustrated that systematically incorporating
625 the rate of ancestral error is crucial for a reliable inference of DFE
626 parameters and α .

627 **Distortions of the SFS by linkage and demography**

628 It has previously been suggested that correcting for the effect
629 of demography – using the observed SFS at neutral sites – can
630 also reduce some of the bias introduced by linkage in the data

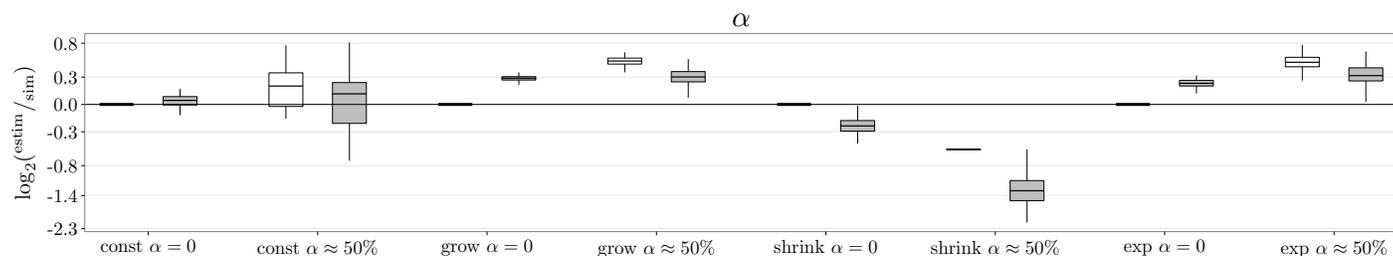


Figure 5 Inference of α (proportion of beneficial substitutions). The figure shows the inference quality for α_{dfe} (inferred from DFE alone, white) and α_{div} (inferred from DFE and divergence data, gray) for four simulated demographic scenarios (detailed in the Supplemental Information) and a deleterious DFE only ($\alpha = 0$, corresponding to DelMSD from Table S1) or a full DFE ($\alpha \approx 50\%$, corresponding to MAMSD from Table S1). For all inference only SFS data was used, and a LRT was performed to compare the full and deleterious only DFE models. The estimated value of α was chosen from the model preferred by the LRT.

631 (Kousathanas and Keightley 2013; Messer and Petrov 2012). We
 632 explored this possibility by simulating different levels of linkage
 633 (see Supplemental Material for details). We found that, indeed,
 634 the r_i parameters could partially correct for the presence of link-
 635 age (Figure S17), with the most pronounced effect on S_d and
 636 b . A LRT for $r_i \neq 1$ (Figure S18) increasingly favored the more
 637 models fitting r_i as the level of linkage increased.

638 For the previous simulations we used a constant population
 639 size. To check that the r_i parameters can also correct for demog-
 640 raphy, we simulated additional data, using different demogra-
 641 phy scenarios (see Supplemental Material for details). When
 642 populations size varies in time, N_e is typically taken to be the
 643 harmonic mean of the different sizes (Kliman *et al.* 2008). While
 644 this might be a good approximation for the neutral sites, the sites
 645 under selection experience a different N_e , which depends on the
 646 strength of selection S (Otto and Whitlock 1997). Therefore, for
 647 these simulations, we do not have a priori knowledge of the N_e
 648 that accurately captures the interaction between selection and
 649 demography, and we could only compare parameters that are in-
 650 dependent of N_e (b , p_b), and α , for which a value can be obtained
 651 by tracking the proportion of adaptive mutations contributing to
 652 divergence in the forward simulations used to generate the data
 653 sets. Like before, we first simulated a deleterious DFE, similar to
 654 previous studies (Boyko *et al.* 2008; Eyre-Walker *et al.* 2006; Eyre-
 655 Walker and Keightley 2009; Keightley and Eyre-Walker 2007).
 656 We found that a LRT correctly detected that $r_i \neq 1$ (Figure S20),
 657 but that the parameters can correct only partially for the effect
 658 of demography (Figure S19). The b parameter was inferred ac-

659 curately when the r_i parameters were estimated. However, the
 660 estimated α was still biased (Figure 5). As no full DFE was in-
 661 ferred, α was calculated from the divergence data. For this, the
 662 same DFE was assumed in the ingroup and outgroup. How-
 663 ever, as the ingroup now underwent variable population size,
 664 its N_e was different from the N_e of the outgroup (which had a
 665 constant size), and therefore the two populations had different
 666 scaled DFEs. This difference could explain the observed bias in
 667 α Eyre-Walker and Keightley (2009) noticed the same effect and
 668 proposed a correction for α . However, their correction requires
 669 the ratio of the N_e s of the two populations, which is typically
 670 not known.

We then investigated if the r_i parameters could also correct
 for demography when a full DFE was simulated (Figure S21).
 The LRT (Figure S22) showed a clear preference for $r_i \neq 1$. When
 inferring the DFE from both polymorphism and divergence data,
 we observed a bias in b and p_b . As before, this was caused by
 the incorrect assumption of a shared DFE between the ingroup
 and outgroup. When only polymorphism data was used for the
 inference, the r_i parameters could accurately correct the estima-
 tion of p_b and b , but the inferred α (estimated via α_{dfe}) was still
 slightly biased. To investigate if this bias was caused by linkage,
 we also ran simulations with reduced linkage (Figure S21), but
 the bias remained.

We investigated if the full or deleterious DFE model is pre-
 ferred for the data simulated under variable population size.
 We found that a LRT consistently preferred the full DFE model
 when the SFS data contained beneficial mutations (Figure S23).

687 Under demographic simulations, the estimated α_{dfe} and α_{div} 724
688 could differ considerably (Figure 5). When only a deleterious 725
689 DFE was simulated, relying on divergence data to estimate α 726
690 can lead to heavily biased estimates. Note that when the popu- 727
691 lation size was halved and a full DFE was simulated, the LRT 728
692 favored the incorrect deleterious DFE model. This indicates that 729
693 the r_i parameter cannot fully capture the demography and in 730
694 this particular simulation, the incorrect model choice can be ex- 731
695 plained by the extra deleterious load incurred by the population 732
696 shrinkage. 733

697 The simulated demographics were the same for both dele- 734
698 terious and full DFE simulations, and therefore the inferred 735
699 r_i parameters on the deleterious and full DFE data should be 736
700 highly correlated. We did, in fact, detect a strong correlation 737
701 (Figure S24). One of the simulations showed no correlation in 738
702 r_i and the LRT preferred the less complex model with $r_i = 1$ 739
703 (Figures S20 and S22, SHRINK). The change in population size 740
704 for this simulation was most likely not strong enough for it to 741
705 leave an appreciable footprint in the data. 742

706 [Galtier \(2016\)](#) is the only study that we are aware of that 743
707 tested if demography can accurately be accounted for when a 744
708 full DFE was simulated. While the estimated α s from [Galtier](#) 745
709 (2016) were somewhat more accurate than what we found, there 746
710 are critical differences between these studies. While we sim- 747
711 ulated a continuous full DFE, [Galtier \(2016\)](#) assumed that all 748
712 beneficial mutations had the same selection coefficient S_b . Nev- 749
713 ertheless, [Galtier \(2016\)](#) inferred a continuous full DFE and used 750
714 equation (8) for calculating α , where the integration limit was 751
715 set to some $S_{\text{adv}} > 0$ instead of 0. The reasoning behind this is 752
716 that mutations with a selection coefficient $S > 0$ that is not very 753
717 large should not be considered advantageous mutations. [Galtier](#) 754
718 (2016) used an arbitrary cutoff at $S_{\text{adv}} = 5$. Note that a different 755
719 cut-off value of S_{adv} would lead to different α s: the smaller S_{adv} , 756
720 the larger the estimated α . 757

721 **Comparison with the dfe-alpha method**

722 We chose to compare our method with dfe-alpha, one of the 760
723 most widely used inference methods for DFE and α . dfe-alpha

was originally developed to infer a deleterious DFE ([Keightley and Eyre-Walker 2007](#)), and it was subsequently extended to estimate α ([Eyre-Walker and Keightley 2009](#)), model a full DFE ([Schneider et al. 2011](#)) and correct α for misattributed polymorphism ([Keightley and Eyre-Walker 2012](#)). While dfe-alpha can infer a full discrete DFE, the method to account for potential errors in the ancestral state described in [Schneider et al. \(2011\)](#) is not implemented in dfe-alpha. At the time when we ran out comparison, we could not find any option in dfe-alpha for accounting for such errors. As we showed that this is crucial for accurate inference (Figure 4 and Figure S11), we therefore chose to run dfe-alpha with a folded SFS, where only a deleterious DFE can be estimated. We then compared with our method when only a deleterious DFE was inferred, where, as before, ϵ was either set to 0, or estimated. Although these comparisons are therefore quite limited in scope, we found that, for simulations with only a deleterious DFE, our method provided better estimates and with lower variance than dfe-alpha (Figure 6 and Figure S25). For these simulations, we also found that, sometimes, dfe-alpha estimated an α that was very large, both on the negative and positive side (Figure S25, DelHB simulation). This seemed to be the result of the correction for the misattributed polymorphism, as the uncorrected α was much closer to the true value (data not shown). This most likely explains the general differences observed between the estimated α from dfe-alpha and our method. When the inference was performed on data simulated with a full DFE, we observed the same type of bias in S_d and b as described before (Figure 6 and Figure S25). When demography and a strictly deleterious DFE were simulated, the estimation was, again, comparable (Figure S26). However, when demography was simulated on top of a full DFE, the bias of b differed between dfe-alpha and our method. This could, perhaps, be explained by the differences between the two methods for accounting for demography: while we used the nuisance parameters r_i , dfe-alpha assumes a strict simplified demographic scenario and only allows the population to undergo one size change in the past.

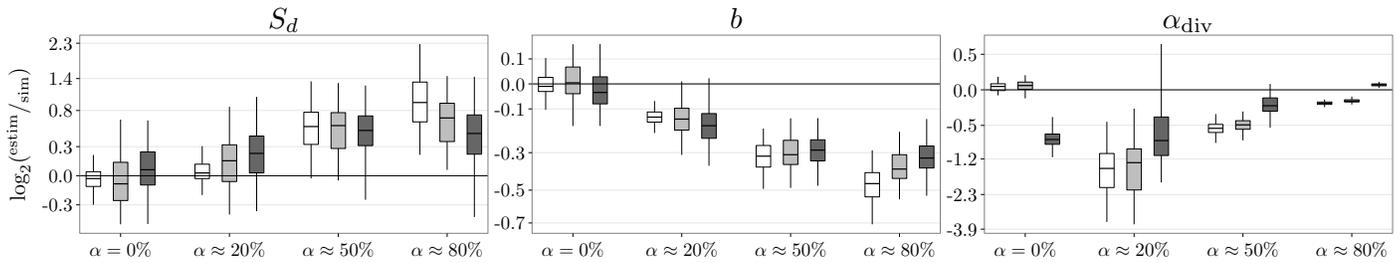


Figure 6 Comparison to dfe-alpha of inference of α (proportion of beneficial substitutions) and deleterious DFE parameters: S_d (mean selection coefficient of deleterious mutations) and b (shape of distribution of deleterious DFE), for four simulated deleterious DFEs (corresponding to DelMSD, LALSD, MAMSD and HAHS from Table S1) with different α s. The DFE parameters are inferred using only polymorphism data, assuming a deleterious DFE, where ϵ is set to 0 and is not estimated (white boxes), or is estimated (light grey boxes). The inference from dfe-alpha is given in the dark grey boxes. The data was simulated with $\epsilon = 0$.

761 Conclusion

762 We have presented a new method to infer the DFE and proportion of advantageous substitutions, α , from polymorphism and
 763 tion of advantageous substitutions, α , from polymorphism and
 764 divergence data. Using our framework, we demonstrated that
 765 inference can be performed using polymorphism data alone,
 766 and that this lead to more accurate inference when the DFE is
 767 not shared between the ingroup and the outgroup. We additionally
 768 illustrated that when the effects of beneficial mutations on
 769 polymorphism data were not modeled, the inferred deleterious
 770 DFE was biased. This bias comes from an increase of mutations
 771 at selected sites that segregate at high frequencies. Methods
 772 ignoring the contribution of beneficial fraction to SFS counts
 773 will tend to infer DFEs that have a larger amount of slightly
 774 deleterious mutations, as this is the best way to account for the
 775 observed data. Therefore, the estimated deleterious DFE had
 776 a much larger mass close to 0 compared to the simulated deleterious
 777 DFE. This, in turn, could be achieved by a larger (more
 778 negative) S_d and a lower b of the Γ distribution used here for
 779 the deleterious DFE. In cases where polymorphism data did not
 780 contain any beneficial mutations, the inference was much more
 781 accurate under a reduced model positing only a deleterious DFE.
 782 We showed that when applying our method, the use of a LRT
 783 comparing a model featuring a full DFE and a deleterious DFE,
 784 would accurately select the reduced model and allow precise
 785 inference of the deleterious DFE. This is an important result, as
 786 it suggests that using a full DFE for inference from SFS data does
 787 not come with a cost when no beneficial mutations contributed

788 to the SFS counts, and that the method does not spuriously infer
 789 presence of beneficial mutations.

In order to correct for demography and other forces that can
 distort the SFS data, such as linkage, we used the so-called nuisance
 parameters r_i s. These parameters have the potential of
 accounting for more complex scenarios without directly modeling
 the underlying changes in population size, and potentially,
 other events such as migration and admixture. This could prove
 more robust than just allowing for one (or two) population size
 changes, as dfe-alpha assumes. However, we did not test the
 behavior of our method under these more complex scenarios
 and the extent of bias in α they might generate.

In order to infer the full DFE, we used the unfolded site frequency
 spectrum (SFS). This requires the identification of the
 ancestral state, which is prone to errors. The errors in the identification
 of the ancestral state can, for example, be accounted for
 by using a probabilistic modeling of the ancestral state (Schneider
 et al. 2011; Gronau et al. 2013). We chose to assume that
 the polymorphism data is composed of a mixture of sites with
 correctly inferred ancestral state and sites with incorrect ancestral
 state. This approach has proved to be efficient for unbiased
 estimation of GC-biased gene conversion (Glémin et al. 2015),
 a weak selection-like process. Here, we also showed that we
 could capture the errors in the identification of ancestral state
 under a general distribution of fitness effects and, as apposed to
 the expectations of Galtier (2016), that the r_i parameters are
 not sufficient to correct for misidentification of ancestral state.

When using the divergence data in the inference, we cor-

816 rected for mutations that were fixed in the sample but that were,
817 in fact, polymorphic in the population. These mutations would
818 incorrectly be counted into the divergence data. Our correction is
819 different than the one used by [Keightley and Eyre-Walker \(2012\)](#),
820 which is implemented in dfe-alpha. We found that this correc-
821 tion can lead dfe-alpha to predict values of α that are extreme,
822 both on the positive and negative side. Our approach showed a
823 much more consistent behavior throughout the simulations.

824 One drawback of the method presented here is that as the
825 sample size n increases, so does the number of required r_i param-
826 eters. Estimating too many parameters could lead to numerical
827 difficulties in finding the optimum. One might expect that muta-
828 tions present in i copies could be distorted to similar extent as
829 mutations that are present in $i - 1$ or $i + 1$ copies. Using this, the
830 number of r_i parameters could be reduced by allowing different
831 consecutive polymorphism counts to share the same r parameter.
832 A model selection procedure, via LRT or AIC, can then be used
833 to decide on the most adequate grouping of the r parameters.

834 Similar to the r_i parameters, both our approach and methods
835 that use probabilistic modeling to account for the identification
836 of ancestral state rely on that the same process applies to both
837 neutral and selected sites. This is probably not the case, as one
838 could expect that the error in the identification of the ancestral
839 state is different for the sites that are under selection. Theoretically,
840 the neutral and selected sites could each have their own
841 ϵ , but this would most likely not be identifiable. Nonetheless,
842 it would be useful to investigate how robust the inference is
843 when neutral and selected sites have different errors in the iden-
844 tification of the ancestral state. One could also put more effort
845 in reducing the misidentification error when obtaining the un-
846 folded from the folded SFS. Such an approach is pursued by
847 [Keightley et al. \(2016\)](#), where the unfolded SFS is obtained by
848 relying on two, instead of one, outgroup populations.

849 All methods that estimate the DFE require an a priori strict
850 division of sites into neutral and selected classes. This is needed
851 to disentangle the effects of selection from other forces, such as
852 demography and misidentification of the ancestral state. It is
853 expected that real data violates this assumptions, and it is not

854 known to how extent this biases the inference. Similar to the
855 ϵ , one could add a contamination error, ϵ_{con} , with which, the
856 observed neutral data would be modeled as a mixture of truly
857 neutral sites and selected sites. However, this parameter would
858 not be identifiable. It would though be interesting to investigate
859 to what extent violations of this assumption bias the inference.

860 Throughout this paper, we used LRTs for model testing. How-
861 ever, inferences with or without divergence data are not compa-
862 rable through LRT or AIC, or any other similar method (as the
863 data are different). A goodness of fit test could be developed,
864 that would investigate how closely the predicted SFS matches
865 the observed one. This could then be used to decide if diver-
866 gence data should be used in the inference or not.

867 Here, we assumed that selection is additive, where fitness of
868 the heterozygote and derived homozygote are $1 + s$ and $1 + 2s$,
869 and the selection coefficient s is fixed in time. This assumption
870 is made by most methods that infer the DFE and α , though
871 some approaches exist for modeling arbitrary dominance or
872 potentially temporal variation/fluctuations in selection coeffi-
873 cients. [Williamson et al. \(2004\)](#), [Huerta-Sanchez et al. \(2008\)](#) and
874 [Gossmann et al. \(2014\)](#) pursue this in more details, illustrating
875 the need for future development accounting for other types of
876 selection regimes.

877 Our general approach can be applied to a wide range of
878 species where the amount and impact of beneficial mutations
879 on patterns of polymorphism and divergence varies widely (as
880 uncovered by [Galtier \(2016\)](#)). Our method allows to accurately
881 detect if beneficial mutations are present in the data, and a LRT
882 can be used for model reduction, to let the data decide if a full
883 or strictly deleterious DFE should be inferred. Importantly, we
884 also show that estimating a full DFE, and thus learning about
885 the property of beneficial mutations and expected amounts of
886 adaptive substitution, is possible without relying on divergence
887 data.

888 **Availability**

889 The source code is available upon request from PT.

890 **Acknowledgments**

891 We would like to thank Nicolas Galtier for useful early discus-
892 sions and comments on the manuscript. This work has been
893 supported by the European Research Council under the Euro-
894 pean Union's Seventh Framework Program (FP7/20072013, ERC
895 grant number 311341).

Literature Cited

- Arndt, P. F., T. Hwa, and D. A. Petrov, 2005 Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *Journal of molecular evolution* **60**: 748–763.
- Arunkumar, R., R. W. Ness, S. I. Wright, and S. C. Barrett, 2015 The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199**: 817–829.
- Bataillon, T. and S. F. Bailey, 2014 Effects of new mutations on fitness: insights from models and data. *Annals of the New York Academy of Sciences* **1320**: 76–92.
- Bataillon, T., J. Duan, C. Hvilsom, X. Jin, Y. Li, L. Skov, S. Glemin, K. Munch, T. Jiang, Y. Qian, *et al.*, 2015 Inference of purifying and positive selection in three subspecies of chimpanzees (*Pan troglodytes*) from exome sequencing. *Genome biology and evolution* **7**: 1122–1132.
- Bataillon, T., T. Zhang, and R. Kassen, 2011 Cost of adaptation and fitness effects of beneficial mutations in *Pseudomonas fluorescens*. *Genetics* **189**: 939–949.
- Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**.
- Bustamante, C. D., R. Nielsen, and D. L. Hartl, 2003 Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theoretical population biology* **63**: 91–103.
- Charlesworth, B., 2015 Causes of natural variation in fitness: evidence from studies of drosophila populations. *Proceedings of the National Academy of Sciences* **112**: 1662–1669.
- Chevin, L.-M., R. Lande, and G. M. Mace, 2010a Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biol* **8**: e1000357.
- Chevin, L.-M., G. Martin, and T. Lenormand, 2010b Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution. *Evolution* **64**: 3213–3231.
- Durrett, R., 2008 *Probability models for DNA sequence evolution*. Springer Science & Business Media.
- Eyre-Walker, A. and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution* **26**: 2097–2108.
- Eyre-Walker, A., M. Woolfit, and T. Phelps, 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**: 891–900.
- Fay, J. C. and C.-I. Wu, 2000 Hitchhiking under positive darwinian selection. *Genetics* **155**: 1405–1413.
- Francioli, L. C., P. P. Polak, A. Koren, A. Menelaou, S. Chun, I. Renkens, C. M. van Duijn, M. Swertz, C. Wijmenga, G. van Ommen, *et al.*, 2015 Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics* .
- Galtier, N., 2016 Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS genetics* **12**.
- Glémin, S., P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier, and L. Duret, 2015 Quantification of gc-biased gene conversion in the human genome. *Genome research* **25**: 1215–1228.
- Golding, G., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol* **1**: 1X–142.
- Gossmann, T. I., D. Waxman, and A. Eyre-Walker, 2014 Fluctuating selection models and McDonald-Kreitman type analyses. *PloS one* **9**.
- Gronau, I., L. Arbiza, J. Mohammed, and A. Siepel, 2013 Inference of natural selection from interspersed genomic elements

- based on polymorphism and divergence. *Molecular biology and evolution* p. mst019.
- Halligan, D. L. and P. D. Keightley, 2009 Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* **40**: 151–172.
- Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eöry, T. M. Keane, D. J. Adams, and P. D. Keightley, 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet* **9**: e1003995.
- Harris, K. and R. Nielsen, 2016 The genetic cost of neanderthal introgression. *Genetics* **203**: 881–891.
- Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786–2787.
- Hill, W. G., 2010 Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **365**: 73–85.
- Hodgkinson, A. and A. Eyre-Walker, 2011 Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* **12**: 756–766.
- Hoffmann, A. A. and C. M. Sgrò, 2011 Climate change and evolutionary adaptation. *Nature* **470**: 479–485.
- Huerta-Sanchez, E., R. Durrett, and C. D. Bustamante, 2008 Population genetics of polymorphism and divergence under fluctuating selection. *Genetics* **178**: 325–337.
- Jacquier, H., A. Birgy, H. Le Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud, *et al.*, 2013 Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences* **110**: 13067–13072.
- Keightley, P. D., J. L. Campos, T. R. Booker, and B. Charlesworth, 2016 Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. *Genetics* **203**: 975–984.
- Keightley, P. D. and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.
- Keightley, P. D. and A. Eyre-Walker, 2010 What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 1187–1193.
- Keightley, P. D. and A. Eyre-Walker, 2012 Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *Journal of molecular evolution* **74**: 61–68.
- Kimura, M., T. Maruyama, and J. F. Crow, 1963 The mutation load in small populations. *Genetics* **48**: 1303.
- Kliman, R., B. Sheehy, and J. Schultz, 2008 Genetic drift and effective population size. *Nature Education* **1**: 3.
- Kousathanas, A. and P. D. Keightley, 2013 A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* **193**: 1197–1208.
- Loewe, L., B. Charlesworth, C. Bartolomé, and V. Nöel, 2006 Estimating selection on nonsynonymous mutations. *Genetics* **172**: 1079–1092.
- Lourenço, J., N. Galtier, and S. Glémin, 2011 Complexity, pleiotropy, and the fitness effect of mutations. *Evolution* **65**: 1559–1571.
- McDonald, J. H., M. Kreitman, *et al.*, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.
- Messer, P. W. and D. A. Petrov, 2012 The McDonald-Kreitman test and its extensions under frequent adaptation: Problems and solutions. arXiv:1211.0060 [q-bio.PE].
- Nielsen, R., 2005 Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**: 197–218.
- Otto, S. P. and T. Lenormand, 2002 Resolving the paradox of sex and recombination. *Nature Reviews Genetics* **3**: 252–261.
- Otto, S. P. and M. C. Whitlock, 1997 The probability of fixation in populations of changing size. *Genetics* **146**: 723–733.
- Piganeau, G. and A. Eyre-Walker, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proceedings of the National Academy of Sciences* **100**: 10335–10340.

- Racimo, F. and J. G. Schraiber, 2014 Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genetics* **10**.
- Sawyer, S. A. and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. *Genetics* **132**: 1161–1176.
- Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* **189**: 1427–1437.
- Sethupathy, P. and S. Hannenhalli, 2008 A tutorial of the poisson random field model in population genetics. *Advances in bioinformatics* **2008**.
- Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright, 2010 Genome-wide evidence for efficient positive and purifying selection in *capsella grandiflora*, a plant species with a large effective population size. *Molecular biology and evolution* **27**: 1813–1821.
- Smith, N. G. and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022–1024.
- Sousa, A., S. Magalhães, and I. Gordo, 2011 Cost of antibiotic resistance and the geometry of adaptation. *Molecular biology and evolution* p. msr302.
- Strasburg, J. L., N. C. Kane, A. R. Raduski, A. Bonin, R. Michellmore, and L. H. Rieseberg, 2011 Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular biology and evolution* **28**: 1569–1580.
- Welch, J., 2006 Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* **173**: 821–837.
- Welch, J. J., A. Eyre-Walker, and D. Waxman, 2008 Divergence and polymorphism under the nearly neutral theory of molecular evolution. *Journal of molecular evolution* **67**: 418–426.
- Williamson, S., A. Fledel-Alon, and C. D. Bustamante, 2004 Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **168**: 463–475.
- Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences* **102**: 7882–7887.
- Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America* **24**: 253.
- Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11**: 367–372.