SUPPORTING INFORMATION:

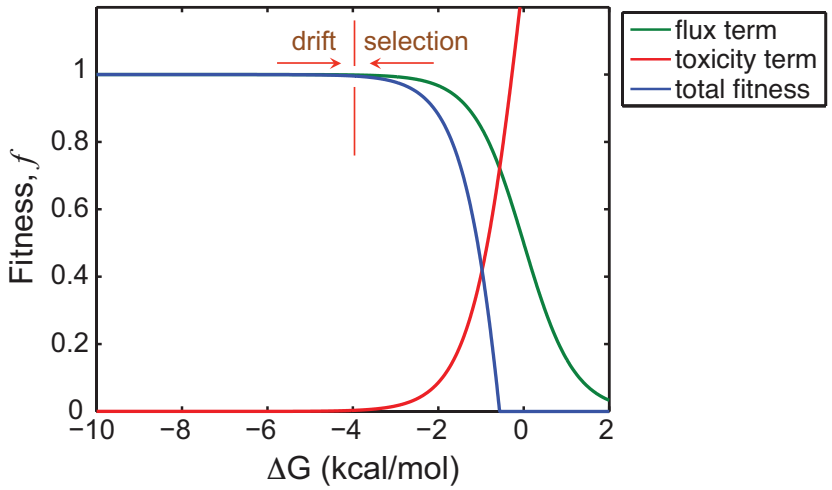
Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions

Adrian W.R. Serohijos and Eugene I. Shakhnovich[†]

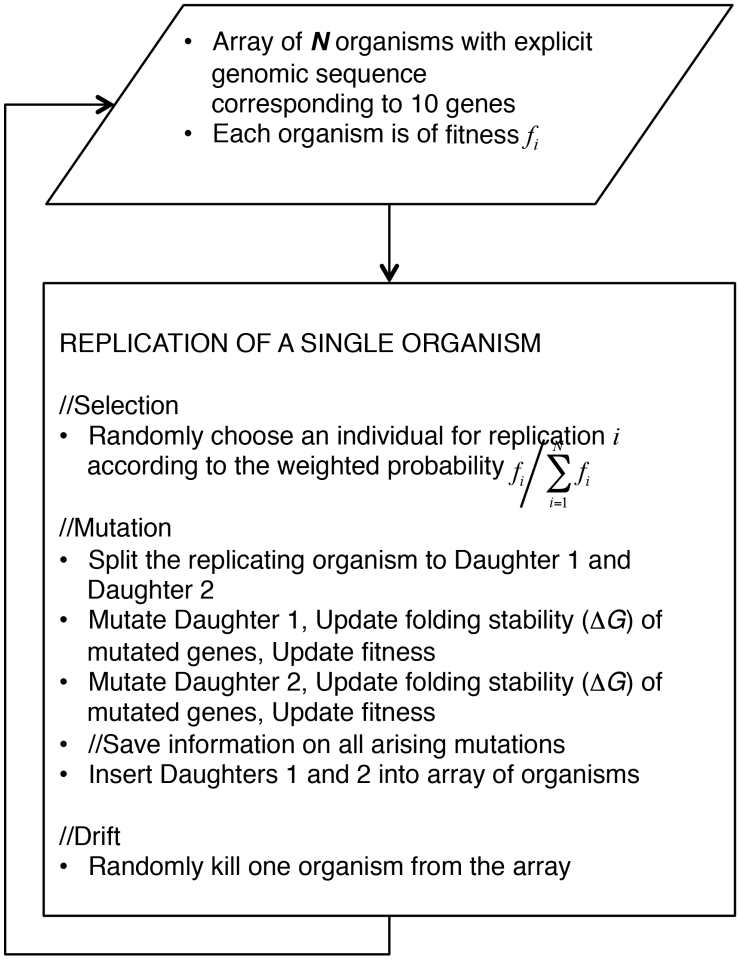
Department of Chemistry and Chemical Biology, Harvard University,

Cambridge, MA USA 02138

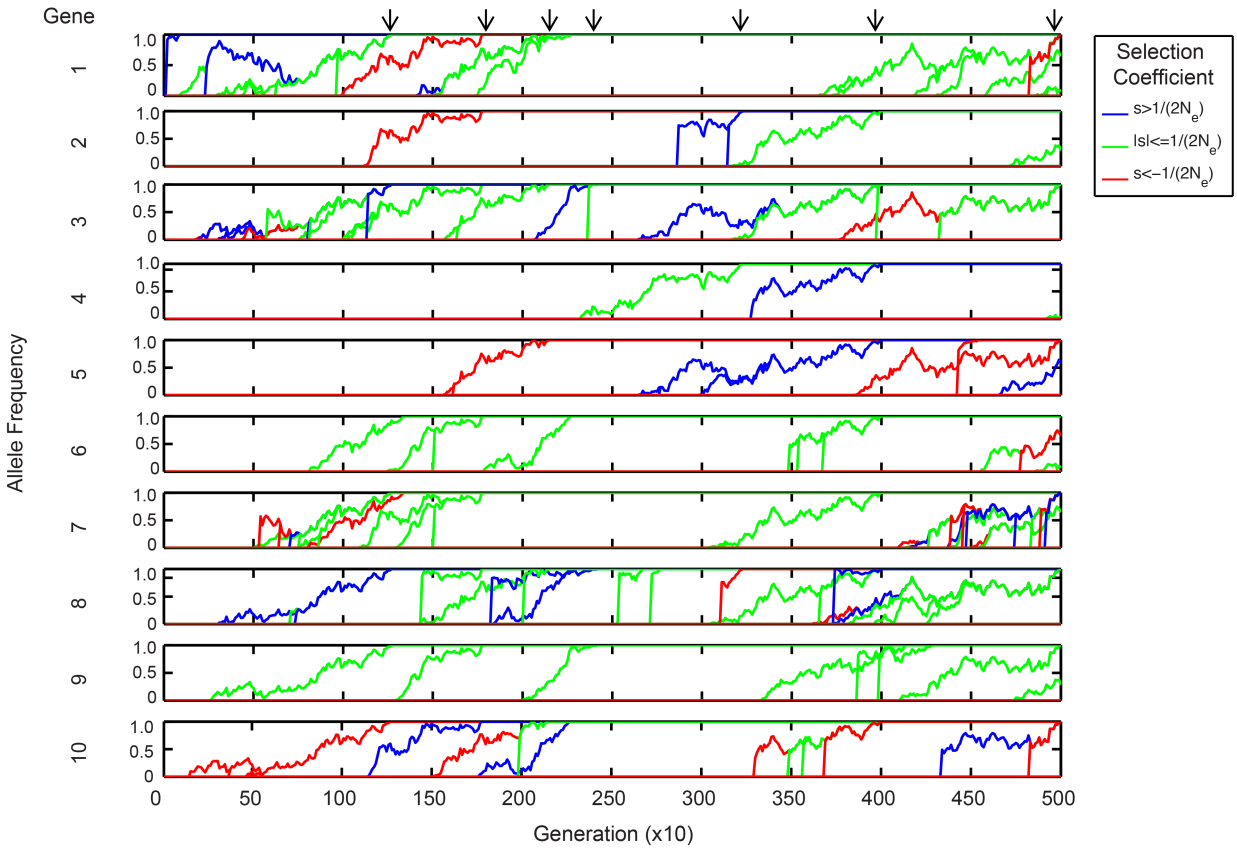
[†]*Correspondence: E.I.S. (shakhnovich@chemistry.harvard.edu)*



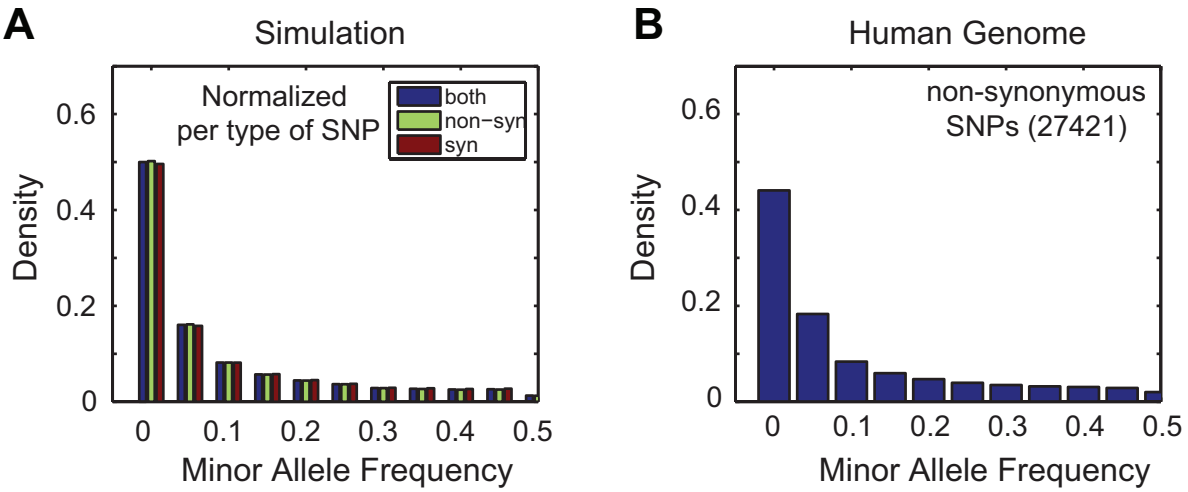
Supplementary Fig. S1. Organismal fitness *f* as a function of folding stability (Eqns. 1 to 3). The stability corresponding to mutation-selection balance is the same as indicated in fig. 4F.



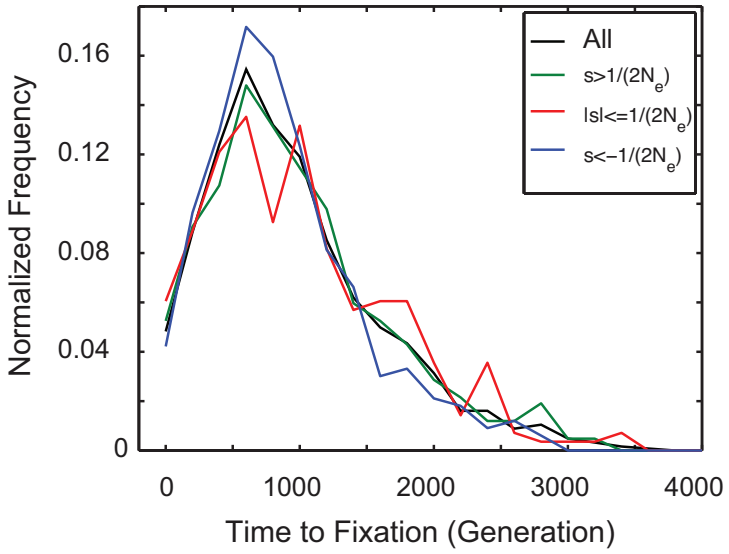
Supplementary Fig. S2. Algorithm of the explicit *in silico* evolution of model organisms in the polyclonal that accounts for mutation, drift, and selection (Methods). Random mutations occur at the primary sequence and can affect the folding stability of the protein product.



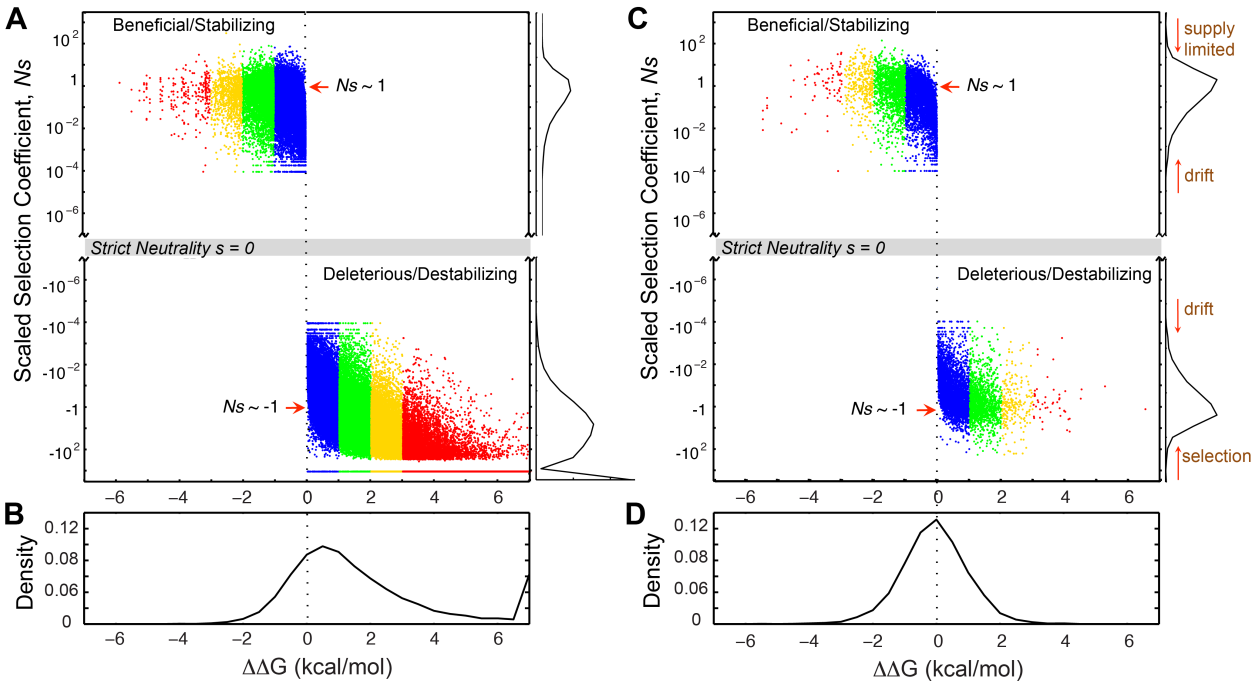
Supplementary Fig. S3. Representative histories of mutations that fixed in the evolving population. Time window is as in figs. 1B-C and 2A-C. Mutations are colored according to their fitness effects.



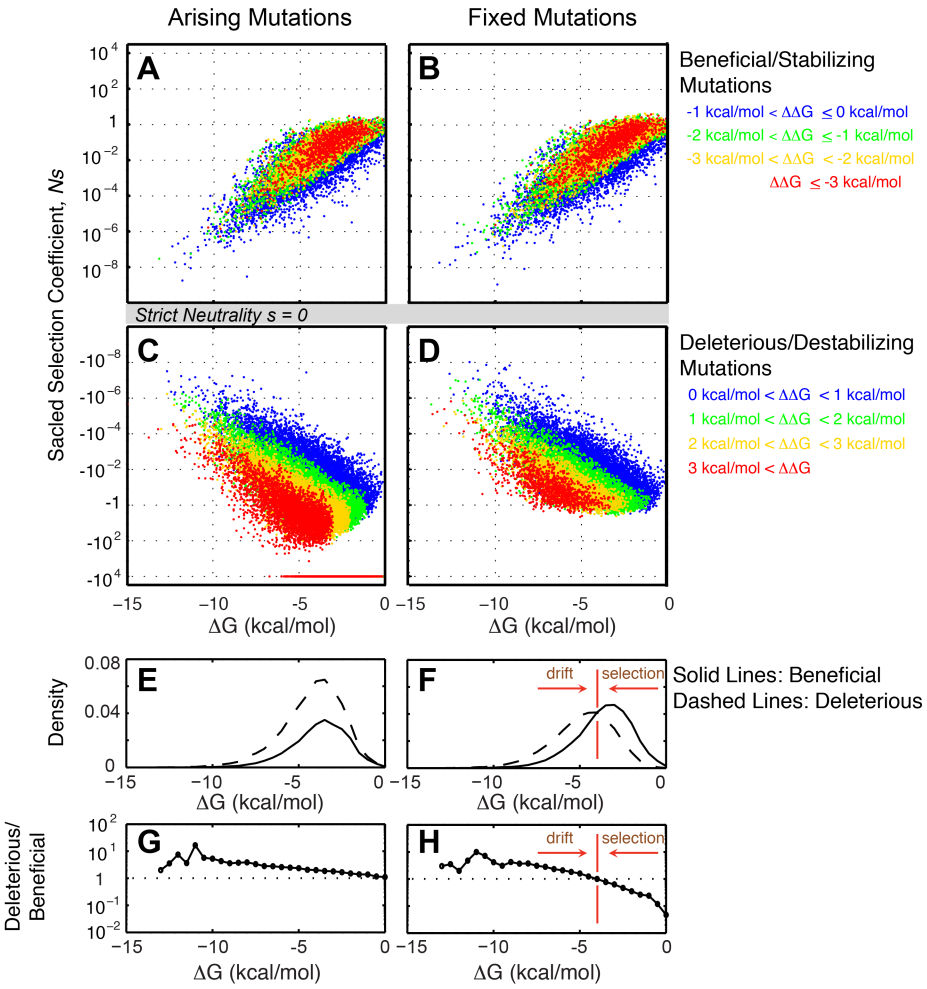
Supplementary Fig. S4. Distribution of minor allele frequencies in simulation **(A)** and in the coding region of the human genome **(B)**.



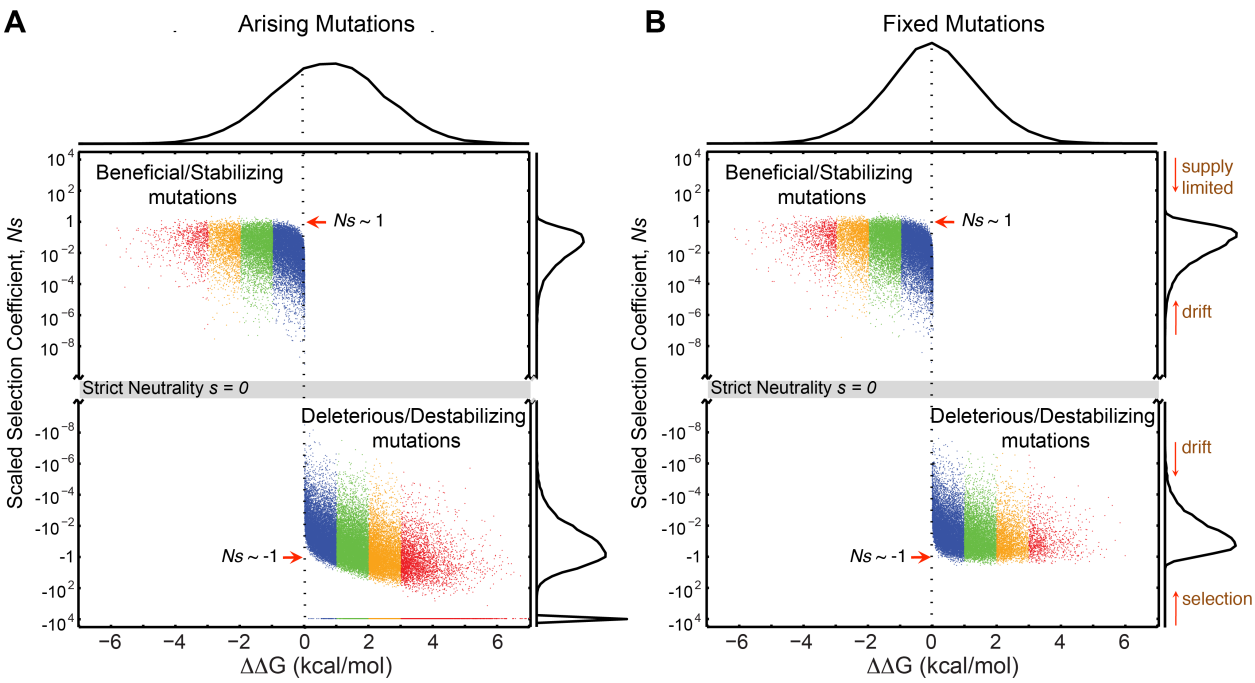
Supplementary Fig. S5. Distribution of fixation of times of among classes of mutations.



Supplementary Fig. S6. Mapping between the distribution of molecular effects of mutations and the distribution of their fitness effects *s* **(Polyclonal regime)**. Data points and coloring are similar to supplementary fig. 6. **(A)** Mapping of ΔΔ*G* to *s* among arising non-synonymous mutations. Most arising mutations have destabilizing or deleterious effect. Histogram on the side panel corresponds to fig. 5B. **(B)** Distribution of ΔΔ*G* for randomly arising non-synonymous mutations (see also ref. (Tokuriki et al, 2007)). **(C)** Similar to panel (A) but for fixed mutations (Also similar to fig. 5A). Data points are as in fig. 5A. The consequence of mutation-selection balance is bimodality in the distribution of *s*. **(D)** Distribution of ΔΔ*G* for fixed non-synonymous mutations (Also similar to fig. 5D).



Supplementary Fig. S7. Influence of epistasis in the mapping between the molecular effect of a mutation $\Delta\Delta G$ to its phenotypic effects s **(Monoclonal regime)**. To serve as control, we performed evolutionary simulations in the monoclonal regime. The monoclonal population consists of N_e=10^4 model cells. The cell is composed of 10^3 genes, each with protein abundances ranging from 10 to 10^6 copies per cell, to recapitulate the broad distribution of abundances in real organisms (Ghaemmaghami et al, 2003; Ishihama et al, 2008). (Detailed method is described in ref. (Serohijos et al, 2012)). Mutations are colored according to the magnitude of the $\Delta\Delta G$ (See leftmost panel for color assignment). **(A)** Arising random beneficial (stabilizing) mutations. **(B)** Fixed beneficial mutations. **(C)** Arising random deleterious (destabilizing) mutations. **(D)** Fixed deleterious mutations. **(E)** Solid line is the distribution of arising beneficial mutations (panel a) while dashed line is the distribution of fixed beneficial mutations (panel B). Each histogram is normalized to the total number of mutations. **(F)** Similar to panel E but for fixed mutations. **(G-H)** Ratio of deleterious to beneficial mutations. Note: For the sake of clarity, we plot only $1/10^4$ (i.e., $1/(N_e)$) of the total number arising mutations sampled randomly.



Supplementary Fig. S8. Mapping between the distribution of molecular effects of mutations and the distribution of their fitness effects *s* **(Monoclonal regime)**. Data points and coloring are similar to fig. 7. **(A)** Mapping of $\Delta\Delta G$ to *s* among arising non-synonymous mutations. Most arising mutations have destabilizing or deleterious effect. Histogram at the right side panel is the distribution of *s*. Histogram at the top is the distribution of $\Delta\Delta G$. **(B)** Similar to panel **A** but for fixed mutations. The magnitude of the selection coefficient of *arising* beneficial mutation (panel **A**) and the *fixed* beneficial and deleterious mutations are sharply bounded by $N|s|\sim 1$.

Supplementary Table S1. Genes in the *in silico* model organism.

Gene ID	PDB ID_Chain ID_Domain ID	Abundance*	Description
1	1IS7_A	2840	GTP cyclohydrolase I
2	1SQL_A	4590	7,8-dihydroneopterin aldolase
3	1K0E_A	1550	P-aminobenzoate synthase component I
4	1I2K	768	Aminodeoxychorismate lyase
5	1I2K_D1	768	Aminodeoxychorismate lyase
6	2BMB_A_D2	4590	Pterin binding enzyme
7	2BMB_A_D3	4590	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)
8	1O5Z_A_D1	2340	Mur ligase family, glutamate ligase domain
9	1O5Z_A_D2	2340	Mur ligase middle domain
10	1AI9_A_D1	1550	Dihydrofolate reductase

*Values are derived from the measured abundance of their orthologs in Yeast.

Supporting References

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. *Nature* **425:** 737-741

Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D (2008) Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* **9:** 102

Serohijos AW, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Reports* **2:** 249-256

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* **369:** 1318-1332