# Towards an evolutionarily appropriate null model: jointly inferring demography and purifying selection

Parul Johri[1,*], Brian Charlesworth[2], & Jeffrey D. Jensen[1,*]

[1]School of Life Sciences, Arizona State University, Tempe, Arizona, USA

[2]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

∗ Corresponding authors:

Parul Johri

Life Sciences C, Room 340

427 East Tyler Mall

Arizona State University

Tempe, AZ 85281

pjohri1@asu.edu

Jeffrey D. Jensen

Life Sciences C, Room 340A

427 East Tyler Mall

Arizona State University

Tempe, AZ 85281

Jeffrey.D.Jensen@asu.edu

## ABSTRACT

The question of the relative evolutionary roles of adaptive and non-adaptive processes has been a central debate in population genetics for nearly a century. While advances have been made in the theoretical development of the underlying models, and statistical methods for estimating their parameters from large-scale genomic data, a framework for an appropriate null model remains elusive. A model incorporating evolutionary processes known to be in constant operation - genetic drift (as modulated by the demographic history of the population) and purifying selection – is lacking. Without such a null model, the role of adaptive processes in shaping within- and between-population variation may not be accurately assessed. Here, we investigate how population size changes and the strength of purifying selection affect patterns of variation at neutral sites near functional genomic components. We propose a novel statistical framework for jointly inferring the contribution of the relevant selective and demographic parameters. By means of extensive performance analyses, we quantify the utility of the approach, identify the most important statistics for parameter estimation, and compare the results with existing methods. Finally, we re-analyze genome-wide population-level data from a Zambian population of *Drosophila melanogaster*, and find that it has experienced a much slower rate of population growth than was inferred when the effects of purifying selection were neglected. Our approach represents an appropriate null model, against which the effects of positive selection can be assessed.


**Keywords**: Background selection, demographic inference, distribution of fitness effects, approximate Bayesian computation

## INTRODUCTION

At the founding of population genetics in the early 20th century, Fisher, Haldane, and Wright developed much of the mathematical and conceptual framework underlying the study of population-level processes dictating variation observed within- and between-species. However, as evidenced by decades of published interactions, they held differing views regarding the relative importance of adaptive vs. non-adaptive processes in driving evolution. As pointed out by Crow (2008), these issues were not really resolved, but "rather they were abandoned in favor of more tractable studies." With the advent of the Neutral Theory (Kimura 1968, 1983; King and Jukes 1969; Ohta 1973), the evolutionary importance of stochastic effects due to finite population size, as earlier advocated by Wright, received renewed attention.

In the following decades, further theoretical developments as well as the availability of large-scale sequencing data have validated the important role of genetic drift (Kimura 1983; Walsh and Lynch 2018). However, subsequent research on the indirect effects of selection on patterns of variability at linked neutral alleles has re-ignited previous debates (Kern and Hahn 2018; Jensen *et al.* 2019). In particular, it remains unclear whether the large class of strongly and weakly deleterious variants hypothesized under the Neutral Theory, and their effects on linked neutral sites (background selection, BGS), are sufficient to explain genome wide patterns of variation, or whether a substantial contribution from the effects of beneficial variants on linked neutral sites (*i.e.*, selective sweeps), is required.

The primary difficulty in answering this question stems from our lack of an appropriate neutral null model - that is, a model incorporating genetic drift as modulated by the demographic history of the population, as well as a realistic distribution of fitness effects summarizing the pervasive effects of both direct and indirect purifying selection. Without a model incorporating these evolutionary processes, which are certain to be occurring constantly in natural populations, it is not feasible to quantify the frequency with which adaptive processes may also be acting to shape patterns of polymorphism and divergence.

It can, however, be difficult to distinguish the individual contributions of positive and purifying selection from demographic factors such as changes in population size, as all of these evolutionary processes may leave similar imprints in population genetic data. For example, both purifying selection and population growth can distort gene genealogies of linked neutral sites in a similar fashion (Charlesworth *et al.* 1993; Kaiser and Charlesworth 2009; O'Fallon *et al.* 2010; Charlesworth 2013; Nicolaisen and Desai 2013), and result in a skewing of the site frequency spectrum (SFS) towards rare variants. In fact, demographic

inference is often performed using either synonymous or intronic sites, which are close to sites in coding regions, but the contribution of the effects of selection at linked sites are generally ignored. Patterns of variation in these regions may be skewed by the effects of either negative selection (Zeng 2013; Ewing and Jensen 2016) or positive selection (Messer and Petrov 2013), and this could strongly affect the accuracy of the inferred demographic model (Ewing and Jensen 2016; Schrider *et al.* 2016). In other words, selection may cause demographic parameters to be mis-estimated in such a way that population size changes are over or under-estimated.

In addition, the extent of BGS can vary considerably across the genome. Although it is understood theoretically to be a function of the number and selective effects of directly selected sites, as well as the rate of recombination (Hudson and Kaplan 1995; Nordborg *et al.* 1996; Charlesworth 1996, 2013), the interaction between these parameters and the underlying demographic history of the population remains poorly understood, even for simple models. Furthermore, existing analytical work (Zeng and Charlesworth 2010b; Zeng 2013; Nicolaisen and Desai 2013) has largely been done under the assumption of demographic equilibrium, and is often restricted to describing mutations of large effect. Thus, weak selection effects (on the order of $|2N_e s| < 10$), which are thought to be common, may not be well captured by these predictions. Furthermore, in regions of low crossing over, interference between this class of mutations may result in even greater distortions of the underlying genealogies (Kaiser and Charlesworth 2009; O'Fallon *et al.* 2010; Good *et al.* 2014).

We first investigate the joint effects of demography, the shape of the distribution of fitness effects (DFE) of deleterious mutations, and the number of selected sites in shaping linked neutral variation. Next, we utilize the decay of background selection effects, by examining regions spanning coding / non-coding boundaries to jointly infer the DFE of the coding region and the demographic history of the population. By performing extensive performance analyses, quantifying both power and error associated with this approximate Bayesian (ABC) approach (Beaumont *et al.* 2002), the method is shown to perform well across arbitrary demographic histories and DFE shapes. Importantly, by utilizing patterns of variation and divergence across coding and non-coding boundaries, this approach avoids the assumption of synonymous site neutrality inherent to MK-style approaches - an assumption that has been shown to be strongly violated in many organisms of interest (Chamary and Hurst 2005; Lynch 2007; Zeng and Charlesworth 2010a; Lawrie *et al.* 2013; Choi and Aquadro 2016; Jackson *et al.* 2017) and which can result in serious mis-inference (Matsumoto *et al.* 2016). In applying this approach to genome-wide data from a Zambian

population of *Drosophila melanogaster*, results show that the Zambian population has experienced a very mild 1.2-fold growth, considerably less than previous estimates which did not account for the BGS-induced skew of the SFS. In addition, we estimate that ~25% of all mutations in exons are effectively neutral in this population, and we find little evidence for wide-spread selection on synonymous sites.

## METHODS

***Simulations:*** SLiM 3.1 (Haller and Messer 2019) was used to simulate a functional element of length $L$, which is flanked by neutral non-functional regions. The functional region experiencing purifying selection is given by a DFE that is modeled as a discrete distribution with four bins (Figure 1a) representing effectively neutral ($|\gamma| < 1$), weakly deleterious ($1 <= |\gamma| < 10$), moderately deleterious ($10 <= |\gamma| < 100$), and strongly deleterious ($100<=|\gamma| >= 10000$) classes of mutations, where $\gamma=2N_e s$, and $s$ is the selection coefficient for homozygous mutations. Semi-dominance is assumed, so that the fitness of mutant heterozygotes is exactly intermediate between the values for the two homozygotes (a dominance coefficient, $h$, of 0.5). Fitness effects are assumed to follow a uniform distribution within each of the four bins. In order to infer the extent of purifying selection, we estimated the fraction of mutations in each bin, referred to as $f_0, f_1, f_2$ and $f_3$, respectively (Figure 1a), such that $0 <= f_i <= 1$, and $\Sigma_i f_i = 1$, for $i = 0, 1, 2$, and 3. In addition, in order to limit the computational complexity, we restricted values of $f_i$ to multiples of 0.05 (*i.e.*, $f_i \in \{0.0, 0.05, 0.10 \ldots 0.95, 1.0\} \forall i$). These constraints allowed us to sample 1,771 different DFE realizations, and to work independently of any arbitrary assumption regarding DFE shape.

***Simulations under demographic equilibrium***: Simulations were performed for 4 different values of $L$ – 0.5 kb, 1kb, 5kb, and 10kb. The intergenic regions were assumed to be 10kb and simulations were restricted to the intergenic region on one side of the functional region. For the purpose of power analyses and testing, we used population-genetic parameter values that approximately resemble *Drosophila* populations. Population size was assumed to be $10^6$ and the recombination rate ($1 \times 10^{-8}$ per site per generation) and mutation rate ($1 \times 10^{-8}$ per site per generation) were constant across the simulated region. Although we have not included gene conversion in this study, it will be an important addition in future studies. The

simulations were performed with 5000 (=$N_{sim}$) diploid individuals and the recombination and mutation rates were scaled proportionally to maintain realistic values of $N_e s$.

We used a burn-in period of 80,000 generations, and an additional 20,000 (=$4N_{sim}$) generations were allowed for neutral evolution. For every set of parameter combination (*i.e.*, $f_0, f_1, f_2$ and $f_3$) we performed 1000 replicate simulations, and summarized both the mean and variance of common summary statistics to perform the subsequent ABC analysis.

***Simulations under non-equilibrium demography***: Simulations with demographic changes were performed specifically to match the details of the *D. melanogaster* genome. A set of 94 exons belonging to the *D. melanogaster* genome were chosen according to certain criteria (see Results). For each exon, simulations were performed using the length of the exon together with 4 kb of flanking intergenic sequence. The mutation rate was conservatively assumed to be $3.0 \times 10^{-9}$ per site per generation (Keightley *et al.* 2014) although somewhat higher mutation rates have been estimated in other studies (Schrider *et al.* 2013; Assaf *et al.* 2017). Ancestral and current population sizes were sampled from a uniform prior between $10^5$-$10^7$ and $f_i \in \{0.0, 0.05, 0.10 \ldots 0.95, 1.0\}$ such that $\Sigma_i f_i = 1$, for $i = 0$, 1, 2, and 3. Nucleotide diversity at 4-fold degenerate sites was found to be 0.019 for the Zambian population of *D. melanogaster*, which would give an estimate of $N_e$ of $1.6 \times 10^6$. A scaling factor of 320 corresponding to $N_e/N_{sim}$ (=$1.6 \times 10^6$ / 5000) was used to perform all simulations with demographic changes. A total of 10 replicates were performed for each exon, resulting in 940 replicates for every parameter combination. These simulations were conducted using the computational resources of Open Science Grid (Pordes *et al.* 2007; Sfiligoi *et al.* 2009).

***Calculation of summary statistics***: First, we fit a logarithmic function to the recovery of nucleotide diversity ($\pi$) around the functional region such that $\pi = slope * \ln(x) + intercept$, where $x$ is the distance of the site from the functional region in base pair. We used the *slope* and *intercept* of the fit to define the number of bases required for a 50%, 75%, and 90% recovery of nucleotide diversity, with 50% and below being defined as the "linked neutral" region and the 50% and above as the "neutral" region. This analysis provides for three non-overlapping regions: (1) functional (experiencing direct selection), (2) linked-neutral (experiencing observable levels of background selection), and (3) neutral (experiencing low / unobservable levels of background selection). The following statistics were calculated for

each of these three types of regions: nucleotide diversity ($\pi$), Watterson's $\theta$, Tajima's $D$, Fay and Wu's $H$ (both absolute and normalized), number of singletons, haplotype diversity, LD-based statistics ($r^2$, $D$, $D$'), and divergence (*i.e.*, number of fixed mutations per site per generation). Simulations for any particular set of parameters were run with 1000 replicates and the mean and variance of the above statistics across replicates were used as summary statistics for ABC. In addition to these variables, six statistics summarizing the characteristics of the recovery of $\pi$ in linked neutral regions were also included as summary statistics. Specifically, $\pi = slope*\ln(distance)+intercept$ was fit and slope, intercept, maximum value of $\pi$, and number of bases required for 50%, 75%, and 90% recovery were calculated and included as summary statistics. Together, these amount to 72 initial summary statistics. All statistics were calculated using the Python package pylibseq (Thornton 2003). The sample size was kept constant at 100 genomes (*i.e.*, 50 diploid individuals). It should be noted that some statistics are strongly dependent on the number of sites used in the calculations, and the size of linked and neutral regions varied for every set of parameter combination, though this effect is captured in the individual prior distributions.

**ABC**: We used an approximate Bayesian (ABC) approach, using the R package, "abc"(Csilléry *et al.* 2012), to co-estimate the DFE characterizing a functional region, as well as the population history characterizing the population in question. We used linear regression (aided by neural net to handle non-linearity) as well as ridge regression to infer posteriors, with a tolerance of 0.05 (*i.e.*, 5% of the total number of simulations are accepted by ABC to estimate the posterior probability of each parameter), a cross-validation set of 100 simulations, and the weighted median as point estimates.

**Ranking of summary statistics**: Ranking of summary statistics was performed separately for both demographic equilibrium and non-equilibrium cases, using two different methods. The first approach consisted of performing Box-Cox transformations on all 72 summary statistics to correct for non-linear relations between statistics and parameters. The squared correlation coefficient, $r^2$, between the transformed statistics and parameters was then used to rank each statistic for every parameter separately and a statistic was considered to be significantly correlated with the parameter if the $p$-value was less than 0.05 (Bonferroni corrected for multiple testing). The second approach involved a modified version of the algorithm proposed by Joyce and Marjoram (2008) for ranking statistics. With this algorithm, we

started with the entire set of 72 statistics. Every statistic was removed from the set and cross-validation using 20 randomly sampled simulations was used to identify the statistic that corresponded to the least error (*i.e.*, the removal of which causes the least reduction in accuracy). The same algorithm was performed iteratively until only two statistics remained. This method was performed for each parameter separately, was replicated 10 times, and the average ranking across these replicates was used to obtain the final ranking. The second approach was extremely time consuming and was thus only used to rank the statistics to infer the DFE under demographic equilibrium.

**Comparison with DFE-alpha:** Simulations were performed under demographic-non-equilibrium models, with 100 replicates of 94 exons each, and ancestral population sizes of 10,000 for all. Functional regions were simulated with 30% neutral sites, which were used to calculate the neutral SFS required by DFE-alpha. *Est_dfe* (Schneider *et al.* 2011) was used on unfolded SFS to perform demographic inference and to infer the deleterious DFE. The proportion of adaptive mutations was fixed at 0.0. Final estimates of the DFE were obtained as $N_w s$ where $N_w$ is the weighted population size inferred by *est_dfe*.

***Drosophila* data application**: Release 5 of the *D. melanogaster* genome assembly (Hoskins *et al.* 2007) and annotation version 5.57 were used, downloaded from ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.57_FB2014_03/gff/. Crossing over rates estimated by Comeron *et al*. (2012) for every exon and flanking intergenic region were obtained from the *D. melanogaster* Recombination Rate Calculator (https://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl) (Fiston-Lavier *et al.* 2010), and explicitly utilized for each specific region considered. These rates were halved to obtain sex-averaged rates of recombination (Campos *et al.* 2017) as all regions were restricted to autosomes. We excluded all genes that have a crossing over rate 5-fold larger or smaller than the average (*i.e.,* we used only genes with a crossing over rate of between 0.44 and 11 cM/Mb). Consensus sequences of all Zambia lines were downloaded from http://www.johnpool.net/genomes.html (Lack *et al.* 2015). IBD tracks and admixture tracks were masked using scripts provided by the same site. Individuals with any known inversions were entirely excluded from the analysis (Kapopoulou *et al.* 2018).

The final set consisted of 76 haploid genomes. PhastCons scores calculated with respect to 15 insect taxa were downloaded from the UCSC genome browser (https://genome.ucsc.edu/). For each of the 94 exons, summary statistics were calculated

using pylibseq (Thornton 2003) for the coding region and for 2kb intergenic regions flanking both sides. In order to exclude sites in intergenic regions that might be under direct selection, a phastCons cutoff score of 0.8 was used to calculate all statistics. That is, sites that had a greater than or equal to 80% probability of being a conserved noncoding element identified by phastCons, were excluded when calculating statistics.

For the purpose of obtaining derived alleles and for calculating branch-specific rates of substitution, we used the ancestral sequence to the *D. melanogaster* genome provided to us by the authors of Kolaczkowski *et al.* (2011). The ancestral sequence reconstruction had been performed by maximum likelihood over 15 insect genomes available in the UCSC genome browser (Karolchik *et al.* 2004). Sites with missing ancestral sequence were excluded from analysis. Branch-specific rates of substitution (also referred to as divergence in this study) were calculated by identifying derived alleles that were fixed in the *D. melanogaster* Zambian population (*i.e.*, polymorphic sites were removed). After excluding sites with missing ancestral information, with IBD and admixture tracks, and which were likely to belong to a non-coding conserved element, we had on average 1062 sites per exon, 556 sites per linked region, and 666 sites per neutral regions.

It should be noted that for the purpose of performing inference using ABC, substitution rates in simulations were calculated per base pair for 25,000 generations. We thus normalized all rates obtained from simulations by the expected neutral substitution rate (*i.e.*, $\mu_{sim}t_{sim} = 320 \times 3\times10^{-9} \times 25000 = 0.024$, where $\mu$ is the mutation rate and $t$ is the number of generations). Divergence estimates from *D. melanogaster* were normalized by an expected neutral substitution rate of $\mu t = 3 \times 10^{-9} \times 21333333$ (the estimated divergence time) = 0.064. In addition, inference was performed using divergence estimates only in the exonic regions. ABC inference for *Drosophila* was performed using the abc package in R, with linear regression aided by neural net with default parameters. Each inference was performed 50 times, and the mean of point estimates obtained were reported as the final estimates of parameters.

**Data and code availability**

The following data will be made publicly available upon acceptance of the manuscript on https://github.com/paruljohri/BGS_Demography_DFE. 1) Aligned sequences of the single-exon genes and their corresponding intergenic regions used in this study, including derived alleles and fixed substitutions. 2) Scripts to calculate statistics from simulations and from

empirical data as well as the code used to perform simulations. 3) Values of all calculated statistics obtained for all parameter combinations.

## RESULTS AND DISCUSSION

**Recovery of nucleotide diversity as predicted under equilibrium:** The nucleotide site diversity ($B$) at neutral sites with linkage to sites experiencing direct purifying selection can be obtained by modifying Equation 6 of Nordborg *et al*. (1996), which is of the form

$$B = \frac{\pi}{\pi_0} \sim \exp\left[-\int\int E(t,z)dzdt\right]$$

where $\pi_0$ is the nucleotide diversity without selection and $\pi$ is the nucleotide diversity with background selection effects. The exponent $E$ is a function of the distribution of heterozygous selection coefficients ($t = hs$) for deleterious mutations and the physical distance ($z$) between the neutral and selected sites. Here, $s$ is the selection coefficient, and $h$ is the dominance coefficient.

For the purpose of the current study, Equation S1a of the SI of Campos and Charlesworth (2019) was modified to model a neutral site outside a gene, and which is a distance $y$ basepairs from the end of the functional region. If the position of a selected site is a distance $x$ basepairs from the end (in the opposite direction), the distance between the two sites is $z = x + y$. The basic equation for the exponent of the BGS function for a given selection coefficient, $E(t)$, was obtained as follows:

$$E(t) = \frac{Ut}{l}\int_o^l \frac{\mathrm{d}x}{[t + (g + r_c y)(1-t) + r_c x(1-t)]^2} \qquad (1)$$

where $U(t)$ is the total mutation rate to deleterious alleles over the entire gene, $l$ *is* the length of the gene in basepairs, $g$ *is* the rate of gene conversion, and $r_c$ is the rate of crossing over per basepair. The crossover map is assumed to be linear, so that the net rate of recombination between the two sites is $g + r_c z$, and $z$ is assumed to be sufficiently large that the effect of gene conversion is independent of $z$.

On integrating Equation 1 with respect to $x$ between 0 and $l$ (see the Appendix for details), the exponent $E$ as a function of the length of selected sites and the distance from the functional element can be obtained:

$$E(t) = \frac{Ut}{r_c l(1-t)}\left\{\frac{1}{[t+(g+r_c y)(1-t)]} - \frac{1}{[t+(g+r_c y)(1-t)+r_c l(1-t)]}\right\}$$

$$= \frac{Ut}{[t+(g+r_c y)(1-t)][t+g(1-t)+r_c(y+l)(1-t)]} \qquad (2)$$

Note that the above equation shows that, if $t$ is small compared with $y$, BGS effects outside the coding region will be minimal.

We can integrate E($t$) over the distribution of selection coefficients as described in the Appendix. The expectation of $E(t)$ is given by the sum of the following two terms:

$$U[r_c l(1-a)]^{-1}\left\{1 + a[(1-a)(t_{i+1}-t_i)]^{-1}\ln\left[\frac{a+(1-a)t_i}{a+(1-a)t_{i+1}}\right]\right\} \qquad (3a)$$

$$-U[r_c l(1-b)]^{-1}\left\{1 + b[(1-b)(t_{i+1}-t_i)]^{-1}\ln\left[\frac{b+(1-b)t_i}{b+(1-b)t_{i+1}}\right]\right\} \qquad (3b)$$

where $a = g + r_c y$ and $b = g + r_c(y + l)$ and the $t_i$'s correspond to the boundary of the discrete bins. For the case when $b \ll 1$, the sum of the two components is approximately equal to:

$$U(t_{i+1}-t_i)^{-1}\ln\left[\frac{b+t_{i+1}}{b+t_i}\right] \qquad (3c)$$

Figure 1b shows the theoretical results as well as those from simulations, for $r = 10^{-6}$, $l$ =1000, $U = l\mu$, $\mu = 10^{-6}$, $g = 0$, $t_0 = 0$, $t_1 = 0.00005$, $t_2 = 0.0005$, $t_3 = 0.005$, and $t_4 = 0.5$. It should be noted that these derivations assume that $N_e t \gg 1$, which is violated by the presence of the weakly deleterious DFE class ($f_1$). Most studies deal with this assumption by ignoring the contribution of mutations with $N_e t < 5$ or 10 (Charlesworth 2013; Elyashiv *et al.* 2016; Torres *et al.* 2019). As expected, we found a significant discordance between the simulated and theoretically predicted values for the slope of the recovery of diversity as $f_1$ increases

(Figure 2c and 2d, Supp Table 1). On including only mutations with $N_e t > 2.5$, the diversity patterns are mostly well explained, even when the DFE is highly skewed towards the weakly deleterious class. In fact, it is interesting to note that a combination of high values of $f_1$ and $f_2$ can result in BGS effects that extend up to ~4 kb, even for very short exons, although the maximum reduction in diversity is around 10-15% (consistent with Charlesworth 2012, Campos and Charlesworth 2019b) .

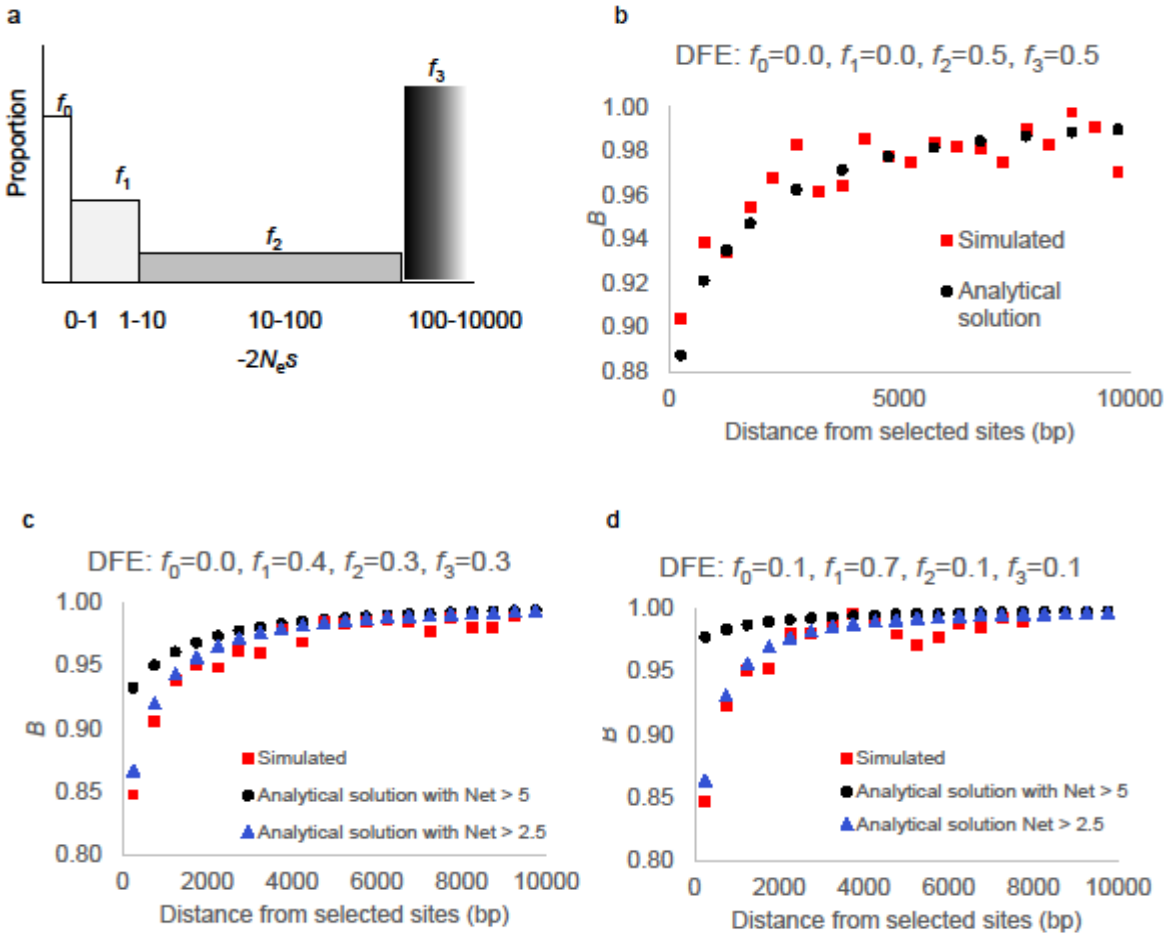



**Figure 1:** (a) An example of a discrete DFE with four classes of mutations. The proportion of each class of mutation, $f_i$, lies between 0 and 1. (b) Nucleotide site diversity relative to the neutral expectation ($B = \pi/\pi_0$) as a function of the distance from the directly selected sites (length 1 kb), as predicted by the analytical solution (black points) and observed in simulations (red points). (c, d) Analytical predictions and simulated values for a DFE with larger contributions from the weakly deleterious class of mutations. Note that, for the analytical solutions, the two classes of results represent cases where mutations with $2N_e t < 5$ (black circles) and $2N_e t < 2.5$ (blue triangles) were ignored.

## Joint effects of demography, the DFE and the number of selected sites on linked neutral variation

While the above results show that the effect of BGS on linked neutral regions can be determined analytically, there are several reasons for investigating background selection effects using simulations. First, the analytical expressions neglect the contribution of very weakly deleterious mutations ($|N_e t| < 2.5$), and do not predict the skew in the SFS. In addition, they assume demographic equilibrium, which is probably not true of natural populations.

***Effects of the shape of the DFE and number of selected sites:*** We first simulated 10kb neutral regions linked to functional regions of varying sizes, 0.5kb, 1kb, 5kb, and 10kb, assuming demographic equilibrium, as shown in Figure 2. By varying the contributions from each bin of selective effects, denoted by $f_0, f_1, f_2$ and $f_3$, it was possible to sample all possible DFE shapes, as described in the Methods section. As expected from equation 3c, the reduction in diversity is non-linearly proportional to the number of selected sites for a given recombination rate. A larger number of selected sites increases both the total reduction in diversity and the slope of the recovery of diversity away from functional regions (Supp Figure 1). The maximum reduction in diversity in the linked neutral regions (immediately adjacent to the functional region), averaged across all DFE realizations, is approximately 8%, 12%, 24%, and 29% for 0.5kb, 1kb, 5kb, and 10kb, respectively. Furthermore, for the chosen recombination rate, the median numbers of base pairs necessary to achieve a 50% recovery in diversity are 955, 1035, 1350, and 1650 bp, respectively, (Figure 2a).

The reduction of nucleotide diversity at closely linked neutral regions was maximized when the proportion of weakly deleterious mutation ($f_1$) and moderately deleterious mutations ($f_2$) was largest (Figure 2b, Supp Table 2). The effect is maximized when purifying selection is weak, allowing mutations to segregate in the population prior to being purged (Campos *et al.* 2017). Although weakly deleterious mutations ($f_1$) only reduce variation slightly, they generate significant distortions in the SFS (Figure 2c), consistent with previous studies (Nordborg *et al.* 1996; Charlesworth 2012; Nicolaisen and Desai 2013). Moderately deleterious mutations ($f_2$) result in the largest reduction in $\pi$, the highest rate of recovery of $\pi$ around functional regions, and the largest skew in the SFS towards rare variants. As expected, the proportion of strongly deleterious mutations ($f_3$) does not greatly affect levels of linked neutral variation, and these mutations skew the SFS only slightly. Further, increasing the number of selected sites results in larger background selection effects for all DFE types,

as is to be expected. It should be noted that these generalizations of BGS effects are dependent on how far from the selected sites the measurements are being considered. For instance, the distance affected by deleterious mutations is expected to be an increasing function of the size of their fitness effects. As we were interested in understanding BGS effects caused by all classes of mutations, we focus our discussion to sites closer to the functional boundary, where all classes of mutation are likely to have an impact.

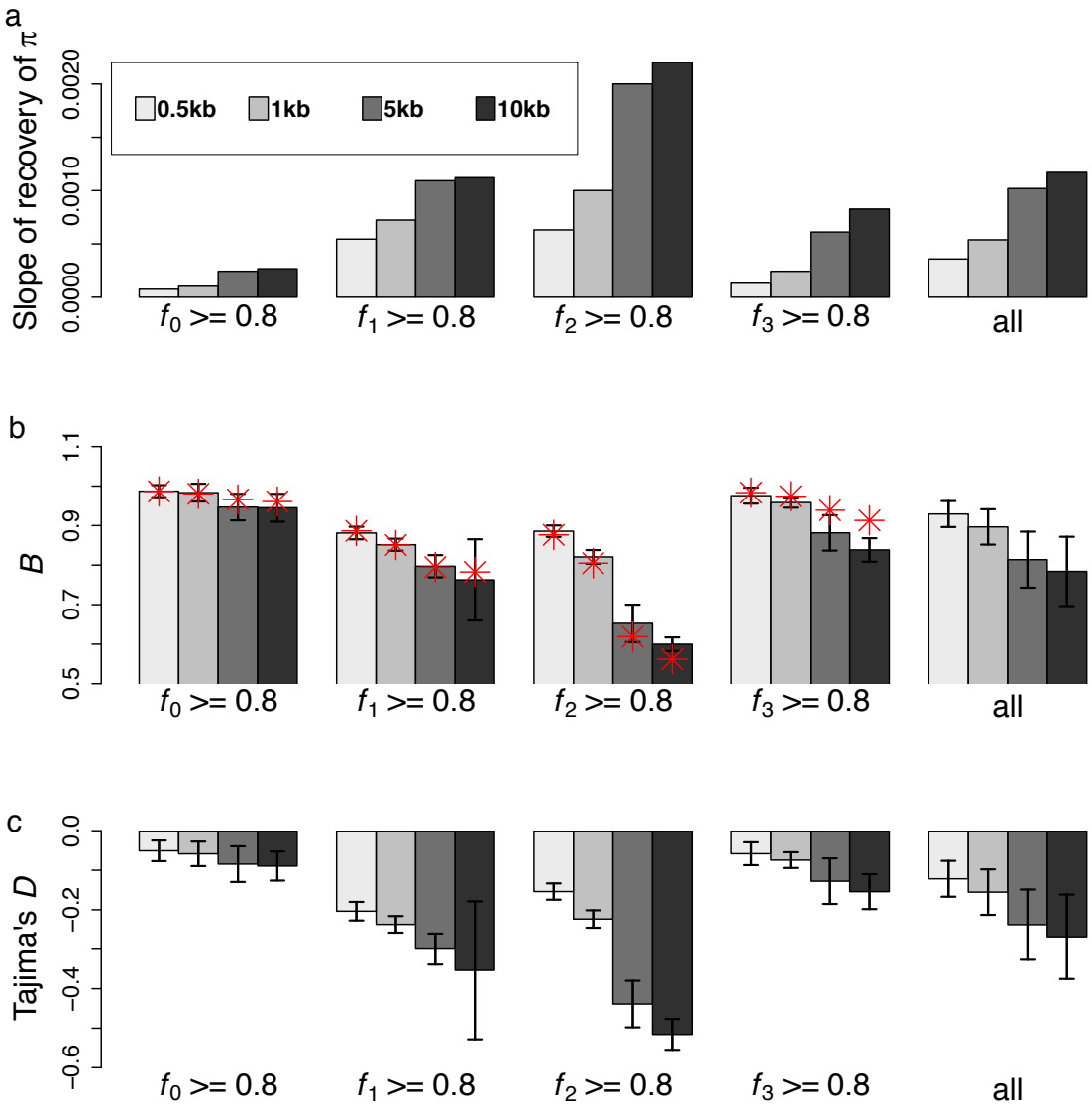



**Figure 2**: Effects of BGS under demographic equilibrium. (a) The slope of the recovery of nucleotide diversity in 10kb linked neutral regions flanking functional regions fitted such that $\pi = slope * \ln(distance\ from\ functional\ region) + intercept$, (b) nucleotide diversity in 500bp linked neutral regions flanking functional regions relative to neutral expectation, and (c) Tajima's $D$ for 500bp linked neutral region flanking functional regions. All of the above are shown for various sizes of functional elements (0.5-10 kb) and DFE shapes. The four DFE shapes considered are $f_i >= 0.8$ for $i$=0,1,2,3, where more than 80% of mutations reside in DFE class $f_i$, such that $\sum f_j = 0.2$, where $j \neq i$. The DFE category "all" represents an average over

all possible DFE shapes. The error bars are 2 × standard deviation. Red points show the analytical prediction for (1) $f_0$=0.85, $f_1$=0.05, $f_2$=0.05, $f_3$=0.05, (2) $f_0$=0.05, $f_1$=0.85, $f_2$=0.05, $f_3$=0.05, (3) $f_0$=0.05, $f_1$=0.05, $f_2$=0.85, $f_3$=0.05, and (4) $f_0$=0.05, $f_1$=0.05, $f_2$=0.05, $f_3$=0.85.

To summarize, at demographic equilibrium, neutral regions linked to functional regions under selection undergo a reduction in diversity and a skew in the site frequency spectrum, both of which depend on the underlying shape of the DFE and the number of directly selected sites (Charlesworth *et al.* 1993; Charlesworth 2013; Campos and Charlesworth 2019). Importantly for the sake of statistical inference, however, the three classes of deleterious mutation ($f_1, f_2, f_3$) behave differently, suggesting the possibility of distinguishing their relative contributions (as discussed in the next section). Furthermore, these results again demonstrate the potentially important role of BGS in shaping patterns of neutral variation, highlighting the danger posed by ignoring these effects when performing demographic inference (see Ewing and Jensen 2016). Additionally, the dramatic difference in the extent of background selection effects as a function of the number of directly selected sites strongly implies the necessity of directly modeling exon sizes in any empirical application.

***Effects of demography and the shape of the DFE on background selection***: We investigated the effects of BGS after recent changes in population size. Populations with the same ancestral population size ($N_{anc}$) either experienced 10-fold exponential growth or contraction in the last $4N_{anc}$ generations and BGS effects were compared to populations that remained in equilibrium throughout, for all possible DFE shapes. Both expansion and contraction result in reduced BGS effects (*i.e.*, there is an increase in $B$ compared to equilibrium), irrespective of the shape of the DFE (Figure 3a, b). This observation suggests that the extent of BGS caused by functional elements may not only be determined by the strength of selection, but rather also by the demographic history of the population. Thus, demographic effects may in principle explain variable inferences among studies of the importance of purifying selection in shaping genome-wide patterns of variation (Cutter and Payseur 2013).

Interestingly however, there is still a significant skew in the SFS at linked neutral sites caused by BGS after a population size change (Figure 3c-e). Thus, in more compact genomes, where background selection is pervasive, this suggests that methods which use the SFS to fit

demographic models may over-estimate growth and either under-estimate population contraction or mis-classify contraction as expansion. It is also interesting to note that BGS effects are largest under demographic equilibrium, such that constant population size is likely to be inferred as population growth.

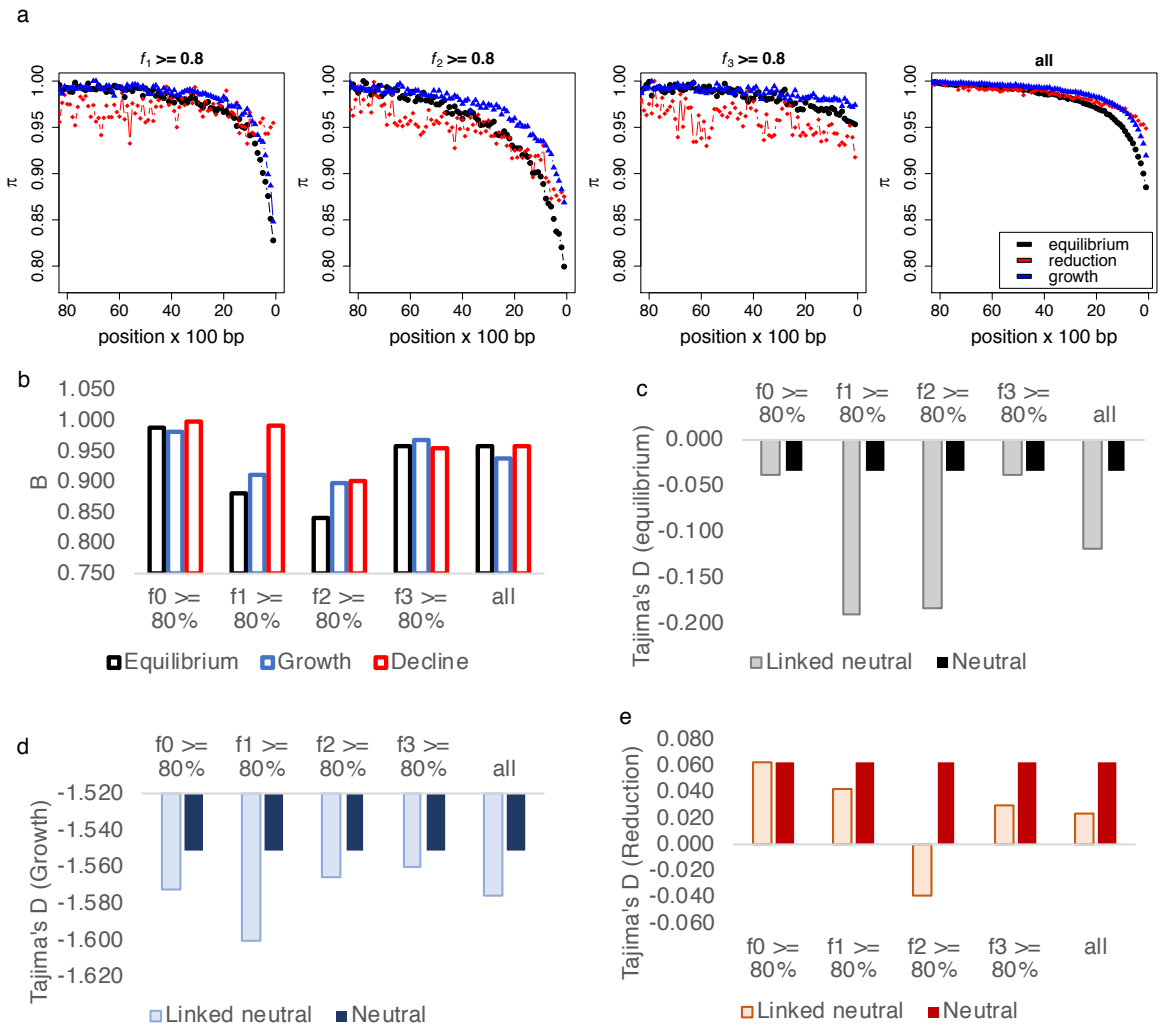



**Figure 3**: Effects of BGS under non-equilibrium demography. (a) The slope of recovery of nucleotide diversity in linked neutral regions for different DFE shapes under equilibrium demography (black), population expansion (blue), and contraction (red). (b) Nucleotide site diversity relative to neutral expectation over 500 bp of linked neutral regions flanking functional regions, for varying DFE shapes and three different demographic models - equilibrium (black), 10-fold exponential expansion (blue), and 10-fold exponential decline (red). (c) Tajima's *D* for the 500 bp linked neutral region flanking the functional region under equilibrium, (d) after a 10-fold expansion, and (e) after a 10-fold population size reduction. The four DFE shapes considered in all panels are $f_i$ >= 80% for $i$=0,1,2,3, where more than 80% of mutations reside in DFE class $f_i$. The DFE category "all" represents an average over all possible DFE shapes. For non-equilibrium demography, $\gamma=2N_{anc}s$, where $N_{anc}$ is the ancestral population size.

**Inference of the DFE under demographic equilibrium**

The next question we investigated was whether the parameters of the DFE can be estimated using the set of summary statistics described in the Methods section. We first determined whether it is possible to distinguish the four different classes of the DFE under demographic equilibrium, using population genomic data and divergence from the closest outgroup species. The simulations involved functional regions of lengths $L$ = 0.5kb, 1kb, 5kb and 10kb, with linked neutral regions of 10kb and a discrete DFE as described previously. An approximate Bayesian (ABC) approach was implemented to quantify our ability to infer the four DFE parameters. The recovery of nucleotide diversity over linked neutral regions was used to calculate the number of bases ($\pi_{50}$) required for diversity to recover to 50% of its maximum value observed (see Methods). The linked neutral region within $\pi_{50}$ base pairs from the functional region was defined as "Linked", and the remainder was defined as "Neutral" (Figure 4a). Statistics were calculated for three regions (Functional, Linked, and Neutral) separately and the means and variances across simulation replicates of each statistic were used to infer the four parameters. The simulation replicates signify independent loci in a genome. In the following sub-sections, we describe the performance of the method and its robustness to various model violations.

***Accuracy of inference*:** All four DFE classes were estimated fairly accurately when using all statistics (Supp Figure 2a). However, under demographic equilibrium, the DFE is inferred much more accurately using statistics from the functional regions alone, thus side-stepping the need for the identification of linked neutral regions (Figure 4b, Supp Figure 2b). In both cases, the accuracy of inference is highest for the neutral class and lowest for $f_2$ (*i.e.,* for moderately deleterious mutations), and improves significantly when the size of the functional region increases (Supp Figure 2). While using only functional regions to perform inference, the absolute difference between the true value and the estimated value of the neutral class is approximately 0.034, 0.030, 0.017, and 0.010 for functional sizes of 0.5kb, 1kb, 5kb and 10 kb. That is, for 1kb regions the method cannot distinguish whether the neutral class of mutations comprise 30% or 33% of the DFE. For the moderately deleterious class this error is larger – 0.077, 0.060, 0.028, and 0.019, respectively. These absolute error values are not surprising, as the $f_i$ in our simulations are multiples of 0.05 out of computational necessity. The accuracy of the estimates can thus can be increased by sampling the parameter space more densely. The accuracy of estimation can also be evaluated using $r^2$ between the true and

estimated values. For instance, for 1kb functional regions, the $r^2$ values for $f_0, f_1, f_2$ and $f_3$ are 0.93, 0.91, 0.89, and 0.87 respectively.

It should be noted that this approach does not distinguish between non-synonymous and synonymous mutations. Indeed, no assumption is made regarding which specific bases are neutral, nearly neutral, or deleterious in the coding region. Thus, this method can be used to estimate the DFE for any type of functional region, as well as to assess the non-neutrality of synonymous sites by comparing their frequency in a given coding region with the occupancy of the $f_0$ class.

***Effect of mis-specification of exon size and recombination rate:*** In view of these results, it is important to consider if accurate estimates depend on correctly specifying precise exon size, or whether it would be sufficient to generate priors assuming, for example, a mean exon length characterizing a genome. To quantify this effect, simulated data sampled from the priors was based on 1kb exons, while the test data were obtained from simulations based on alternative exon sizes. The error in inference of the DFE increases as the difference between exon sizes of the priors and that of the true sizes are increased (Supp Figure 3), with the highest in the moderately deleterious class ($f_2$), although when exon sizes are sufficiently large, mis-specification of exon-size does not strongly impact performance. A similar approach was used to determine if the presence of another functional region (also 1kb in size) separated by an intron or intergenic region would skew inference. As expected, smaller intron sizes result in stronger mis-inference than larger ones, and intronic/ intergenic sizes larger than 4 kb performed essentially as well as those with no nearby functional exon (Supp Figure 4). Moreover, a two-fold difference between assumed and actual recombination rates resulted in inflation of error dramatically (Supp Figure 5 and 6). Informatively, the direction of bias generated differs by DFE class (Supp Figure 6). For example, when true recombination rates are half of those assumed, the inferred weakly deleterious class is greatly inflated. As this class of mutations most strongly skews the linked neutral SFS, this mis-inference presumably arises from an attempt to fit stronger linked effects by inferring a higher proportion of mutations in this class, whereas in reality the increased BGS effects are being generated by fewer recombination events than are assumed.

These results highlight the importance of taking into account the specific exonic-intronic-intergenic structure of a particular genomic region of interest, nearby functional regions and the specific recombination rate. Although any configuration of these details may

be directly simulated, an alternative approach is simply to group exons of like size across a genome, and further reduce these to a group that is devoid of neighboring functional regions.

**Joint inference of purifying selection and demography, under non-equilibrium conditions**

Based on the above results demonstrating that details of exon sizes and recombination rates are essential for accurate inference, we explicitly modeled both exon sizes and recombination rates when examining our ability to jointly infer demographic changes along with the DFE. As our example involved an African population of *D. melanogaster*, we chose single-exon genes that had more than 4kb non-coding regions flanking both sides and whose exon sizes were between 500-2000bp. For this specific set of 94 exons, we simulated functional regions with precise exon sizes linked to 4kb neutral regions and utilized the previously inferred local recombination rate for each exonic region in question. For every parameter combination, we performed 10 replicates of each of the 94 exon sizes (resulting in a total of 940 replicates per parameter combination), with their respective recombination rates and exon sizes, and summarized the resulting mean and variance of summary statistics.

Models of exponential population size expansion and contraction assumed various ancestral population size ($N_{anc}$) and current population size ($N_{cur}$), which were both sampled uniformly between $10^5 - 10^7$, as in previous studies (Duchen *et al.* 2013; Arguello *et al.* 2019). As earlier work has inferred the duration of the expansion in Zambian populations to be of the order of $\sim N_e$ generations, the time duration was scaled down and fixed to $N_{sim}$ (=5000 generations) in order to attempt to infer both historical and current population sizes. Thus, for this framework, we evaluated the estimates of six parameters: $f_0$, $f_1$, $f_2$, $f_3$, $N_{anc}$, and $N_{cur}$.

***Accuracy of joint inference:*** Encouragingly, the results demonstrated an ability to successfully co-estimate the DFE and both ancestral and current population sizes, using the set of coding and linked non-coding summary statistics described above (Figure 4c). Under non-equilibrium demography, the error in inference of the strongly deleterious class of mutations is larger. The absolute differences between true and estimated values were 0.019, 0.027, 0.033, and 0.034 for the four DFE classes, respectively; the errors in ancestral and current sizes were 10.1% and 7.3% respectively. The $r^2$ between the true and estimated values of $f_0$, $f_1$, $f_2$, $f_3$, $N_{anc}$, and $N_{cur}$ were 0.97, 0.97, 0.95, 0.95, 0.99, and 0.99, respectively.

Nonetheless, the performance of the full 6-parameter estimation procedure is good, without relying on the usual step-wise approach of first utilizing putatively neutral sites to infer a demographic history, and then fixing that demographic history in order to estimate DFE parameters. Interestingly, joint estimation is almost as accurate when using statistics only from functional regions (Figure 4d), although it inflates the errors in the estimates of $f_2$ and $f_3$. The absolute differences between the true and estimated values of $f_0, f_1, f_2, f_3$ were 0.015, 0.025, 0.054, and 0.049, respectively, while the error in estimates of population sizes increases to 23% and 8% for $N_{anc}$ and $N_{cur}$, respectively. Although the error in ancestral population size is quite large if only functional regions are used to co-estimate all six parameters, the accuracy of inference can be improved significantly by adding more replicate simulations of each parameter set to the ABC framework.

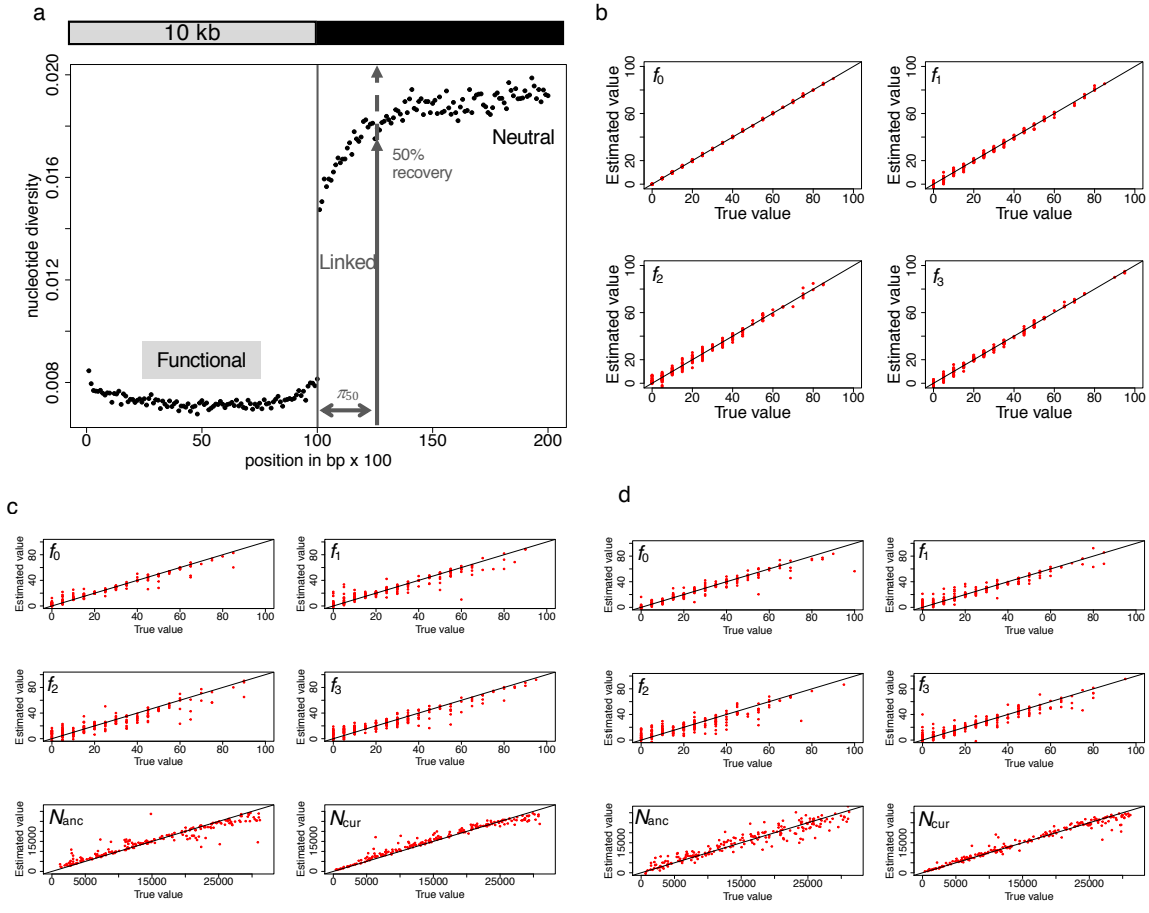



**Figure 4**: (a) Calculation of summary statistics across functional, linked and neutral regions. (b) Accuracy of estimation (cross validation) of the four classes of the DFE using statistics for functional regions only (size 1kb), under equilibrium demography. (c) Joint estimation of population size changes and the DFE using all statistics. (d) Joint estimation of population size changes and the DFE using statistics for functional regions only. The true proportions of mutations in each DFE class and $N_{anc}$, $N_{cur}$ are given on the X-axes, while the estimated

values are given on the Y-axes. Parameters are indicated on the upper left corner for each plot. Each dot represents one out of 200 different parameter combinations, sampled randomly from the entire set of simulations.

**Statistics that are important for distinguishing different classes of the DFE and demography**

As it is important to understand which statistics may be necessary to distinguish between the effects of demography and the different classes of the DFE, two different approaches were used to rank statistics by their importance. First, statistics were simply ranked by their regression coefficient with respect to each parameter separately. Non-linear relationships were taken into account by using Box-Cox transformation, as suggested by Wegmann *et al.* (2009). With stationary population size, most of the top predictors of the fraction of neutral ($f_0$) and strongly deleterious ($f_3$) sites are statistics summarizing the functional region (Supp Table 3). The top four statistics for each parameter are displayed in Supp Figure 7. In addition, a modified method of Joyce and Marjoram (2008) was also employed to rank statistics (Supp Table 4) for equilibrium demography.

As expected, statistics that correlate most strongly with the fraction of neutral mutations are levels of divergence and the fraction of high frequency derived alleles, as summarized by $\theta_H$ (Fu 1995; Fay and Wu 2000) in functional regions. As the weakly deleterious class of mutations generate BGS effects at closely linked sites, statistics in the functional and linked region are most strongly correlated with $f_1$. This also correlates most with $H'$ in functional regions, a statistic that contrasts the proportion of high frequency derived variants with those of derived variants segregating at intermediate frequency (Fay and Wu 2000). Although this statistic was designed to identify selective sweeps, which may result in a larger proportion of high frequency derived alleles, it is highly predictive of the fraction of weakly deleterious class of mutations in the absence of positive selection. As shown previously, larger $f_1$ generates a stronger skew in the linked neutral SFS towards rare variants and is thus also reflected in values of Tajima's $D$ in the linked neutral region. Measures of linkage disequilibrium in the functional and linked neutral regions are also correlated with the weakly deleterious class of mutations.

Because the moderately deleterious class of mutations generates BGS effects that extend for larger distances than the more weakly selected class, the strongest correlates of

this class are generally statistics from the neutral region furthest from the directly selected sites. All the different summaries of the SFS - $\theta_W$, $\theta_\pi$, and $\theta_H$ - correlate with this parameter, as well as the total reduction in linked neutral diversity (given by the intercept of the regression fit of $\pi = slope*\ln(distance) + intercept$, where $\pi$ is the diversity in linked neutral regions). The strongly deleterious class of mutations is correlated with the number of singletons and $\theta_W$, which is highly sensitive to singletons.

A similar analysis was performed on simulations under models of demographic non-equilibrium. Here, the DFE parameters are significantly correlated only with the statistics for functional regions (Supp Table 5 and 6). As expected intuitively, the statistics most highly correlated with the two demographic parameters are for the neutral linked regions. Ancestral population sizes correlate most with statistics that capture high frequency derived alleles in linked neutral as well as functional regions, as these represent older mutations; current population sizes correlate most with statistics that summarize LD. The same is true when ranked statistics are obtained only from functional regions. Because the class of moderately deleterious mutations and ancestral populations sizes are correlated with overlapping sets of statistics, the estimation of these two parameters is partially confounded. As such, LD-based statistics are essential in distinguishing between demography and purifying selection, and in distinguishing between ancestral and current population sizes. In addition, although the variances and means of the statistics are highly correlated, the variances play a more important role in estimating current population sizes.

**Comparison with DFE-alpha**

Although there are no other programs that simultaneously co-estimate both demographic and selection parameters, we compared the performance of our method to the step-wise approach of DFE-alpha (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Schneider *et al.* 2011), a program used widely for the inference of the DFE. DFE-alpha assumes that synonymous sites are neutral and uses their site frequency spectrum to infer changes in demography. Conditional on the inferred demography and under the assumption that the deleterious selection coefficients follow a given distribution (generally gamma), the program infers the shape and rate parameter of the assumed distribution. We simulated demographic equilibrium, 2-fold population growth and 2-fold population contraction, and inferred the change in population size as well as the DFE using both ABC and DFE-alpha. Because DFE-alpha uses neutral sites to infer demography, in all cases we simulated a DFE

consisting of ~30% neutral mutations, which were used as a proxy for synonymous sites. These simulations were performed exactly as described previously for non-equilibrium conditions. Exons sizes between 500-2000 bp with flanking 4 kb linked neutral regions were simulated with recombination rates specific to the selected 94 exons (and a total of 940 replicates for every parameter combination). DFE-alpha performs slightly better than ABC if the true DFE is indeed gamma distributed (Supp Figure 8) although our method is able to infer the DFE with very similar accuracy.

For a discrete DFE which is skewed towards highly deleterious mutations, DFE-alpha and ABC perform with similar accuracy. However, our method performs better if the DFE is skewed towards slightly deleterious mutations (*i.e.*, class $f_1$) as shown in Supp Figure 9. It is important to note that, for the purpose of this comparison, simulations were run with numbers of directly selected sites between 500–2000 bp and ensuring that 30% of mutations were neutral, as the neutral mutations were required to estimate demography by DFE-alpha. Under these conditions, background selection results in a relatively small skew in the neutral SFS (see Campos and Charlesworth 2019).

As noted previously, a potential advantage of the methodology proposed here is that, by simultaneously estimating selection and demography, one is not required to make any assumptions about the neutrality of synonymous sites. We evaluated this feature by simulating a scenario where ~33% of the assumed neutral sites were actually experiencing weak direct selection. As weak purifying selection generates a larger fraction of rare variants than stronger selection, programs based on neutrality would be likely to falsely infer growth. As expected, DFE-alpha inferred 2-fold growth under demographic equilibrium, and in fact inferred slight growth even for a 2-fold contraction (Figure 5). The resulting DFE over-estimated the fraction of neutral mutations and under-estimated the fraction of weakly deleterious mutations. As noted previously, such mis-inference will increase with the density of selected sites. Our ABC approach, however, accurately estimated the proportion of neutral mutations present in the selected region (Figure 5), illustrating the importance of such joint inference.

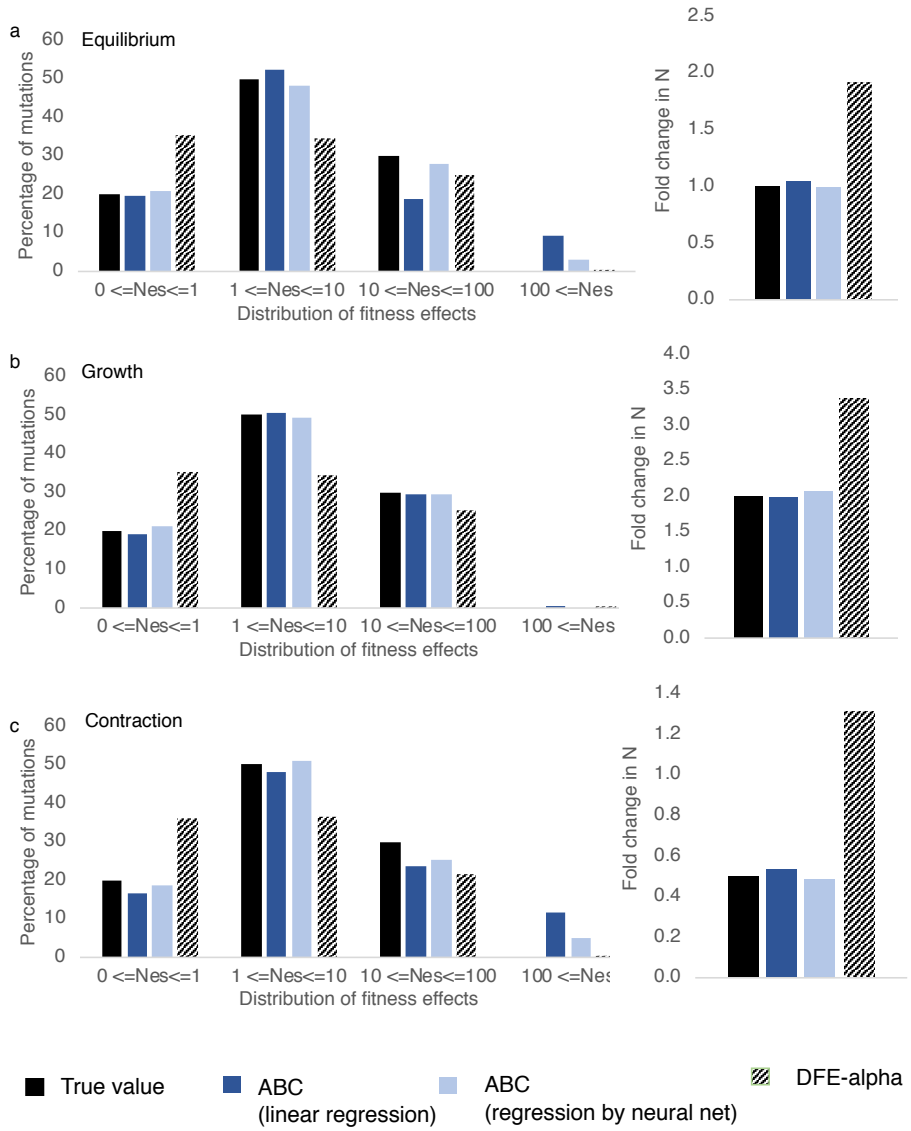



**Figure 5**: Comparison of the performance of the proposed ABC approach in the current study with DFE-alpha, under (a) demographic equilibrium, (b) exponential growth, and (c) exponential decline. In all cases, 30% of sites were assumed to be synonymous, out of which 33% were weakly selected. Solid black bars are the true simulated values, dark blue bars give the ABC performance using ridge regression, and light blue bars give the ABC performance using linear regression aided by neural nets. Patterned bars show the performance of DFE-alpha. A total of 998,300 sites were analyzed in the functional region for each parameter combination, with approximately 332,767 representing synonymous and 665,533 representing nonsynonymous sites.

## Application to *Drosophila melanogaster*

The proposed method is well-suited to compact genomes, in which most sites may be experiencing purifying selection, but is computationally intensive for large genomic regions.

When simulating small genomic regions, the presence of nearby coding regions that are not included in the models can generate additional BGS effects and thus bias inference. We thus restricted our analyses to protein-coding exons in the *D. melanogaster* genome between 500 to 2000 bp in length that are single exon genes, and are flanked on both sides by intergenic regions larger than 4 kb (with the latter two criteria chosen to avoid strong effects of linkage with other nearby functional elements). It should be noted that any genic structure could readily be chosen for inference by directly simulating the associated details when constructing the priors - we have simply chosen this realization in order to provide an illustrative application.

The recombination rates of both the 5′ and 3′ flanking intergenic regions are highly correlated (Supp Figure 10) and span a considerable magnitude (Supp Figure 11), with a mean rate of 2.21 cM/Mb (*i.e.,* the average recombination rate for these chosen single exon genes is very near the autosomal genome-wide average of 2.32 cM/Mb). We also verified that this set of genes was not unusual with regards to genome-wide coding sequence divergence (Supp Figure 12). Furthermore, because sites in intergenic regions in *D. melanogaster* may also experience direct selection (Halligan and Keightley 2006; Casillas *et al.* 2007), we used phastCons scores to exclude intergenic sites that may potentially be functionally important. All sites with a phastCons score larger than 0.8 were excluded (Siepel *et al.* 2005). Table 1 provides the observed summary statistics for each region class, where intergenic sites that had a greater than or equal to 80% probability of belonging to a conserved element (*i.e*., with phastCons score >= 0.8) were excluded. It should be noted that, although there does not appear to be a large difference between divergence (*i.e.*, number of fixed substitutions specific to *D. melanogaster*) in exonic vs intergenic regions, this observation is consistent with previous studies (Table 1 in Andolfatto 2005). In addition, because we have restricted our analyses to sites where the ancestor of *D. melanogaster* could be predicted with high confidence, our analyses may be skewed towards more conserved sites, potentially resulting in lower divergence in intergenic regions. Previous estimates of divergence at 4-fold degenerate sites have been estimated to be roughly 0.05-0.06 (Halligan and Keightley 2006; Langley *et al.* 2012; Charlesworth *et al.* 2018), while that in coding regions to be 0.023 (Langley *et al.* 2012). Although our estimates are lower than previous estimates, this discrepancy is well explained by the larger number of individuals used to subtract polymorphic sites in this study (Supp Table 7). At a sample size of 1 individual (corresponding to pairwise divergence), our estimates of divergence at 4-fold degenerate sites is 0.05 and in coding regions is 0.023, consistent with previous studies. In addition, a very

similar reduction between pairwise divergence and polymorphism-adjusted divergence is observed in simulated data (Supp Table 8).

Interestingly, although previous studies have inferred ~2-fold growth in the Zambian population of *D. melanogaster* (Li and Stephan 2006; Laurent *et al.* 2011; Duchen *et al.* 2013; Kapopoulou *et al.* 2018; Arguello *et al.* 2019), we infer only a 1.2-fold growth, with an ancestral $N_e$ of 1,225,393 and current $N_e$ of 1,357,760. In contrast to previous studies (Keightley and Eyre-Walker 2007; Huber *et al.* 2017), we infer a much larger proportion of mildly deleterious mutations and a smaller proportion of highly deleterious mutations (see Figure 6), with $f_0 = 24.7\%$, $f_1 = 49.4\%$, $f_2 = 3.9\%$, and $f_3 = 21.9\%$ - but this reflects the fact that our procedure includes the possibility of selection on synonymous sites. As we have inferred the DFE for a select class of single exon genes, genes which have slightly higher divergence than average (Supp Figure 12), it is possible that these exons are experiencing weaker purifying selection compared to the genome-wide mean. Furthermore, because we have obtained the DFE of both coding sequences and UTR regions, 4-fold degenerate sites represent 12% of all sites, while UTR regions comprise 29% of all sites. Previous studies have estimated that roughly 6-10% of all mutations at non-synonymous sites may be effectively neutral. Thus, assuming that all 4-fold degenerate sites are neutral, ~40% of UTR regions are neutral (Andolfatto 2005; Campos *et al.* 2017), and ~6-10% of nonsynonymous mutations are neutral, we expect $f_0$ to be ~27-30%. Encouragingly, we infer $f_1$ =25%. This observation implies that the majority of synonymous sites are not experiencing direct selection, consistent with previous results for *D. melanogaster* (Jackson *et al.* 2017).

In order to confirm whether our inferred parameters explain the observed *D. melanogaster* data, we simulated 10 replicates of each of the 94 exons using the parameter estimates, and evaluated whether the mean of the observed *D. melanogaster* values are in the 5% tails of the distribution of statistics obtained via simulations. Our parameter estimates result in a very good fit to empirical *D. melanogaster* population data (Figure 6, Supp Figure 13) for all three categories - functional (*i.e.*, exonic), linked (*i.e.*, non-coding region adjacent to exons) and neutral (*i.e.,* non-coding region adjacent to the linked region). Our parameter estimates fail, however, to explain the observed Tajima's *D* values (linked region $p = 0.011$, neutral region $p = 0.010$) and divergence (linked region $p = 0.029$, neutral region $p$=0.0) in intergenic regions – though both are well fit in functional regions.

As both positive selection in exons and purifying selection in non-coding regions could partially drive these patterns, we investigated both of these model violations. Non-coding regions flanking 2kb of the selected exons (which were used to perform inference)

were found to have 777 sites that had phastCons scores greater than or equal to 0.8, with a mean and median length of 25 and 15 bp, respectively. We therefore simulated conserved elements in non-coding regions that were 20bp in length, uniformly distributed, and which made up 40% of the flanking neutral sites (*i.e.*, 800 sites in total). Conserved elements were simulated with weak ($f_1$ =100%), moderate ($f_2$ =100%) and strong ($f_3$ =100%) purifying selection separately. Upon masking these sites, as was done in our *Drosophila* data analysis, there was no observed difference in the distribution of all statistics (Supp Figure 14), suggesting that background selection caused by small conserved elements does not significantly affect our inference, and in fact does not alter the fit of our inferred model to the data. Interestingly, without masking sites – that is, allowing sites that experience direct weak purifying selection to remain in the flanking sequence - our model is much better able to explain a lower Tajima's *D* and divergence in intergenic regions (Supp Figure 15). Thus, it appears likely that unaccounted-for weak purifying selection across multiple sites in intergenic regions could contribute to the discrepancy between statistics generated by our model and those observed in the data.

Next, we simulated positive selection under 4 different scenarios - representing rare and strong (1% of all mutations in exonic regions are beneficial with $2N_es = 1000$), common and strong (5% of mutations in exonic regions are beneficial with $2N_es = 1000$), common and weak (5% of mutations in exonic regions are beneficial with $2N_es = 10$) and rare and weak (1% of mutations in exonic regions are beneficial with $2N_es = 10$) selection. Interestingly, we find that, although strong positive selection, whether common or rare, better explains the lower Tajima's *D* values in intergenic regions, it also drastically alters the distribution of most other statistics, resulting in an overall much poorer fit (Supp Figure 16, 17). For instance, common and strong positive selection reduces $\Theta_H$ by an order of magnitude relative to our fitted model and drastically increases the variance while decreasing the mean of haplotype diversity. In contrast to strong positive selection, weakly positively selected mutations do not alter the distribution of Tajima's *D* in intergenic regions, but do slightly increase $\theta_H$ in functional regions, which improves the fit to the observed data (Supp Figure 18, 19). In addition, all cases of positive selection significantly increase divergence in functional regions. For comparison, we also simulated the two scenarios of positive selection used by Lange and Pool (2018) - 0.2% of all mutations are beneficial with $2N_es$ =60, and 0.00013% of all mutations are beneficial with $2N_es$ =10000. As the frequency of positively selected alleles is lower in these scenarios, there was no observed difference between the distribution of statistics resulting from including or excluding positive selection (Supp Figure

20, 21). Thus, if the frequency of strongly positively selected mutations is much lower than 1%, our estimates of both demography and DFE shape should be unbiased, and the beneficial fixations would be virtually undetectable. Future studies will further investigate the ability of such an approach to quantify the occupancy of a beneficial mutational class.

**Table 1**: Statistics calculated for the 94 single-exon genes including their 3′ flanking intergenic sequences, for 76 individuals (devoid of any inversion) in the African Zambian population. Sites with phastCons scores higher than 0.8 were excluded. Functional refers to exons, linked refers to intergenic region (~ 1kb) adjacent to exons and neutral refers to intergenic regions further away from exons that are adjacent to linked regions and ~1kb in size (Figure 4a). Derived alleles were identified by polarizing alleles with respect to the ancestral sequence of *D. melanogaster* obtained from ancestral reconstruction over 15 insect species.

| | mean | | | standard deviation | | |
|---|---|---|---|---|---|---|
| | **functional** | **linked** | **neutral** | **functional** | **linked** | **neutral** |
| $\pi$ | 0.0083 | 0.0106 | 0.0107 | 0.0039 | 0.0042 | 0.0038 |
| $\theta_W$ | 0.0120 | 0.0166 | 0.0162 | 0.0045 | 0.0053 | 0.0049 |
| $\theta_H$ | 0.0088 | 0.0098 | 0.0097 | 0.0054 | 0.0053 | 0.0056 |
| $H'$ | -0.0633 | 0.0871 | 0.1169 | 0.5371 | 0.4118 | 0.3829 |
| **Tajima's *D*** | -1.0537 | -1.1469 | -1.1103 | 0.5338 | 0.4874 | 0.4694 |
| **Singleton density** | 0.0215 | 0.0303 | 0.0307 | 0.0086 | 0.0116 | 0.0117 |
| **Haplotype diversity** | 0.9711 | 0.9680 | 0.9762 | 0.0452 | 0.0458 | 0.0444 |
| $r^2$ | 0.0328 | 0.0364 | 0.0363 | 0.0109 | 0.0136 | 0.0128 |
| ***D*** | 0.0005 | 0.0005 | 0.0006 | 0.0009 | 0.0010 | 0.0012 |
| **Branch-specific divergence** | 0.01378 | 0.0156 | 0.0159 | 0.0075 | 0.0077 | 0.0071 |

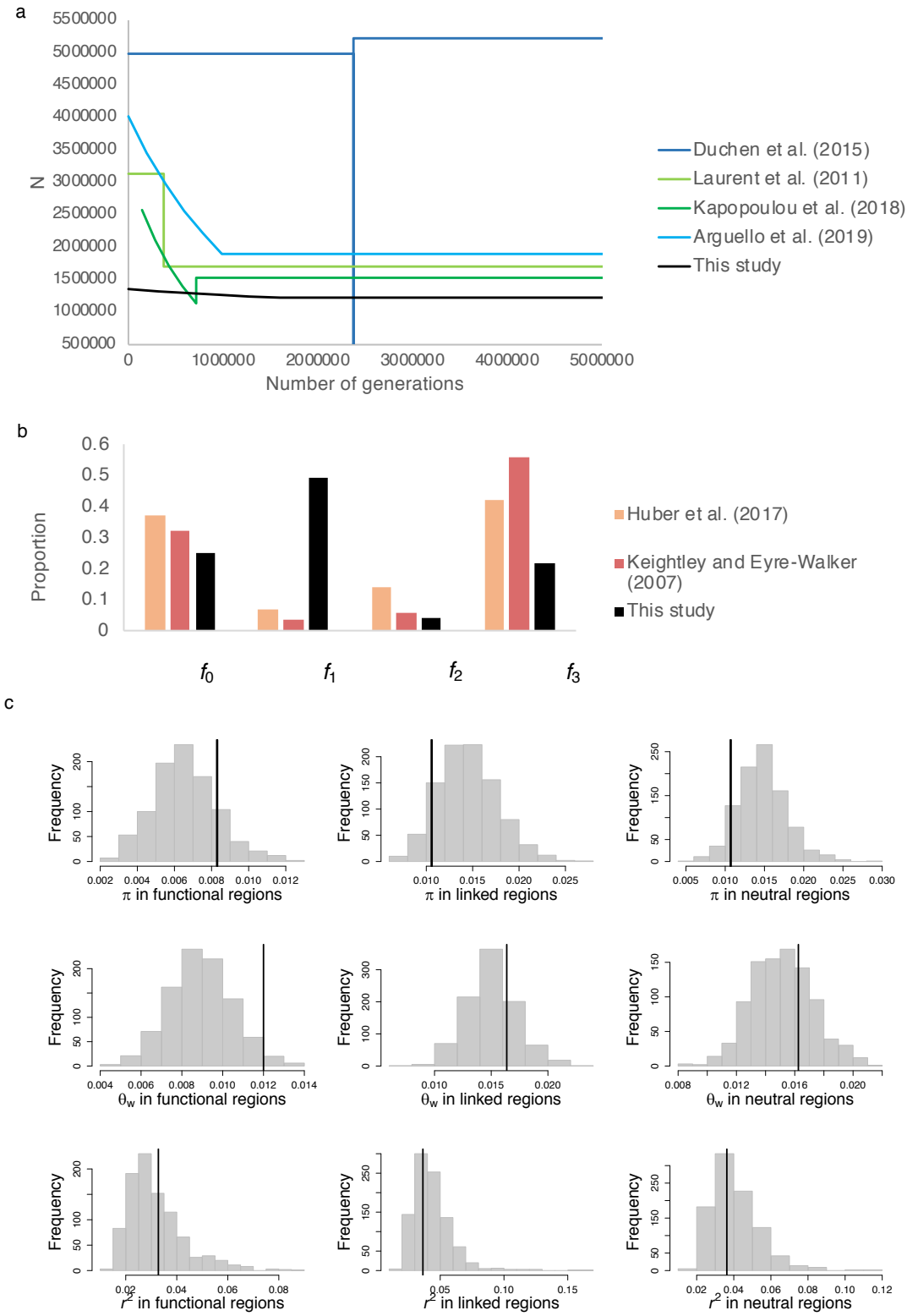



**Figure 6**: Joint inference of demography and purifying selection in the Zambian population of *D. melanogaster*. (a) Demographic model inferred from previous studies (colored lines) and from the current study (black lines). (b) The distribution of fitness effects of deleterious mutations at coding regions (including synonymous and non-synonymous sites) as inferred by previous studies of other populations (colored bars) and at exonic sites of single-exons

genes as inferred by the current study (black bars). The X-axis is given by $f_0$: $0<|2N_es|<1$, $f_1$: $1<|2N_es|<10$, $f_2$: $10<|2N_es|<100$, and $f_3$: $100<|2N_es|<10000$. For the previous studies, the DFE shown in this figure includes the fraction of synonymous sites in the neutral $f_0$ class. (c) Distribution of key summary statistics ($\pi$, $\theta_W$, $r^2$) in functional, linked and neutral regions upon simulating 100 replicates of 94 exons each under the inferred parameters. The vertical lines represent the values of the statistics obtained from 76 individuals of *D. melanogaster* from Zambia, after excluding non-coding sites with phastCons score >= 0.8.

## CONCLUSION

Independent of any dogmatic stance regarding the roles of adaptive vs. non-adaptive explanations for observed levels and patterns of DNA sequence variation and divergence, it has been widely accepted that natural populations are not at demographic equilibrium, but are often characterized by fluctuating population sizes and other demographic perturbations. Additionally, a rich empirical and experimental literature has clarified the pervasive importance of purifying selection in eliminating the constant input of deleterious variants. It has also been well-demonstrated that ignoring direct effects of purifying selection and its impact on linked sites can strongly bias demographic inference (Ewing and Jensen 2016), and that ignoring demographic effects biases estimates of parameters of selection (Jensen *et al.* 2005; Thornton and Jensen 2007; Crisci *et al.* 2012, 2013). Yet, despite agreement that these processes are certain to be occurring constantly in populations and shaping patterns of variation and evolution, the construction of a statistical approach capable of simultaneously estimating parameters of the concerned processes has proven challenging. Here we provide one such approach, for which we demonstrate an ability to co-estimate the parameters of a generalized DFE along with those underlying the population history.

By fitting a four-parameter DFE model that includes weak, intermediate and strong purifying selection, as well as neutrally evolving sites, this approach avoids two common, and potentially perilous, assumptions: 1) synonymous sites are not assumed to be neutral, consistent with a growing body of literature (Chamary and Hurst 2005; Lynch 2007; Zeng and Charlesworth 2010a; Lawrie *et al.* 2013; Choi and Aquadro 2016; Jackson *et al.* 2017), and 2) the DFE is not assumed to follow a specific parameterized distribution, such as the widely-used gamma distribution.

Our results demonstrate that it is possible to jointly infer the deleterious DFE and past demographic changes using an ABC framework, by including various summary statistics

capturing aspects of the SFS, linkage disequilibrium and divergence, compared between coding and flanking non-coding sequence. Ancestral population sizes and the frequency of the most deleterious classes of the DFE are estimated with relatively low accuracy, whereas the current population sizes and the neutral mutation class are estimated with high accuracy. In addition, we demonstrated that, if synonymous sites are indeed experiencing substantial purifying selection, existing programs such as DFE-alpha will over-estimate recent growth and under-estimate the proportion of mildly deleterious mutations. Importantly, the approach proposed here performs equally well regardless of whether synonymous sites are neutral or selected. However, our approach continues to assume the neutrality of flanking non-coding regions, though putatively conserved sites were masked, and the impact of that masking on inference was thoroughly assessed via simulation.

Because we make no assumptions about which sites in the functional region of interest are neutral, it is in principle possible to estimate the DFE for any functional element using this methodology, including regulatory elements or functional regions with interdigitated sites experiencing direct selection. The results further suggest that the accurate co-estimation of these parameters is possible using only functional regions. Such an approach may be extremely useful in genomes for which it is difficult to characterize putatively neutral sites, as well as for compact genomes in which non-coding regions may be limited.

This approach can in principle be applied to any organism and functional class of interest, although power analyses suggest the utility of prior knowledge of the boundaries of functional regions and recombination rates. Here we have provided an illustrative example in *D. melanogaster*. The results suggest that the Zambian population has been largely stable in size, and that exonic regions have a large proportion of mildly deleterious mutations. Although this result might seem surprising, the DFE inferred by the current method provides the distribution of selective effects over all sites, including synonymous sites and sites in UTRs. Hence, in comparing the DFE estimated in the current study with previous estimates of the neutral class of mutations, it appears unnecessary to invoke widespread selection on synonymous sites in *D. melanogaster.* This result is largely consistent with most previous studies (Akashi 1995; Jackson *et al.* 2017). For instance, our estimate of the strength of purifying selection acting on synonymous sites in the Zambian population is in line with earlier estimates for African populations (Zeng and Charlesworth 2010a; Jackson *et al.* 2017).

In addition to the proposed inference framework, we have derived an analytical expression for the reduction in variation caused by background selection at neutral sites

outside functional regions for the case of a discrete DFE, making it feasible to obtain analytical predictions for any chosen DFE. Not only does a discrete DFE provide flexibility in inference, it may also be a more realistic representation of the true DFE (Kousathanas and Keightley 2013; Bank *et al.* 2014b). Although gamma distributions represent a reasonable fit to the DFE inferred from genome-wide studies (Eyre-Walker and Keightley 2007), the DFE will be mis-inferred if the true distribution is multimodal (Kousathanas and Keightley 2013), as has been observed widely (*e.g.*, in yeast (Bank *et al.* 2014a), viruses (Sanjuán 2010), and *E.coli* (Jacquier *et al.* 2013)). In addition, the best fitting parameterized continuous distribution appears to be extremely specific to the particular dataset being tested, and most alternative distributions fit the data nearly as well as the best fitting-distribution (Huber *et al.* 2017; Kim *et al.* 2017). The discrete DFE proposed here thus reduces the number of necessary assumptions, and has been shown to perform well in the plausible scenario in which common assumptions are indeed violated (*e.g.*, if the true DFE is not gamma-distributed). Analytical results under demographic equilibrium and simulations under demographic non-equilibrium stress that the number of selected sites and the specific shape of the DFE (for instance the presence of mildly and moderately deleterious mutations) both decrease linked neutral variation around functional regions more than previously appreciated, and skew the SFS even when there is no reduction in diversity. Such variation in exon lengths and DFE shapes across a genome can increase variance of statistics in linked neutral regions, which could contribute to false positives when detecting positive selection using outlier approaches.

There are at least two important caveats worth considering, which will be the subject of future study. The first concerns the estimates of ancestral and current effective population sizes. As the effective population size varies across the genome in a fashion correlated with local recombination rates (Becher *et al.* 2020, in press), the estimates provided here ought to be viewed as a mean across the loci in question. While we have improved upon the common assumption of a singular genome-wide value by directly modeling each locus-specific recombination rate when performing inference, the general importance of this effect in demographic modeling remains in need of further study. The second concerns the inference of selection. This study represents a proof-of-concept in demonstrating that such simultaneous inference of demography and the DFE is feasible, thereby avoiding common assumptions underlying a step-wise inference approach. While this interplay of genetic drift and purifying selection is in fact sufficient alone to fit all features of the data (consistent with previous claims: Comeron 2014, 2017; Harris *et al.* 2018; Jensen *et al.* 2019), this is not the

same as claiming that positive selection is not also occurring. As our simulation results demonstrate, the addition of rare, weakly beneficial mutations is consistent with the data, though the inclusion of these parameters does not result in an improved fit. The question is less about presence/absence, than it is about statistical identifiability. Conversely, the addition of a strongly beneficial mutational class was found to be inconsistent with observed data. In order to investigate this further, future work will evaluate the ability to co-estimate a beneficial class of fitness effects within this framework. It should also be noted that the example chosen to highlight our approach focuses on only a subset of genes in the *D. melanogaster* genome, and there is no expectation that the observed DFE in this class will necessarily be universal across all coding regions in the population under consideration. In fact, means of scaled selection coefficients of deleterious mutations have been shown to be negatively correlated with divergence at nonsynonymous sites (Campos *et al.* 2017). Importantly however, this general inference approach accounting for these two dominant processes will be a valuable tool in future genomic scans, and this appropriate null is anticipated to greatly reduce the notoriously high false-positive rates associated with the identification of positively selected loci.

## ACKNOWLEDGEMENTS

We would like to thank Rebecca Harris for discussions related to this project. This research was conducted using resources provided by Research Computing at Arizona State University (http://www.researchcomputing.asu.edu) and the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. We especially thank Lauren Michael and Christina Koch from the Open Science Grid, for all of their efforts to provide technical assistance. This work was funded by National Institutes of Health grant R01GM135899 to JDJ.

## DISCLOSURE DECLARATION

The authors declare no conflicts of interests.

# APPENDIX

## Derivation of the analytical expression for the reduction in diversity due to background selection generated by a discrete distribution of fitness effects under demographic equilibrium

Because we model a discrete DFE with four fixed bins, with $t$ being uniformly distributed within each bin, the definite integral for each bin is the integral of E($t$) with respect to $t$. We thus have:

$$\int_0^1 E(t)dt = \frac{f_0}{t_1-t_0}\int_{t_0}^{t_1} E(t)dt + \frac{f_1}{t_2-t_1}\int_{t_1}^{t_2} E(t)dt + \frac{f_2}{t_3-t_2}\int_{t_2}^{t_3} E(t)dt \quad + \frac{f_3}{t_4-t_3}\int_{t_3}^{t_4} E(t)dt \qquad (A1)$$

where the $t_i$'s correspond to the boundary of the discrete bins. The first integral is over $t$ such that 0 <= $2N_es$<= 1, the second over $t$ such that 1 <= $2N_es$ <= 10, the third such that 10<=$2N_es$<=100 and fourth as 100<=$2N_es$<=10000. In our case, $t_0 = 0$, $t_1 = 0.00005$, $t_2 = 0.0005$, $t_3 = 0.005$, and $t_4 = 0.5$. While this mirrors the DFE considered here, the same procedure can be done for any set of bins for a given DFE.

Integrating $E$ over a uniform distribution between $t_0$ and $t_1$, with probability density $(t_1 - t_0)^{-1}$, where $a = g + r_c y$ and $b = g + r_c(y + l)$, we have:

$$\int \frac{t\,dt}{(1-t)[a+t(1-a]} = \int\{\frac{t}{(1-t)} + \frac{t(1-a)}{[a+t(1-a)]}\}dt \qquad (A2)$$

The second integral on the right-hand side of this equation can be evaluated by substituting $u = a + t(1 - a)$ for $t$, with $t = (u - a)/(1 - a)$ and $dt = du/(1 - a)$. This gives:

$$(1-a)\int \frac{t(1-a)}{[a+t(1-a)]}dt = (1-a)^{-1}\int u^{-1}(u-a)du = (1-a)^{-1}[u - a\ln(u)] \qquad (A3)$$

With the change in variable, the normalizing factor for the probability density function is now $(u_1 - u_0)^{-1} = (1 - a)^{-1}(t_1 - t_0)^{-1}$. The contribution of this component to the expectation of $E(t)$ over the uniform distribution yields equation (3b) of the main text. A similar expression can be written for the integral of $-t/[(1 - t)[b + t(1- b)]$ in the first line of equation (2). When adding this to the integral of $t/[(1 - t)[a + t(1- a)]$, the integrals involving $1/(1 - t)$ cancel out, so this term simply contributes the following term to the expectation of $E(t)$, yielding equation (3b). The expectation of $E(t)$ is the sum of these two terms. $E(t)$ can

also be numerically integrated over a definite interval as specified above, with constant values as chosen in our simulations: $r = 10^{-6}$, $l = 500$ or 1000 or 5000 or 10000, $U = l \times \mu$, $\mu = 10^{-6}$, and $g = 0$.

## REFERENCES

Akashi H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139: 1067–1076.

Andolfatto P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149–1152. https://doi.org/10.1038/nature04107

Arguello J. R., S. Laurent, and A. G. Clark, 2019 Demographic history of the human commensal *Drosophila melanogaster*. Genome Biol Evol 11: 844–854. https://doi.org/10.1093/gbe/evz022

Assaf Z. J., S. Tilk, J. Park, M. L. Siegal, and D. A. Petrov, 2017 Deep sequencing of natural and experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new mutations. Genome Res. 27: 1988–2000. https://doi.org/10.1101/gr.219956.116

Bank C., R. T. Hietpas, A. Wong, D. N. Bolon, and J. D. Jensen, 2014a A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics 196: 841–852. https://doi.org/10.1534/genetics.113.156190

Bank C., G. B. Ewing, A. Ferrer-Admettla, M. Foll, and J. D. Jensen, 2014b Thinking too positive? Revisiting current methods of population genetic selection inference. Trends in Genetics 30: 540–546. https://doi.org/10.1016/j.tig.2014.09.010

Beaumont M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in Population Genetics. Genetics 162: 2025–2035.

Becher H., B. C. Jackson, and B. Charlesworth, 2020 Patterns of genetic variability in genomic regions with low rates of recombination. bioRxiv 739888 (in press). https://doi.org/10.1101/739888

Campos J. L., L. Zhao, and B. Charlesworth, 2017 Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. PNAS 114: E4762–E4771. https://doi.org/10.1073/pnas.1619434114

Campos J. L., and B. Charlesworth, 2019 The effects on neutral variability of recurrent selective sweeps and background selection. Genetics 212: 287–303. https://doi.org/10.1534/genetics.119.301951

Casillas S., A. Barbadilla, and C. M. Bergman, 2007 Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol. Biol. Evol. 24: 2222–2234. https://doi.org/10.1093/molbev/msm150

Chamary J., and L. D. Hurst, 2005 Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biology 6: R75. https://doi.org/10.1186/gb-2005-6-9-r75

Charlesworth B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Charlesworth B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. 68: 131–149.

Charlesworth B., 2012 The effects of deleterious mutations on evolution at linked sites. Genetics 190: 5–22. https://doi.org/10.1534/genetics.111.134288

Charlesworth B., 2013 Background Selection 20 Years onThe Wilhelmine E. Key 2012 Invitational Lecture. J Hered 104: 161–171. https://doi.org/10.1093/jhered/ess136

Charlesworth B., J. L. Campos, and B. C. Jackson, 2018 Faster-X evolution: Theory and evidence from Drosophila. Molecular Ecology 27: 3753–3771. https://doi.org/10.1111/mec.14534

Choi J. Y., and C. F. Aquadro, 2016 Recent and long term selection across synonymous sites in *Drosophila ananassae*. J Mol Evol 83: 50–60. https://doi.org/10.1007/s00239-016-9753-9

Comeron J. M., R. Ratnappan, and S. Bailin, 2012 The many landscapes of recombination in *Drosophila melanogaster*. PLOS Genetics 8: e1002905. https://doi.org/10.1371/journal.pgen.1002905

Comeron J. M., 2014 Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLOS Genetics 10: e1004434. https://doi.org/10.1371/journal.pgen.1004434

Comeron J. M., 2017 Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Phil. Trans. R. Soc. B 372: 20160471. https://doi.org/10.1098/rstb.2016.0471

Crisci J. L., Y.-P. Poh, A. Bean, A. Simkin, and J. D. Jensen, 2012 Recent progress in polymorphism-based population genetic inference. J. Hered. 103: 287–296. https://doi.org/10.1093/jhered/esr128

Crisci J. L., Y.-P. Poh, S. Mahajan, and J. D. Jensen, 2013 The impact of equilibrium assumptions on tests of selection. Front. Genet. 4: 235. https://doi.org/10.3389/fgene.2013.00235

Crow J. F., 2008 Mid-century controversies in population genetics. Annual Review of Genetics 42: 1–16. https://doi.org/10.1146/annurev.genet.42.110807.091612

Csilléry K., O. François, and M. G. B. Blum, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution 3: 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Cutter A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics 14: 262–274. https://doi.org/10.1038/nrg3425

Duchen P., D. Živković, S. Hutter, W. Stephan, and S. Laurent, 2013 Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics 193: 291–301. https://doi.org/10.1534/genetics.112.145912

Elyashiv E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, *et al.*, 2016 A genomic map of the effects of linked selection in *Drosophila*. PLOS Genet 12: e1006130. https://doi.org/10.1371/journal.pgen.1006130

Ewing G. B., and J. D. Jensen, 2016 The consequences of not accounting for background selection in demographic inference. Molecular Ecology 25: 135–141. https://doi.org/10.1111/mec.13390

Eyre-Walker A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nature Reviews Genetics 8: 610–618. https://doi.org/10.1038/nrg2146

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. Mol Biol Evol 26: 2097–2108. https://doi.org/10.1093/molbev/msp119

Fay J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Fiston-Lavier A.-S., N. D. Singh, M. Lipatov, and D. A. Petrov, 2010 *Drosophila melanogaster* recombination rate calculator. Gene 463: 18–20. https://doi.org/10.1016/j.gene.2010.04.015

Fu Y.-X., 1995 Statistical properties of segregating sites. Theoretical Population Biology 172–197.

Good B. H., A. M. Walczak, R. A. Neher, and M. M. Desai, 2014 Genetic diversity in the interference selection limit. PLOS Genetics 10: e1004222. https://doi.org/10.1371/journal.pgen.1004222

Haller B. C., and P. W. Messer, 2019 SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol 36: 632–637. https://doi.org/10.1093/molbev/msy228

Halligan D. L., and P. D. Keightley, 2006 Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res 16: 875–884. https://doi.org/10.1101/gr.5022906

Harris R. B., A. Sackman, and J. D. Jensen, 2018 On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. PLOS Genetics 14: e1007859. https://doi.org/10.1371/journal.pgen.1007859

Hoskins R. A., J. W. Carlson, C. Kennedy, D. Acevedo, M. Evans-Holm, *et al.*, 2007 Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. Science 316: 1625–1628. https://doi.org/10.1126/science.1139816

Huber C. D., B. Y. Kim, C. D. Marsden, and K. E. Lohmueller, 2017 Determining the factors driving selective effects of new nonsynonymous mutations. PNAS 114: 4465–4470. https://doi.org/10.1073/pnas.1619508114

Hudson R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics 141: 1605–1617.

Jackson B. C., J. L. Campos, P. R. Haddrill, B. Charlesworth, and K. Zeng, 2017 Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in *Drosophila*. Genome Biol Evol 9: 102–123. https://doi.org/10.1093/gbe/evw291

Jacquier H., A. Birgy, H. L. Nagard, Y. Mechulam, E. Schmitt, *et al.*, 2013 Capturing the mutational landscape of the beta-lactamase TEM-1. PNAS 110: 13067–13072. https://doi.org/10.1073/pnas.1215206110

Jensen J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–1410. https://doi.org/10.1534/genetics.104.038224

Jensen J. D., B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch, *et al.*, 2019 The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution 73: 111–114. https://doi.org/10.1111/evo.13650

Joyce P., and P. Marjoram, 2008 Approximately sufficient statistics and bayesian computation. Statistical Applications in Genetics and Molecular Biology 7. https://doi.org/10.2202/1544-6115.1389

Kaiser V. B., and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends in Genetics 25: 9–12. https://doi.org/10.1016/j.tig.2008.10.009

Kapopoulou A., S. P. Pfeifer, J. D. Jensen, and S. Laurent, 2018 The demographic history of African *Drosophila melanogaster*. Genome Biol Evol 10: 2338–2342. https://doi.org/10.1093/gbe/evy185

Karolchik D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, *et al.*, 2004 The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32: D493-496. https://doi.org/10.1093/nar/gkh103

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. https://doi.org/10.1534/genetics.107.080663

Keightley P. D., R. W. Ness, D. L. Halligan, and P. R. Haddrill, 2014 Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196: 313–320. https://doi.org/10.1534/genetics.113.158758

Kern A. D., and M. W. Hahn, 2018 The neutral theory in light of natural selection. Mol Biol Evol 35: 1366–1371. https://doi.org/10.1093/molbev/msy092

Kim B. Y., C. D. Huber, and K. E. Lohmueller, 2017 Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206: 345–361. https://doi.org/10.1534/genetics.116.197145

Kimura M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626. https://doi.org/10.1038/217624a0

Kimura M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.

King J. L., and T. H. Jukes, 1969 Non-Darwinian Evolution. Science 164: 788–798. https://doi.org/10.1126/science.164.3881.788

Kolaczkowski B., A. D. Kern, A. K. Holloway, and D. J. Begun, 2011 Genomic differentiation between temperate and tropical australian populations of *Drosophila melanogaster*. Genetics 187: 245–260. https://doi.org/10.1534/genetics.110.123059

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. https://doi.org/10.1534/genetics.112.148023

Lack J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, *et al.*, 2015 The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics 199: 1229–1241. https://doi.org/10.1534/genetics.115.174664

Lange J. D., and J. E. Pool, 2018 Impacts of recurrent hitchhiking on divergence and demographic inference in *Drosophila*. Genome Biol Evol 10: 1882–1891. https://doi.org/10.1093/gbe/evy142

Langley C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, *et al.*, 2012 Genomic variation in natural populations of Drosophila melanogaster. Genetics 192: 533–598. https://doi.org/10.1534/genetics.112.142018

Laurent S. J. Y., A. Werzner, L. Excoffier, and W. Stephan, 2011 Approximate bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Molecular Biology and Evolution 28: 2041–2051. https://doi.org/10.1093/molbev/msr031

Lawrie D. S., P. W. Messer, R. Hershberg, and D. A. Petrov, 2013 Strong purifying selection at synonymous sites in *D. melanogaster*. PLOS Genetics 9: e1003527. https://doi.org/10.1371/journal.pgen.1003527

Li H., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in Drosophila. PLOS Genetics 2: e166. https://doi.org/10.1371/journal.pgen.0020166

Lynch M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, Massachusetts.

Matsumoto T., A. John, P. Baeza-Centurion, B. Li, and H. Akashi, 2016 Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Mol Biol Evol 33: 1580–1589. https://doi.org/10.1093/molbev/msw027

Messer P. W., and D. A. Petrov, 2013 Frequent adaptation and the McDonald–Kreitman test. PNAS 110: 8615–8620. https://doi.org/10.1073/pnas.1220835110

Nicolaisen L. E., and M. M. Desai, 2013 Distortions in genealogies due to purifying selection and recombination. Genetics 195: 221–230. https://doi.org/10.1534/genetics.113.152983

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

O'Fallon B. D., J. Seger, and F. R. Adler, 2010 A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Mol Biol Evol 27: 1162–1172. https://doi.org/10.1093/molbev/msq006

Ohta T., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98. https://doi.org/10.1038/246096a0

Pordes R., D. Petravick, B. Kramer, D. Olson, M. Livny, *et al.*, 2007 The open science grid. J. Phys.: Conf. Ser. 78: 012057. https://doi.org/10.1088/1742-6596/78/1/012057

Sanjuán R., 2010 Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci 365: 1975–1982. https://doi.org/10.1098/rstb.2010.0063

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. https://doi.org/10.1534/genetics.111.131730

Schrider D. R., D. Houle, M. Lynch, and M. W. Hahn, 2013 Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194: 937–954. https://doi.org/10.1534/genetics.113.151670

Schrider D. R., A. G. Shanku, and A. D. Kern, 2016 Effects of linked selective sweeps on demographic inference and model selection. Genetics 204: 1207–1223. https://doi.org/10.1534/genetics.116.190223

Sfiligoi I., D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, *et al.*, 2009 The pilot way to grid resources using glideinWMS, pp. 428–432 in *2009 WRI World Congress on Computer Science and Information Engineering*,.

Siepel A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15: 1034–1050. https://doi.org/10.1101/gr.3715005

Thornton K., 2003 Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19: 2325–2327. https://doi.org/10.1093/bioinformatics/btg316

Thornton K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750. https://doi.org/10.1534/genetics.106.064642

Torres R., M. G. Stetter, R. Hernandez, and J. Ross-Ibarra, 2019 The temporal dynamics of background selection in non-equilibrium populations. bioRxiv. https://doi.org/10.1101/618389

Walsh B., and M. Lynch, 2018 *Evolution and Selection of Quantitative Traits*. Oxford University Press.

Wegmann D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov Chain Monte Carlo without likelihood. Genetics 182: 1207–1218. https://doi.org/10.1534/genetics.109.102509

Zeng K., and B. Charlesworth, 2010a Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. J Mol Evol 70: 116–128. https://doi.org/10.1007/s00239-009-9314-6

Zeng K., and B. Charlesworth, 2010b The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics 186: 1411–1424. https://doi.org/10.1534/genetics.110.122150

Zeng K., 2013 A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity 110: 363–371. https://doi.org/10.1038/hdy.2012.102

## SUPPLEMENTARY FIGURES AND TABLES

**Supp Table 1**: Prediction of diversity in linked neutral regions for two different DFE realizations, as predicted by equation 3a and 3b. Expected diversity is 0.02.

| DFE | Length of coding: | Distance from selected site --> | | |
|---|---|---|---|---|
| | | **100 bp** | **10 kb** | **100 kb** |
| $f_0 = 0; f_1 = 0, f_2 = 50, f_3 = 50$ | **1 kb** | 0.0174857 | 0.019805 | 0.0199807 |
| | **5 kb** | 0.0150363 | 0.0192131 | 0.0199054 |
| | **10 kb** | 0.0141392 | 0.0187112 | 0.0198152 |
| $f_0 = 0; f_1 = 0; f_2 = 80; f_3 = 20$ | **1 kb** | 0.0160713 | 0.0197472 | 0.0199892 |
| | **5 kb** | 0.0124679 | 0.0190189 | 0.0199474 |
| | **10 kb** | 0.0113456 | 0.0184565 | 0.0198979 |

**Supp Table 2**: Reduction in neutral and linked neutral diversity as a function of the DFE - as illustrated by considering DFE realizations in which one class is largely over-represented, for different exon lengths (0.5kb, 1kb, 5kb, and 10kb). The expected diversity under neutrality is 0.02.

| | $f_0$ >= 80% | $f_1$ >= 80% | $f_2$ >= 80% | $f_3$ >= 80% |
|---|---|---|---|---|
| **Neutral Diversity (0.5 kb)** | 0.01995 | 0.01982 | 0.01958 | 0.01975 |
| **Linked neutral diversity (0.5kb)** | 0.01976 | 0.01809 | 0.01795 | 0.01891 |
| **Slope of recovery (0.5kb)** | 0.00007 | 0.00054 | 0.00063 | 0.00013 |
| **Neutral Diversity (1 kb)** | 0.01992 | 0.01972 | 0.01926 | 0.01951 |
| **Linked neutral diversity (1kb)** | 0.01968 | 0.01756 | 0.01674 | 0.01906 |
| **Slope of recovery (1kb)** | 0.00010 | 0.00072 | 0.00100 | 0.00024 |
| **Neutral Diversity (5 kb)** | 0.01972 | 0.01934 | 0.01760 | 0.01814 |
| **Linked neutral diversity (5 kb)** | 0.01915 | 0.01634 | 0.01318 | 0.01714 |
| **Slope of recovery (5kb)** | 0.00024 | 0.00109 | 0.00200 | 0.00061 |
| **Neutral Diversity (10kb)** | 0.01960 | 0.01909 | 0.01673 | 0.01709 |
| **Linked neutral diversity (10kb)** | 0.01903 | 0.01615 | 0.01208 | 0.01579 |
| **Slope of recovery (10kb)** | 0.00027 | 0.00112 | 0.00220 | 0.00083 |

**Supp Table 3**: Statistics ranked by their importance in predicting the DFE classes under equilibrium using the correlation coefficients between the statistics and parameters.

| **Statistics ranked for $f_0$** | $r^2$ | **Statistics ranked for $f_1$** | $r^2$ | **Statistics ranked for $f_2$** | $r^2$ | **Statistics ranked for $f_3$** | $r^2$ |
|---|---|---|---|---|---|---|---|
| func_div_sd | 0.962 | func_hprime_m | 0.632 | neu_thetaw_m | 0.477 | func_numSing_m | 0.850 |
| func_thetah_sd | 0.962 | func_rsq_sd | 0.511 | neu_thetapi_m | 0.463 | func_numSing_sd | 0.748 |
| func_thetah_m | 0.960 | link_tajimasd_m | 0.456 | pi_intercept | 0.425 | func_thetaw_m | 0.460 |
| func_div_m | 0.958 | link_Dprime_m | 0.441 | neu_thetah_m | 0.347 | func_Dprime_sd | 0.432 |
| func_thetapi_sd | 0.896 | func_Dprime_sd | 0.384 | pi_slope | 0.340 | func_thetaw_sd | 0.397 |
| func_rsq_m | 0.873 | pi_max | 0.357 | link_thetaw_m | 0.320 | func_hapdiv_m | 0.334 |
| func_Dprime_m | 0.866 | func_D_sd | 0.297 | func_tajimasd_m | 0.309 | func_rsq_sd | 0.317 |
| func_thetapi_m | 0.855 | func_tajimasd_sd | 0.280 | func_rsq_m | 0.278 | pi_max | 0.297 |
| func_D_m | 0.847 | func_hprime_sd | 0.271 | link_thetapi_m | 0.252 | func_hapdiv_sd | 0.295 |
| func_tajimasd_m | 0.828 | link_thetah_m | 0.236 | neu_div_m | 0.240 | link_tajimasd_m | 0.267 |
| func_hapdiv_sd | 0.731 | pi_slope | 0.226 | func_Dprime_m | 0.235 | neu_rsq_m | 0.264 |
| func_thetaw_sd | 0.729 | link_thetapi_m | 0.219 | func_hapdiv_m | 0.230 | func_thetapi_m | 0.258 |
| func_hapdiv_m | 0.683 | func_numSing_m | 0.200 | link_div_m | 0.229 | link_hapdiv_sd | 0.252 |
| func_thetaw_m | 0.663 | func_numSing_sd | 0.185 | link_thetah_m | 0.222 | link_rsq_sd | 0.234 |
| func_hprime_sd | 0.578 | link_div_m | 0.183 | func_hapdiv_sd | 0.220 | link_D_sd | 0.230 |
| func_hprime_m | 0.553 | pi_intercept | 0.165 | func_thetapi_m | 0.215 | link_hapdiv_m | 0.218 |
| func_D_sd | 0.426 | link_hapdiv_sd | 0.159 | func_thetapi_sd | 0.210 | link_Dprime_sd | 0.217 |
| pi_intercept | 0.407 | neu_rsq_m | 0.143 | func_D_m | 0.195 | pi_slope | 0.217 |
| neu_thetaw_m | 0.404 | func_D_m | 0.141 | func_thetah_m | 0.181 | func_thetapi_sd | 0.208 |
| func_tajimasd_sd | 0.381 | link_rsq_sd | 0.126 | func_D_sd | 0.175 | link_D_m | 0.205 |
| neu_thetapi_m | 0.357 | link_hapdiv_m | 0.125 | func_tajimasd_sd | 0.171 | neu_D_m | 0.200 |
| pi_slope | 0.345 | neu_D_m | 0.120 | func_thetah_sd | 0.168 | link_Dprime_m | 0.191 |
| func_numSing_sd | 0.339 | link_Dprime_sd | 0.119 | func_div_m | 0.165 | func_div_m | 0.187 |
| link_thetapi_m | 0.326 | link_D_sd | 0.117 | func_div_sd | 0.159 | link_hprime_sd | 0.183 |
| link_thetaw_m | 0.320 | func_Dprime_m | 0.114 | func_thetaw_sd | 0.158 | link_tajimasd_sd | 0.182 |
| link_thetah_m | 0.315 | link_hprime_sd | 0.108 | func_thetaw_m | 0.148 | pi_intercept | 0.173 |
| func_numSing_m | 0.279 | func_tajimasd_m | 0.104 | func_hprime_sd | 0.136 | func_div_sd | 0.173 |
| link_div_m | 0.259 | link_thetaw_m | 0.103 | func_rsq_sd | 0.128 | link_rsq_m | 0.172 |
| func_rsq_sd | 0.255 | link_div_sd | 0.099 | neu_tajimasd_m | 0.091 | link_div_sd | 0.164 |
| neu_thetah_m | 0.251 | link_tajimasd_sd | 0.098 | link_tajimasd_m | 0.081 | func_thetah_m | 0.163 |
| link_Dprime_m | 0.220 | neu_D_sd | 0.096 | link_Dprime_m | 0.059 | link_thetapi_m | 0.156 |
| link_tajimasd_m | 0.196 | link_D_m | 0.090 | func_Dprime_sd | 0.044 | link_div_m | 0.155 |
| neu_div_m | 0.171 | neu_rsq_sd | 0.086 | neu_Dprime_m | 0.043 | link_thetah_m | 0.154 |
| neu_numSing_m | 0.076 | link_numSing_sd | 0.081 | link_numSing_m | 0.024 | neu_rsq_sd | 0.153 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| neu_Dprime_m | 0.053 | neu_thetaw_m | 0.076 | neu_numSing_m | 0.023 | func_thetah_sd | 0.153 |
| neu_tajimasd_m | 0.049 | link_rsq_m | 0.076 | func_numSing_sd | 0.021 | link_thetah_sd | 0.144 |
| func_Dprime_sd | 0.028 | link_thetaw_sd | 0.072 | neu_hapdiv_m | 0.011 | neu_D_sd | 0.142 |
| link_numSing_m | 0.027 | link_thetah_sd | 0.072 | link_Dprime_sd | 0.010 | link_numSing_sd | 0.140 |
| neu_rsq_m | 0.009 | func_thetaw_m | 0.062 | neu_thetaw_sd | 0.010 | link_thetaw_sd | 0.134 |
| link_thetapi_sd | 0.008 | link_thetapi_sd | 0.061 | neu_thetapi_sd | 0.009 | link_thetapi_sd | 0.127 |
| neu_tajimasd_sd | 0.008 | func_rsq_m | 0.058 | link_D_m | 0.008 | func_tajimasd_sd | 0.104 |
| link_thetah_sd | 0.007 | func_hapdiv_m | 0.053 | link_D_sd | 0.008 | link_thetaw_m | 0.101 |
| neu_Dprime_sd | 0.007 | neu_hprime_sd | 0.047 | neu_numSing_sd | 0.007 | neu_numSing_m | 0.098 |
| neu_D_m | 0.007 | neu_div_sd | 0.043 | link_rsq_sd | 0.007 | func_D_sd | 0.094 |
| link_thetaw_sd | 0.006 | neu_Dprime_sd | 0.039 | link_tajimasd_sd | 0.006 | neu_div_sd | 0.091 |
| link_rsq_m | 0.005 | neu_thetapi_m | 0.037 | link_hapdiv_sd | 0.005 | neu_hprime_sd | 0.078 |
| neu_hprime_sd | 0.005 | neu_numSing_m | 0.035 | link_div_sd | 0.005 | neu_tajimasd_sd | 0.075 |
| neu_rsq_sd | 0.005 | neu_tajimasd_sd | 0.034 | link_rsq_m | 0.005 | neu_Dprime_sd | 0.073 |
| link_hapdiv_m | 0.005 | neu_thetah_sd | 0.033 | func_hprime_m | 0.005 | neu_thetah_sd | 0.065 |
| link_D_m | 0.004 | func_thetah_sd | 0.033 | link_hprime_sd | 0.005 | neu_thetaw_sd | 0.053 |
| link_numSing_sd | 0.003 | func_thetaw_sd | 0.030 | neu_thetah_sd | 0.004 | neu_thetapi_sd | 0.053 |
| neu_D_sd | 0.003 | link_hprime_m | 0.029 | func_numSing_m | 0.003 | neu_thetaw_m | 0.050 |
| link_hprime_m | 0.003 | func_div_sd | 0.028 | link_hapdiv_m | 0.002 | neu_numSing_sd | 0.037 |
| neu_div_sd | 0.003 | neu_thetah_m | 0.025 | neu_div_sd | 0.002 | neu_hapdiv_m | 0.033 |
| link_D_sd | 0.002 | func_hapdiv_sd | 0.024 | neu_rsq_m | 0.001 | func_rsq_m | 0.028 |
| link_rsq_sd | 0.002 | neu_thetapi_sd | 0.024 | link_numSing_sd | 0.001 | neu_tajimasd_m | 0.019 |
| neu_hapdiv_m | 0.002 | func_thetah_m | 0.023 | pi_numbp50 | 0.001 | func_hprime_sd | 0.016 |
| link_tajimasd_sd | 0.001 | neu_thetaw_sd | 0.020 | pi_numbp75 | 0.001 | func_hprime_m | 0.014 |
| pi_max | 0.001 | func_div_m | 0.020 | pi_numbp90 | 0.001 | neu_Dprime_m | 0.014 |
| link_hapdiv_sd | 0.001 | neu_div_m | 0.018 | link_thetah_sd | 0.001 | func_Dprime_m | 0.012 |
| link_hprime_sd | 0.001 | neu_hapdiv_m | 0.015 | neu_rsq_sd | 0.001 | neu_thetapi_m | 0.012 |
| neu_thetapi_sd | 0.001 | neu_numSing_sd | 0.015 | link_thetaw_sd | 0.000 | func_D_m | 0.011 |
| link_Dprime_sd | 0.001 | neu_Dprime_m | 0.008 | link_thetapi_sd | 0.000 | link_hprime_m | 0.009 |
| neu_hapdiv_sd | 0.000 | link_numSing_m | 0.003 | link_hprime_m | 0.000 | neu_thetah_m | 0.005 |
| link_div_sd | 0.000 | neu_tajimasd_m | 0.003 | neu_hprime_m | 0.000 | link_numSing_m | 0.004 |
| neu_hprime_m | 0.000 | func_thetapi_m | 0.002 | neu_D_m | 0.000 | neu_div_m | 0.004 |
| neu_thetah_sd | 0.000 | neu_hapdiv_sd | 0.002 | pi_max | 0.000 | neu_hapdiv_sd | 0.004 |
| neu_numSing_sd | 0.000 | func_thetapi_sd | 0.001 | neu_Dprime_sd | 0.000 | pi_numbp50 | 0.003 |
| neu_thetaw_sd | 0.000 | pi_numbp50 | 0.001 | neu_hprime_sd | 0.000 | pi_numbp75 | 0.003 |
| pi_numbp50 | 0.000 | pi_numbp75 | 0.001 | neu_D_sd | 0.000 | pi_numbp90 | 0.003 |
| pi_numbp75 | 0.000 | pi_numbp90 | 0.001 | neu_hapdiv_sd | 0.000 | neu_hprime_m | 0.001 |
| pi_numbp90 | 0.000 | neu_hprime_m | 0.000 | neu_tajimasd_sd | 0.000 | func_tajimasd_m | 0.001 |

**Supp Table 4**: Statistics ranked by their importance in predicting the DFE classes under equilibrium using a modified algorithm of Joyce and Marjoram (2008) and by averaging the ranking across 10 replicates for each parameter separately.

| **Statistics ranked for $f_0$** | **Avg rank** | **Statistics ranked for $f_1$** | **Avg rank** | **Statistics ranked for $f_2$** | **Avg rank** | **Statistics ranked for $f_3$** | **Avg rank** |
|---|---|---|---|---|---|---|---|
| func_div_m | 1.2 | func_thetah_m | 4.8 | func_thetaw_m | 2.8 | func_thetaw_m | 4.3 |
| func_thetah_m | 4.7 | func_thetaw_m | 6.1 | func_thetah_m | 4.7 | func_thetah_m | 5.2 |
| func_thetaw_m | 8.5 | link_thetaw_m | 7.9 | link_thetaw_m | 6.2 | neu_thetapi_m | 8.2 |
| neu_thetapi_m | 13.8 | neu_thetapi_m | 9.6 | neu_thetapi_m | 8 | link_thetaw_m | 10.6 |
| func_hprime_m | 14.7 | link_thetapi_m | 11 | func_numSing_m | 9.1 | func_numSing_m | 12.8 |
| func_tajimasd_m | 17.6 | neu_thetaw_m | 14.2 | link_thetapi_m | 11.4 | neu_thetaw_m | 12.8 |
| link_hapdiv_m | 18.7 | func_numSing_m | 17.4 | link_thetah_m | 12.7 | link_thetapi_m | 14.3 |
| link_thetaw_m | 18.7 | link_hprime_m | 18.8 | neu_thetah_m | 18.9 | func_hprime_m | 15.8 |
| func_rsq_m | 18.9 | pi_max | 19.9 | func_Dprime_sd | 21.5 | link_hprime_m | 18.3 |
| link_thetapi_m | 19.8 | pi_numbp90 | 20.6 | func_thetapi_m | 21.5 | pi_intercept | 18.5 |
| link_tajimasd_m | 20 | func_thetapi_m | 20.7 | link_div_sd | 22.7 | func_Dprime_sd | 19.5 |
| func_numSing_m | 20.3 | func_div_m | 22 | func_div_sd | 23.2 | pi_max | 19.8 |
| neu_rsq_m | 20.7 | pi_numbp50 | 23 | pi_intercept | 23.8 | link_thetah_m | 20 |
| link_rsq_m | 21 | link_thetah_m | 23.2 | pi_numbp90 | 24.1 | pi_numbp90 | 20.1 |
| func_hapdiv_m | 23 | func_hprime_m | 23.9 | pi_slope | 24.4 | neu_Dprime_sd | 21 |
| neu_numSing_m | 23 | pi_intercept | 24.3 | neu_Dprime_sd | 24.6 | pi_numbp75 | 22.1 |
| func_D_m | 25.3 | func_Dprime_sd | 24.7 | link_Dprime_sd | 24.8 | func_thetapi_m | 24.2 |
| link_thetah_m | 25.3 | func_div_sd | 24.8 | neu_thetaw_m | 24.8 | neu_hprime_m | 24.5 |
| func_thetapi_m | 26.2 | pi_slope | 26.5 | link_hprime_m | 25.6 | link_hapdiv_sd | 24.7 |
| link_hprime_m | 26.7 | func_tajimasd_m | 26.9 | func_numSing_sd | 26.9 | pi_numbp50 | 24.9 |
| func_Dprime_sd | 26.8 | neu_hprime_m | 27.5 | link_D_sd | 27.1 | neu_D_sd | 25 |
| link_D_m | 27.3 | func_D_sd | 27.7 | pi_numbp75 | 27.4 | neu_thetah_m | 26.3 |
| neu_hprime_m | 28 | link_div_sd | 27.8 | func_hprime_m | 28.3 | func_D_sd | 26.4 |
| pi_numbp75 | 28.4 | neu_thetah_m | 29.7 | pi_max | 28.9 | func_tajimasd_m | 26.5 |
| pi_intercept | 28.8 | link_D_sd | 31.8 | link_numSing_sd | 29 | pi_slope | 26.5 |
| func_div_sd | 29.1 | neu_rsq_m | 32 | pi_numbp50 | 29 | link_div_sd | 28.6 |
| link_numSing_m | 29.6 | link_tajimasd_m | 32.9 | neu_D_sd | 30.1 | neu_div_sd | 29.2 |
| neu_D_m | 29.7 | neu_div_sd | 33.6 | func_rsq_sd | 30.4 | func_div_sd | 30.1 |
| neu_thetah_m | 30 | pi_numbp75 | 34.1 | neu_rsq_sd | 31.7 | func_rsq_sd | 31.1 |
| pi_max | 30.7 | link_rsq_sd | 34.5 | neu_div_sd | 32.2 | neu_rsq_sd | 32.7 |
| neu_rsq_sd | 31.5 | neu_D_sd | 34.6 | func_D_sd | 32.7 | neu_tajimasd_sd | 33.6 |
| neu_thetaw_m | 32.5 | func_D_m | 34.9 | func_tajimasd_m | 33.6 | link_rsq_sd | 33.7 |
| func_Dprime_m | 36 | link_Dprime_sd | 35.6 | neu_hprime_m | 34.6 | neu_hapdiv_sd | 33.7 |
| func_D_sd | 36.5 | func_rsq_m | 35.9 | func_hapdiv_sd | 35.2 | link_Dprime_sd | 34.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pi_numbp90 | 38.2 | link_hapdiv_m | 36 | func_thetah_sd | 36.1 | link_D_sd | 37.2 |
| link_Dprime_sd | 38.4 | func_numSing_sd | 36.4 | link_hapdiv_sd | 36.4 | func_numSing_sd | 37.6 |
| link_hapdiv_sd | 38.6 | neu_Dprime_sd | 37.5 | link_thetapi_sd | 36.4 | neu_numSing_sd | 38.8 |
| func_hapdiv_sd | 38.9 | link_rsq_m | 39.4 | neu_hapdiv_sd | 36.5 | link_tajimasd_m | 40.8 |
| pi_numbp50 | 39.1 | neu_hapdiv_sd | 40.7 | func_hprime_sd | 38 | neu_thetah_sd | 41.2 |
| neu_D_sd | 39.2 | neu_numSing_sd | 41.3 | func_tajimasd_sd | 38.1 | link_thetaw_sd | 41.3 |
| link_rsq_sd | 40.3 | neu_numSing_m | 42.1 | link_hprime_sd | 40.7 | func_hapdiv_sd | 41.7 |
| link_div_sd | 40.6 | link_D_m | 42.7 | func_div_m | 41.1 | link_numSing_sd | 42 |
| func_rsq_sd | 41.9 | func_Dprime_m | 43 | neu_tajimasd_sd | 41.1 | func_thetapi_sd | 42.8 |
| pi_slope | 43.6 | link_thetaw_sd | 44.3 | link_rsq_sd | 41.7 | link_hprime_sd | 42.8 |
| link_numSing_sd | 43.8 | link_hapdiv_sd | 44.4 | neu_numSing_sd | 42.1 | func_thetaw_sd | 44.8 |
| neu_Dprime_sd | 43.8 | neu_hprime_sd | 44.7 | link_tajimasd_sd | 42.7 | func_rsq_m | 44.9 |
| neu_hapdiv_sd | 44.2 | func_tajimasd_sd | 44.9 | neu_thetaw_sd | 43.2 | neu_hprime_sd | 45 |
| link_tajimasd_sd | 44.5 | link_hprime_sd | 45.1 | neu_thetapi_sd | 43.6 | func_Dprime_m | 45.1 |
| func_numSing_sd | 44.7 | func_hapdiv_sd | 45.4 | func_D_m | 44.6 | func_hprime_sd | 45.3 |
| link_Dprime_m | 44.7 | func_rsq_sd | 45.5 | neu_hprime_sd | 44.8 | func_tajimasd_sd | 46.5 |
| neu_numSing_sd | 45 | func_thetaw_sd | 46.7 | link_div_m | 46.5 | link_tajimasd_sd | 46.8 |
| neu_hapdiv_m | 45.9 | link_tajimasd_sd | 46.7 | link_Dprime_m | 47.7 | func_div_m | 47.3 |
| neu_tajimasd_m | 46.9 | neu_rsq_sd | 46.7 | link_hapdiv_m | 48.5 | link_thetah_sd | 48.2 |
| func_tajimasd_sd | 47.2 | neu_thetah_sd | 47.4 | func_Dprime_m | 49.7 | func_thetah_sd | 48.4 |
| link_D_sd | 47.9 | link_Dprime_m | 47.5 | neu_div_m | 50 | func_D_m | 48.8 |
| neu_div_sd | 48.4 | func_thetah_sd | 47.8 | link_rsq_m | 50.1 | link_D_m | 48.9 |
| func_thetaw_sd | 49.2 | neu_tajimasd_sd | 47.9 | neu_Dprime_m | 50.1 | neu_thetaw_sd | 49.6 |
| link_thetaw_sd | 49.6 | link_numSing_sd | 48.2 | func_thetaw_sd | 50.5 | link_div_m | 50.1 |
| neu_tajimasd_sd | 49.8 | func_hprime_sd | 49.2 | func_thetapi_sd | 51.1 | neu_D_m | 50.4 |
| func_thetapi_sd | 51.3 | neu_thetaw_sd | 49.4 | link_thetaw_sd | 51.3 | neu_div_m | 51.1 |
| func_hprime_sd | 52.7 | neu_D_m | 49.5 | link_thetah_sd | 51.4 | link_rsq_m | 51.3 |
| neu_Dprime_m | 52.8 | func_hapdiv_m | 49.9 | func_hapdiv_m | 51.8 | neu_rsq_m | 54 |
| neu_thetaw_sd | 55 | link_thetapi_sd | 50.1 | neu_D_m | 52.9 | link_thetapi_sd | 54.2 |
| neu_thetah_sd | 56.3 | link_div_m | 51.2 | neu_thetah_sd | 53 | neu_thetapi_sd | 54.8 |
| neu_hprime_sd | 56.7 | func_thetapi_sd | 51.7 | func_rsq_m | 53.3 | neu_Dprime_m | 56.5 |
| neu_div_m | 57.1 | link_thetah_sd | 52 | link_tajimasd_m | 54.1 | link_Dprime_m | 57.4 |
| link_hprime_sd | 57.5 | neu_thetapi_sd | 53.8 | link_D_m | 56.5 | neu_tajimasd_m | 58.4 |
| link_thetah_sd | 58.6 | neu_div_m | 56.4 | neu_rsq_m | 58.3 | func_hapdiv_m | 59.4 |
| neu_thetapi_sd | 59.5 | neu_tajimasd_m | 57.5 | neu_numSing_m | 63.4 | link_numSing_m | 61.2 |
| link_div_m | 60 | link_numSing_m | 57.9 | neu_tajimasd_m | 65.3 | neu_numSing_m | 62.5 |
| func_thetah_sd | 60.6 | neu_Dprime_m | 59.9 | link_numSing_m | 65.6 | link_hapdiv_m | 62.6 |
| link_thetapi_sd | 62.5 | neu_hapdiv_m | 64.3 | neu_hapdiv_m | 67.9 | neu_hapdiv_m | 69 |

**Supp Table 5**: Ranking of statistics under demographic non-equilibrium. Statistics significantly correlated with parameters of the DFE when statistics from all regions are used and when only functional statistics are used for ranking. Significance was evaluated with $p < 0.05$ with Bonferonni correction.

| **Ranking using all statistics** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Statistics ranked for $f_0$** | **$r^2$** | **Statistics ranked for $f_1$** | **$r^2$** | **Statistics ranked for $f_2$** | **$r^2$** | **Statistics ranked for $f_3$** | **$r^2$** |
| func_div_m | 0.893 | func_hprime_m | 0.273 | func_div_sd | 0.125 | func_numSing_sd | 0.205 |
| func_div_sd | 0.850 | func_tajimasd_sd | 0.121 | func_div_m | 0.125 | func_thetaw_sd | 0.200 |
| func_thetah_sd | 0.632 | func_hprime_sd | 0.099 | func_thetapi_sd | 0.117 | func_thetaw_m | 0.180 |
| func_thetapi_sd | 0.612 | func_rsq_sd | 0.082 | func_thetapi_m | 0.114 | func_numSing_m | 0.180 |
| func_thetah_m | 0.585 | func_Dprime_sd | 0.079 | func_thetaw_sd | 0.107 | func_div_m | 0.117 |
| func_thetapi_m | 0.556 | func_tajimasd_m | 0.077 | func_thetaw_m | 0.098 | func_hapdiv_sd | 0.114 |
| func_thetaw_sd | 0.473 | func_div_m | 0.065 | func_thetah_sd | 0.094 | func_div_sd | 0.108 |
| func_thetaw_m | 0.407 | func_Dprime_m | 0.060 | func_hapdiv_sd | 0.092 | func_hapdiv_m | 0.106 |
| func_Dprime_m | 0.343 | func_div_sd | 0.059 | func_thetah_m | 0.085 | func_thetapi_sd | 0.104 |
| func_tajimasd_m | 0.325 | func_numSing_sd | 0.056 | func_hapdiv_m | 0.079 | func_thetapi_m | 0.102 |
| func_hprime_m | 0.286 | func_numSing_m | 0.056 | func_Dprime_m | 0.075 | func_Dprime_sd | 0.077 |
| func_hapdiv_sd | 0.284 | func_thetah_m | 0.053 | func_tajimasd_m | 0.072 | func_thetah_sd | 0.074 |
| func_hapdiv_m | 0.209 | func_thetah_sd | 0.048 | func_rsq_m | 0.045 | func_tajimasd_sd | 0.062 |
| func_numSing_sd | 0.143 | func_D_sd | 0.034 | func_numSing_sd | 0.029 | func_thetah_m | 0.061 |
| func_rsq_m | 0.142 | func_hapdiv_m | 0.020 | func_numSing_m | 0.020 | func_rsq_sd | 0.023 |
| func_hprime_sd | 0.116 | func_thetapi_sd | 0.015 | func_D_m | 0.015 | func_hprime_sd | 0.010 |
| func_numSing_m | 0.102 | func_D_m | 0.014 | func_hprime_sd | 0.015 | func_D_sd | 0.008 |
| func_D_m | 0.081 | func_hapdiv_sd | 0.010 | func_rsq_sd | 0.010 | func_rsq_m | 0.005 |
| func_rsq_sd | 0.057 | func_thetapi_m | 0.009 | func_D_sd | 0.007 | | |
| func_D_sd | 0.033 | func_rsq_m | 0.009 | | | | |
| func_tajimasd_sd | 0.023 | func_thetaw_m | 0.008 | | | | |
| | | func_thetaw_sd | 0.006 | | | | |
| | | | | | | | |
| **Ranking using statistics calculated from functional regions.** | | | | | | | |
| **Statistics ranked for $f_0$** | **$r^2$** | **Statistics ranked for $f_1$** | **$r^2$** | **Statistics ranked for $f_2$** | **$r^2$** | **Statistics ranked for $f_3$** | **$r^2$** |
| func_div_m | 0.89 | func_hprime_m | 0.27 | func_div_sd | 0.12 | func_numSing_sd | 0.20 |
| func_div_sd | 0.85 | func_tajimasd_sd | 0.12 | func_div_m | 0.12 | func_thetaw_sd | 0.20 |
| func_thetah_sd | 0.63 | func_hprime_sd | 0.10 | func_thetapi_sd | 0.12 | func_thetaw_m | 0.18 |
| func_thetapi_sd | 0.61 | func_rsq_sd | 0.08 | func_thetapi_m | 0.11 | func_numSing_m | 0.18 |
| func_thetah_m | 0.59 | func_Dprime_sd | 0.08 | func_thetaw_sd | 0.11 | func_div_m | 0.12 |
| func_thetapi_m | 0.56 | func_tajimasd_m | 0.08 | func_thetaw_m | 0.10 | func_hapdiv_sd | 0.11 |
| func_thetaw_sd | 0.47 | func_div_m | 0.06 | func_thetah_sd | 0.09 | func_div_sd | 0.11 |
| func_thetaw_m | 0.41 | func_Dprime_m | 0.06 | func_hapdiv_sd | 0.09 | func_hapdiv_m | 0.11 |
| func_Dprime_m | 0.34 | func_div_sd | 0.06 | func_thetah_m | 0.08 | func_thetapi_sd | 0.10 |
| func_tajimasd_m | 0.32 | func_numSing_sd | 0.06 | func_hapdiv_m | 0.08 | func_thetapi_m | 0.10 |
| func_hprime_m | 0.29 | func_numSing_m | 0.06 | func_Dprime_m | 0.07 | func_Dprime_sd | 0.08 |
| func_hapdiv_sd | 0.28 | func_thetah_m | 0.05 | func_tajimasd_m | 0.07 | func_thetah_sd | 0.07 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| func_hapdiv_m | 0.21 | func_thetah_sd | 0.05 | func_rsq_m | 0.04 | func_tajimasd_sd | 0.06 |
| func_numSing_sd | 0.14 | func_D_sd | 0.03 | func_numSing_sd | 0.03 | func_thetah_m | 0.06 |
| func_rsq_m | 0.14 | func_hapdiv_m | 0.02 | func_numSing_m | 0.02 | func_rsq_sd | 0.02 |
| func_hprime_sd | 0.12 | func_thetapi_sd | 0.01 | func_D_m | 0.02 | func_hprime_sd | 0.01 |
| func_numSing_m | 0.10 | func_D_m | 0.01 | func_hprime_sd | 0.01 | func_D_sd | 0.01 |
| func_D_m | 0.08 | func_hapdiv_sd | 0.01 | func_rsq_sd | 0.01 | func_rsq_m | 0.01 |
| func_rsq_sd | 0.06 | func_thetapi_m | 0.01 | func_D_sd | 0.01 | func_Dprime_m | 0.00 |
| func_D_sd | 0.03 | func_rsq_m | 0.01 | | | | |
| func_tajimasd_sd | 0.02 | func_thetaw_m | 0.01 | | | | |
| | | func_thetaw_sd | 0.01 | | | | |

**Supp Table 6**: Ranking of statistics when distinguishing between demography and purifying selection. Statistics significantly correlated with parameters of demography when statistics from all regions are used, and when only functional statistics are used for ranking. Significance was evaluated with $p < 0.05$ with Bonferonni correction.

| **Ranking using all statistics** | | | | **Ranking using statistics calculated from functional regions.** | | | |
|---|---|---|---|---|---|---|---|
| **Statistics ranked for *N*anc** | $r^2$ | **Statistics ranked for *N*cur** | $r^2$ | **Statistics ranked for *N*anc** | $r^2$ | **Statistics ranked for *N*cur** | $r^2$ |
| neu_thetah_m | 0.99 | neu_hapdiv_m | 0.92 | func_thetapi_m | 0.230 | func_rsq_sd | 0.659 |
| link_thetah_m | 0.99 | link_hapdiv_m | 0.91 | func_thetah_m | 0.229 | func_rsq_m | 0.618 |
| neu_thetapi_m | 0.93 | link_numSing_m | 0.84 | func_thetapi_sd | 0.196 | func_numSing_m | 0.577 |
| link_thetapi_m | 0.93 | neu_numSing_m | 0.84 | func_thetaw_sd | 0.192 | func_D_sd | 0.557 |
| link_thetapi_sd | 0.93 | neu_rsq_m | 0.83 | func_thetah_sd | 0.179 | func_tajimasd_sd | 0.487 |
| link_thetah_sd | 0.93 | link_rsq_m | 0.82 | func_thetaw_m | 0.159 | func_numSing_sd | 0.482 |
| neu_thetah_sd | 0.91 | neu_rsq_sd | 0.79 | func_Dprime_m | 0.136 | func_Dprime_sd | 0.435 |
| neu_thetapi_sd | 0.91 | link_rsq_sd | 0.79 | func_tajimasd_m | 0.135 | func_D_m | 0.374 |
| neu_thetaw_sd | 0.90 | neu_hprime_sd | 0.78 | func_hapdiv_sd | 0.121 | func_hprime_sd | 0.368 |
| link_thetaw_sd | 0.90 | link_hprime_sd | 0.78 | func_hapdiv_m | 0.078 | func_tajimasd_m | 0.275 |
| neu_thetaw_m | 0.72 | link_hapdiv_sd | 0.76 | func_hprime_m | 0.069 | func_Dprime_m | 0.216 |
| link_thetaw_m | 0.71 | neu_hapdiv_sd | 0.74 | func_numSing_sd | 0.042 | func_hapdiv_m | 0.165 |
| link_div_sd | 0.49 | link_D_sd | 0.68 | func_rsq_m | 0.041 | func_thetaw_m | 0.157 |
| neu_div_sd | 0.45 | func_rsq_sd | 0.66 | func_Dprime_sd | 0.032 | func_hapdiv_sd | 0.129 |
| neu_Dprime_m | 0.45 | link_numSing_sd | 0.64 | func_numSing_m | 0.025 | func_thetaw_sd | 0.070 |
| link_Dprime_m | 0.44 | neu_numSing_sd | 0.64 | func_rsq_sd | 0.012 | func_hprime_m | 0.061 |
| link_tajimasd_m | 0.43 | link_tajimasd_sd | 0.64 | func_tajimasd_sd | 0.009 | func_div_m | 0.022 |
| neu_tajimasd_m | 0.43 | neu_D_sd | 0.64 | func_D_m | 0.008 | func_div_sd | 0.017 |
| neu_Dprime_sd | 0.41 | func_rsq_m | 0.62 | | | func_thetapi_m | 0.008 |
| link_Dprime_sd | 0.38 | neu_tajimasd_sd | 0.62 | | | | |
| link_hprime_m | 0.35 | neu_div_m | 0.59 | | | | |
| neu_hprime_m | 0.34 | link_D_m | 0.58 | | | | |
| link_div_m | 0.31 | func_numSing_m | 0.58 | | | | |
| neu_div_m | 0.30 | link_div_m | 0.57 | | | | |
| neu_numSing_sd | 0.25 | func_D_sd | 0.56 | | | | |
| link_numSing_sd | 0.25 | neu_D_m | 0.56 | | | | |
| func_thetapi_m | 0.23 | func_tajimasd_sd | 0.49 | | | | |
| func_thetah_m | 0.23 | func_numSing_sd | 0.48 | | | | |
| func_thetapi_sd | 0.20 | func_Dprime_sd | 0.44 | | | | |

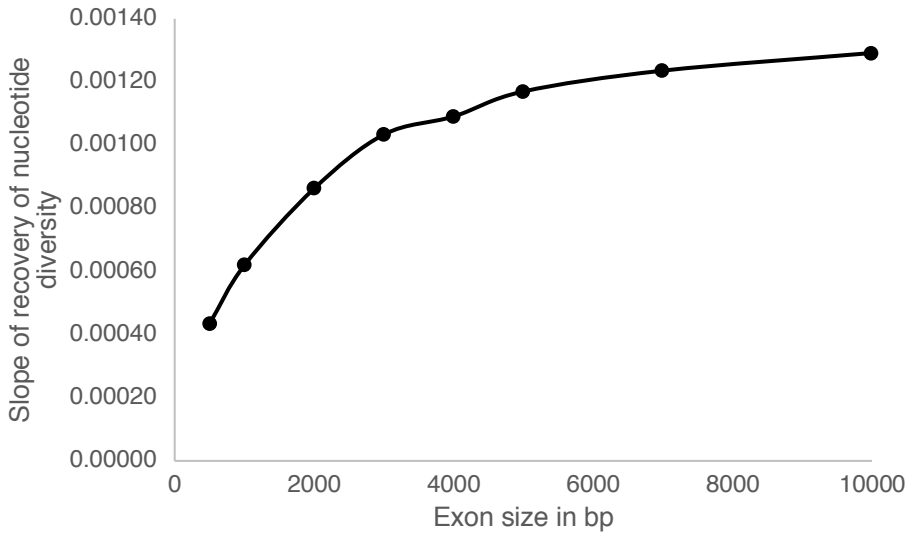
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| neu_tajimasd_sd | 0.19 | link_tajimasd_m | 0.43 | | | | |
| func_thetaw_sd | 0.19 | neu_tajimasd_m | 0.43 | | | | |
| func_thetah_sd | 0.18 | link_Dprime_sd | 0.39 | | | | |
| link_tajimasd_sd | 0.16 | neu_Dprime_sd | 0.38 | | | | |
| func_thetaw_m | 0.16 | func_D_m | 0.37 | | | | |
| func_Dprime_m | 0.14 | func_hprime_sd | 0.37 | | | | |
| func_tajimasd_m | 0.14 | link_Dprime_m | 0.35 | | | | |
| func_hapdiv_sd | 0.12 | neu_Dprime_m | 0.35 | | | | |
| neu_numSing_m | 0.11 | neu_hprime_m | 0.34 | | | | |
| link_numSing_m | 0.11 | link_hprime_m | 0.34 | | | | |
| link_rsq_sd | 0.08 | func_tajimasd_m | 0.27 | | | | |
| neu_rsq_sd | 0.08 | link_thetaw_m | 0.22 | | | | |
| func_hapdiv_m | 0.08 | func_Dprime_m | 0.22 | | | | |
| link_rsq_m | 0.08 | neu_thetaw_m | 0.21 | | | | |
| func_hprime_m | 0.07 | func_hapdiv_m | 0.16 | | | | |
| neu_rsq_m | 0.07 | func_thetaw_m | 0.16 | | | | |
| func_numSing_sd | 0.04 | neu_div_sd | 0.15 | | | | |
| func_rsq_m | 0.04 | link_div_sd | 0.14 | | | | |
| func_Dprime_sd | 0.03 | func_hapdiv_sd | 0.13 | | | | |
| func_numSing_m | 0.03 | func_thetaw_sd | 0.07 | | | | |
| func_rsq_sd | 0.01 | func_hprime_m | 0.06 | | | | |
| link_D_m | 0.01 | link_thetaw_sd | 0.04 | | | | |
| link_hapdiv_sd | 0.01 | neu_thetaw_sd | 0.03 | | | | |
| link_hapdiv_m | 0.01 | link_thetapi_m | 0.03 | | | | |
| func_tajimasd_sd | 0.01 | neu_thetapi_m | 0.03 | | | | |
| func_D_m | 0.01 | func_div_m | 0.02 | | | | |
| neu_D_m | 0.01 | func_div_sd | 0.02 | | | | |
| | | neu_thetapi_sd | 0.01 | | | | |
| | | func_thetapi_m | 0.01 | | | | |
| | | neu_thetah_sd | 0.01 | | | | |
| | | link_thetapi_sd | 0.01 | | | | |

**Supp Table 7**: The rate of fixed differences (*i.e.*, polymorphism-adjusted divergence) for different site types in *D. melanogaster*, where different numbers of individuals from the Zambia population were used to identify the set of polymorphic sites.

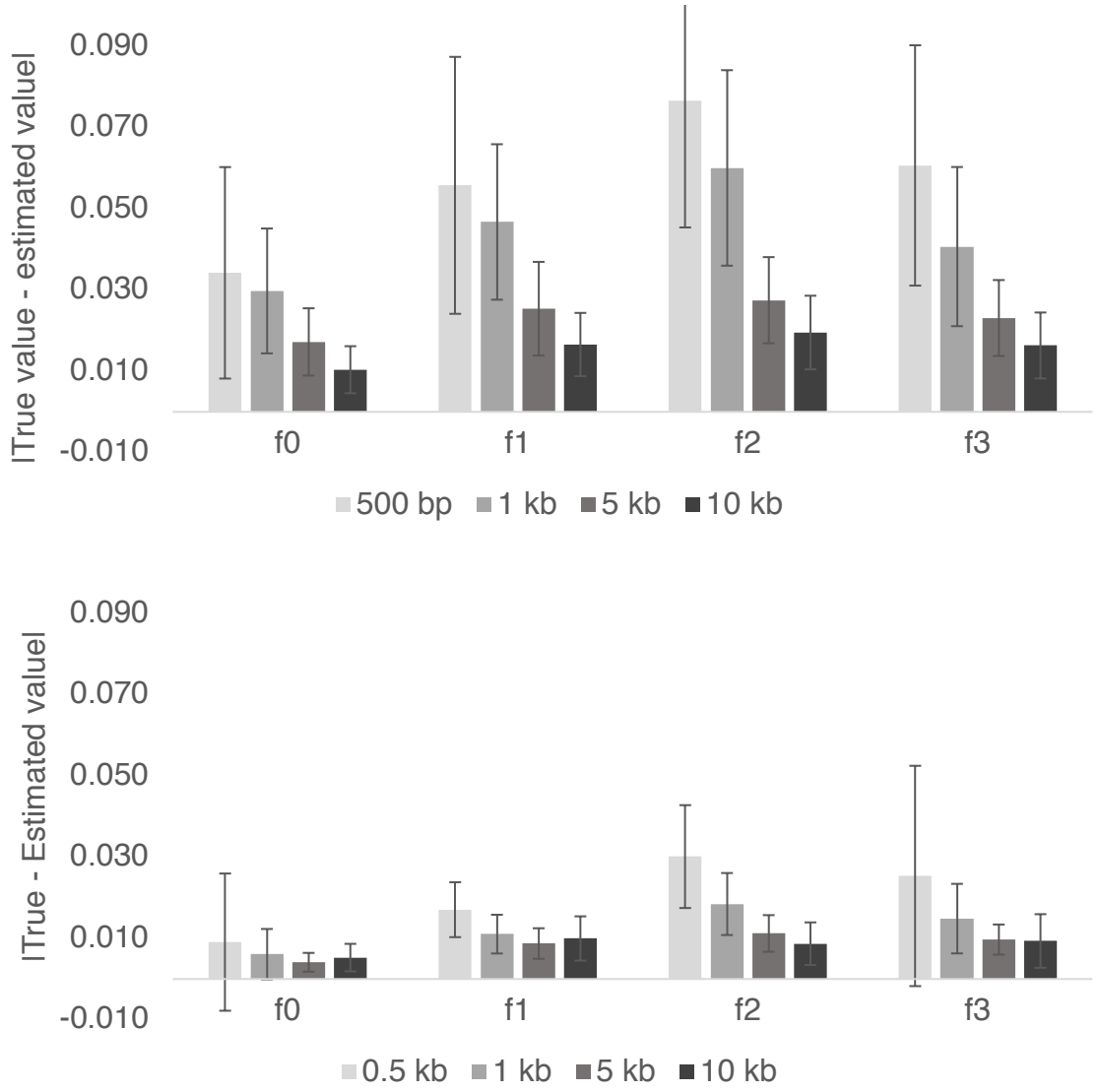
| | **Sample size:** | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **5** | **15** | **30** | **76** |
| **exon** | 0.0238 | 0.0198 | 0.0170 | 0.0160 | 0.0159 | 0.0153 |
| **coding** | 0.0228 | 0.0182 | 0.0157 | 0.0146 | 0.0141 | 0.0135 |
| **4-fold degenerate** | 0.0497 | 0.0423 | 0.0349 | 0.0316 | 0.0311 | 0.0300 |
| **0-fold degenerate** | 0.0182 | 0.0123 | 0.0108 | 0.0102 | 0.0098 | 0.0094 |

**Supp Table 8**: The increase in divergence values obtained when calculating pairwise divergence (corresponding to sample size of 1) relative to when a much larger number of individuals are used to calculate the rate of fixed differences with exclusion of polymorphic sites.

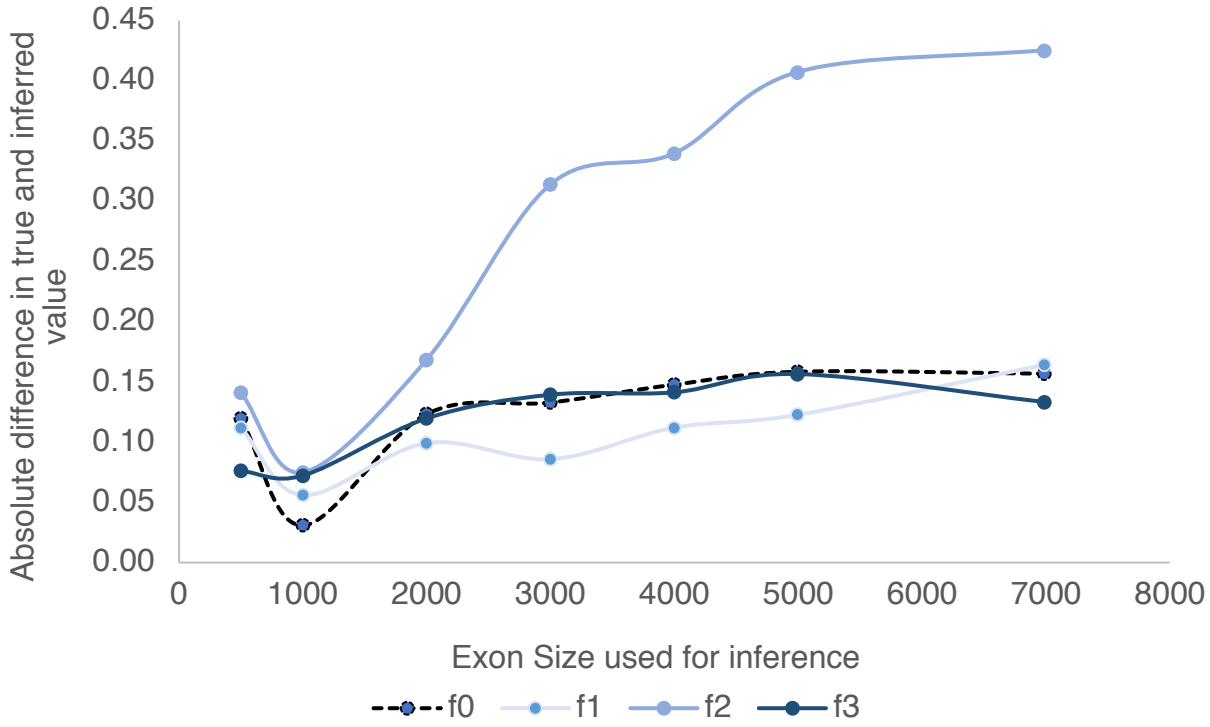
| | **Sample size:** | | |
|---|---|---|---|
| | **1** | **76** | **100** |
| ***D. melanogaster* exon** | 1.551 | 1.000 | |
| ***D. melanogaster* 4-fold degenerate** | 1.658 | 1.000 | |
| **Simulated exon** | 1.736 | | 1.000 |
| **Simulated neutral** | 1.634 | | 1.000 |



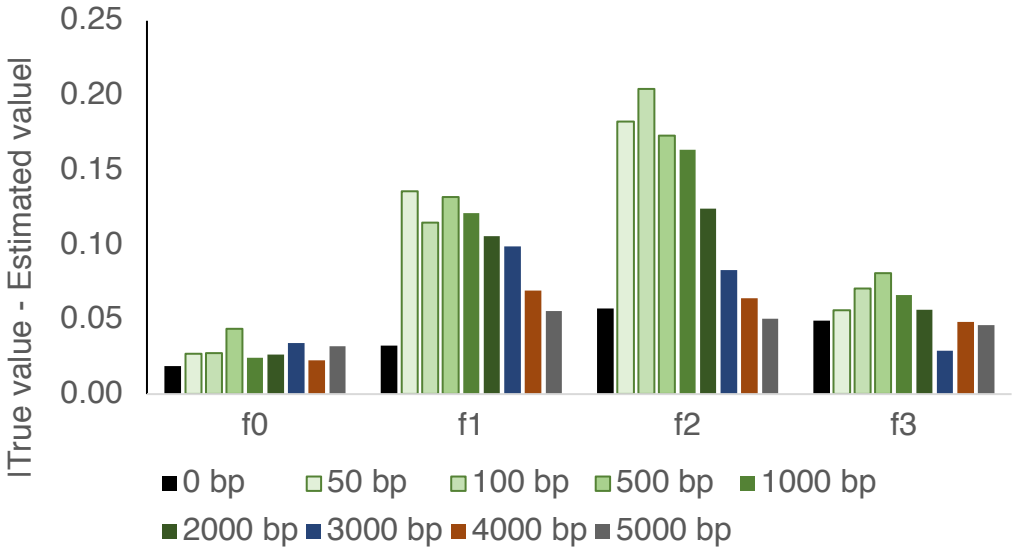

**Supp Figure 1:** Increase in the slope of recovery of diversity near functional regions of varying sizes. Larger values of slope represent a steeper recovery, concordant with larger reduction in diversity observed in the non-coding region.



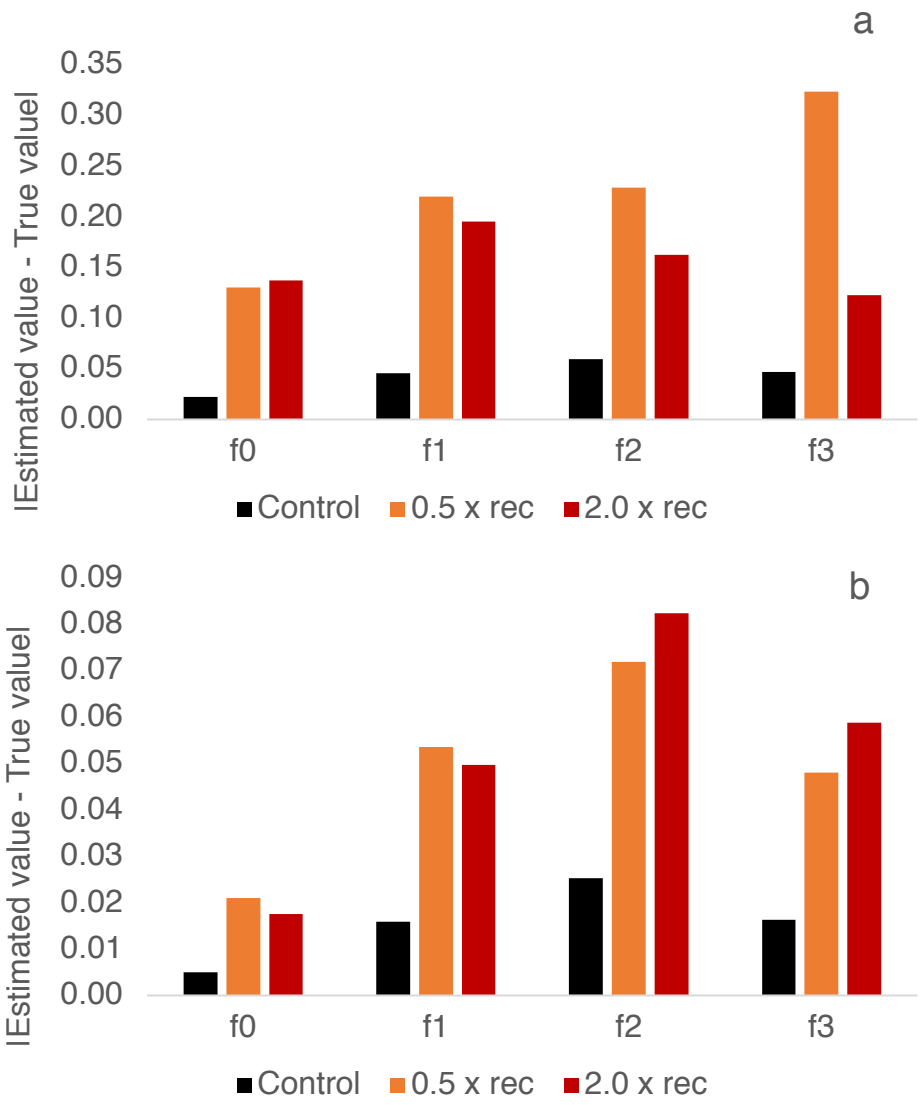

**Supp Figure 2**: Absolute difference between true and inferred value of parameters characterizing the DFE for 0.5 kb, 1 kb, 5 kb, and 10 kb functional regions. The upper panel displays the error in inference when using all statistics, while the lower uses only functional regions.



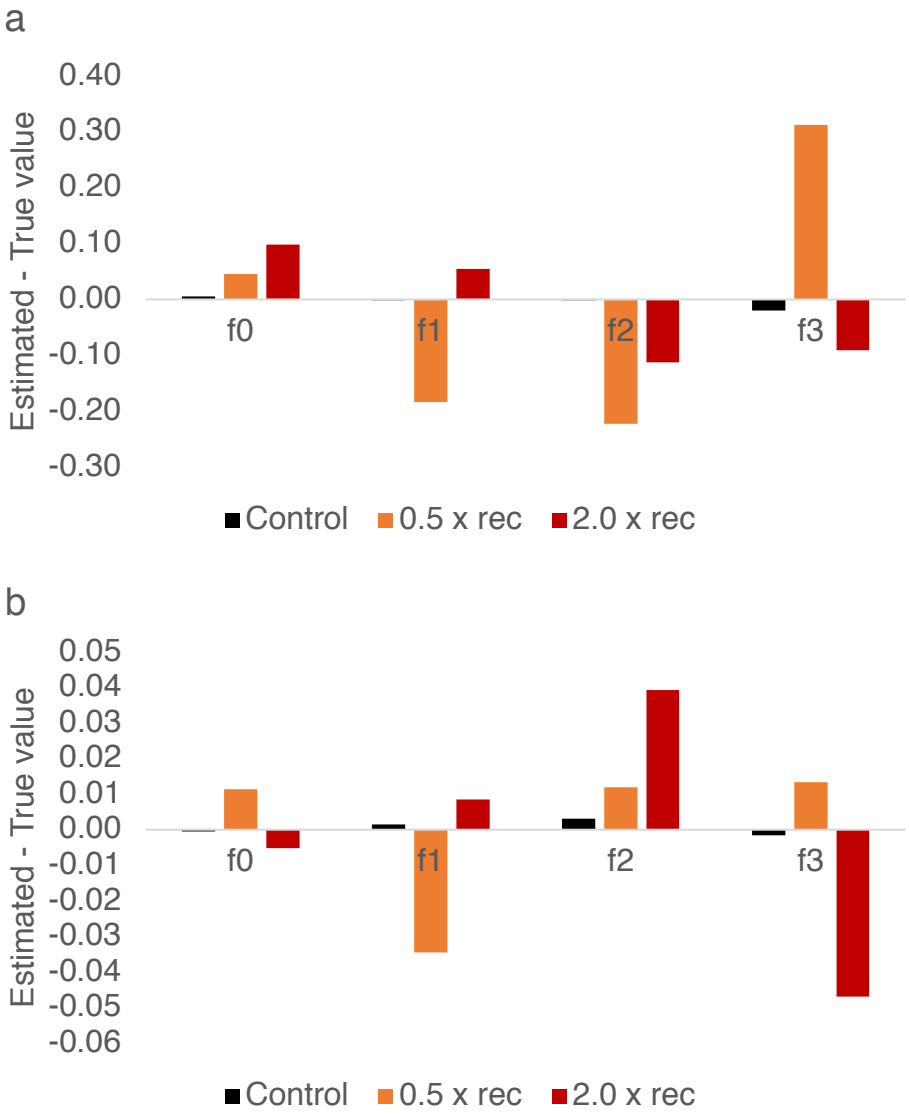

**Supp Figure 3:** Decrease in accuracy of inference for different DFE classes as the exon size assumed for inference is mis-specified. In this figure, the assumed exon size was 1kb, and the X-axis gives the true exon size.



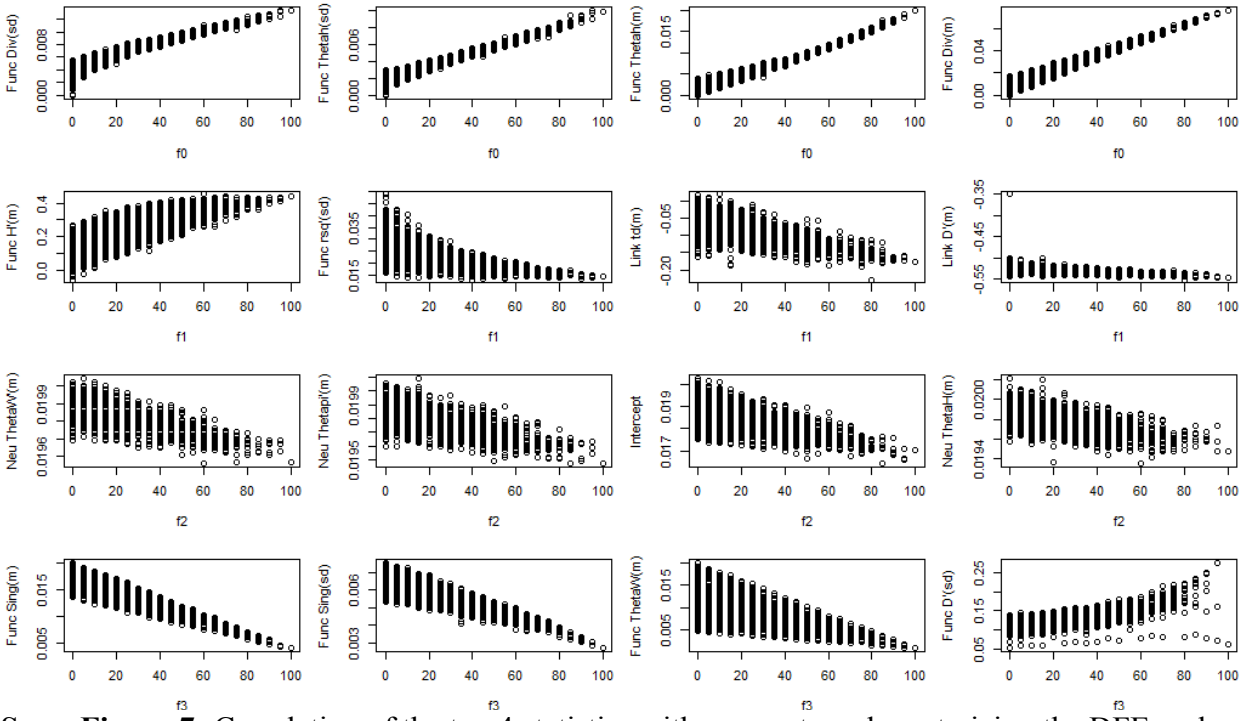

**Supp Figure 4**: Mis-inference of DFE in the presence of an additional unaccounted for 1 kb functional region near the target 1kb exon used for inference. The intron/ intergenic distance between the two exons varies from 50-5000 bp, as shown by different colored bars. "0 bp" represents the negative control where there is no additional 1 kb exon present.



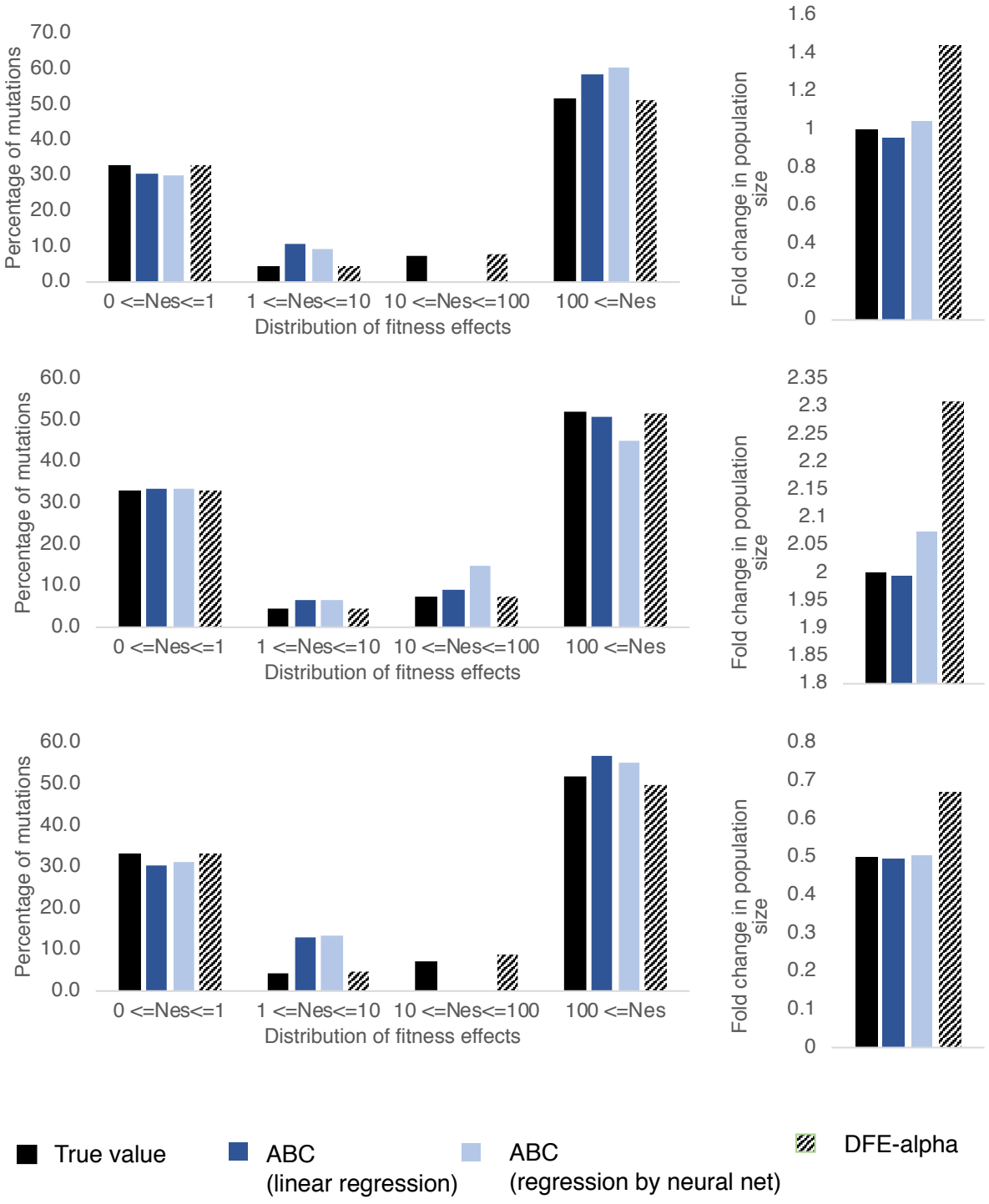

**Supp Figure 5**: Absolute difference between the true and estimated value of the DFE class, when the true recombination rate is half of that assumed for inference (orange) and when the true value is twice that assumed for inference (red), using a) all statistics and b) statistics only pertaining to the functional region.



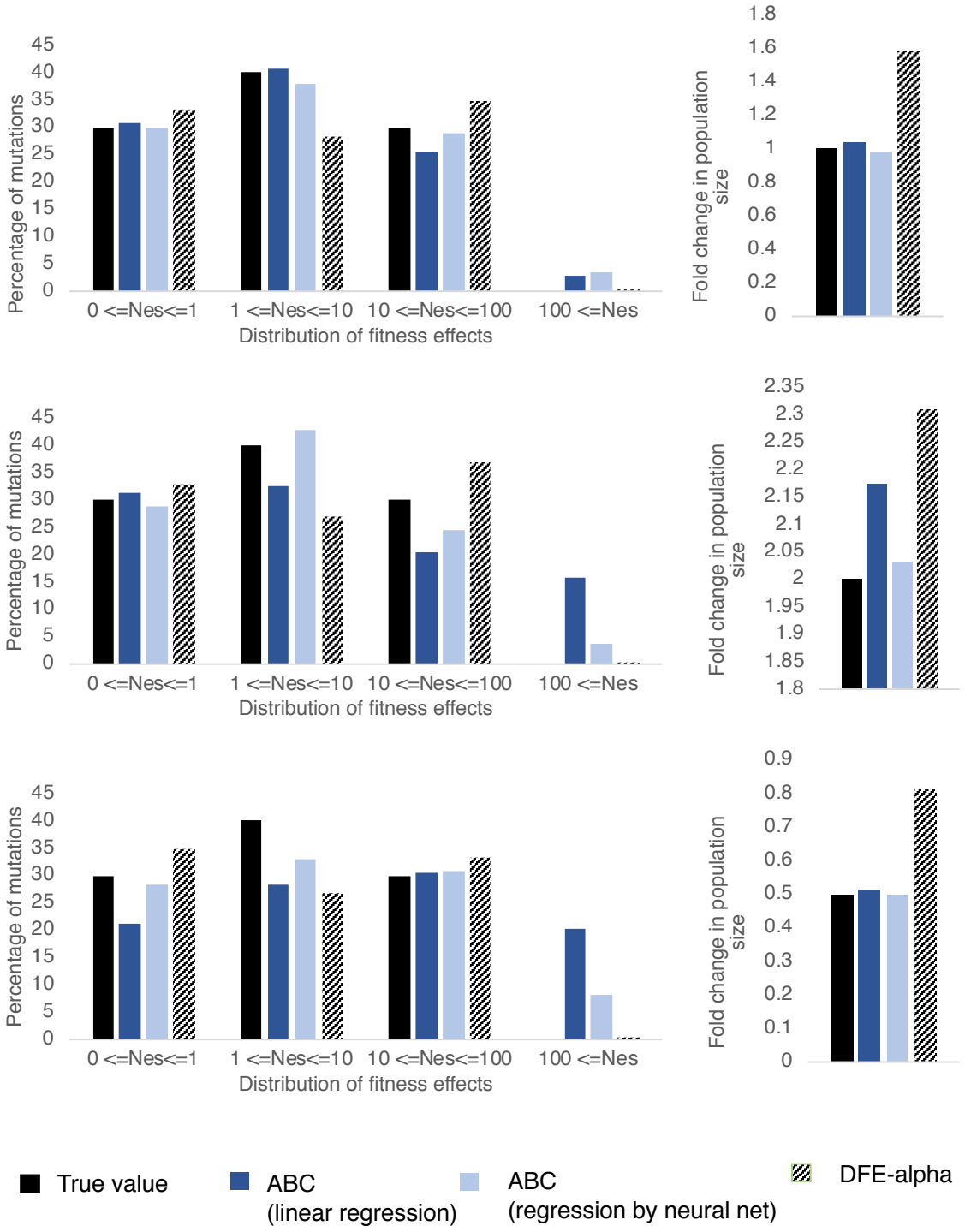

**Supp Figure 6**: Following Supp Figure 5, the direction of bias in inference of the DFE classes upon mis-specification of the recombination rate, using a) all statistics and b) statistics only pertaining to the functional region.



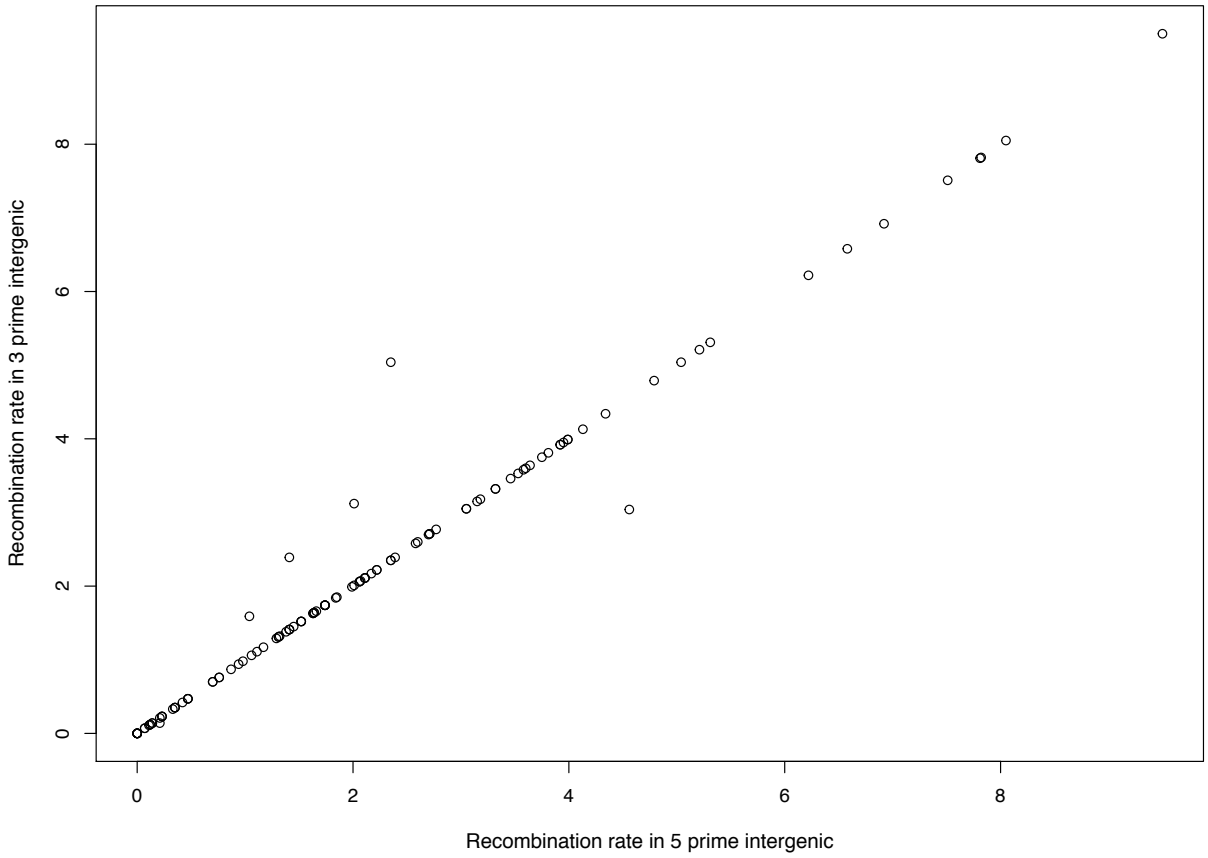

**Supp Figure 7:** Correlation of the top 4 statistics with parameters characterizing the DFE under demographic equilibrium. "Func" corresponds to the functional region, "Link" to the immediately linked region and "Neu" to the less linked region.



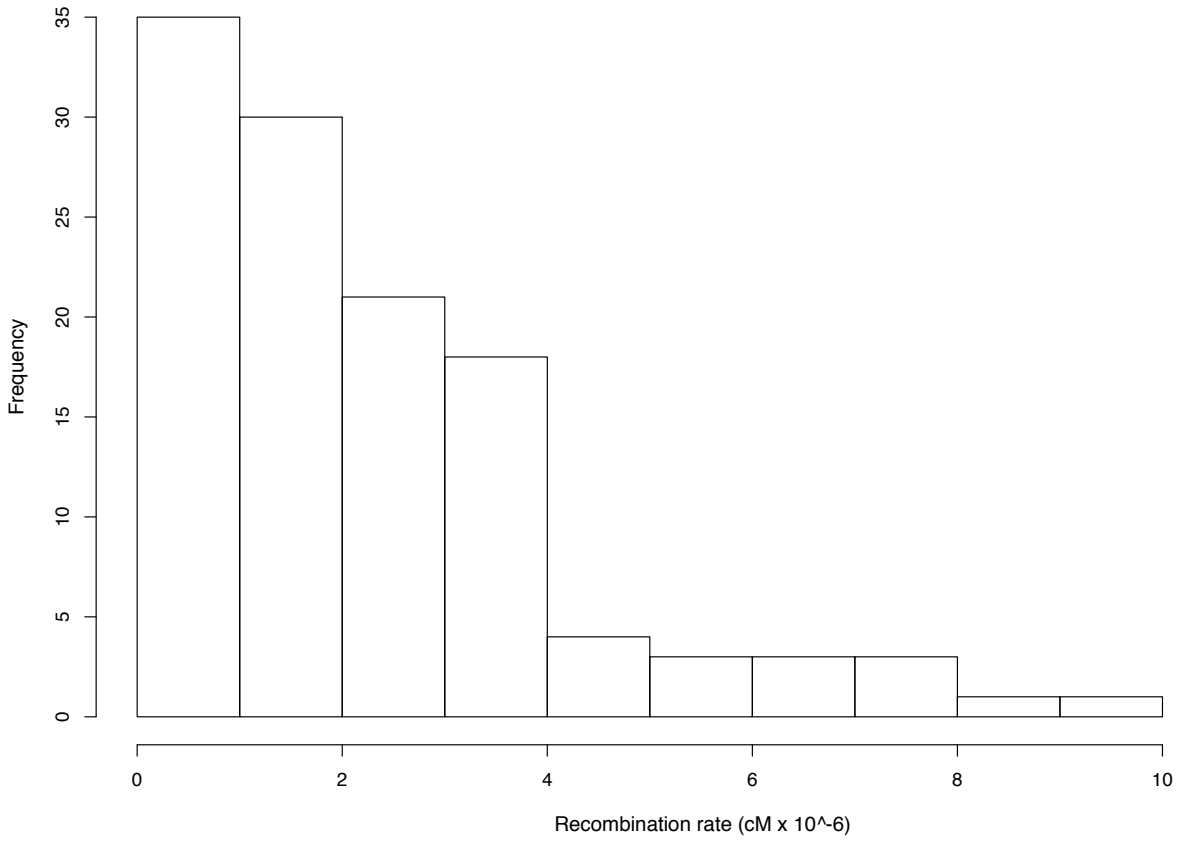

**Supp Figure 8**: Inference of demography and the DFE by the approach proposed here and DFE-alpha, when the true shape of the DFE is gamma distributed, for equilibrium (top panel), growth (middle panel), and decline (bottom panel). Solid black bars show the true value simulated, dark blue bars show our ABC performance using ridge regression, light blue bars show the ABC performance using linear regression aided by neural net. Patterned bars show the performance of DFE-alpha.



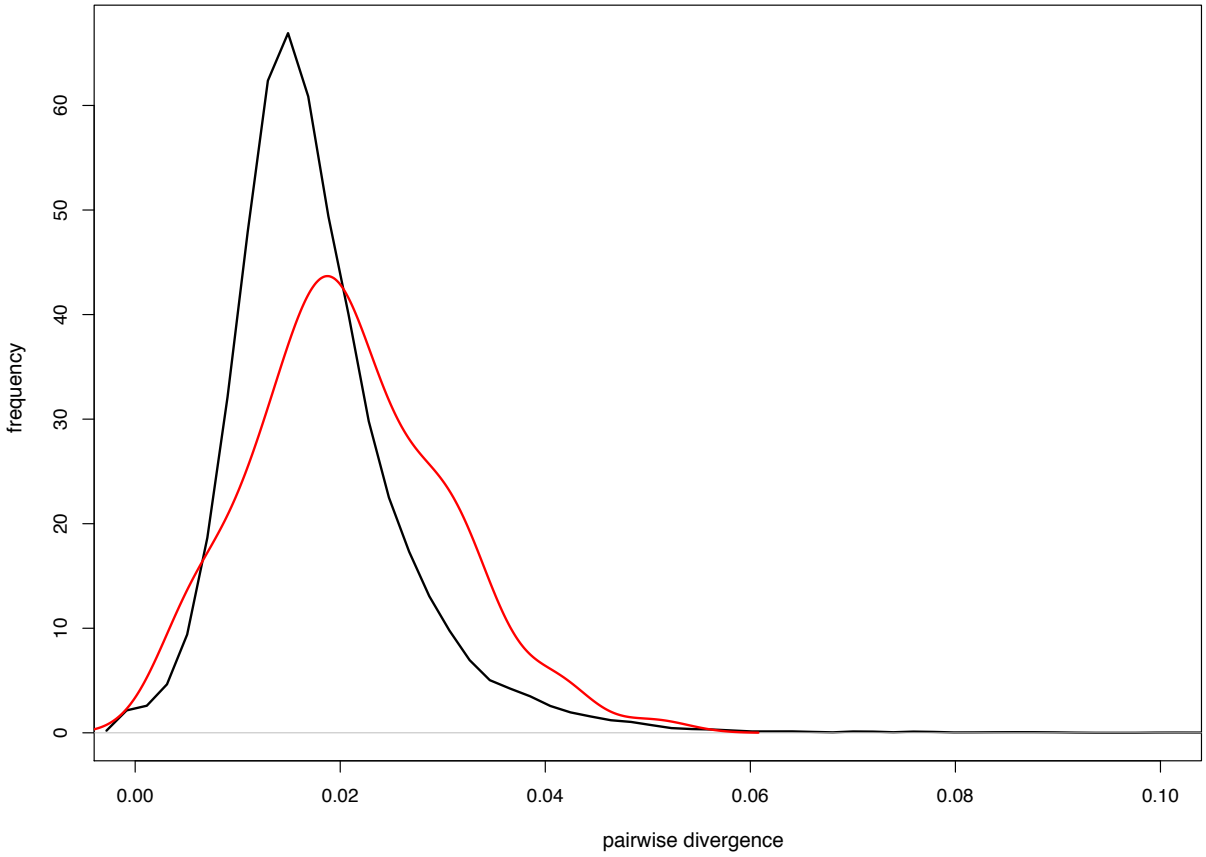

**Supp Figure 9**: Inference of demography and the DFE when the true shape of the DFE is discrete and skewed towards slightly deleterious class of mutations, for equilibrium (top panel), growth (middle panel), and decline (bottom panel). Solid black bars show the true value simulated, dark blue bars show the ABC performance using ridge regression, light blue bars show the ABC performance using linear regression aided by neural net. Patterned bars show the performance of DFE-alpha.



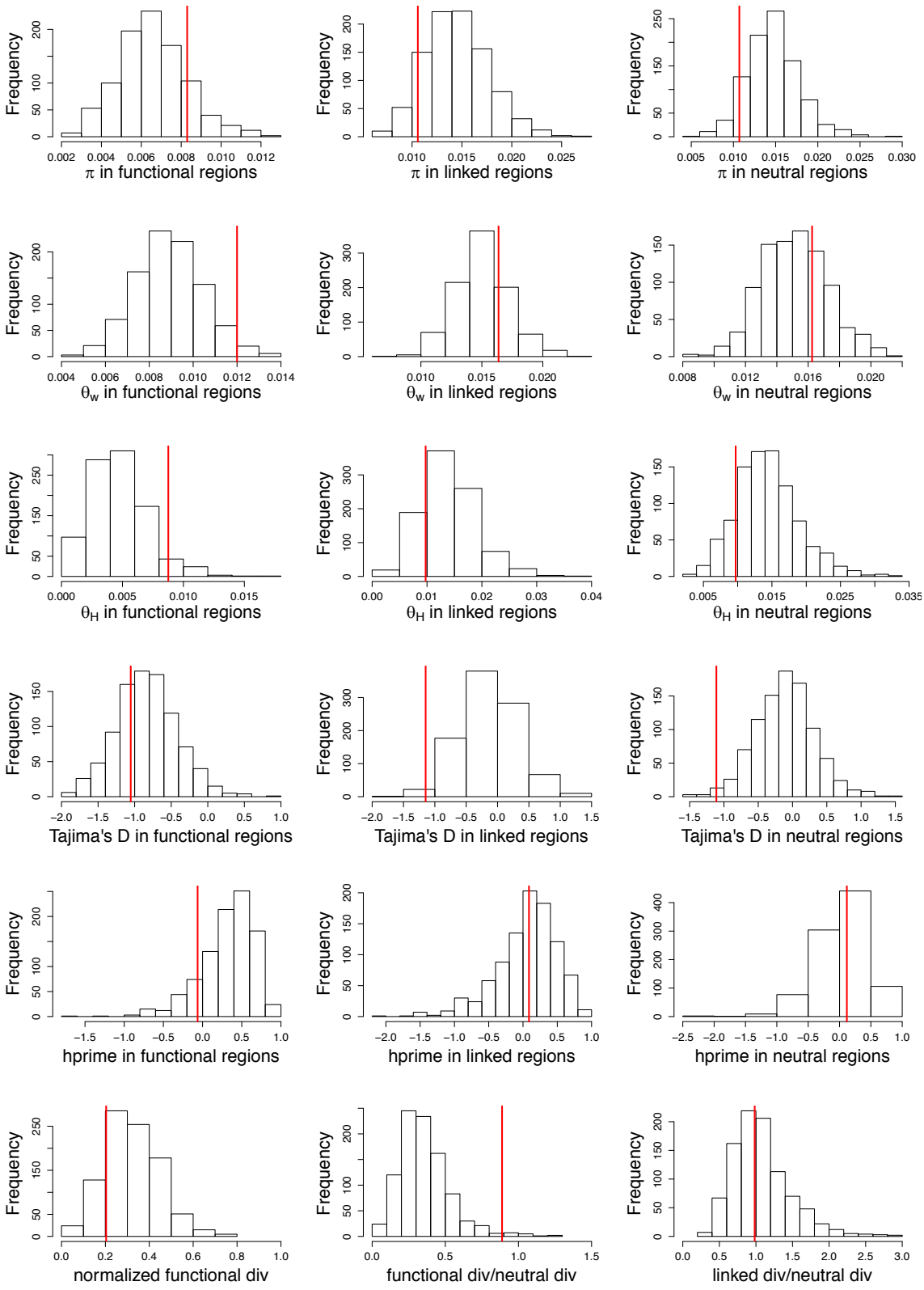

**Supp Figure 10**: Correlation of recombination rates at 5 prime flanking intergenic versus that in 3 prime flanking intergenic of all 94 exons chosen for analysis in *D. melanogaster*.

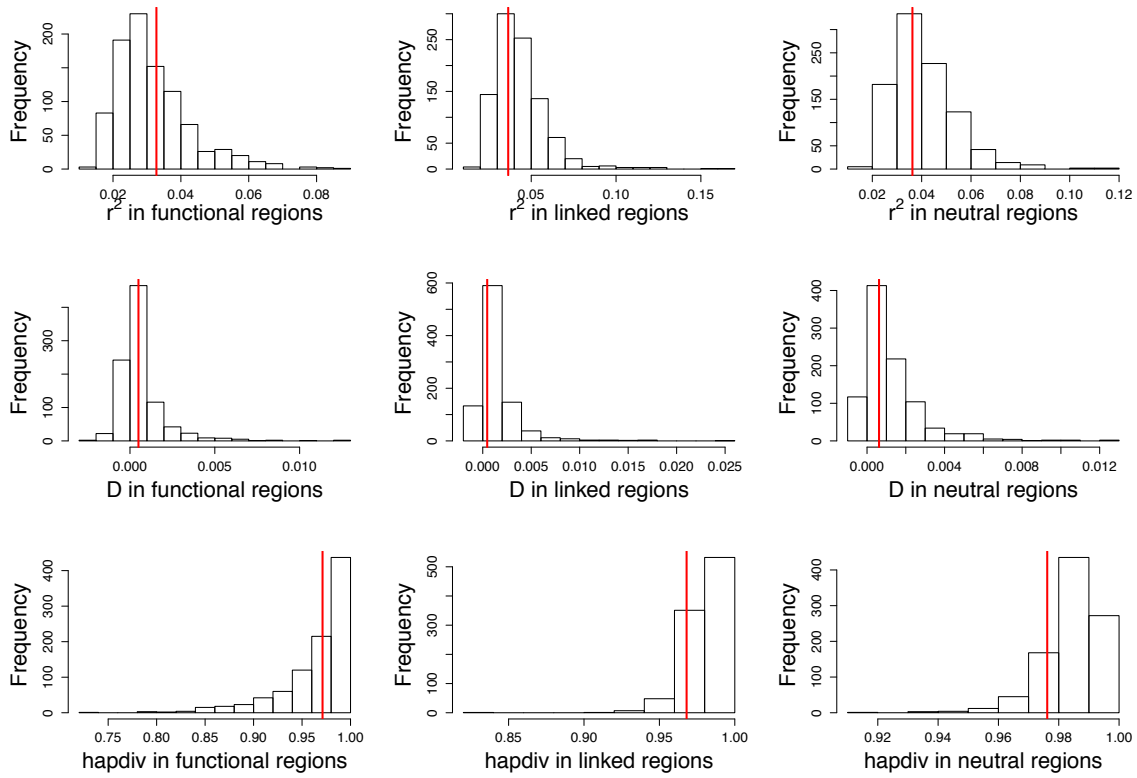



**Supp Figure 11**: Distribution of the rate of recombination in cM/Mb for all 94 exons selected for analysis in *D. melanogaster*.

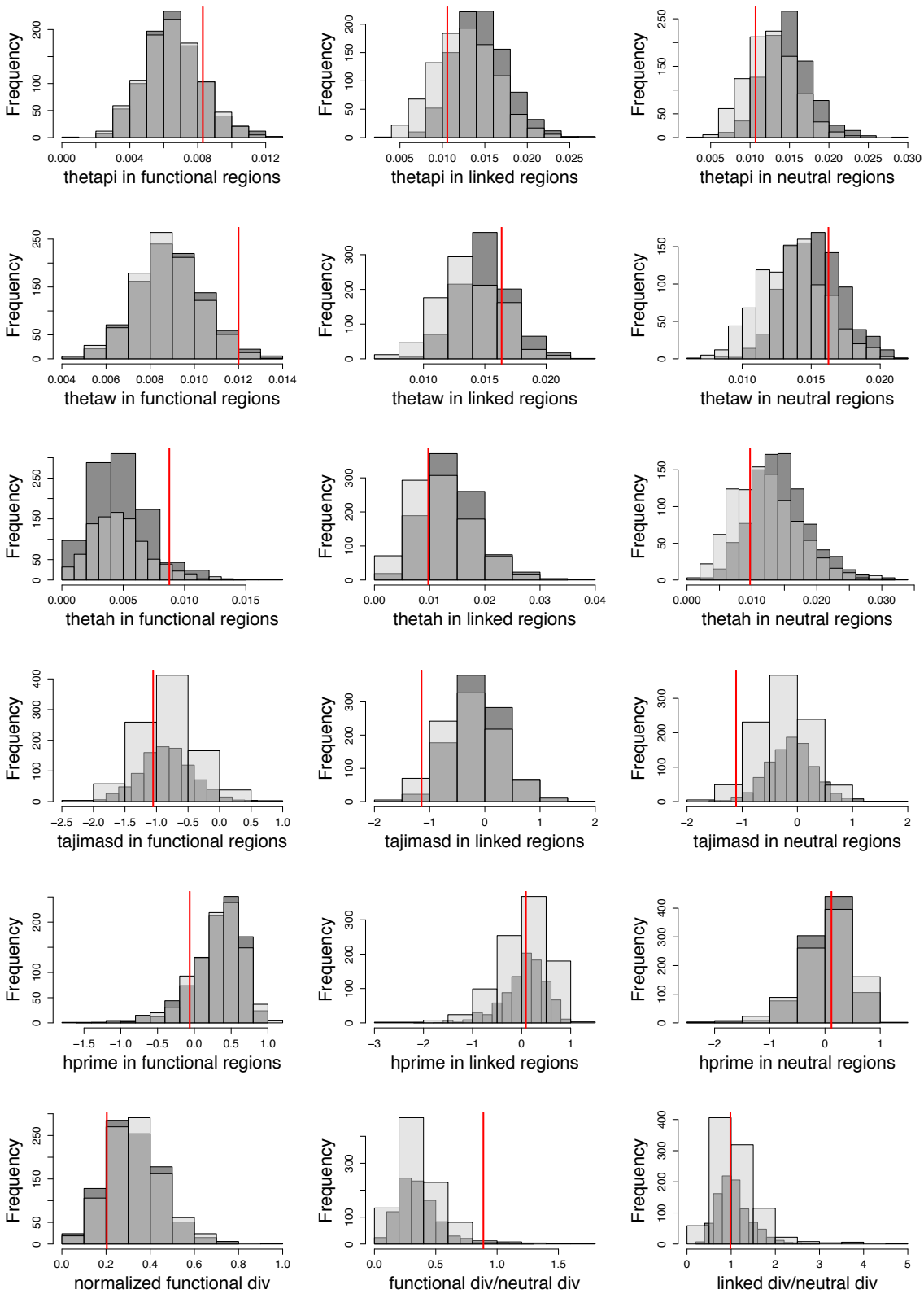


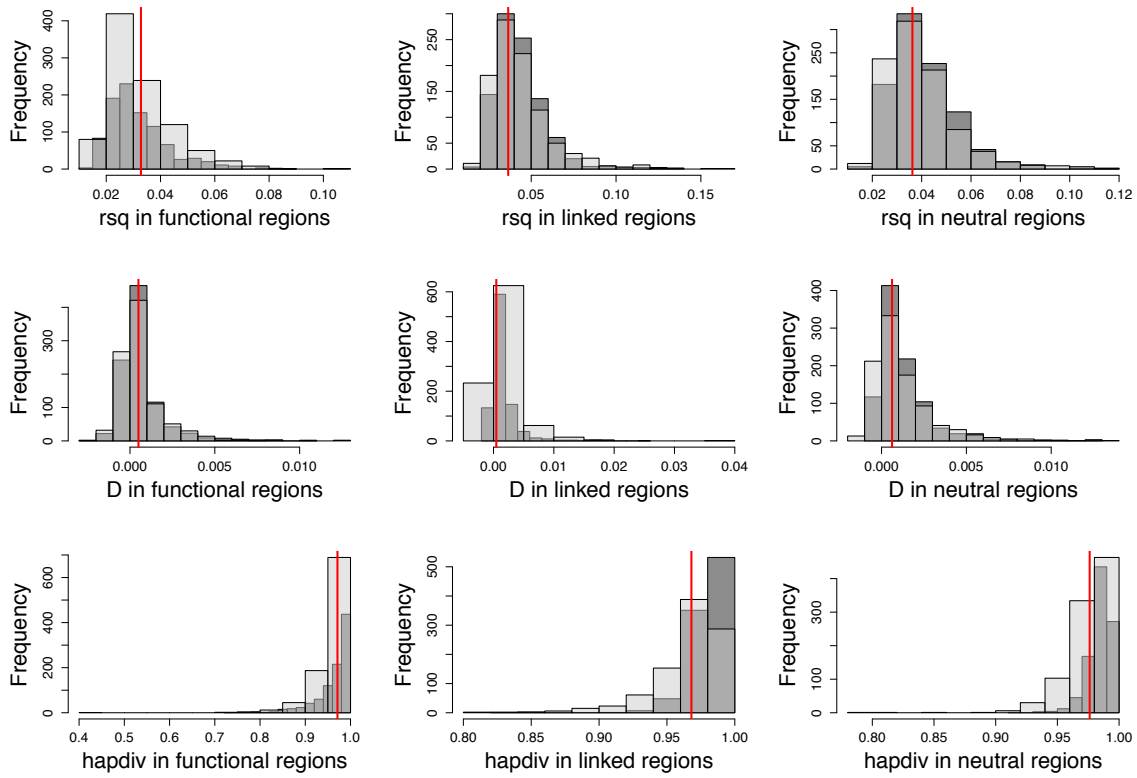

**Supp Figure 12**: Distribution of divergence per site of single-exon genes that have flanking intergenic regions larger than 4 kb (in red), and for all genes (in black), from *D. melanogaster*.


Frequency
π in functional regions
π in linked regions
π in neutral regions
θw in functional regions
θw in linked regions
θw in neutral regions
θH in functional regions
θH in linked regions
θH in neutral regions
Tajima's D in functional regions
Tajima's D in linked regions
Tajima's D in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

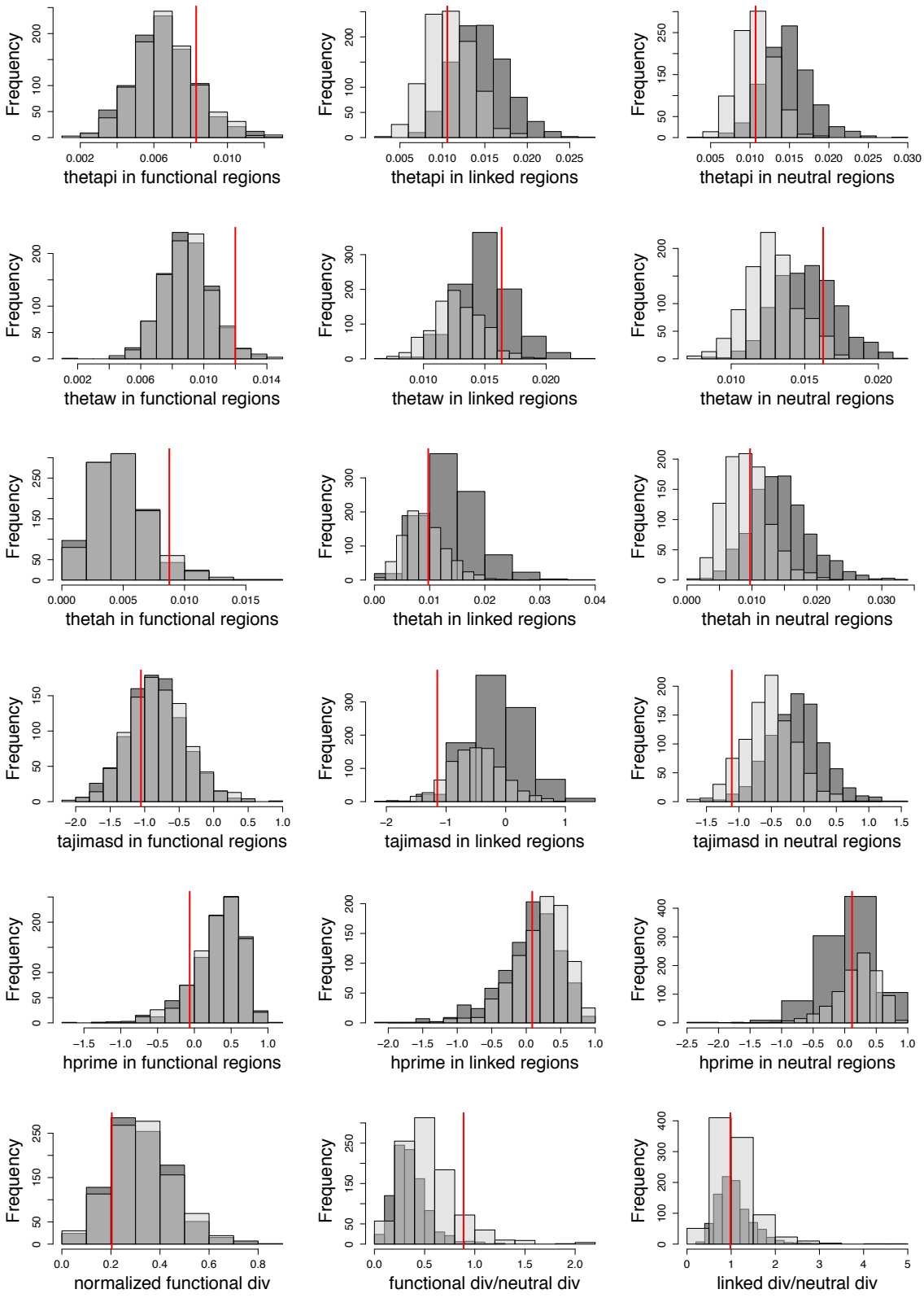


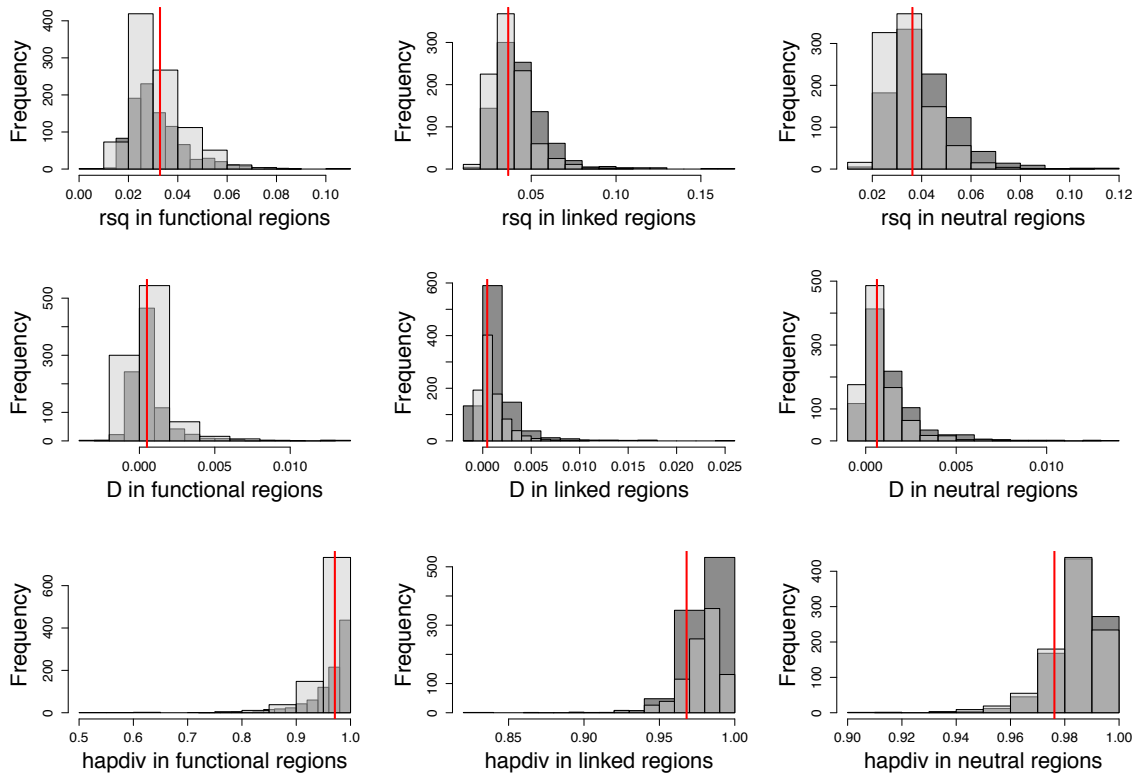

**Supp Figure 13**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Red lines indicate the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

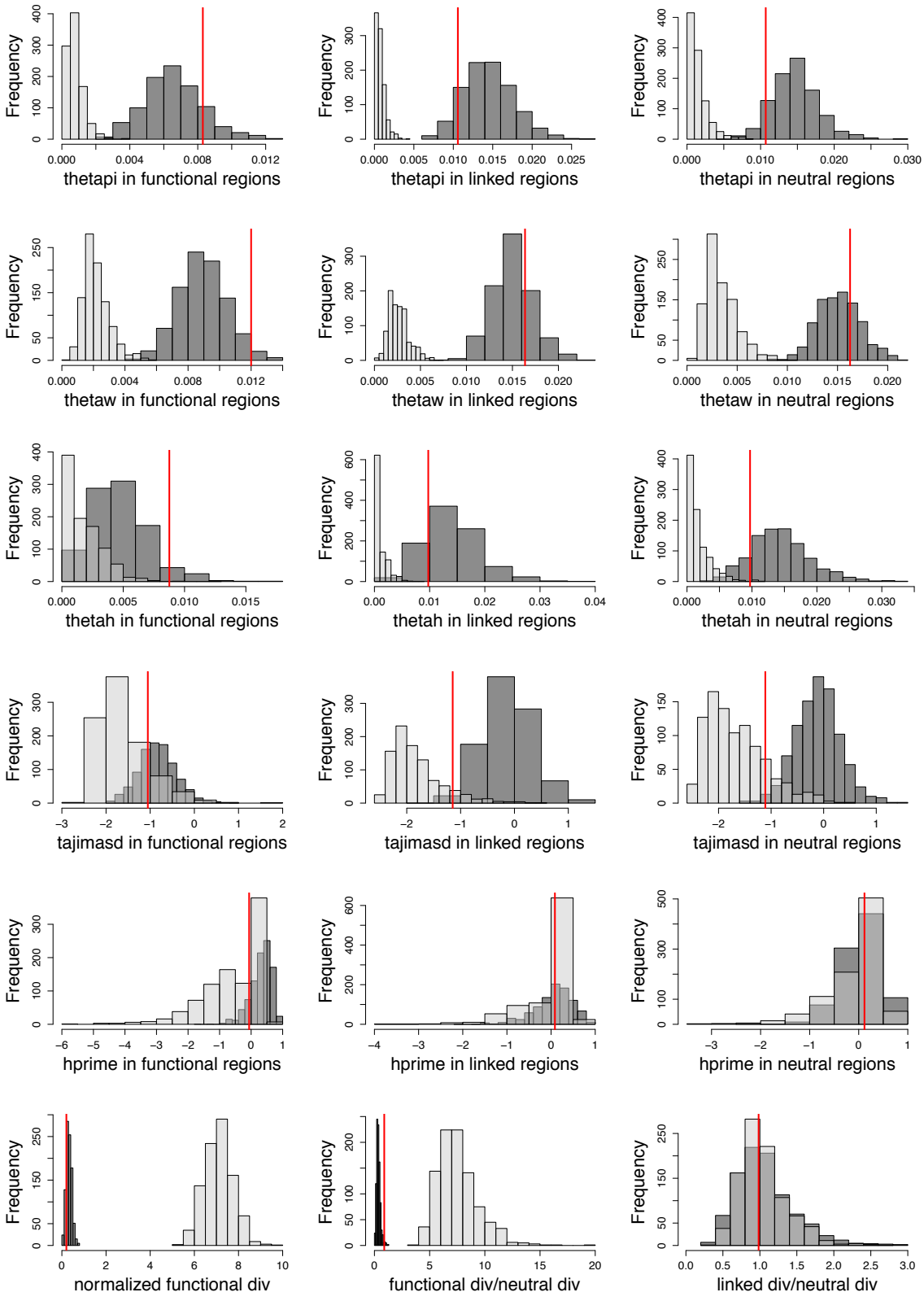


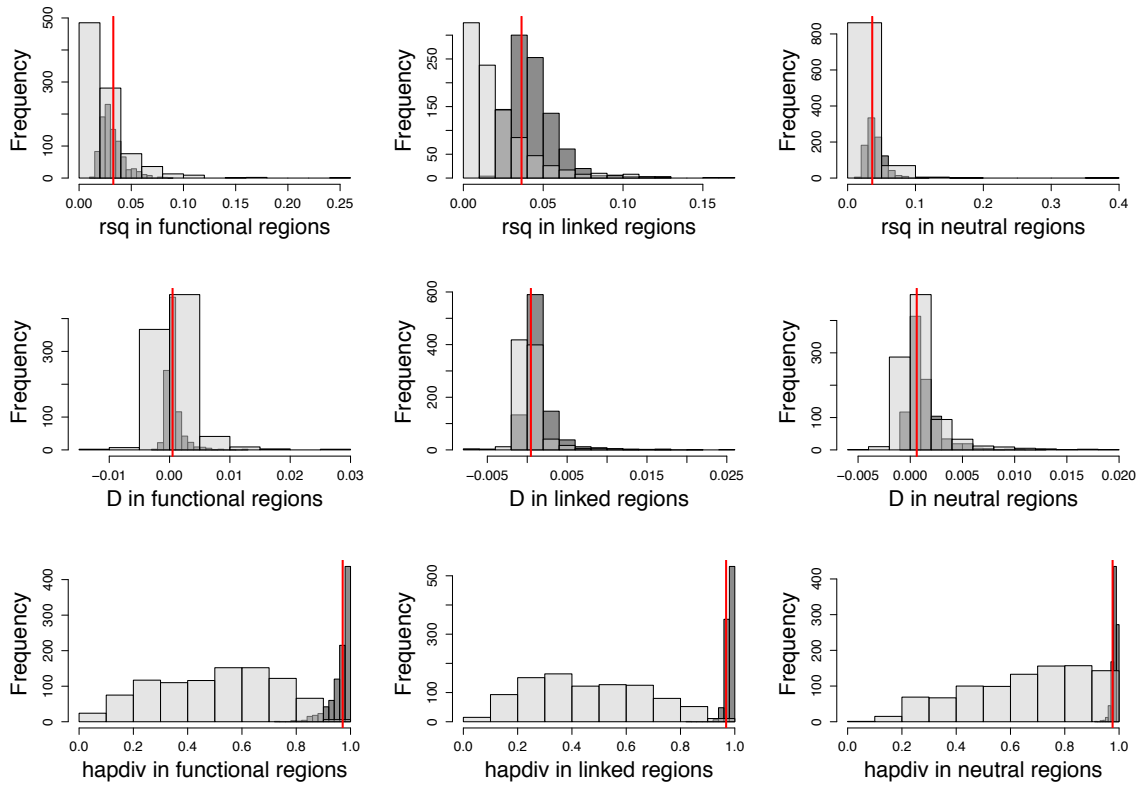

**Supp Figure 14**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). In this case, conserved elements that represent 40% of non-coding regions were simulated to experience purifying selection with the class of mutations that result in strongest BGS effects ($-100<2N_e s<-10$) and these sites were masked while calculating statistics. Red line indicates the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no selection on non-coding regions and light grey bars represent simulations with selection on non-coding regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

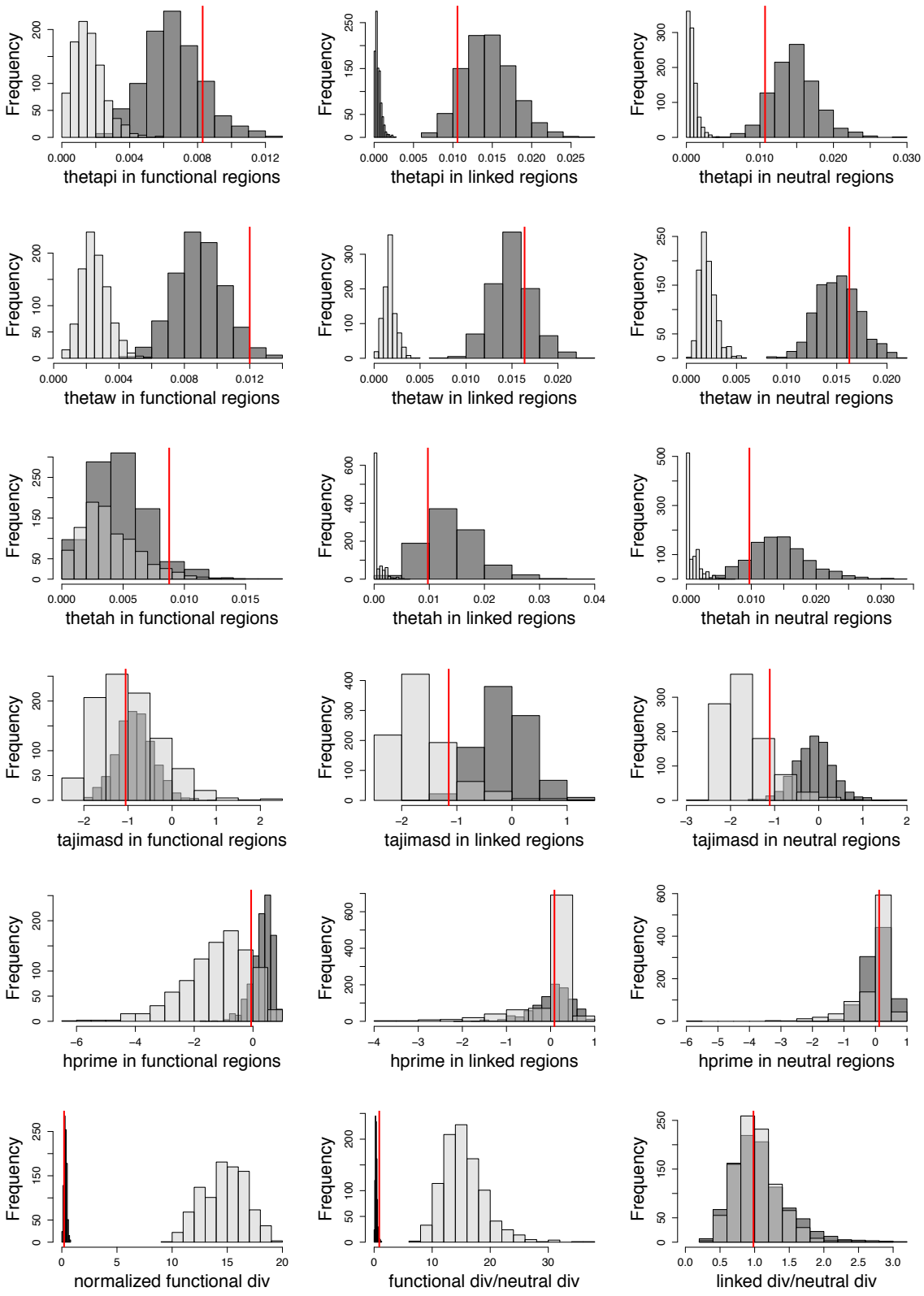


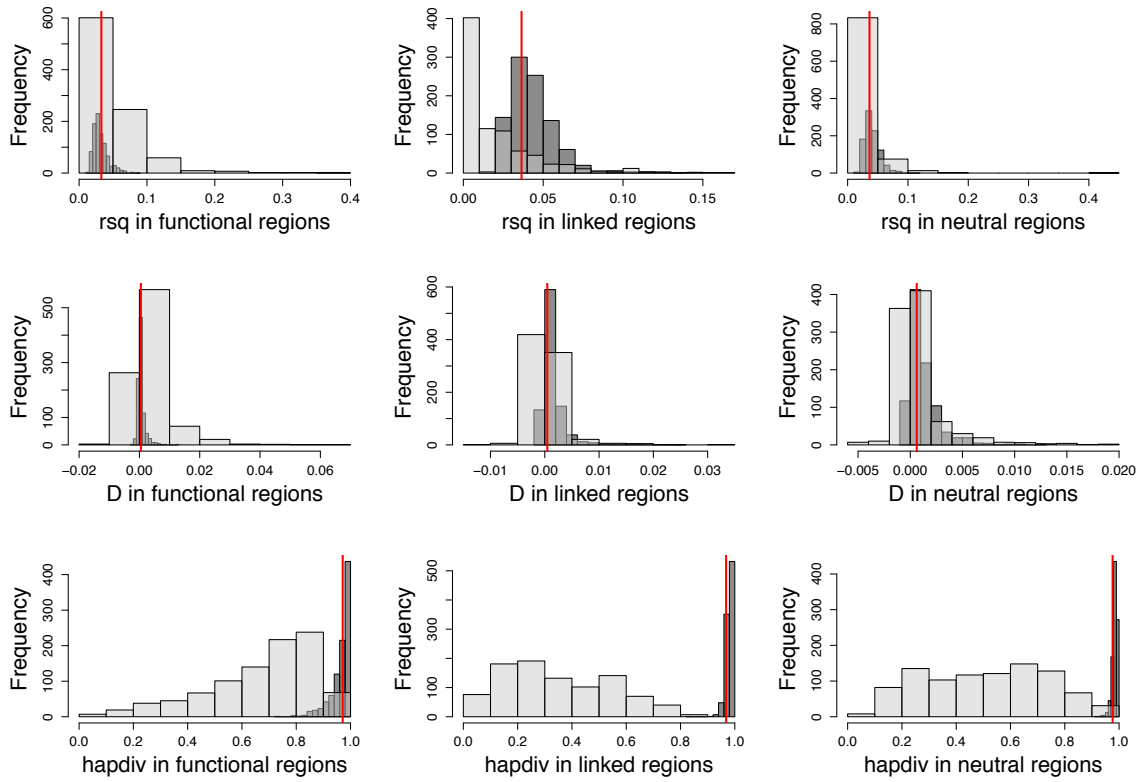

**Supp Figure 15**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). In this case, conserved elements that represent 40% of non-coding regions were simulated to experience weak purifying selection ($-10<2N_es<-1$) and these sites were included while calculating statistics. Red line indicates the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no selection on non-coding regions and light grey bars represent simulations with selection on non-coding regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

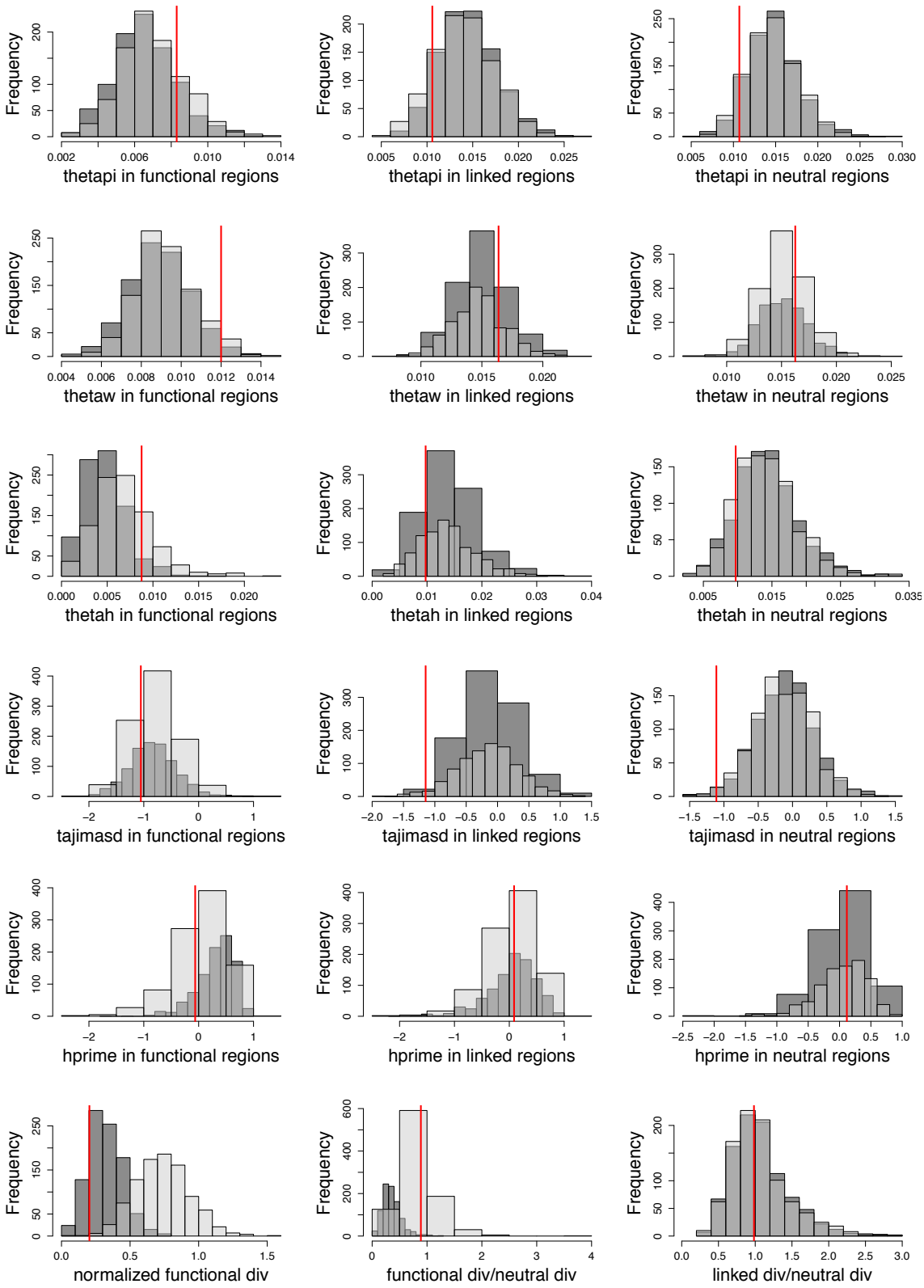


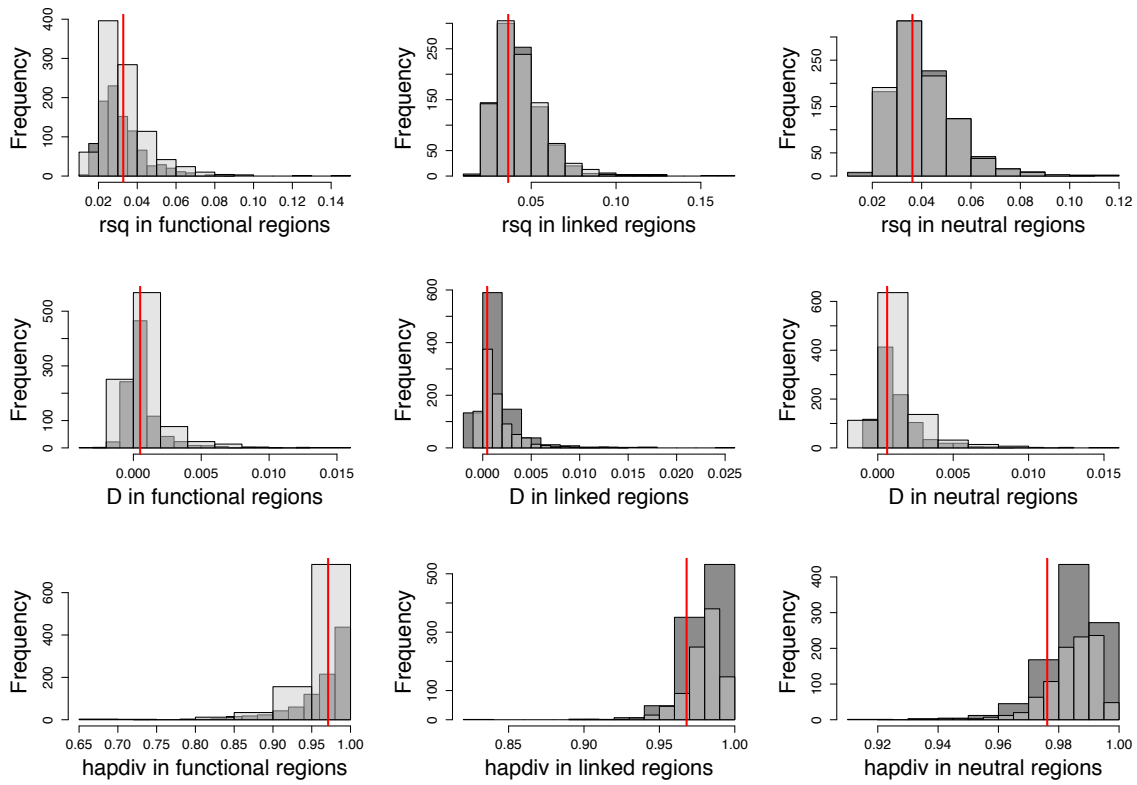

**Supp Figure 16**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25, f_1=0.49, f_2=0.04, f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Functional regions were simulated to experience rare (1%) and strong positive selection ($2N_{anc}s = 1000$). Red lines indicate the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

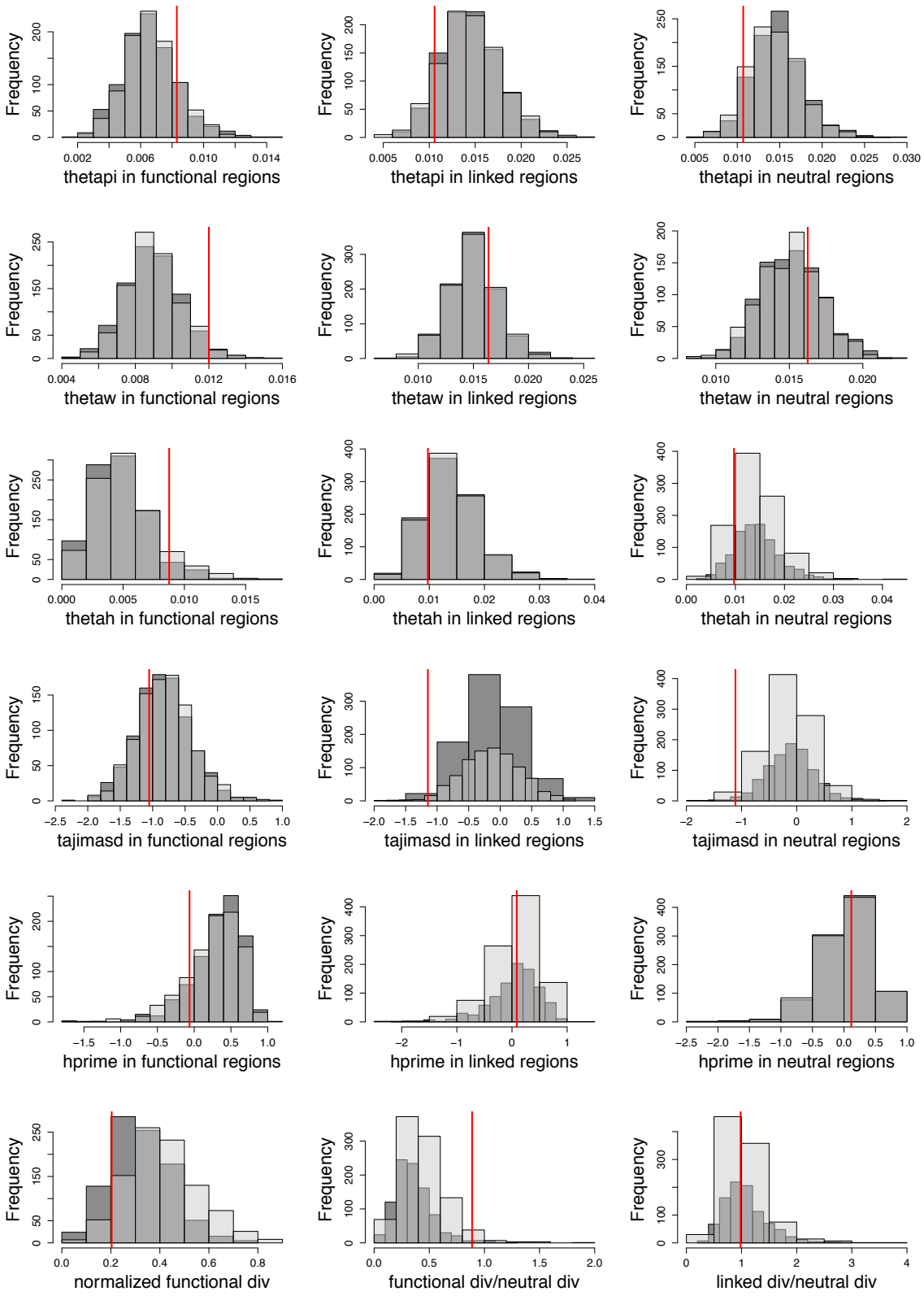


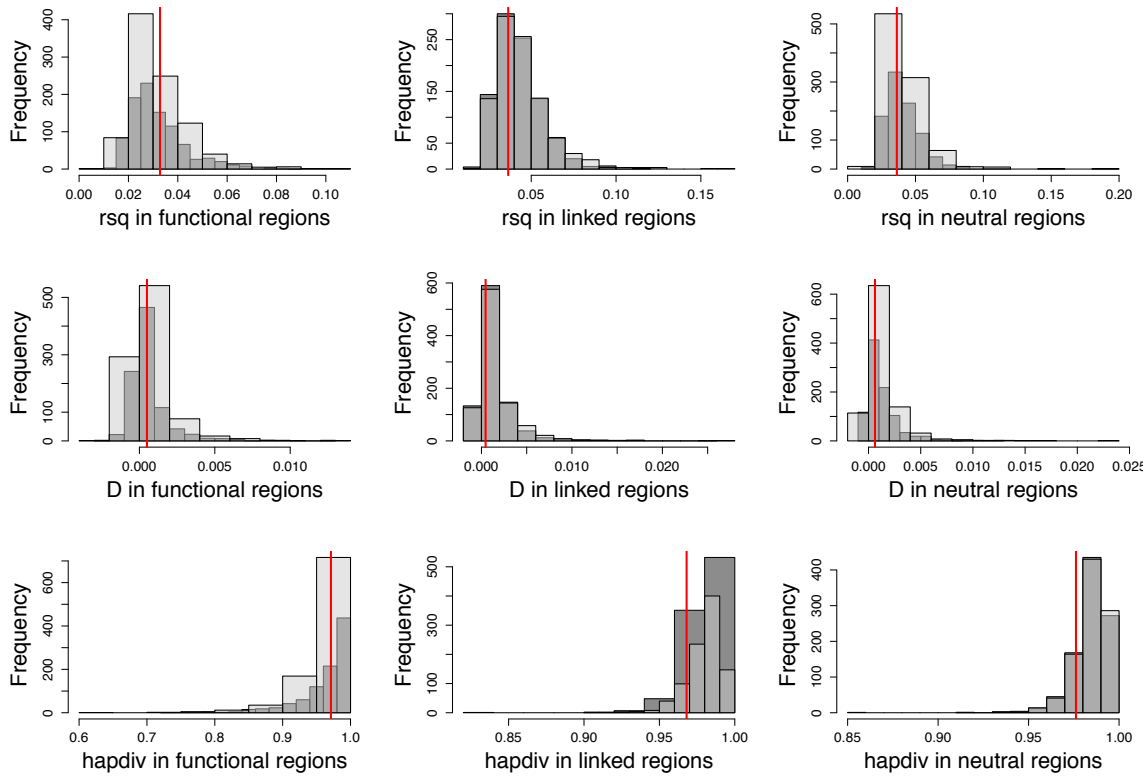

**Supp Figure 17**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Functional regions were simulated to experience common (5%) and strong positive selection ($2N_{anc}s = 1000$). Red line indicates the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

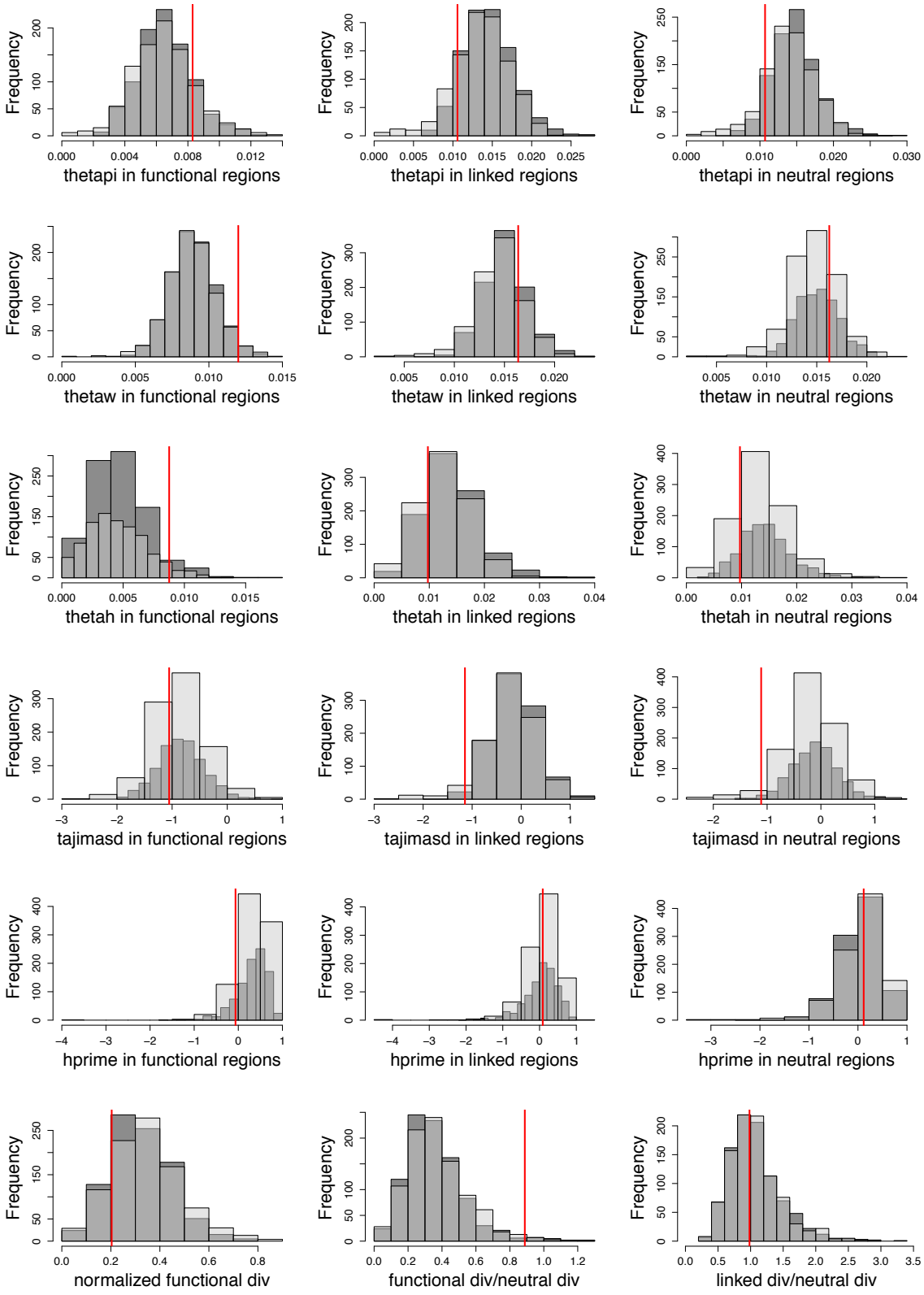


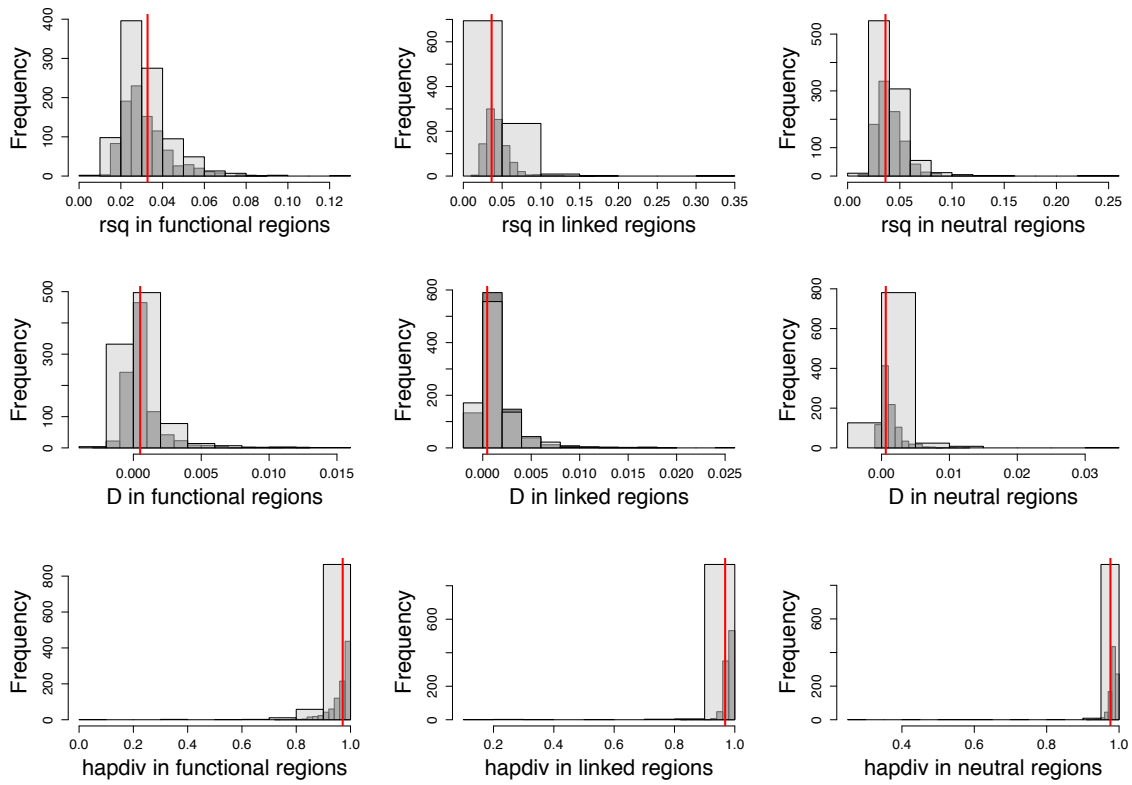

**Supp Figure 18**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Functional regions were simulated to experience common (5%) and weak positive selection ($2N_{anc}s = 10$). Red lines indicate the value observed in 76 individuals of *Drosophila melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div

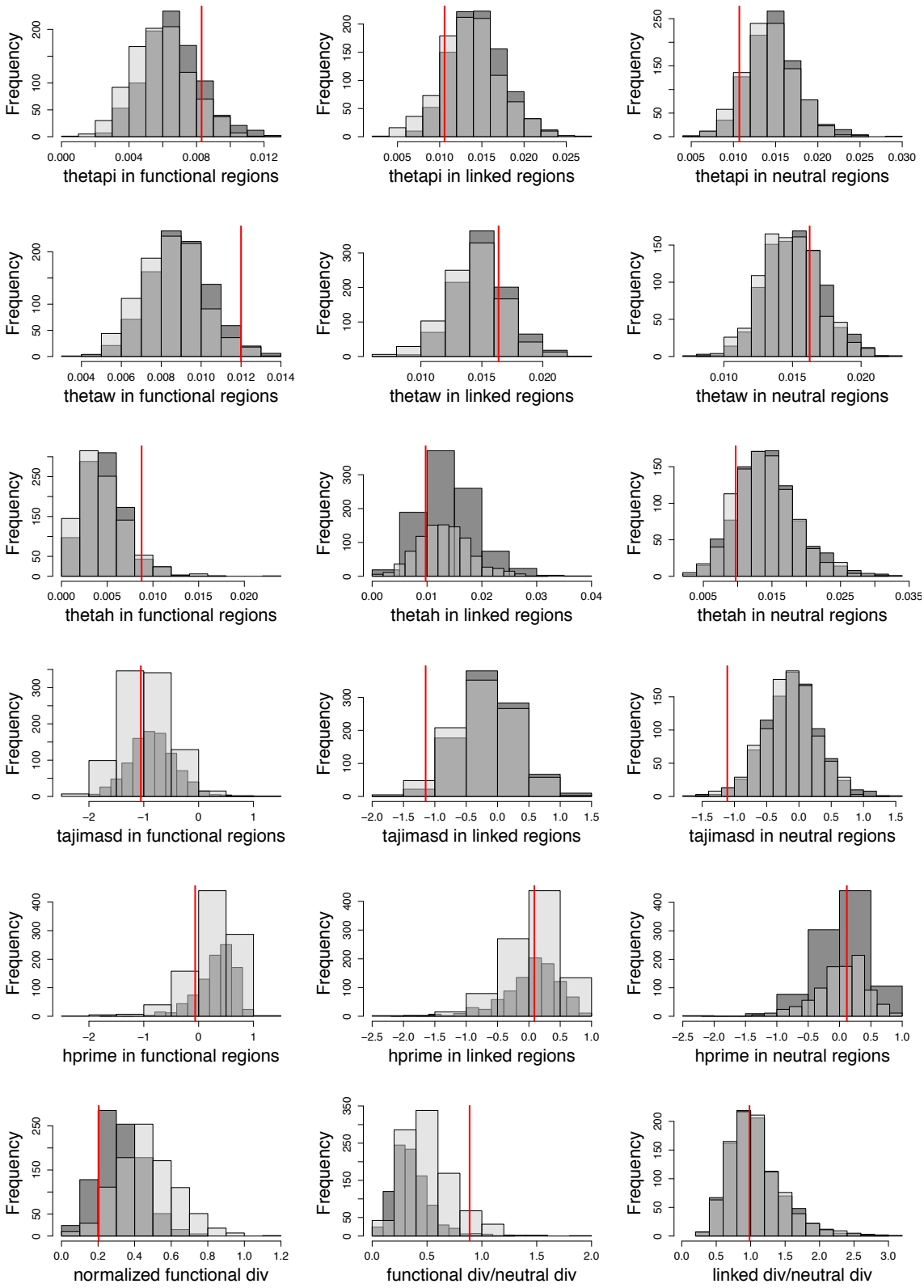


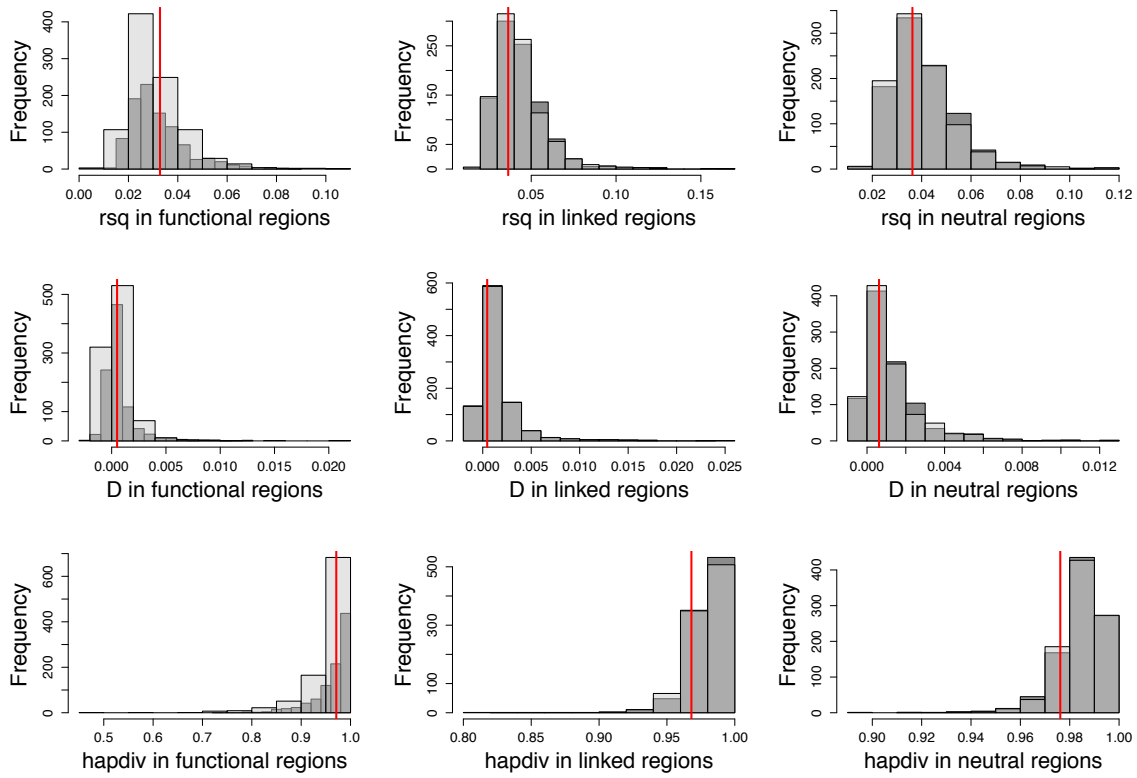

**Supp Figure 19**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25, f_1=0.49, f_2=0.04, f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Functional regions were simulated to experience rare (1%) and weak positive selection ($2N_{anc}s = 10$). Red lines indicate the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div




**Supp Figure 20**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25, f_1=0.49, f_2=0.04, f_3=0.22, N_{anc}=$ 1,225,393, $N_{cur}$ = 1,357,760). Functional regions were simulated to experience rare ($1.28 \times 10^{-4}$ %) and strong positive selection ($2N_{anc}s = 10000$) as in Lange and Pool (2018). Red lines indicate the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.


Frequency
thetapi in functional regions
thetapi in linked regions
thetapi in neutral regions
thetaw in functional regions
thetaw in linked regions
thetaw in neutral regions
thetah in functional regions
thetah in linked regions
thetah in neutral regions
tajimasd in functional regions
tajimasd in linked regions
tajimasd in neutral regions
hprime in functional regions
hprime in linked regions
hprime in neutral regions
normalized functional div
functional div/neutral div
linked div/neutral div




**Supp Figure 21**: Distribution of summary statistics calculated from 94 exons simulated with 100 replicates each using our inferred model (*i.e.*, $f_0 = 0.25$, $f_1=0.49$, $f_2=0.04$, $f_3=0.22$, $N_{anc}=$ 1,225,393, $N_{cur} = 1{,}357{,}760$). Functional regions were simulated to experience rare (0.2%) and weak positive selection ($2N_{anc}s = 60$) as in Lange and Pool (2018). Red lines indicate the value observed in 76 individuals of *D. melanogaster* from Zambia, after excluding sites with phastCons score >= 0.8. Dark grey bars represent no positive selection and light grey bars represent simulations with positive selection in functional regions.