# polyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species

# Supplementary Information

# SFS data simulation setups

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table S1 Properties of DFEs assumed when simulating SFS datasets | | | | | | | |
|  | **DFE type** | **DFE abbreviation** | ***Sd*** | ***b*** | ***pb*** | ***Sb*** | ***α*** |
| **deleterious DFEs** | high *Sd* | DelHSD | -10 000 | 0.4 | 0 | - | 0 |
| medium *Sd* | DelMSD | -400 | 0.4 | 0 | - | 0 |
| **full**  **DFEs** | medium *α*, medium *Sd* | MAMSD | -400 | 0.4 | 0.02 | 4 | 0.53 |
| high *α*, high *Sb* | HAHSB | -400 | 0.4 | 0.02 | 16 | 0.81 |
|  | | | | | | | |

To investigate the type I error rate and the power of our tests to detect DFE heterogeneity, we relied on simulated SFS data obtained from exome-like regions using SFS\_CODE (Hernandez 2008). Among the 18 DFE scenarios used to validate the statistical performance of the polyDFE method (Tataru et al. 2017) to estimate DFE and use LRTs for the presence of beneficial mutations, we explored here data generated under four possible types of DFEs (summarized in Table S1). Each DFE consists of a mixture between a gamma and exponential distributions: a new mutation is deleterious with probability1 - *pb,* and has a scaled selection coefficient drawn from a reflected gamma distribution with mean *Sd < 0* and shape *b;* a new mutation is beneficial with probability *p*b*,* and has a scaled selection coefficient drawn from an exponential distribution with mean *S*b *> 0.* Two DFEs comprise only deleterious mutations but have different intensity of purifying selection as measured by*Sd* (DelHSD, DelMSD, Table S1). Two other DFEs contain a fraction *p*b of beneficial mutations with different mean scaled positive selective coefficient *S*b. The different combinations of the four parameters that specify the simulated DFEs generate different expected amounts of adaptive evolution as measured by the parameter  (from 0 to 0.81, Table S1). Note that it is not a parameter that we vary per se. It is instead fully determined by the parameters of each DFE (see Tataru et al. 2017 for details).

For each DFE type, we simulated 100 replicate datasets comprising SFS data containing 10.8 Mb of exonic sites in 10 diploid individuals (20 haplotypes). Each dataset was simulated using a scaled mutation rate of 0.001. The simulated data considered here do not contain misidentification of the ancestral state or departure from a stable demography (i.e. population size was assumed to be constant).

# Details of inference on simulated data

The inference was performed on pairs of simulated SFS dataset under models M1-M4 presented in Table 1 in the main text, where these SFS were either simulated using the same underlying DFE (for investigating type I error) or using two different DFEs (for investigating the power to detect differences in DFEs, see below). When fitting the data using polyDFEv2.0, the ancestral misidentification error *ε* was fixed to 0 and not estimated, while the demography nuisance parameters (*ri*s) were estimated, as it has been shown that fitting these accounts for SFS distortions induced by demographic history (relative to the SFS expectation in a stable population) but also the presence of linkage in the data (Tataru et al. 2017). The scaled mutation rate parameter was assumed constant across the data. To ensure that the likelihood function was reliably maximized, we performed 10 runs of the BFGS optimization algorithm, with randomly starting values for each dataset analyzed. We also provided the simulated values of the parameters to polyDFE as an additional starting point, to ensure that a failure in finding the simulated values as the true optimum was not caused by a failure of the BFGS algorithm to reach the global optimum.

# Evaluating the type I error and power of likelihood ratio tests for DFE invariance

polyDFE’s statistical performance for estimating DFE parameters has already been validated using datasets simulated under a broad range of DFE types and demographic scenarios (Tataru et al. 2017), including the four DFE types listed in Table S1. Therefore, we focus here on the performance of two likelihood ratio tests (LRT) comparing models M1 vs M2 and M3 vs M4 (see Table 1 in the main text). The percentage of p-values that are below the 5% threshold gives the type I error (when a pair of datasets were simulated using the same underlying DFE) and the power (when a pair of datasets were simulated using different underlying DFEs).

When performing a LRT, the distribution of the LRT statistic under the null model (invariant DFE) is expected to asymptotically converge towards a χ2 distribution. This χ2 approximation is expected to be valid when the sample size, here the total number of counts in the SFS, becomes large. For both of the LRTs we investigate here (M1 vs M2 and M3 vs M4), the difference in number of fitted parameters between the models is the number of parameters that characterize each DFE. Here we use a DFE parametrization with 4 independent parameters (*Sd, b, pb* and *Sb* see Table S1) and fit either a single DFE (4 parameters in M1 or M3) or two distinct DFEs (8 parameter in M2 or M4). Note that polyDFE implements different functional forms for the DFE (see Tataru et al. 2017 for details). The difference in the number of fitted parameters when testing for DFE invariance should always be equal to the number of parameters characterizing the DFE.

When using datasets of approximately 10Mb of sites each, we expect that the number of counts in each SFS is large enough for a χ2distribution with 4 degrees of freedom to be accurate when calculating a p-value for the LRT (Figure S1) for both model comparisons (models M1 vs M2 and M3 vs M4). When comparing models M1 and M2, the empirical null distribution of the LRT statistic is poorly approximated by the χ2 distribution (Figure S2A), even when the amount data is large (10Mb), which leads to a high type I error (Figure S1). However, the χ2 approximation performs much better for the LRT comparing models M3 and M4 (Figure S2A). This suggests that model M1 may not be a reliable null model for determining if the DFE is shared or not between datasets.

The generative model implemented in polyDFE assumes unlinked SNPs and uses nuisance parameters to model distortions of SFS counts relative to expectations in a stable Wright Fisher population. Future work should investigate whether this strategy can fully account for the intricate interaction between selection, genetic drift and linkage.

When smaller datasets are analyzed, the empirical null distribution of the LRT statistic is no longer reliably approximated by a χ2 distribution (Figure S2). Therefore, to investigate how power changes as a function of the size of the dataset (Figure S4), we relied on the empirical (simulated) null distribution of the LRT statistic to obtain the p-values. Note that in this case the type I error is automatically calibrated at its nominal value, here 5%. To avoid using the χ2 distribution to approximate the distribution of LRTs when testing for DFE invariance among datasets of small size, one can simulate multiple datasets under the inferred null model, infer the parameters for these datasets using the null and alternative models and obtain an empirical null distribution for the LRT statistic. The simulation can be performed directly from polyDFE.

# Details of the inference on the chimpanzee datasets

For illustration, we analyzed a chimpanzee dataset (described in Bataillon et al. 2015) and tested: (1) if the central and eastern chimpanzee subspecies shared the same DFE, (2) if autosomes and X chromosome share the same DFE in each chimpanzee subspecies. To do so, we fitted the data in polyDFEv2.0 using models M3 and M4 (Figure S3). The inference was performed on the SFS data only (divergence counts to an outgroup were not fitted). We deliberately fit SFS data only because we do not have any means to know a priori if the intensity of selection remained constant along the divergence branch. That assumption is necessary when using divergence data, and it can bias inference if it is violated by the data (Tataru et al. 2017). SFS data was fitted using a DFE model comprising both deleterious and beneficial mutations, and both ancestral SNP misidentification error *ε* and demography nuisance parameters (*ri*s) were estimated jointly with the DFE parameters. Mutation rates were assumed constant across the data (this assumption does not affect the DFE estimation, see Tataru et al. 2017 for details). To ensure that the likelihood function was reliably maximized, we performed 10 runs of the BFGS algorithm, with randomly starting values.

# References

1. Bataillon T, Duan J, Hvilsom C, Jin X, Li Y, Skov L, Glemin S, Munch K, Jiang T, Qian Y, et al. 2015. Inference of purifying and positive selection in three subspecies of chimpanzees (*Pan troglodytes*) from exome sequencing. Genome Biol Evol 7:1122–1132.
2. Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103–1119.

Figure S1 Violin plot of p-values for LRTs for DFEs heterogeneity. DFE types used for simulations are indicated on the x-axis and detailed in Table S1. LRTs are obtained by comparing models M1 and M2 (solid line) or models M3 and M4 (dashed line). The p-value of each LRT is computed assuming a χ2 distribution. All LRTs are performed on pairs of SFS data containing approximately 10MB of sites. (A) Type I error: polyDFEv2.0 was run on pairs of datasets simulated using the same DFE (black outline). (B) Power: polyDFEv2.0 was run on pairs of datasets simulated using different DFEs (red outline). Dots indicate the median. The horizontal lines show the 5% threshold and a percentage in each violin plot indicates the percentage of p-values that are under the 5% threshold. For visualization reasons, p-values smaller than 10-10 have been set to 10-10.

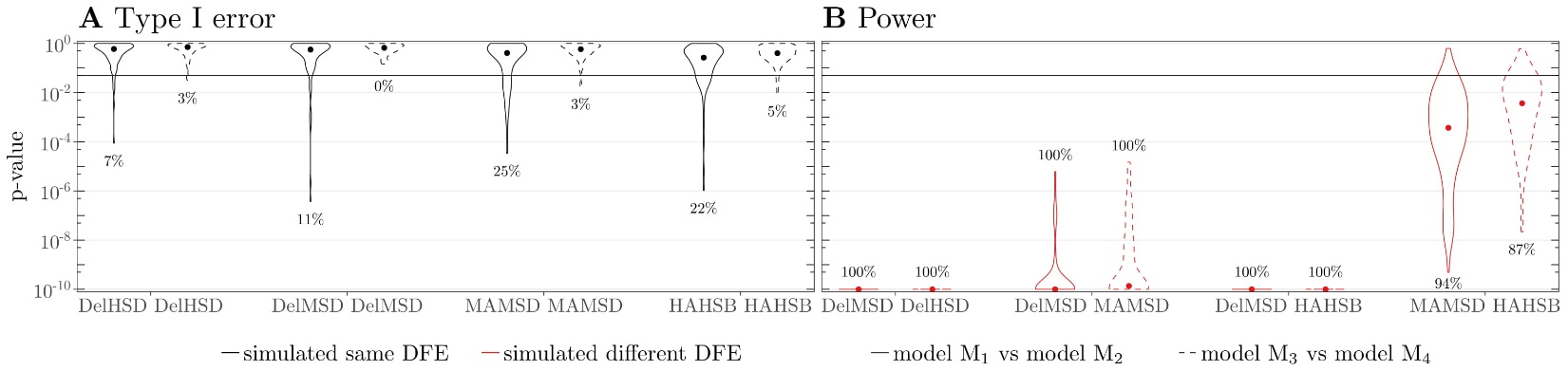
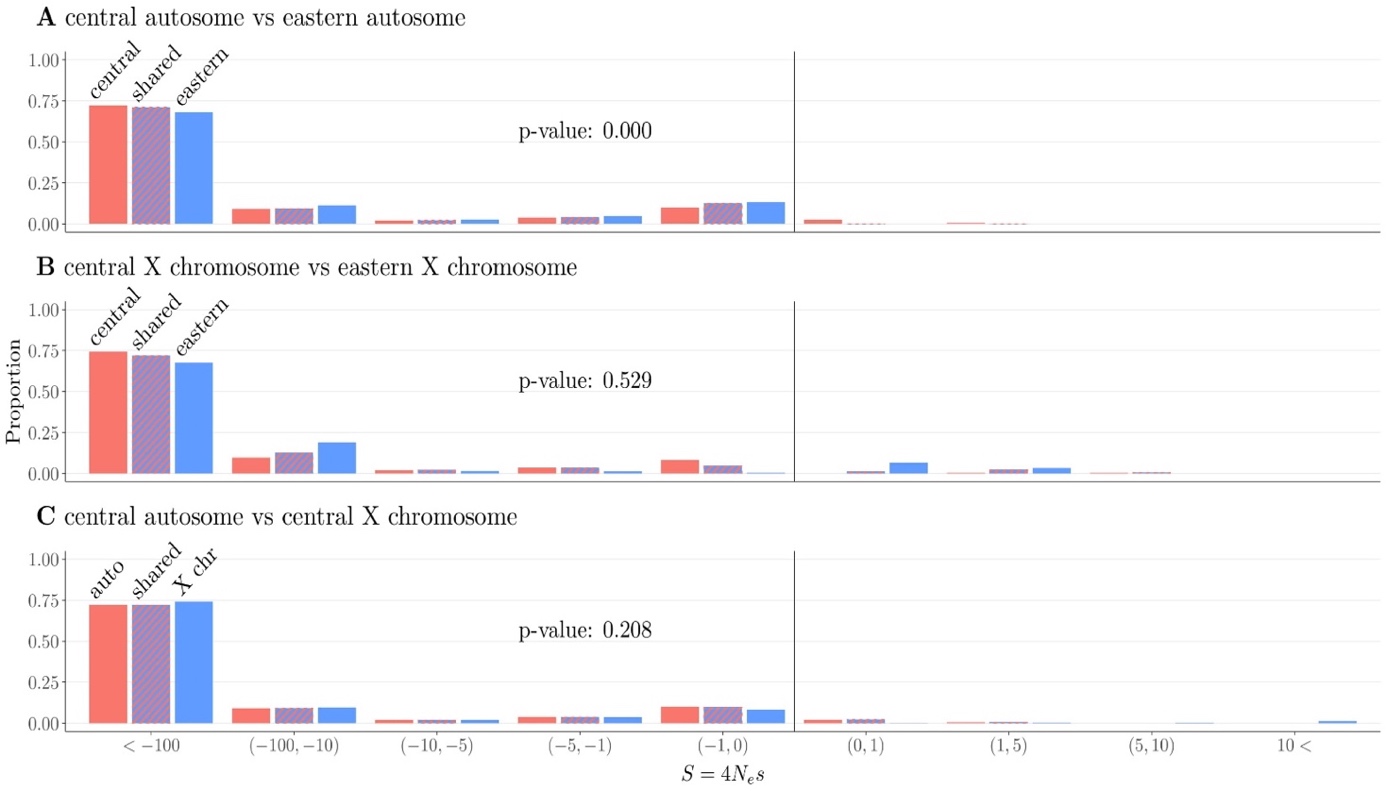
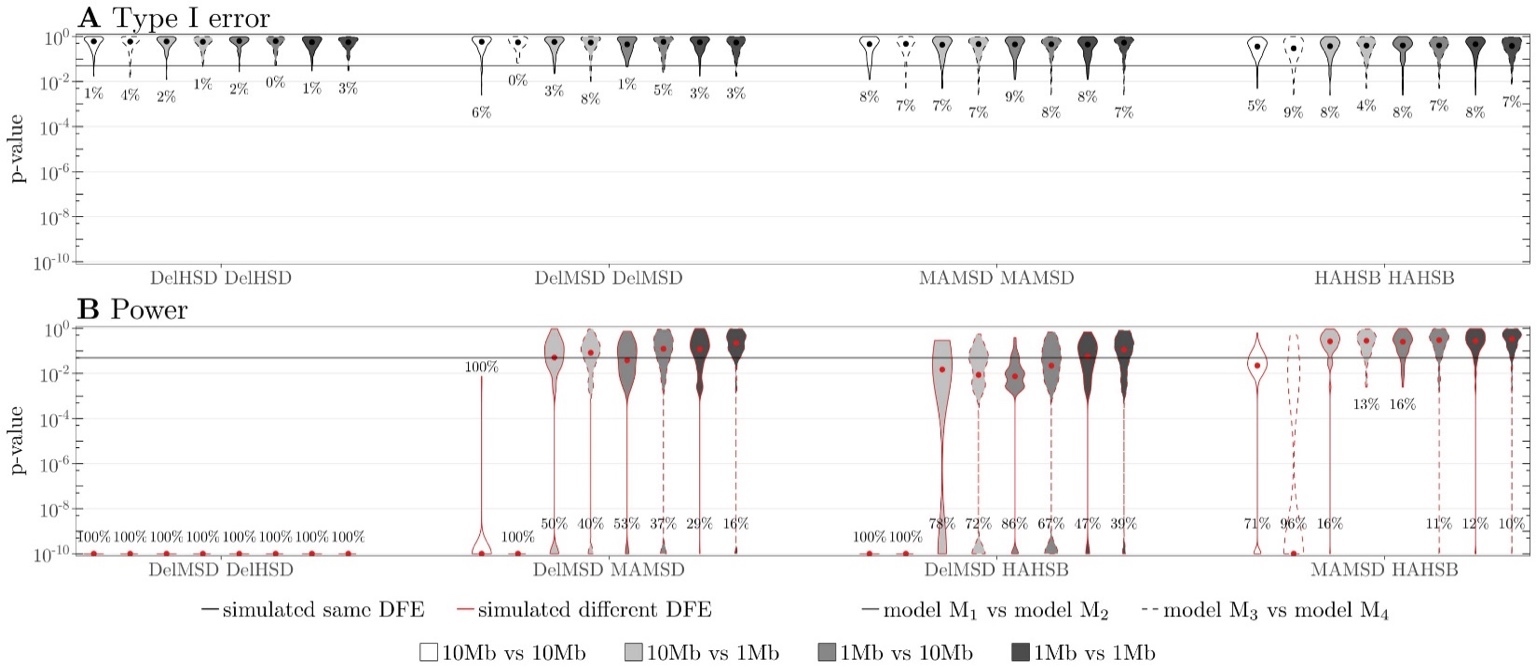


Figure S2 Empirical null distribution of the LRT statistic used for each LRT and their 2 approximation. The empirical distribution of each LRT (M1 versus M2 andM3 versus M4) is obtained from all datasets that were simulated assuming the same underlying DFE, regardless of the actual simulated DFE (Table S1). When using datasets of different size for each SFS (based on 1 or 10 Mb), we obtain different empirical distributions (A to D). The χ2 distribution (with 4 degrees of freedom) expected to be asymptotically a good approximation for each LRT, is given in gray. The vertical dotted line shows the 5% cutoff from the χ2 distribution, while the numbers in the plot indicate the percentage of LRT statistics that are above this cutoff, for each empirical distribution. Note that in all figures panels (A to D) we aggregated the LRT statistic from the LRT computed on pairs of datasets that shared the DFE, regardless of the actual simulated DFE.



*Figure S3* Inferred DFE on the chimpanzee SFS datasets. The DFE model fitted comprises both deleterious mutations (negative scaled selection coefficients *S*) and beneficial mutation effects (positive *S*). The DFE was either inferred under model M4, fitting DFE parameters independently for each dataset (red and blue), or under model M3, assuming shared DFE parameters for both datasets (hashed red and blue). The p-values were obtained from LRTs comparing models M3 and M4 using a χ2 with 4 degrees of freedom as null distribution. Inference was performed using a continuous DFE (mixture of gamma and exponential, as detailed in the SFS simulation setup section) but for graphical convenience DFEs are discretized in classes of *S*.



*Figure S4* Violin plot of p-values for LRTs testing for DFEs heterogeneity. DFEs used for simulation of dataset are indicated on the x-axis and detailed in Table S1. LRTs are obtained by comparing models M1 and M2 (solid line) or models M3 and M4 (dashed line). The p-value for each LRT performed is obtained by relying on the empirical null distribution (Figure S2). Grey shading indicates the size of each SFS dataset analyzed when performing the LRT. (A) Type I error: polyDFEv2.0 was run on pairs of datasets simulated using the same DFE (black outline). (B) Power: polyDFEv2.0 was run on pairs of datasets simulated using different DFEs (red outline). Dots indicates the median. The horizontal lines show the 5% threshold and percentages in each violin plot indicate the percentage of p-values that are below the 5% threshold. For visualization reasons, p-values smaller than 10-10 have been set to 10-10..