Supplemental Material

**Supplemental Text**

**Text S1: Program Validation- comparison with DFE-alpha**

To validate our program’s methodology, we used a separate program, DFE-alpha to check for consistency between an observed SFS data set and the expected SFS generated from our model (Eyre-Walker and Keightley 2009). Where DFE-alpha explicitly models demography, our program applies a nonparametric approach, skewing the expected SFS with frequency-dependent correction factors. Likewise our program takes a nonparametric approach to modeling the DFE, using a categorical approach as opposed to DFE-alpha’s gamma distribution. Thus, the method detailed in this paper – hereafter referred by the acronym NPC for brevity – is a nonparametric categorical DFE inference engine as opposed to the parametric DFE-alpha.

We ran DFE-alpha and NPC over the three spectra: Zambia-preferred, Zambia-full, and DGRP-preferred. For each comparison, we chose a SFS pair (i.e. 4D SFS and corresponding short-intron SFS) whose inferred DFEs were the closest to the values reported in Table 1. For DFE-alpha, we estimated a 3-epoch demographic model on the short-intron SFS and estimated the gamma DFE on the 4D SFS. Similarly, for NPC, the frequency-dependent correction factors were estimated on the short-intron SFS while the categorical distribution parameters were estimated on the 4D SFS. When either DFE-alpha or NPC infer an evolutionary model via maximum-likelihood on site frequency spectra, they also generate what the expected spectra should look like under that model. To validate that our NPC method effectively models the real data, we then re-ran DFE-alpha on the expected SFS generated from the NPC.

As shown in Table ST1 below, DFE-alpha infers very similar selection and demographic parameters on the theoretical SFS generated by the NPC as it does on the real data. One discrepancy is in Zambia-all where DFE-alpha infers a much longer generation time spent in epoch 2 in the theoretical SFS as compared to the observed SFS. But this has a minor impact on the overall inferred effective population size and an even more negligible influence on the DFE inference. Thus, overall, the theoretical predictions from the NPC model are consistent with and effectively capture the demographic and selection signal in the real data.

|  |  |  |
| --- | --- | --- |
| Zambia-preferred  | Observed SFS | NPC Expected SFS |
| (gamma DFE) shape | 0.1095 | 0.1136 |
| (gamma DFE) scale | 74.33 | 67.24 |
| (3-epoch) N1 | 100 | 100 |
| (3-epoch) N2 | 80 | 80 |
| (3-epoch) T2 | 2349.29 | 2215.038 |
| (3-epoch) N3 | 170 | 170 |
| (3-epoch) T3 | 20.87 | 19.79 |
| (3-epoch) Nw | 90.67 | 90.17 |
|  |
| Zambia-all  | Observed SFS | NPC Expected SFS |
| (gamma DFE) shape | 0.050 | 0.050 |
| (gamma DFE) scale | 81.92 | 81.51 |
| (3-epoch) N1 | 100 | 100 |
| (3-epoch) N2 | 60 | 60 |
| (3-epoch) T2 | 434.89 | 1831.11 |
| (3-epoch) N3 | 150 | 150 |
| (3-epoch) T3 | 13.31 | 11.22 |
| (3-epoch) Nw | 70.58 | 67.83 |
|  |
| DGRP-preferred  | Observed SFS | NPC Expected SFS  |
| (gamma DFE) shape | 0.050 | 0.050 |
| (gamma DFE) scale | 3543.08 | 3584.23 |
| (3-epoch) N1 | 100 | 100 |
| (3-epoch) N2 | 60 | 60 |
| (3-epoch) T2 | 25.81 | 21.91 |
| (3-epoch) N3 | 100 | 100 |
| (3-epoch) T3 | 5 | 5 |
| (3-epoch) Nw | 95.1 | 95.82 |

Table ST1. Program validation: Comparison of the inferred DFE-alpha parameters between the true data (observed Zambia-preferred, Zambia-all, DGRP-preferred SFS) and the expected SFS generated by NPC for each observed SFS. Shape and scale are the parameters of the gamma distribution inferred for the DFE. N1-3 are the scaled population sizes for the 3-epoch model while T2,3 are the number of generations at each size. Nw is the overall weighted effective population size of the 3-epoch model.

 **Text S2: Shape-only analysis**

We performed a maximum likelihood (ML) estimation of the strength and amount of selection on synonymous sites using only polymorphic sites. We call this the “shape-only” ML model because it relies solely on deviations in the shape of the SFS and does not use polymorphism-level information. While this is less powerful than using the density of polymorphism - particularly with respect to inferring strong purifying selection - it also sidesteps issues of mutation rate variation between classes. Comparing the “shape-only” analysis with the maximum-likelihood equation presented in the methods section “Maximum-likelihood estimation of selection parameters from SFS”: for this “shape-only” analysis, p(0|...) is dropped in equation 2 and the remaining product is divided by the probability of seeing a polymorphism rather than by all sites.

In this analysis we tested both the full datasets and the subset of synonymous sites found in preferred codons (the most frequent codon per amino acid). We hypothesized that preferred codons were under stronger purifying selection than unpreferred codons. For both the Zambia and the DGRP full datasets, we found no evidence for selection. However, for the subset of 4D sites in preferred codons we did find evidence for selection, estimating that 29% (95% bootstrap CI: 26-34) of sites were under purifying selection at Nes = −3 (95% bootstrap CI: 1-6) (Zambia population). For the DGRP population, the selection estimates were similar (25% at Nes = −3); however, the confidence intervals were much larger (18%−89% at Nes = −Inf −0) and the selection model was not significantly better than neutrality. One factor potentially contributing to the large confidence intervals for DGRP is reduced power due to fewer polymorphic sites (less than one-half the number of polymorphisms as Zambia). The detection of purifying selection in the set of 4D sites in preferred codons suggested that 1) synonymous sites in preferred codons were under greater purifying selection than the genomic average and 2) we were underpowered to detect this selection in the full dataset due to a low total proportion of sites under selection.

 The expected polymorphism ratio for Zambia preferred codons, based on the shape-only ML estimate of 29% sites under selection at Nes = −3, is 0.07 (95% CI estimate: 0.03-0.11). This expected polymorphism ratio of 0.07 was significantly lower than the observed polymorphism ratio of 0.19, suggesting that the shape-only ML model does not fully explain the data. As this model does not use polymorphism density and the population sampling is still not deep enough, multiple non-neutral selection categories cannot be supported by the data and strong selection is obscured in these inferences. Thus, in order to make a full accounting of the selection strength affecting 4D sites, we only used the full ML model that looked both at the shape of the SFS and the density of the polymorphism in the main results.

**Supplemental Figures**



Figure S1: Short intron polymorphism as a function of distance from the 4D site.



Figure S2: The effect of limiting the analysis to a single 4D site per SI matched control, compared with the original matching that allows multiple 4D sites per SI match (average of 3 per SI site). Each point/group of bars represents one dataset (DGRP, DGRP preferred=D.pre, DGRP unpreferred=D.unp, Zambia, Zambia preferred=Z.pre, Zambia unpreferred=Z.unp). A) Polymorphism ratio for each dataset comparing the original 4D matching with single 4D matching per SI site (black points). The resulting dataset is further sampled to recover the same trinucleotide proportions as the original dataset (grey points). Solid lines: linear regression. Dashed black line: y=x. B) The number of sites in each dataset as a result of matching method and additional trinucleotide proportion sampling.

Figure S3: Maximum-likelihood estimates using a two category model (neutral, deleterious) for inference when the data was simulated with a three category model (neutral, weakly deleterious, strongly deleterious). The weakly deleterious class is s1, the strongly deleterious class is s2. Each point is the maximum-likelihood estimate of the strength of selection and proportion of sites for the single inferred deleterious category. Inferring selection parameters on a single deleterious selection category suffers from the tension that the stronger it infers the strength of selection to be, the fewer sites it can assign to the deleterious category (e.g. circles vs triangles). Meanwhile, the inferred proportion of sites under selection in this category is mostly governed by the strength of the simulated selection in s1 (e.g. blue vs red). Estimates of the inferred proportion of sites under selection for blue simulations are approximately the total proportion of simulated sites in s1+s2, while for red simulations they are approximately the proportion of simulated sites in s2. These patterns can be seen in the real data as well. In Table 1, compare the results of n+s and n+s+s models for DGRP preferred and one sees a similar relationship between the two category and three category estimations as for the blue triangle in the figure above. Some insight of the underlying DFEs can be gleaned from the results. For instance, the results show that there are overall more sites under selection in the scenarios in plot B) versus plot A). Within each plot, the results show that blue scenarios have overall stronger selection than red scenarios and that triangles have stronger selection than circles - but crucially not that the total proportion of sites under selection is actually the same. Conversely, it is easy to see, given the above, that different three category models could be constructed to have similar inferred two category parameters (not shown). Having as many inferred selection categories as can be statistically supported helps properly characterize the features of the true, underlying distribution of fitness effects.





Figure S4: Selection strength power and bias analyses. A) Maximum strength of selection detectable (significantly distinguishable from lethality) and B) minimum strength of selection detectable (significantly distinguishable from neutral) as a function of the number of sites analyzed (in 1000’s of sites), for a range of proportions of sites under selection, for a population in mutation/selection equilibrium. C,D,F,G) Effect of demographic model on ML estimates of the proportion of selected sites and strength of selection. Simulated populations have a theta of 0.035. Across the 2M simulated sites, there are two selection categories (neutral, deleterious) where the “Expected Proportion” and “Expected -Nes” define the proportion of sites and strength of selection in the deleterious category. C,F) Mutation/selection equilibrium model. Many of the points are overlapping, as the ML estimates of proportions of sites for the range of Nes tested are nearly identical. Likewise are the ML estimates of Nes for the range of proportions of sites tested. D,E,G) Bottleneck/growth model: the simulated population undergoes a bottleneck then experiences rapid, exponential growth. The fitness of a mutation is held constant over the simulation (constant s) and thus the effective selection coefficient, Nes, rises and falls with the changing population size. D,E) the “Final” - Nes of a simulation denotes the ML Observed - Nes of a simulated set of sites with a given “Initial” selection strength after undergoing a bottleneck/growth demographic event - e.g. a mutation with an initial -Nes of 75, after the bottleneck/growth is inferred to have had an overall effective selection coefficient of 65 while mutations with starting out with a -Nes of 1 effectively evolve with a -Nes of 1.4 over the demographic changes. Thus in the x-axis of plot G), the “Expected” Nes is this “Final” Nes - i.e. what we expect the inferred Nes to be under the demographic scenario if there had been no neutral category in the test sites of the simulation. In panels D) and G), several of these datasets had a “lethal” selection class best-fit ML model (e.g. Nes =42 and Nes =65) and these are not represented in the plots.



Figure S5: Correlation between the polymorphism ratio and the proportion of sites under selection. A) Polymorphism ratio for simulated datasets of a range of selection strengths and proportion of sites under selection. B) Observed polymorphism ratio and ML estimated sites under selection in the Zambia or DGRP datasets. Each point represents a different subset of the data (different functional classes tested). Solid line is a linear regression (R2 = 0.65).



Figure S6: Polymorphism ratio by frequency of preferred codons (FOP). Error bars represent two standard error.



Figure S7: Polymorphism ratio by recombination rate. Error bars represent two standard error.



Figure S8: Selection by functional class (DGRP). A) Extent of strong purifying selection as measured by the polymorphism ratio for each class of site (grey) and the dataset excluding the focal class sites (white). The number of sites in a focal class is listed below the corresponding bar. The red dashed line is the polymorphism ratio for the full dataset. Error bars represent two standard error. B): Relative proportion of synonymous sites under strong purifying selection due to splicing, CUB, or transcription factor (TF) binding.

**Supplemental Tables**

|  |  |
| --- | --- |
| Bin number | Frequency bin |
| 1 | 1/N |
| 2 | 2/N : 3/N |
| 3 | 4/N : 7/N |
| 4 | 8/N : 15/N |
| 5 | 16/N : (N/2 – 15)/2 |
| 6 | (N/2 – 15)/2 + 1 : N/2  |

Table S1: Six-bin free-alpha model.

|  |  |  |
| --- | --- | --- |
|  | **1 selection strength model** | **2 selection strength model** |
| **S1** | -0.25,-1.25,-2.5,-7.5,-12.5,-25,-100,-162.5 | -0.25,-1.25,-2.5,-7.5 |
| **S2** | NA | -12.5,-25,-100,-162.5 |
| **Proportion\_S1** | 0.1,0.5,0.8 | 0.2,0.7 |
| **Proportion\_S2** | NA | 0.1,0.2 |
| **Proportion\_lethal** | 0.01,0.1 | 0.01,0.1 |
| **Theta** | 0.01,0.05 | 0.01,0.05 |

Table S2: Seed values for ML estimation.