Supplementary Information for

Mutation bias shapes the spectrum of adaptive substitutions

Alejandro V. Cano, Hana Rozhoňová, Arlin Stoltzfus, David McCandlish and Joshua L. Payne

David McCandlish
E-mail: mccandlish@cshl.edu
Joshua L. Payne
E-mail: joshua.payne@env.ethz.ch

This PDF file includes:

Supplementary text
Figs. S1 to S5
Tables S1 to S3
SI References

Supporting Information Text

Model description

Our motivation is to develop and apply a phenomenological model that allows us to define a statistic that quantifies the influence of the mutation spectrum on the spectrum of adaptive substitutions. We focus on the most general and widely available data on adaptive genetic change: Single-nucleotide changes that alter amino acids, i.e., single-nucleotide missense changes. The standard genetic code specifies a set of 354 different types of single-nucleotide missense changes defined by a starting codon and an ending amino acid. For genomes that use the standard genetic code, any given episode of adaptation involving missense changes induces a distribution of adaptive substitution events over these 354 types, which we refer to as the spectrum of adaptive substitutions.

Because our goal is to model the observed number of counts for each type of adaptive substitution, we use negative binomial regression (1), which is a type of generalized linear model that is often employed for modeling count data. It is appropriate because the 354 mutational types are discrete and the substitution events that correspond to each type occur independently of one another. The general form of the model is

$$\log \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \log(\text{exposure}) + \beta\mathbf{x},$$

where Y is a vector of response variables, $\mathbf{x}$ are the explanatory variables, β_0 is the logarithm of the constant of proportionality, and exposure quantifies differences in the potential to observe each type of response. In our case, Y is the number of substitution events of each type, $\mathbf{x}$ is the logarithm of the mutation rate, and the exposure is given by codon frequency, which controls for the number of times each codon appears in protein-coding regions of the genome. Our model then takes the form

$$\log \mathbb{E}[\mathbf{n}(c,a)|\log(\mu(c,a))] = \beta_0 + \log(f(c)) + \beta \log \mu(c,a),$$

where $\mathbf{n}$ is the spectrum of adaptive substitutions (i.e., $\mathbf{n}(c,a)$ is the number of substitution events from codon c to amino acid a), $\mu(c,a)$ is the mutation rate from codon c to amino acid a, and $f(c)$ is the frequency of codon c in the genome. The coefficient β is a single statistic that captures the influence of the mutation spectrum on the spectrum of adaptive substitutions. The expected range of β is from 0 to 1: If $\beta = 0$, the mutation spectrum has no influence on the spectrum of adaptive substitutions. If $\beta = 1$, the mutation spectrum has a proportional influence on the spectrum of adaptive substitutions. Values of β between 0 and 1 represent an intermediate influence.

The above model only describes the expected number of counts for each type of substitution, however to fit the parameters of the model by maximum likelihood we must specify a full distribution for $\mathbf{n}(c,a)$. One common choice would be to assume that these counts are Poisson distributed (i.e. Poisson regression). However, Poisson regression assumes that the variance in the counts data is equal to the mean. In our data, we instead observe overdispersion, i.e. that the variance is larger than the mean. Such overdispersion is a common problem in Poisson regression. The standard solution is to instead use negative binomial regression, a more general model that allows the variance to be different from the mean (1). In the main text, we therefore use negative binomial regression to model the influence of the mutation spectrum on the spectrum of adaptive substitutions.

Meaning of key terms

Several important terms used in our study, such as “mutation”, have meanings that are interpreted differently in different parts of the scientific community (2). Moreover, our study design requires additional precision in being able to describe genetic and evolutionary changes, for example distinguishing a possible beneficial change to the genome of an organism from a realized instance where a heritable change of that type arises in a particular individual. In order to avoid any terminological ambiguity, we therefore provide formal definitions for these key terms below.

adaptive substitution An adaptive substitution is an evolutionary change in a population or sub-population, where each adaptive substitution is understood (in the present context) to result from an event of mutational introduction and an episode of selective enrichment that raises the mutant allele to a frequency close to 1.

event An instance of change, having a particular time and place of occurrence, is an event. Compare to path or type. Here we assume that events occur independently from each other, and distinguish e.g. the number of times a particular mutational variant is observed from the number of distinct mutational events that introduced that variant into the population.

missense (nonsynonymous) In the literature of molecular evolution, codon changes that alter the amino acid are missense changes, and this class of change is often called “non-synonymous” (e.g., in the dN / dS literature) although technically non-synonymous changes include both missense and nonsense changes.

mutation A mutation is a heritable change to the genetic material in an individual lineage. The process of such change is also called mutation. The product of a mutational change is also called “a mutation” or a “mutant allele,” and in population genetics this kind of usage is often extended to refer generally to derived alleles, e.g., the “concurrent mutations” regime refers to mutant alleles segregating concurrently in a population.

mutational type An event of evolutionary or mutational change can be assigned to a variety of mutational types or categories defined by a class of starting states (e.g., ATG codons) and a class of ending states (e.g., TTG codons). Here we are mainly focused on the 6 (reversible) nucleotide-to-nucleotide types and the 354 codon-to-amino-acid types. We use “path” for a specific kind of mutational type (see path).

path For the purposes of describing observed data sets for specific organisms, a path is a mutational type defined by a specific genomic site and a codon-to-amino-acid change. Parallel or recurrent events within a dataset are events that take place along the same path.

spectrum A set of intensities or frequencies over some space of possibilities (e.g. a collection of different types of mutations) is a spectrum.

	Data		Neg. binomial regression		Prediction model		Spectrum elements	
Study	Paths	Events	β	p_β	Correlation	p_{corr}	Non-zero elements	Entropy
Basel (3)	126	2319	0.83 ± 0.27	0.002	0.15	0.005	78	0.53
Manson (4)	168	2094	0.84 ± 0.27	0.001	0.17	0.002	80	0.52

Table S1. Separately analyzing the adaptive events from the two meta-analyses of antibiotic resistance substitutions in *M. tuberculosis* yields qualitatively similar results to analyzing them together. Shown are the observed numbers of paths and events, the mutation coefficient β (with standard error) and its p-value, the Pearson's correlation between observed and predicted spectra of adaptive substitutions and its p-value, as well as the number of non-zero elements of the spectrum of adaptive substitutions and the entropy of the spectrum of adaptive substitutions.

	Influence of ti/tv ratio			Influence of rest of mutation spectrum			Model comparison	
Species	$\beta_{\mathrm{ti/tv}}$	95% CI	$p_{\beta_{\mathrm{ti/tv}}}$	β_{rest}	95% CI	$p_{\beta_{\mathrm{rest}}}$	log likelihood	p_{LRT}
S. cerevisiae	0.79 ± 0.13	$[0.53, 1.05]$	$< 10^{-8}$				-1266.15	$< 10^{-16}$
	0.85 ± 0.11	$[0.63, 1.07]$	$< 10^{-16}$	1.25 ± 0.11	$[1.03, 1.47]$	$< 10^{-16}$	-1156.48	
E. coli	0.80 ± 0.17	$[0.46, 1.14]$	$< 10^{-5}$				-1109.90	$< 10^{-6}$
	0.85 ± 0.17	$[0.51, 1.19]$	$< 10^{-6}$	1.28 ± 0.26	$[0.77, 1.80]$	$< 10^{-6}$	-1083.80	
M. tuberculosis	0.84 ± 0.33	$[0.19, 1.50]$	0.01				-1233.32	0.03
	0.89 ± 0.32	$[0.26, 1.52]$	0.01	0.80 ± 0.36	$[0.09, 1.51]$	0.02	-1228.67	

Table S2. The entire mutation spectrum provides better model fits than just the transition-transversion ratio. Shown are the regression coefficient of the transition-transversion ratio $\beta_{\mathrm{ti/tv}}$ (with standard error), its 95% confidence interval and its p-value, the regression coefficient associated to the rest of the mutation spectrum β_{rest} (with standard error), its 95% confidence interval and its p-value, as well as the p-value of the likelihood ratio test comparing both models p_{LRT}, which indicates that the more complex model including the full mutation spectrum provides a significantly better fit than the simpler model that only includes the transition-transversion ratio.

	Only codon frequencies		Complete model	
Species	Correlation [CI]	p_{corr}	Correlation [CI]	p_{corr}
S. cerevisiae	$0.36\ [0.25, 0.44]$	$< 10^{-11}$	$0.68\ [0.62, 0.73]$	$< 10^{-16}$
E. coli	$0.31\ [0.22, 0.40]$	$< 10^{-9}$	$0.41\ [0.31, 0.49]$	$< 10^{-14}$
M. tuberculosis	$0.10\ [-0.0004, 0.2059]$	0.05	$0.16\ [0.05, 0.26]$	0.003

Table S3. A model using codon frequencies and the mutation spectrum provides better predictions than a model using only codon frequencies ($\beta = 0$). Shown are the correlation coefficients for the two models, with 95 % confidence intervals and p-values.

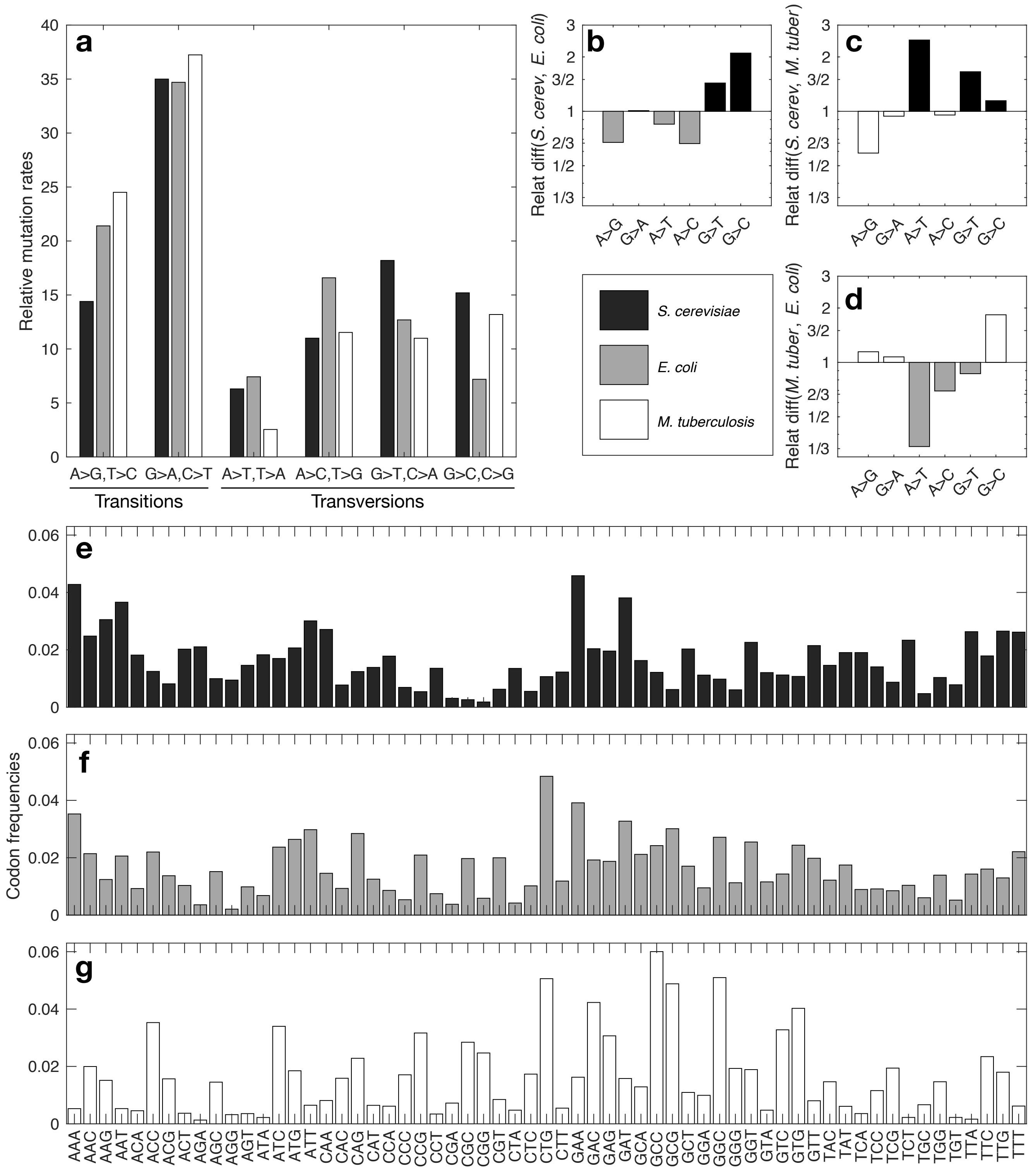

Fig. S1. Empirical mutation spectra and codon frequencies. (a) Bar plots of the empirical mutation spectra for *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. Bar color indicates the species; see legend. (b-d) Relative difference in mutation rates per mutation type, Relat diff$(b, a) = b/a$. Bar color indicates the species with the higher mutation rate for each mutation type. The vertical axis is logarithmically scaled for visual clarity. (e-g) Bar plots of the empirical codon frequencies for (e) *S. cerevisiae*, (f) *E. coli*, and (g) *M. tuberculosis*.

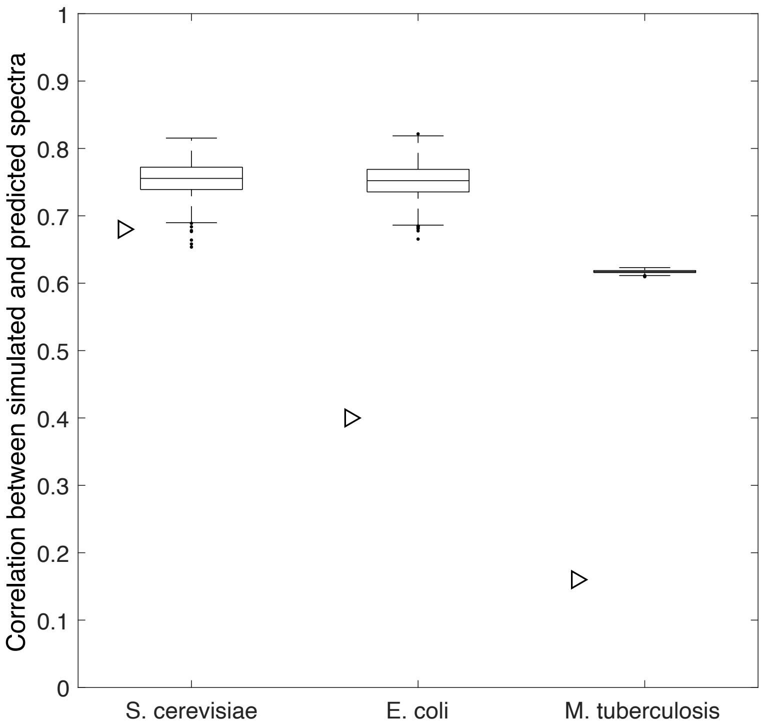

Fig. S2. The correlation between predicted and simulated spectra of adaptive substitutions depends on mutational target size, even under origin-fixation dynamics. The distribution of correlations between predicted and simulated spectra of adaptive substitutions using the codon frequencies, mutation spectra, and number of non-zero elements in the spectrum of adaptive substitutions are shown for *S. cerevisiae*, *E. coli*, and *M. tuberculosis*. Data pertain to 10^3 simulations. Triangles show the correlations reported in Table 1, for reference.

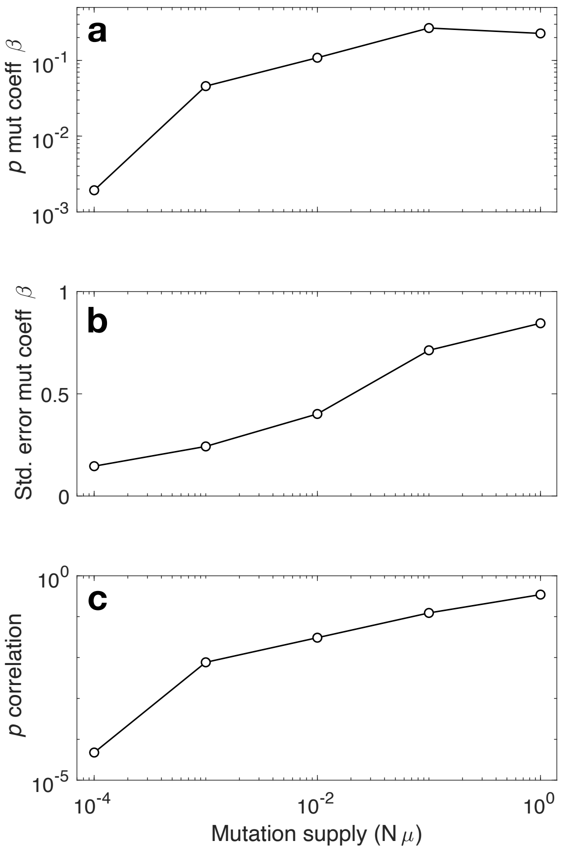

Fig. S3. High mutation supply diminishes the influence of mutation bias on adaptive evolution. The a) average p-value and b) standard error of the mutation coefficient β, and c) the average p-value of the correlation between predicted and simulated spectra of adaptive substitutions are shown in relation to mutation supply $N\mu$. Data pertain to those shown in Figs. 4a-c.

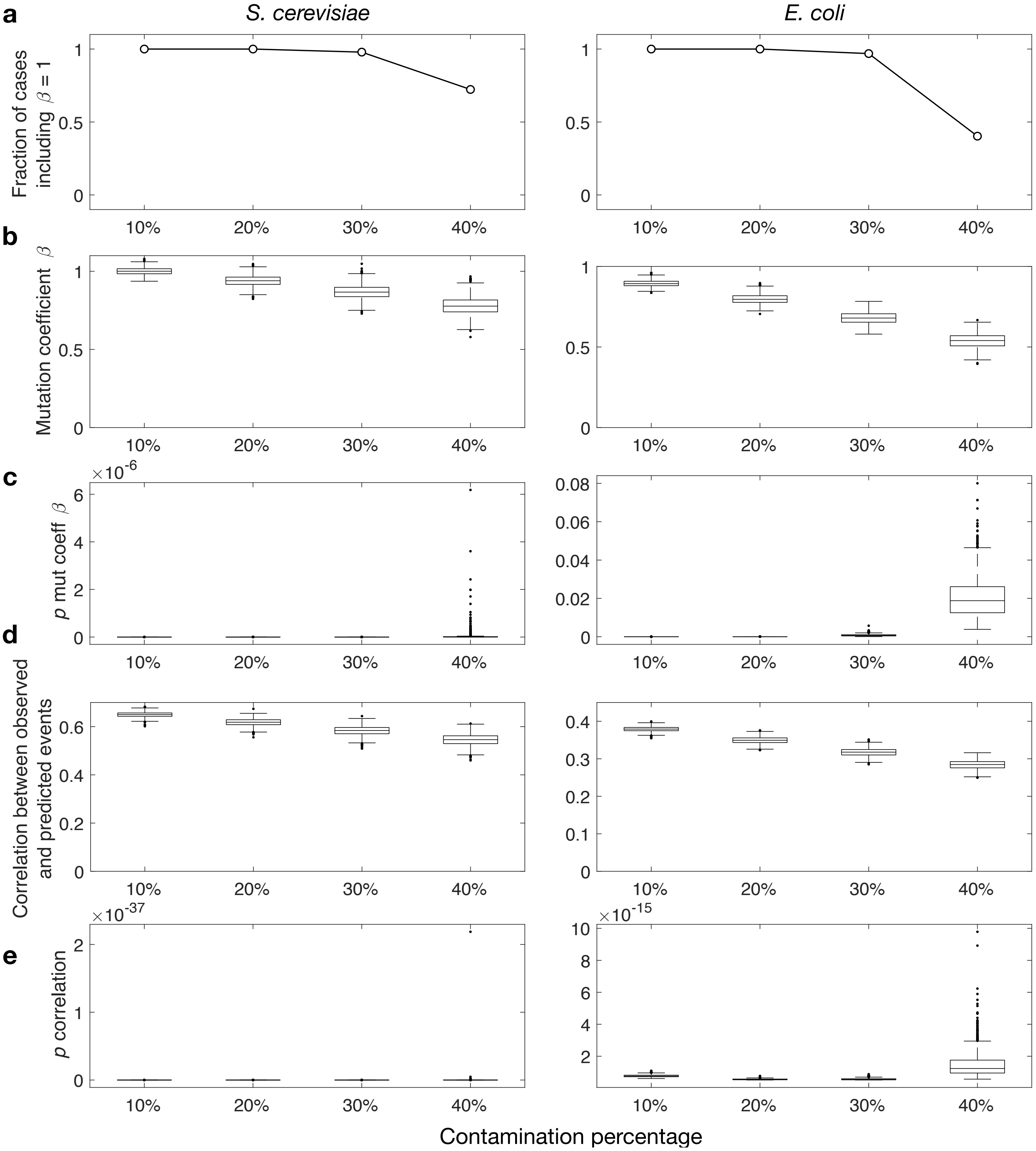

Fig. S4. Contamination analysis supports the influence of mutation bias on adaptation. (a) Fraction of simulated data sets in which the confidence interval includes $\beta = 1$. (b) Inferred mutation coefficients β, (c) *p*-values of the regression coefficients β, (d) Pearson's correlation coefficients between observed and predicted spectra of adaptive substitutions, and (e) the *p*-values of the correlation coefficients, are all shown in relation to the percentage of substitutions randomly removed from the data sets of adaptive substitutions.

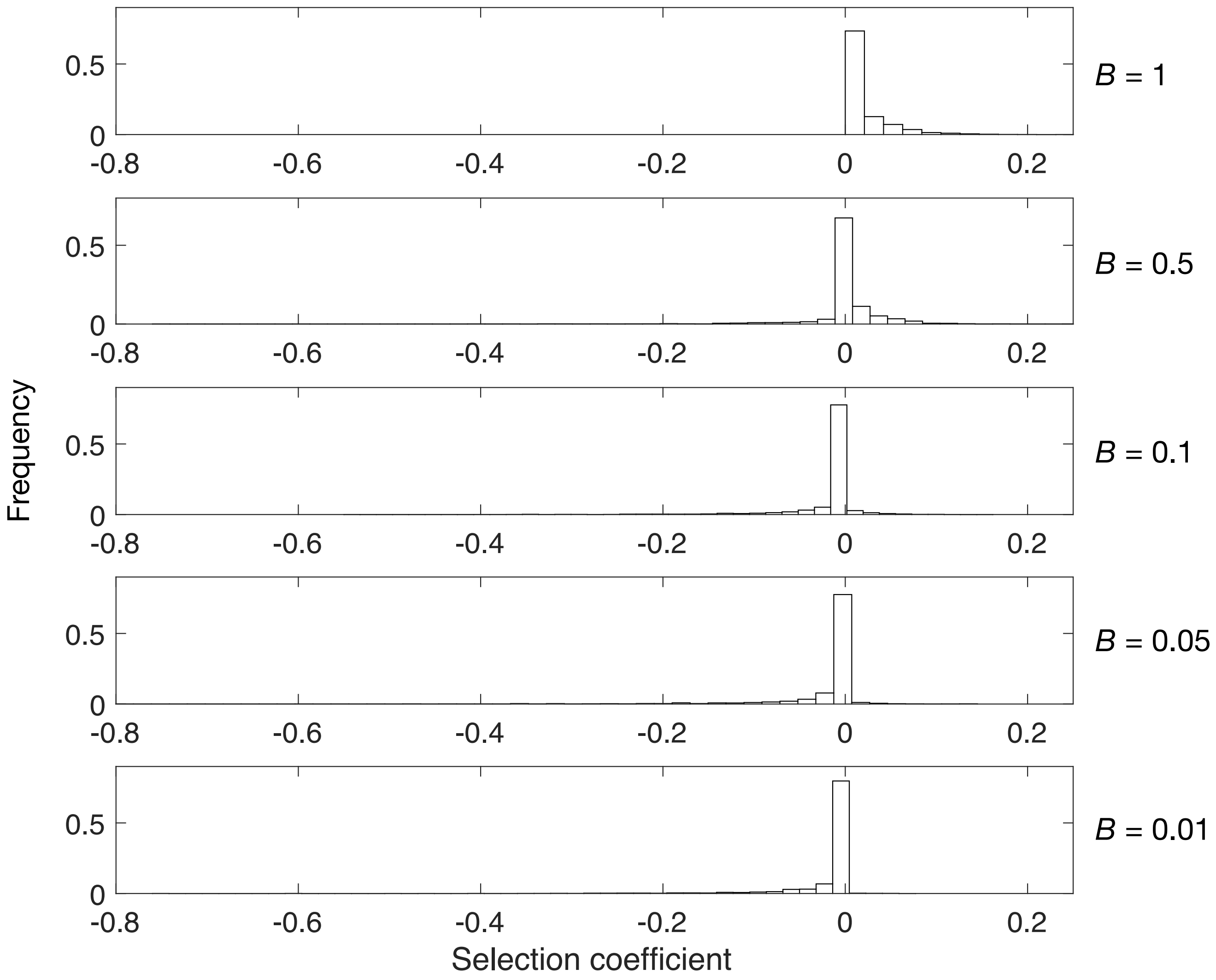

Fig. S5. Distributions of fitness effects. Representative distributions of fitness effects used in the evolutionary simulations for five different proportions of beneficial mutations B.

References

1. P McCullagh, J Nelder, *Generalized Linear Models, Second Edition*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. (Taylor & Francis), (1989).
2. R Karki, D Pandya, RC Elston, C Ferlini, Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC medical genomics* **8**, 1–7 (2015).
3. JL Payne, et al., Transition bias influences the evolution of antibiotic resistance in *Mycobacterium tuberculosis*. *PLoS Biol.* **17** (2019).
4. A Manson, et al., Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* **49**, 395–402 (2017).