A comprehensive, high-resolution map of a gene's fitness landscape

Elad Firnberg[1], Jason W. Labonte[1], Jeffrey J. Gray[1], and Marc Ostermeier[1]

[1]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218 USA

Figures S1-S15

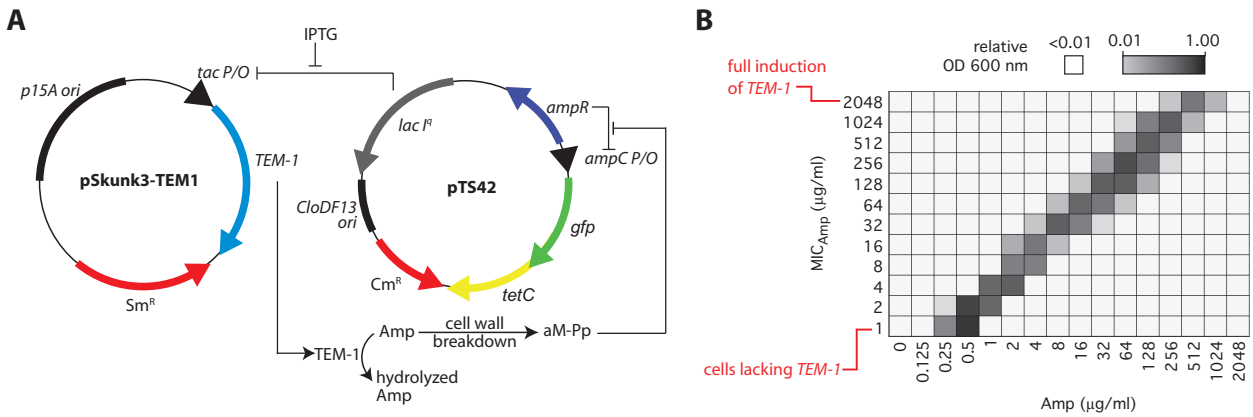


Fig. S1. Bacterial band-pass filter for beta-lactamase activity. (A) Essential components of the genetic circuit for band-pass selection (Sohka, et al. 2009). In the absence of sufficient cellular TEM-1 β-lactamase activity for hydrolysis of ampicillin (Amp), cell wall synthesis is compromised and cells cannot proliferate. Cell wall breakdown results in the accumulation of aM-pentapeptide (aM-Pp), which induces the *ampC* promoter via interactions with AmpR (Dietz, et al. 1997; Valtonen, et al. 2002) resulting in the production of TetC (which confers tetracycline resistance) and the green fluorescent protein (GFP). The level of Amp necessary to induce *ampC* is lower than the level that prevents the growth of *E. coli* cells (Valtonen, et al. 2002). Thus, cells that hydrolyze Amp too efficiently cannot grow in the presence of Tet. As a result, in the presence of tetracycline and Amp, cells will proliferate only if they possess an intermediate amount of Amp hydrolysis activity. The level of Amp hydrolysis activity necessary for growth increases linearly with Amp concentration (Valtonen, et al. 2002). *TEM-1* expression is regulated through IPTG-induction of the LacI-repressed *tac* promoter. (B) Demonstration of band-pass selection for BLA activity in *E. coli* SNO301 cells (Sohka, et al. 2009). The growth of cells in liquid LB media containing 20 µg/ml Tet was detected by measuring the OD at 600 nm of the liquid culture and is presented in the form of a heat map as a function of Amp concentration (x-axis) for cells with different levels of β-lactamase activity (y-axis, as measured by the minimum inhibitory concentration (MIC) for Amp).

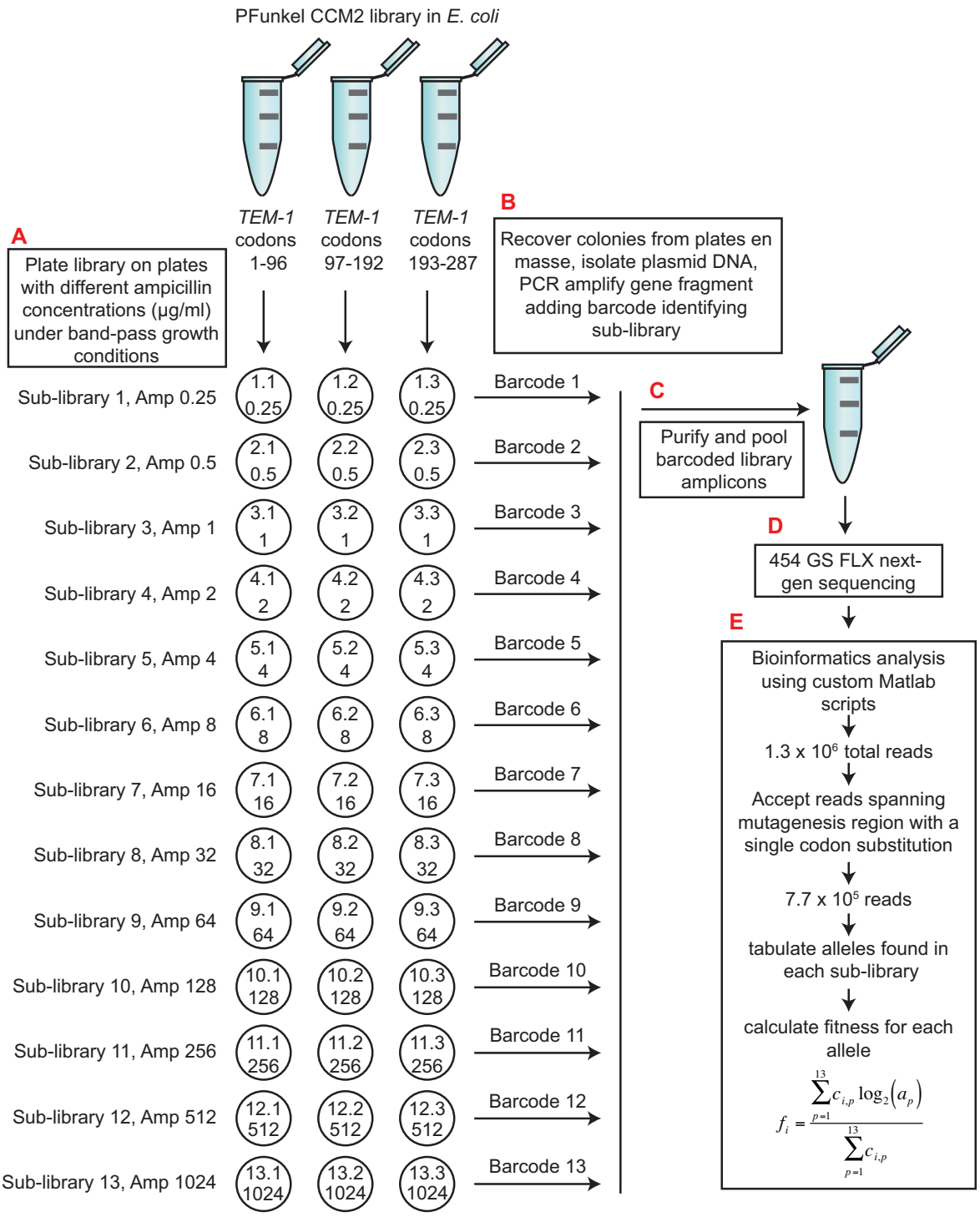


Fig. S2. System for measuring gene fitness of *TEM-1* alleles. The CCM2 library in strain SNO301 (separate libraries for each third of the *TEM-1* gene) was plated on media containing Tet and different concentrations of Amp. PCR products with barcodes identifying the plating conditions were subjected to 454 GS FLX DNA sequencing and the counts of alleles on each of the growth conditions was used to quantify fitness (see Methods).

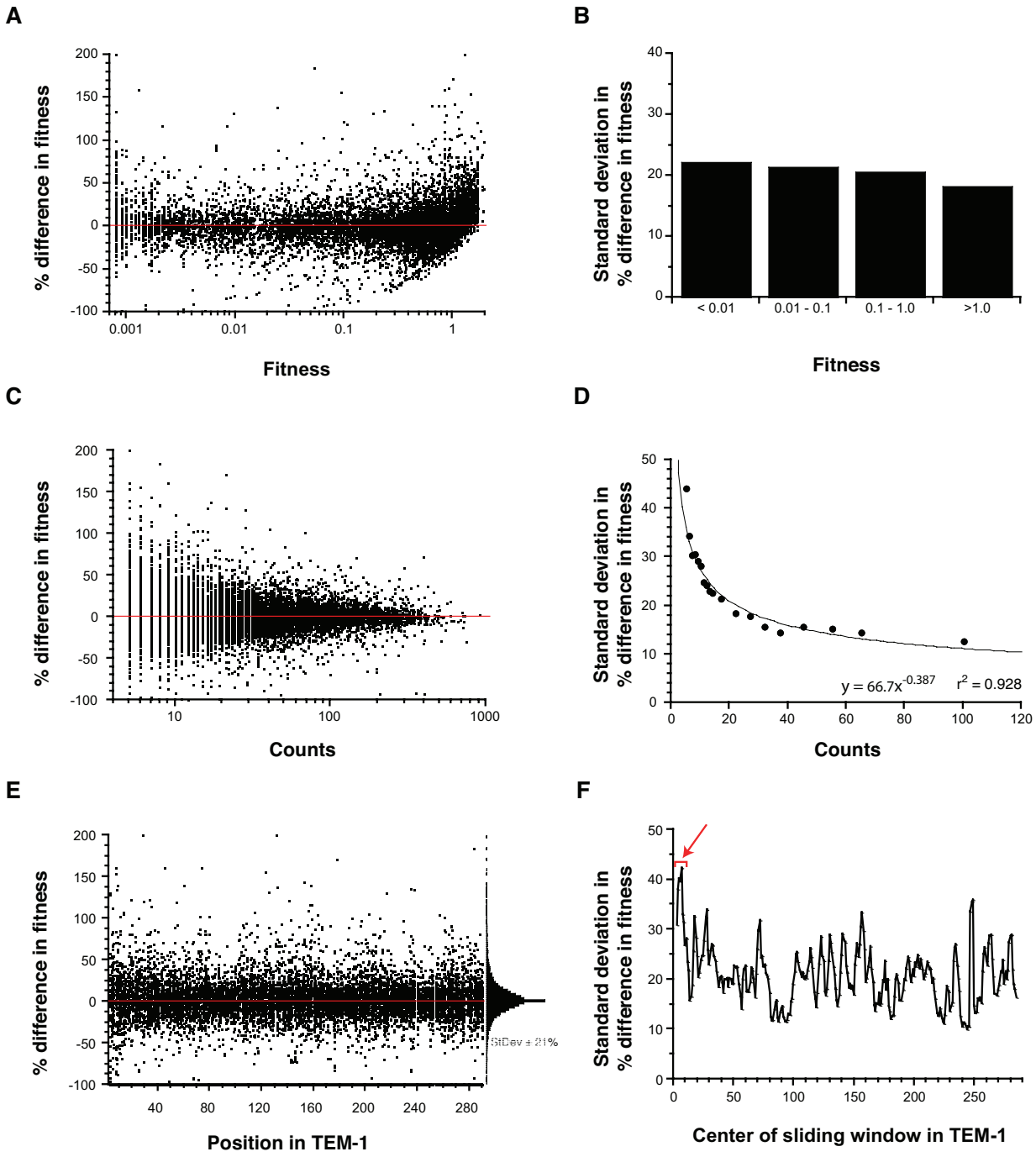


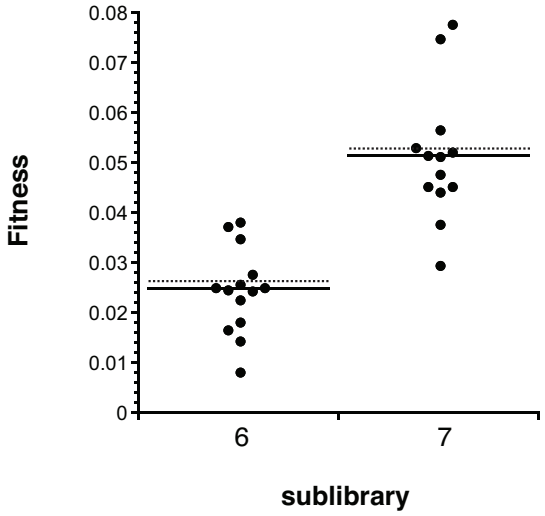
Fig. S3. Distribution of synonymous effects. We determined the percent difference in gene fitness between the fitness of allele *i* with a mutation at codon *j* of the gene and the mean fitness of all alleles with a synonymous mutation at codon *j* (including allele *i*). (A) The distribution of percent fitness difference as a function of fitness indicates that the fitness measurement is equally precise at low and high fitness values. (B) This observation is further illustrated by the standard deviation of percent fitness differences for different fitness ranges. (C) The percent fitness difference decreases with the number of times an allele is encountered in the deep sequencing experiment (i.e. the

counts). (D) The standard deviation in percent fitness difference as a function of counts, defined here, was used as an upper limit of error in the fitness measurements. (E) The percent fitness difference is fairly uniform across the *TEM-1* sequence. (F) However, a broader distribution in the first ten codons of the gene is apparent. A sliding window of three positions was used.

A

Sub-library 6	Sub-library 7
I13E	M68S
E37H	D85R
Y46A	L91R
D50*	I127A
K55P	N136Y
D85I	N136F
I127S	L137N
L137W	L138T
P145Y	G144R
D176R	P145L
T180Q	D176V
T180N	T180I
T181V	R244E
W210S	

B



C

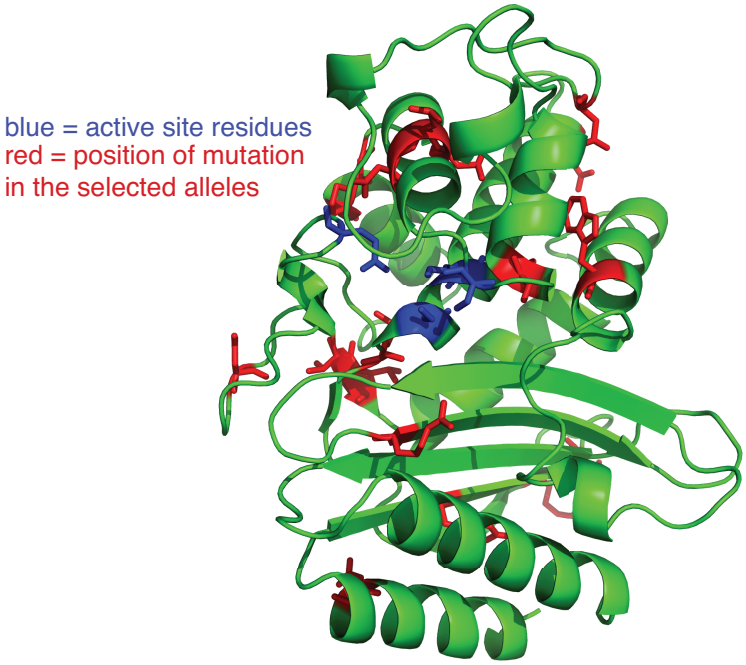


Fig. S4. Randomly selected members of sub-libraries 6 and 7. These 27 members were selected on plates with 8 μg/ml Amp (sub-library 6) or 16 Amp (sub-library 7). (A) Mutations in the 27 alleles (each allele had one mutation). * indicates the mutation is to the amber stop codon (UAG). (B) Gene fitness values. The solid line indicates the mean of the randomly selected members. The dotted line indicates the expected mean of the sub-library based on the Amp concentration used in the genetic selection to obtain the sub-library. (C) The distribution of mutational sites on the structure of TEM-1 (Fonze, et al. 1995). Red indicates the mutational sites and blue indicates the four key active site residues (S70, K73, S130 and E166).

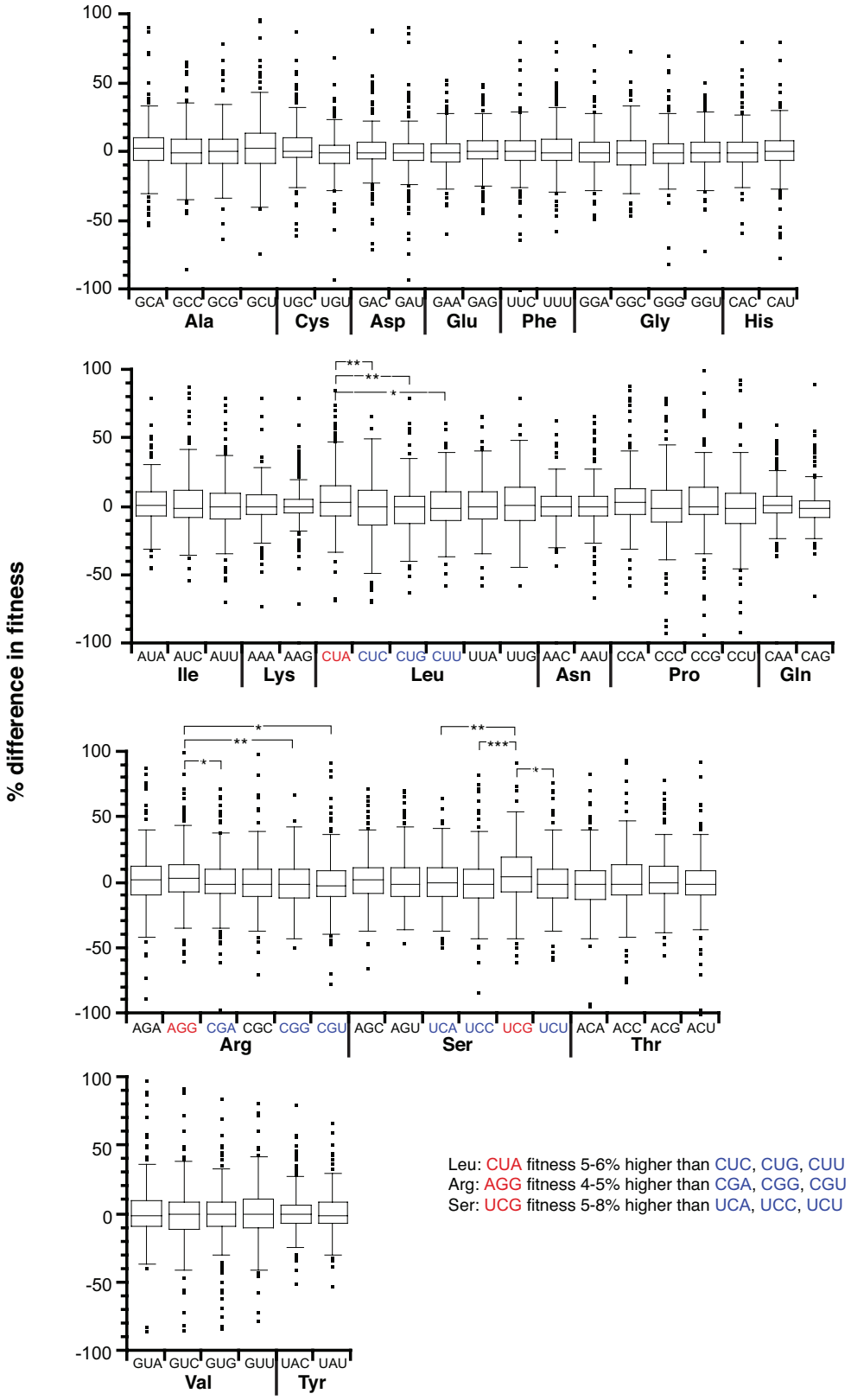


Fig. S5. (legend on next page)

Fig. S5. Global gene fitness effects of codon usage in *TEM-1*. The percent fitness differences of 14,055 synonymous substitutions among the 15,167 alleles with fitness measurements were analyzed as a function of the codon substituted. The global bias along the entire *TEM-1* gene for any particular codon, if there is any, is on the order of 5% or less. Our analysis for global codon bias is more likely to be able to identify smaller differences in mean fitness for codon sets with a greater number of codons. Thus, although we identify about 5% fitness differences between certain codons for Leu, Arg, and Ser, the fact that these are the three amino acids with six codons made the identification of significant but small differences in these codon sets more likely. Equal differences may or may not exist within other codon sets but we lack the statistical significance to identify them. *P* values were determined by Student's *t*-test.

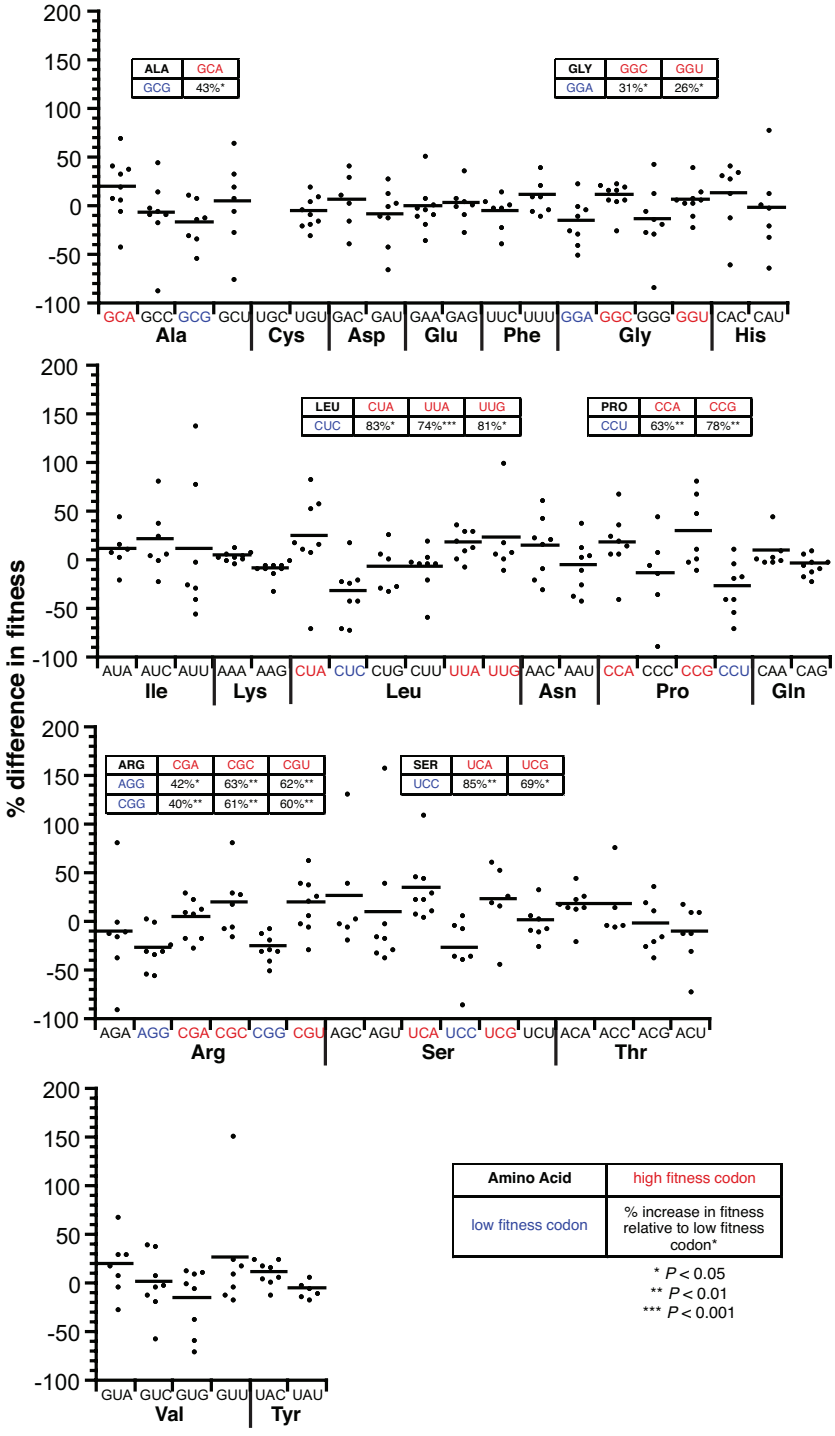


Fig. S6. Gene fitness effects of codon usage at positions 2-10 in *TEM-1*. The percent fitness differences of synonymous substitutions within positions 2-10 in *TEM-1* were analyzed as a function of the codon substituted. *P* values were determined by Student's *t*-test.

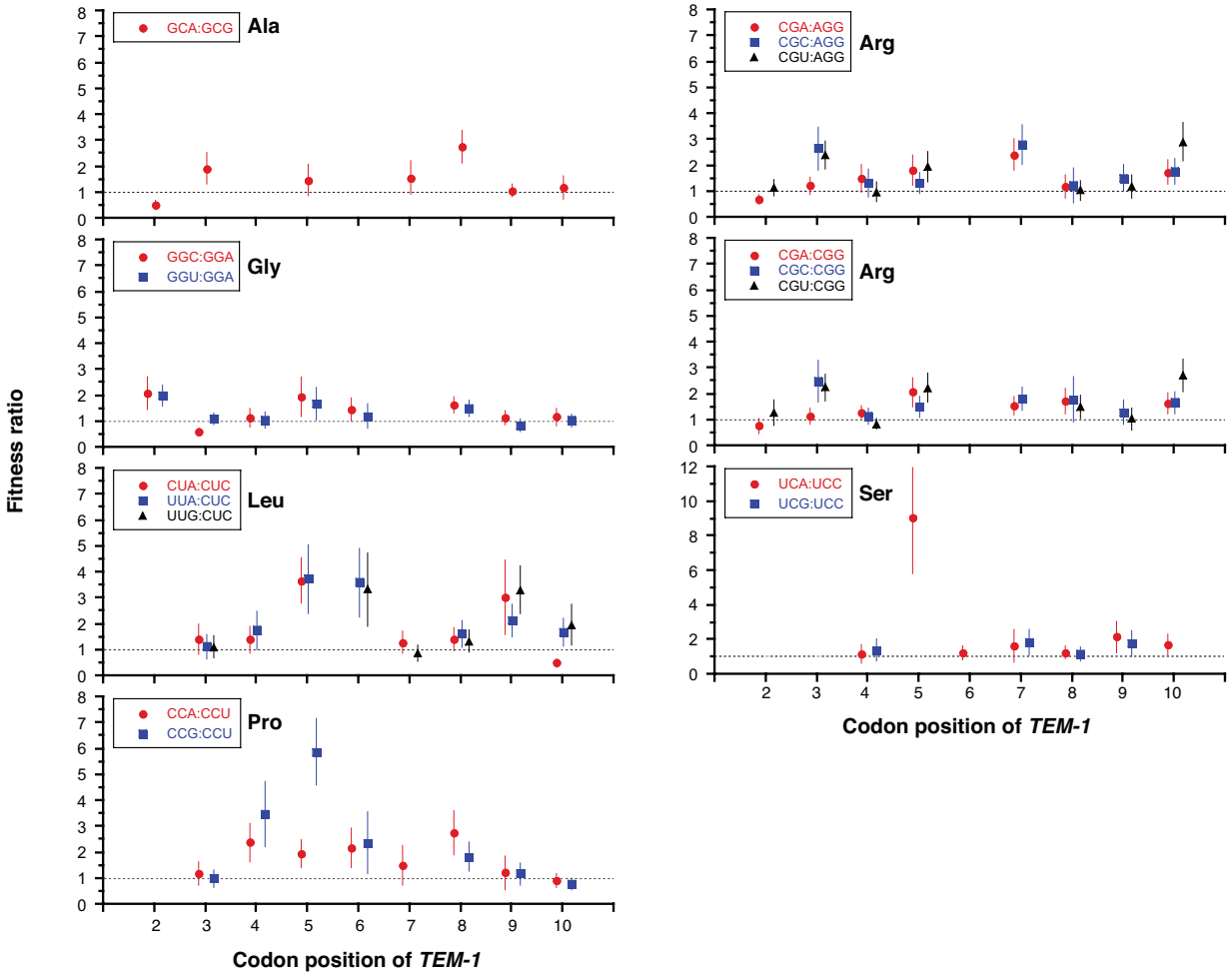


Fig. S7. Positional dependence of synonymous fitness effects at positions 2-10 in *TEM-1*. For the synonymous codon pairs determined to have significant differences in fitness within positions 2-10 of the gene (Fig. S6), the ratio of fitnesses for the two alleles is shown as a function of position in the gene.

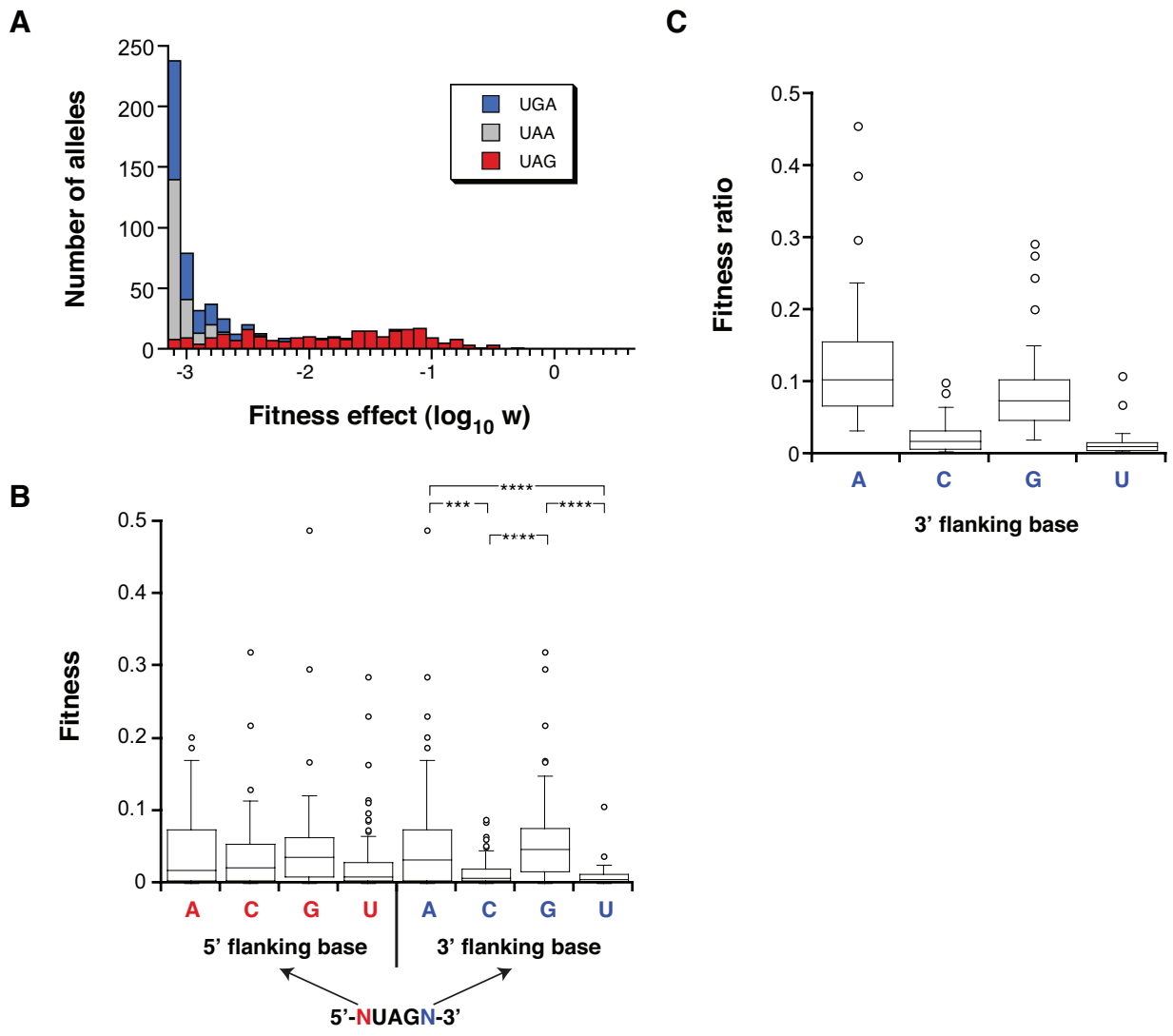


Fig. S8. Gene fitness effects of nonsense mutations in *TEM-1*. (A) The DFE of nonsense mutations as a function of the three nonsense codons. Gene fitness values are presented on a log scale with 0 corresponding to the fitness of *TEM-1*. (B) The efficiency of nonsense suppression at UAG is higher if an A or a G is at the 3' flanking position. *** $P = 0.0002$, **** $P < 0.0001$ by Wilcoxon–Mann–Whitney test. (C) The efficiency of nonsense suppression at UAG is strongly determined by the 3' flanking nucleotide. The fitness ratio compares the fitness of an allele with a mutation to UAG to the fitness of an allele with a missense mutation to glutamine at the same position. Only glutamine missense mutations with $w > 0.25$ were considered. The median efficiencies were 10.4% (3' A) 1.8% (3' C), 7.5% (3' G) and 1.1% (3' U). Differences between A/G and C/U have a P value of <0.0001 (Wilcoxon–Mann–Whitney test).

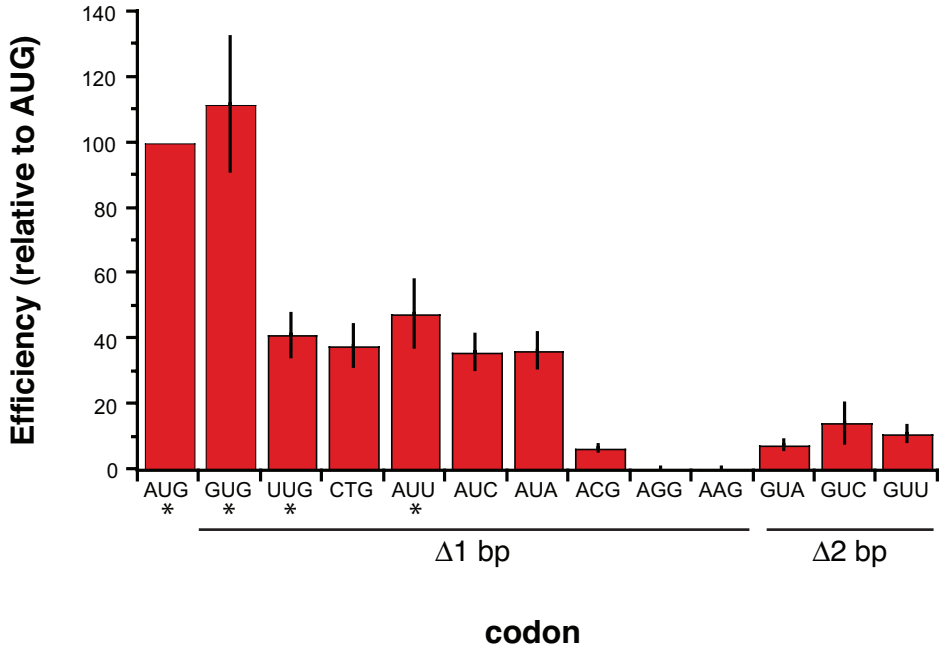


Fig. S9. Relative efficiency at which select codons serve as initiation codons. The efficiency was determined by dividing the gene fitness of the allele with the indicated codon at position 1 in the gene by the fitness of *TEM-1* (i.e. with AUG in position 1). Asterisks indicate known native initiation codons in the *E. coli* genome. All codons that differ from AUG by 1 bp are shown as are the three codons that differ by more than 1 bp that exhibited >1% efficiency.

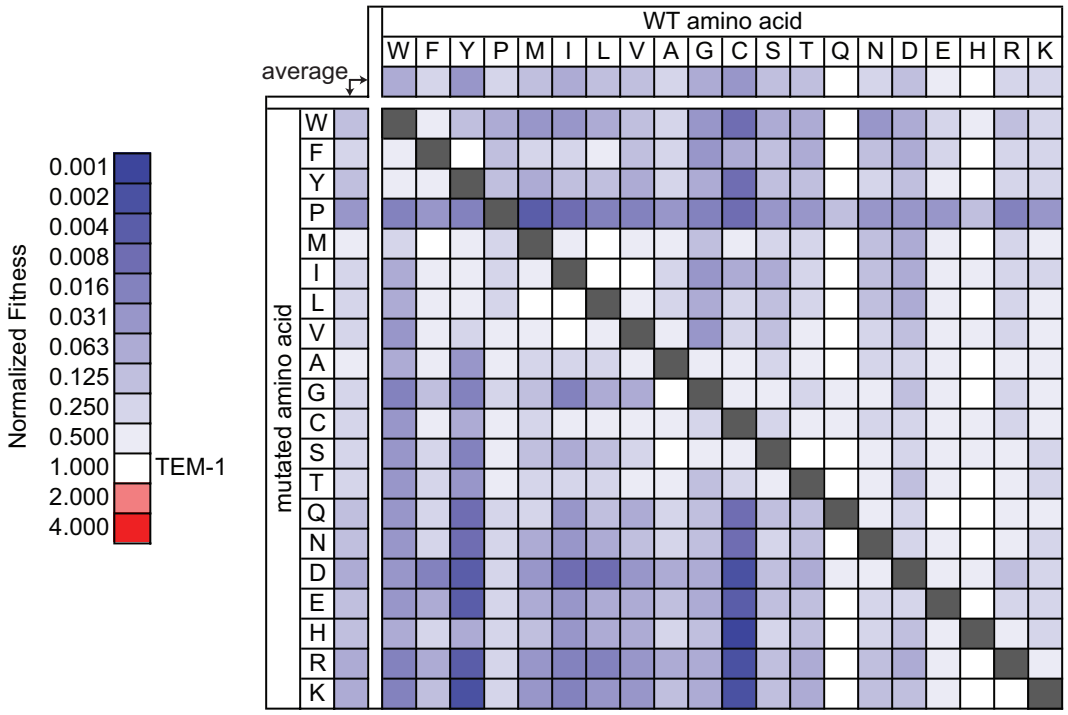


Fig. S10. Amino acid substitution matrix for TEM-1. The heat map indicates the average protein fitness for the indicated substitution. Tabulated data for this heat map is provided in Data S3.

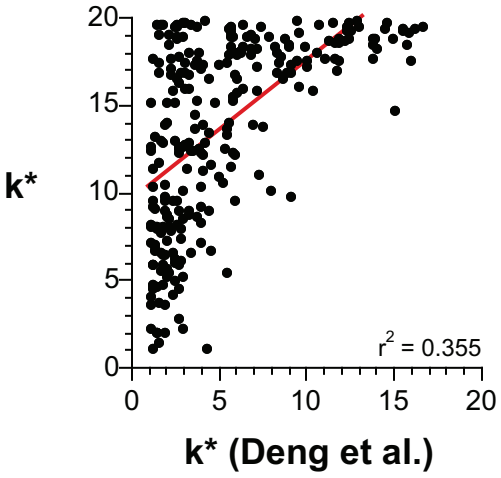


Fig. S11. Comparison of k* of this study with that determined from *TEM-1* alleles with multiple mutations by Deng et al. (Deng, et al. 2012).

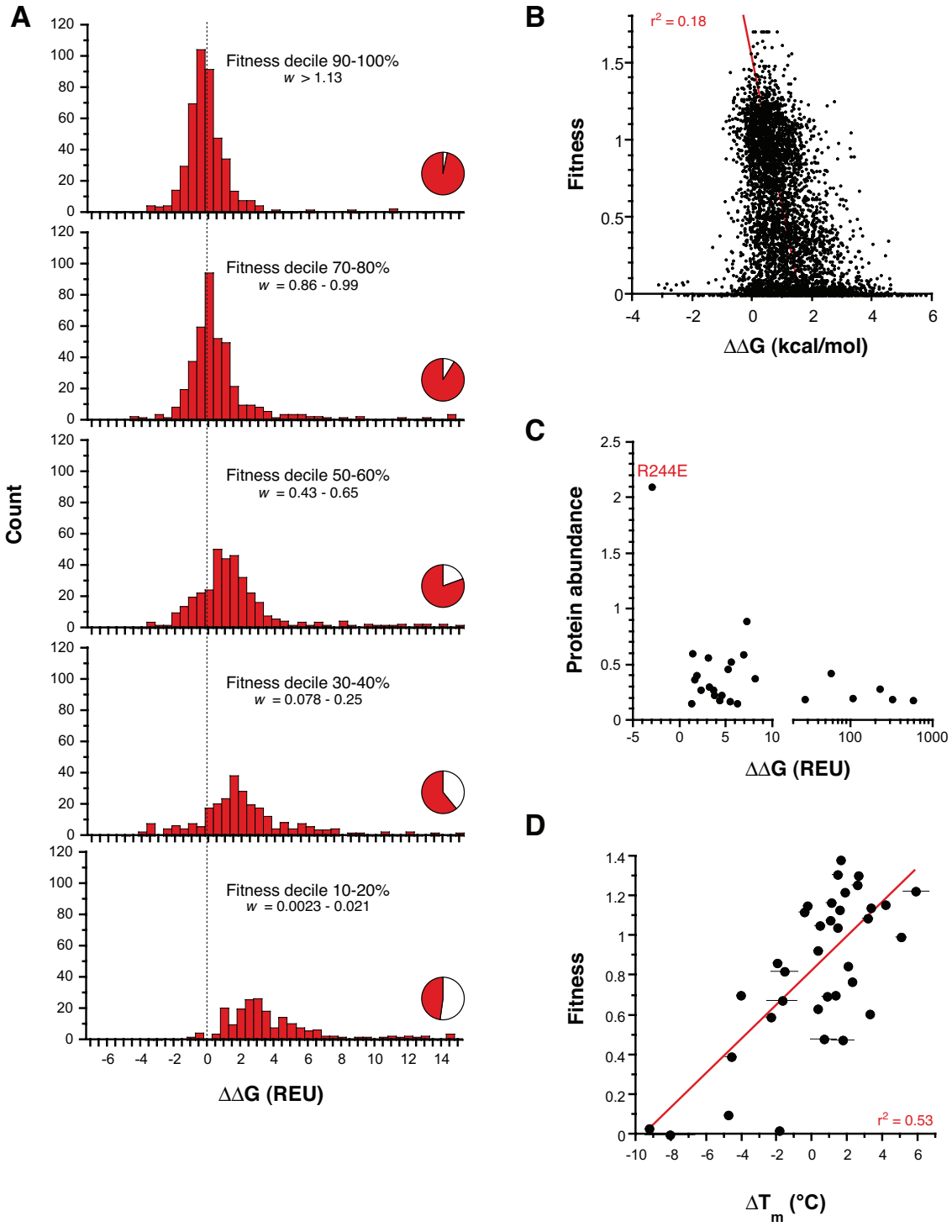


Fig. S12. The correlation between protein fitness and protein stability. (A) Distribution of predicted ΔΔG (by Rosetta) for select fitness deciles of 4783 missense mutations of TEM-1 (i.e. the data of Fig. 5A). In general, proteins with reduced fitness are predicted to have decreased stability. The fitness decile (i.e. 90-100% indicates the fittest 10%)

and the decile's corresponding fitness range are indicated. The pie graphs indicate in red the fraction of ΔΔG values for a decile that are <15 Rosetta energy units (REU). Limitations in the accurate prediction of $\Delta\Delta G$ (Potapov, et al. 2009) including the constraints on backbone movement contribute to the high $\Delta\Delta G$ values of some variants. (B) Protein fitness is shown as a function of change in ΔG as predicted by PoPMuSiC (Dehouck, et al. 2011) for 4783 missense mutations of TEM-1. (C) Protein abundance as a function of predicted ΔΔG (by Rosetta) for 26 randomly selected alleles with low fitness (i.e. the alleles of Fig. S4 with the exception of I13E). Protein abundance is expressed relative to TEM-1. All alleles were predicted to lose thermodynamic stability, the exception being R244E, which had a 2.1 higher protein abundance than TEM-1 that is likely the result of the increase stability. (D) Protein fitness as a function of experimentally measured change in the melting temperature (T_m). We compared the fitness and melting temperature (T_m) of 36 TEM-1 mutants (Bershtein, et al. 2008; Brown, et al. 2010; Deng, et al. 2012; Guillaume, et al. 1997; Kather, et al. 2008; Raquet, et al. 1995; Wang, et al. 2002a, b). The least fit sequences tended to be those few sequences that had lost more than 4 °C in T_m. The dataset is biased towards stabilizing mutations obtained by functional selection that generally had a <2-fold effect on fitness.

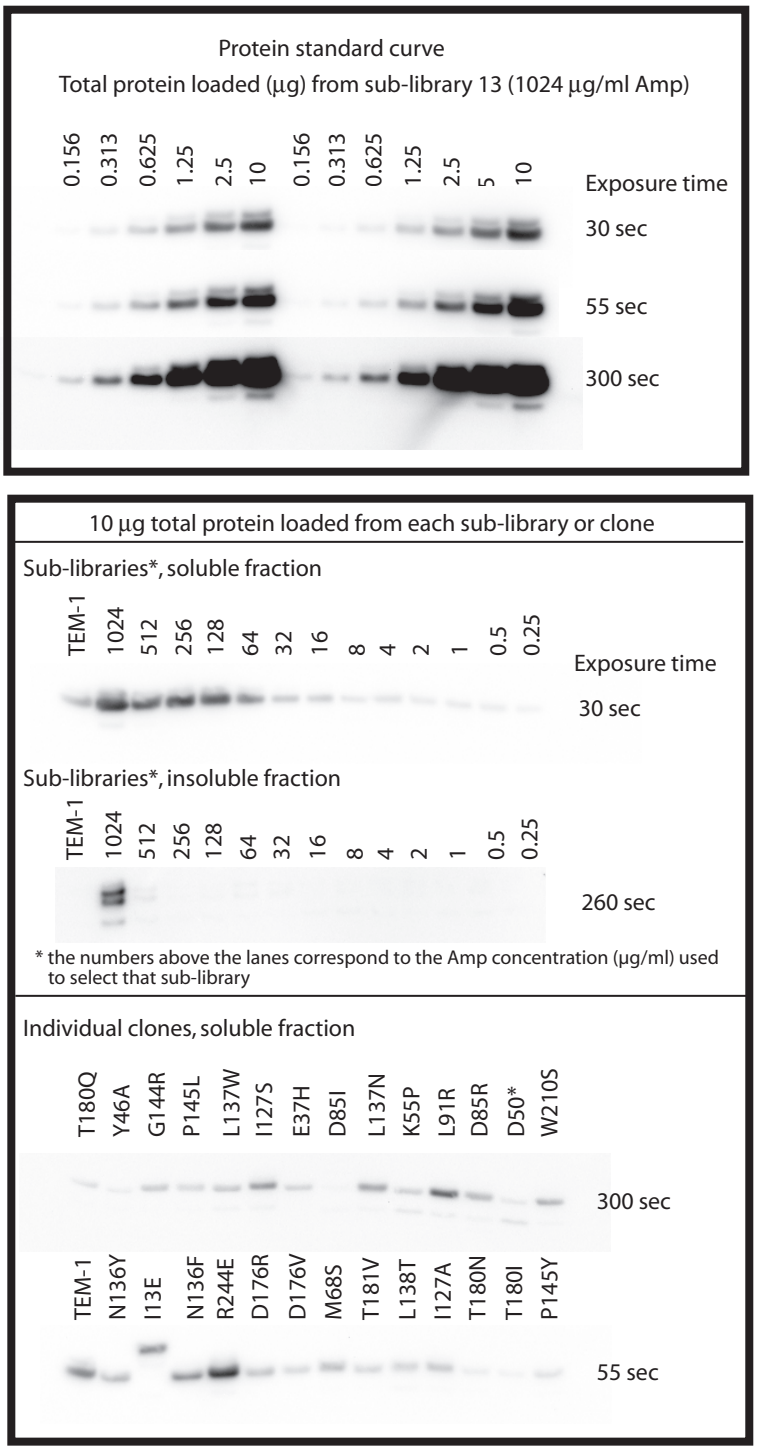


Fig. S13. Representative western blots from protein abundance quantification. The I13E allele (a mutation in the signal sequence) has a higher molecular weight band that corresponds to the size of the protein if the signal sequence has not been removed.

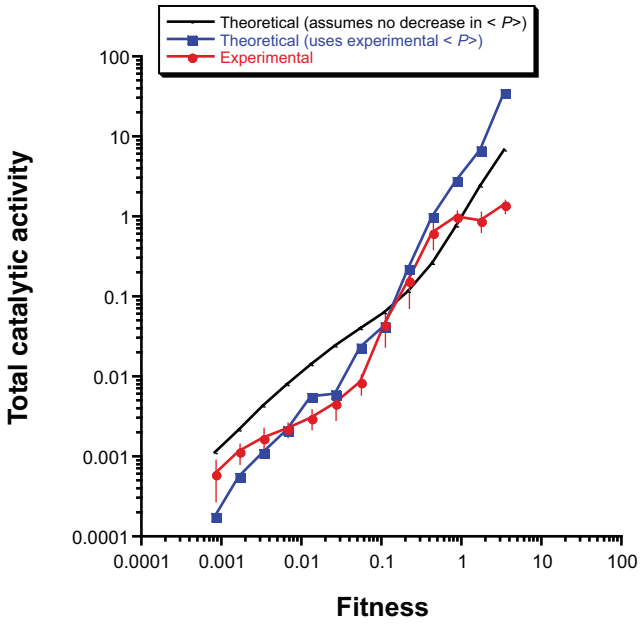


Fig. S14. Expected relationship between protein fitness for Amp resistance and total cellular catalytic activity as measured by nitrocefin hydrolysis. The slight sigmoidal relationship is illustrated by the theoretical calculation that assumes that protein abundance does not change with fitness (black line). By instead using the actual measures of protein abundance (blue squares), the inflection point shifts to lower values and the theoretical curve more closely matches the experimental data (red circles). The calculations generating both theoretical curves are described in Extended Experimental Procedures.

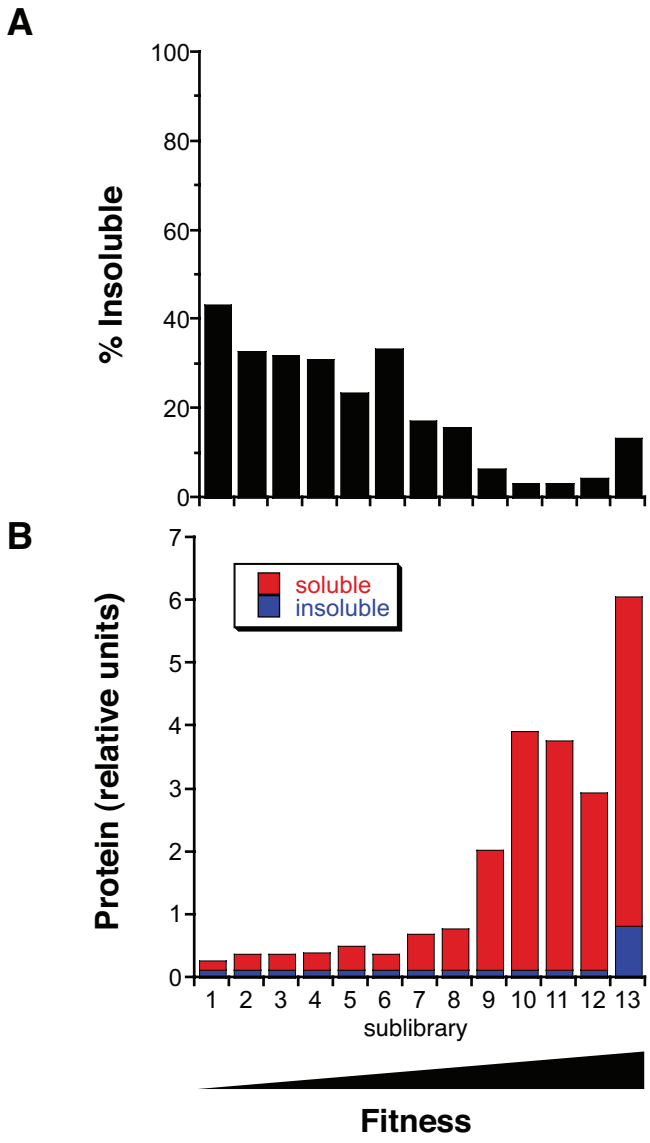


Fig. S15. A decrease in protein fitness is not accompanied by an increase in insoluble TEM-1. (A) The percentage of TEM-1 protein in an insoluble state increased as fitness decreased. (B) However, this trend results from a decrease in the soluble TEM-1 and not from an increase in insoluble TEM-1. A representative western blot of the insoluble fraction is shown in Fig. S13.