

SUPPLEMENTARY MATERIALS FOR

**Impact of translational error-induced and error-free misfolding on the rate of  
protein evolution**

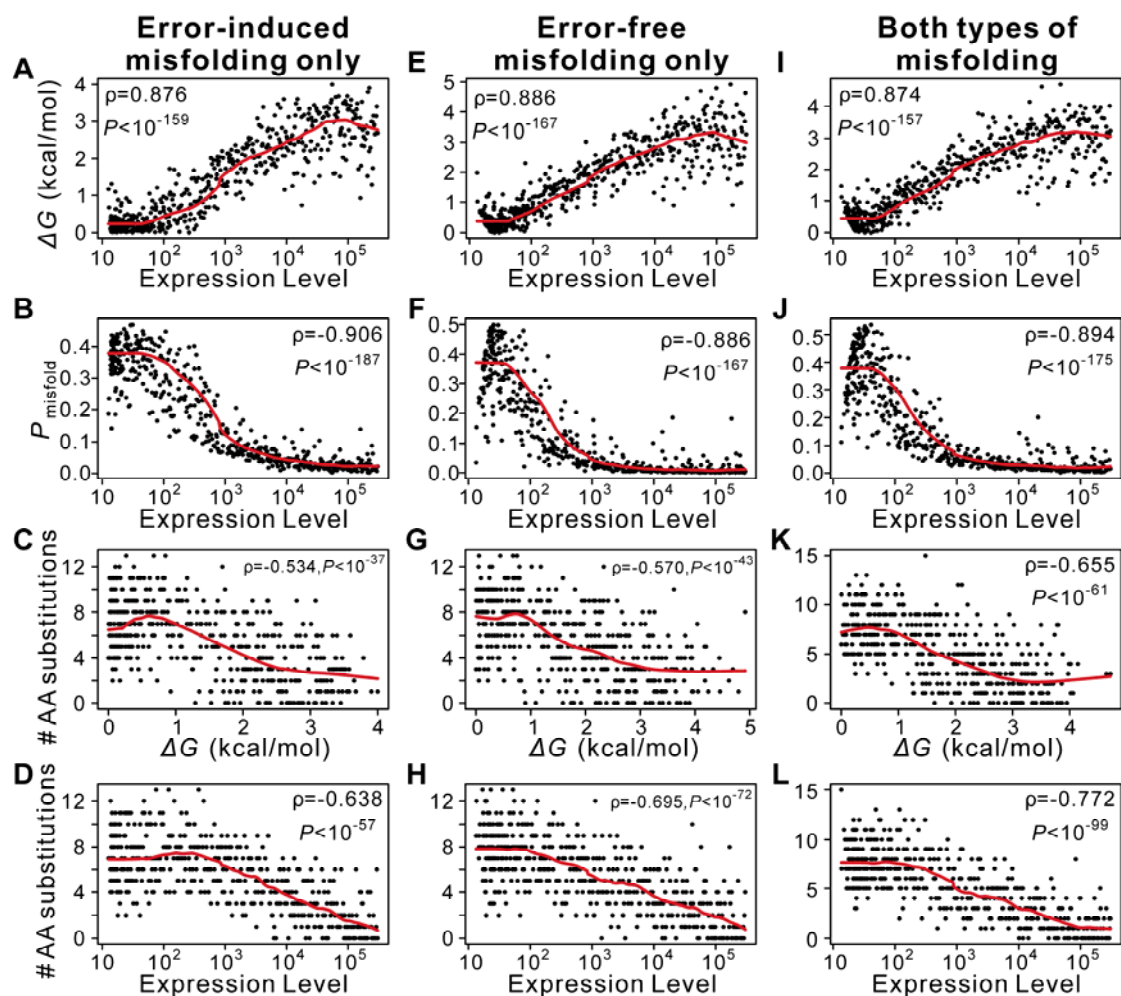
Jian-Rong Yang, Shi-Mei Zhuang, and Jianzhi Zhang

Supplementary materials include:

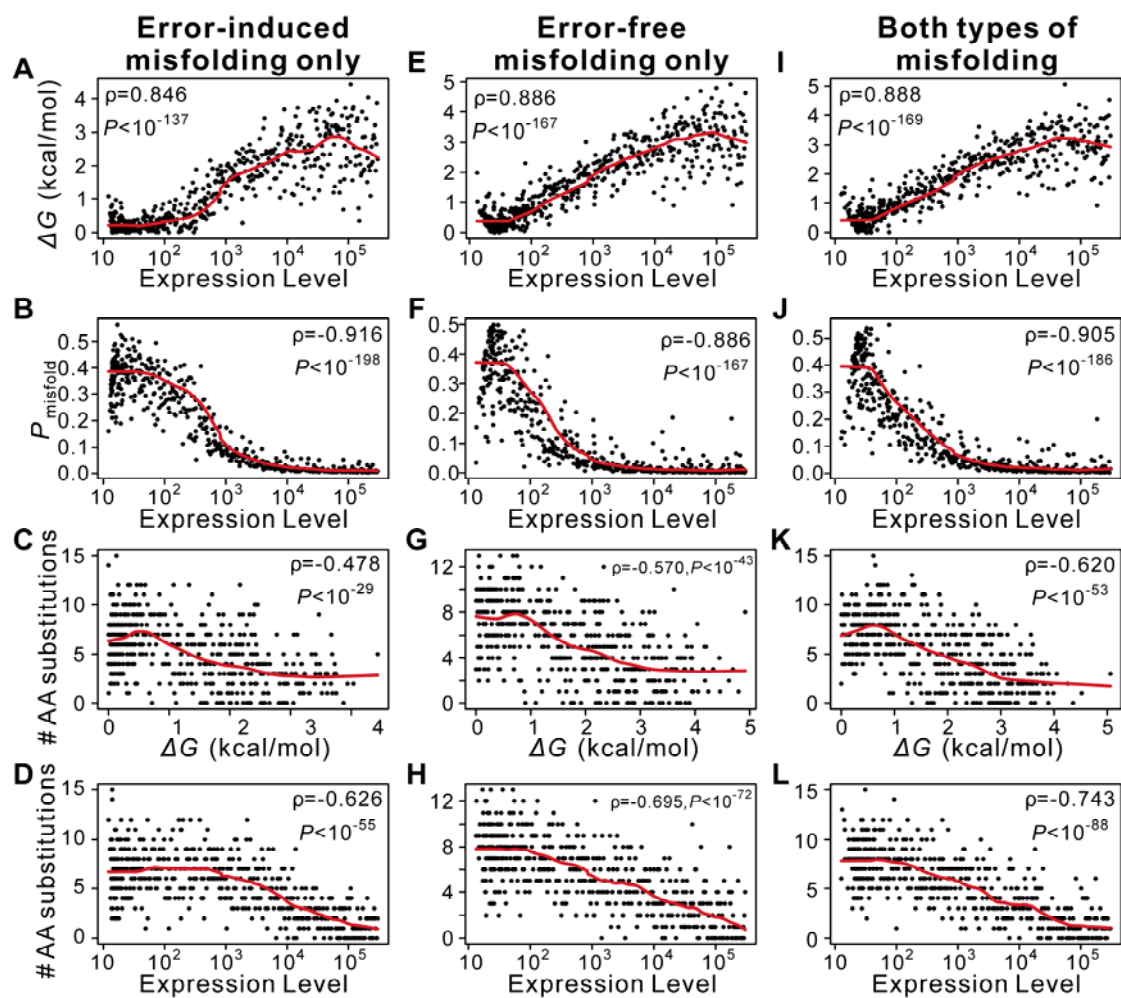
Table S1

Fig. S1-Fig. S8

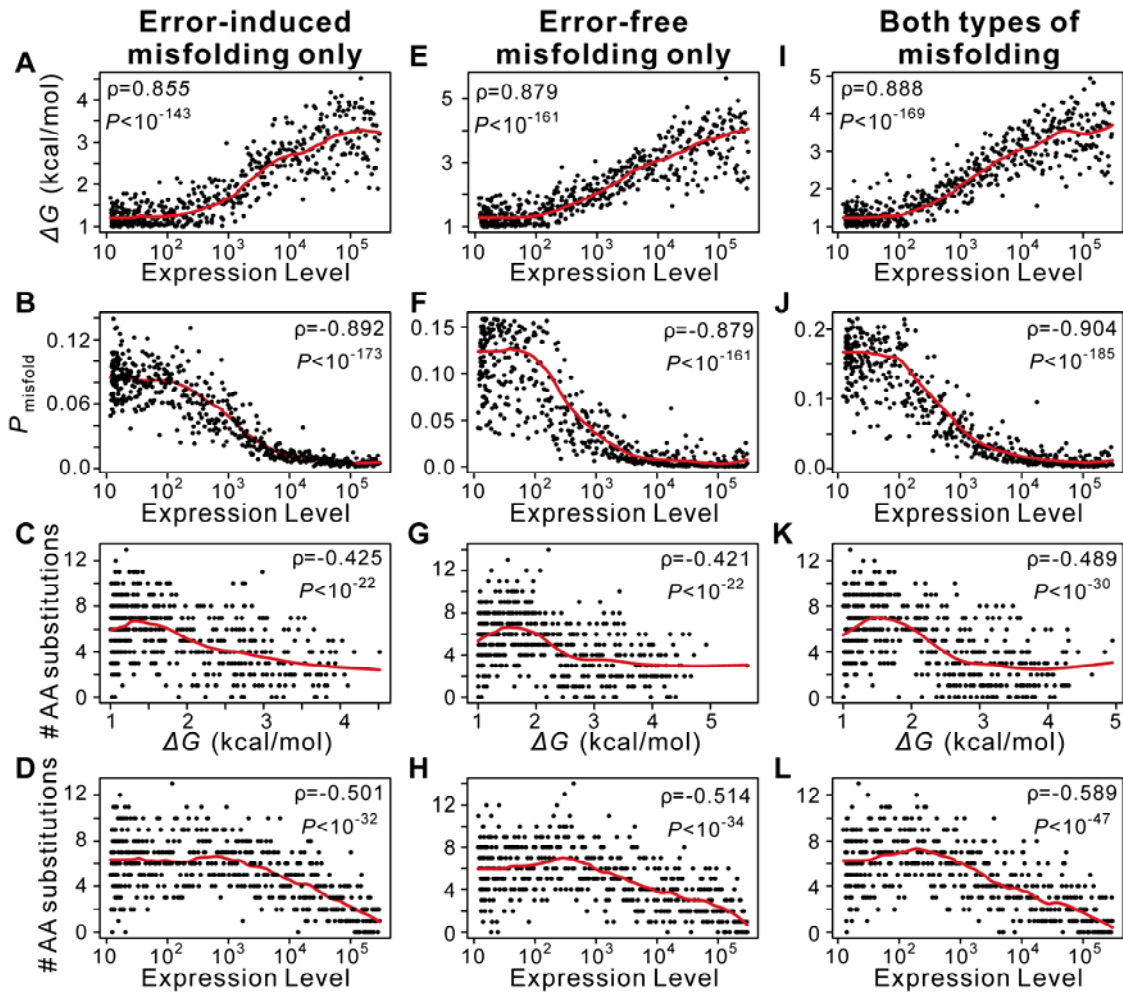




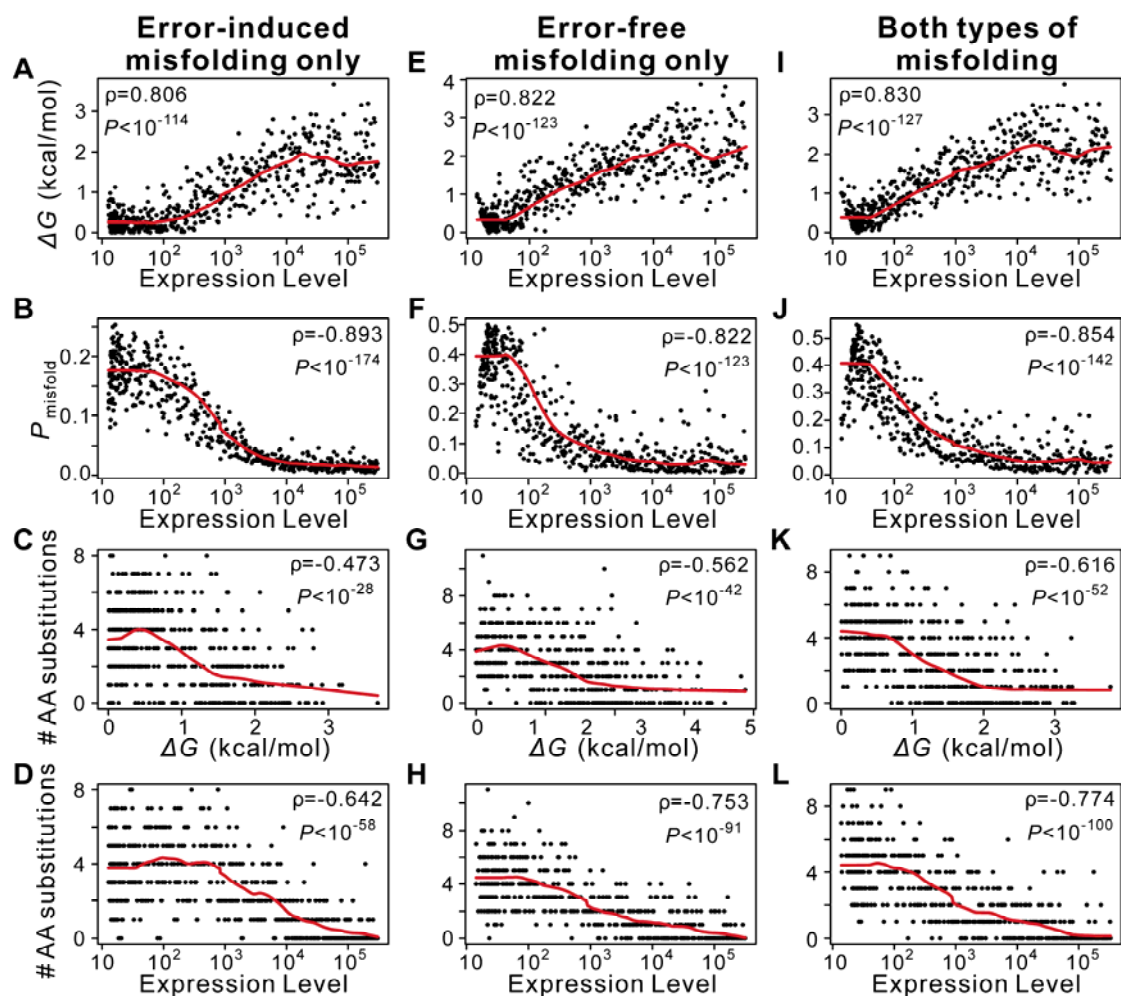
**Fig. S1.** A molecular-level evolutionary simulation for examining the roles of error-induced and error-free misfolding in generating the anticorrelation between protein expression level and evolutionary rate. All are the same as in Fig. 2, except that the mistranslation rate ratio between preferred and non-preferred codons is set at 0.5 (0.2 in Fig. 2).



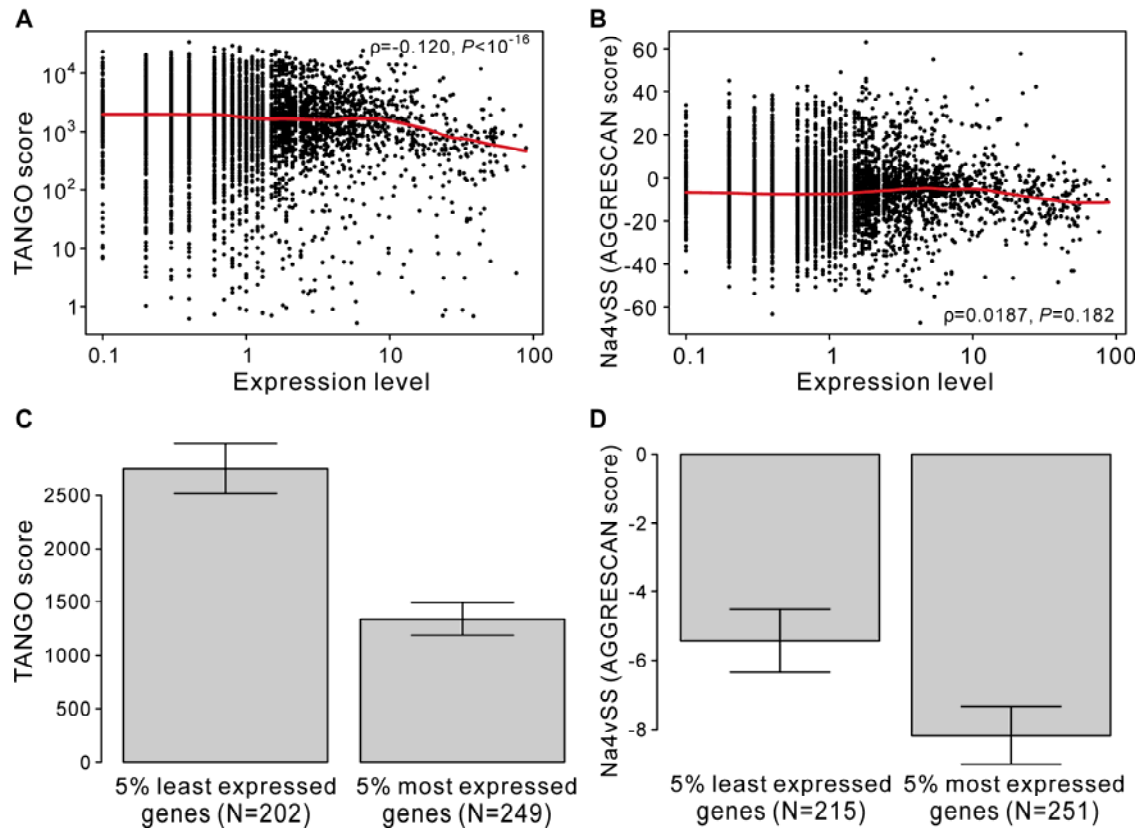
**Fig. S2.** A molecular-level evolutionary simulation for examining the roles of error-induced and error-free misfolding in generating the anticorrelation between protein expression level and evolutionary rate. All are the same as in Fig. 2, except that the mistranslation rate ratio between preferred and non-preferred codons is set at 0.1 in the simulations (0.2 in Fig. 2).



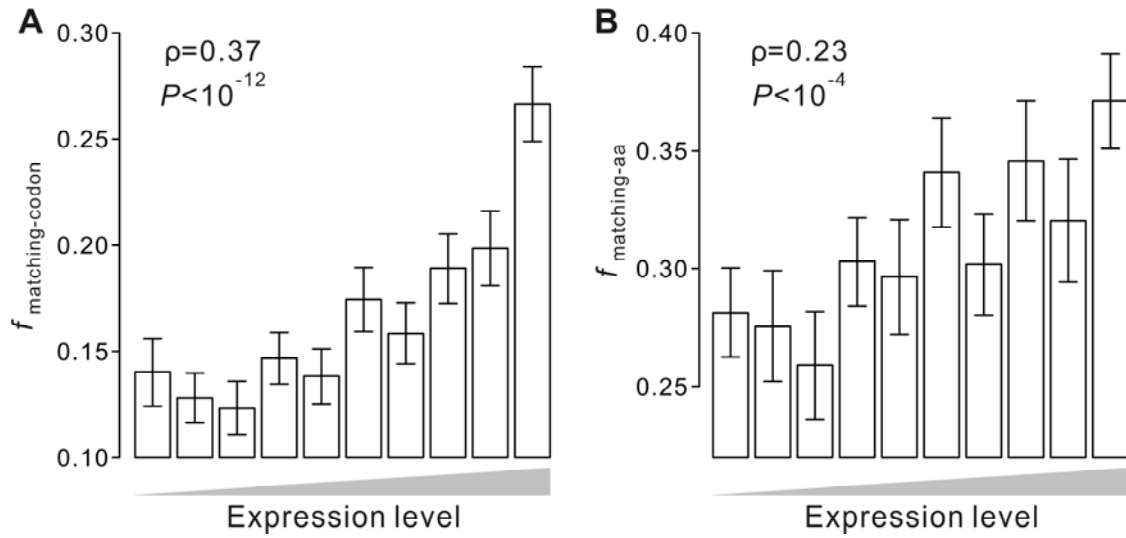
**Fig. S3.** A molecular-level evolutionary simulation for examining the roles of error-induced and error-free misfolding in generating the anticorrelation between protein expression level and evolutionary rate. All are the same as in Fig. 2, except that  $\Delta G_{\text{min}}$  is set at 1 in the simulations (0 in Fig. 2).



**Fig. S4.** A molecular-level evolutionary simulation for examining the roles of error-induced and error-free misfolding in generating the anticorrelation between protein expression level and evolutionary rate. All are the same as in Fig. 2, except that protein length is shortened to 16 amino acids in the simulations (25 in Fig. 2).

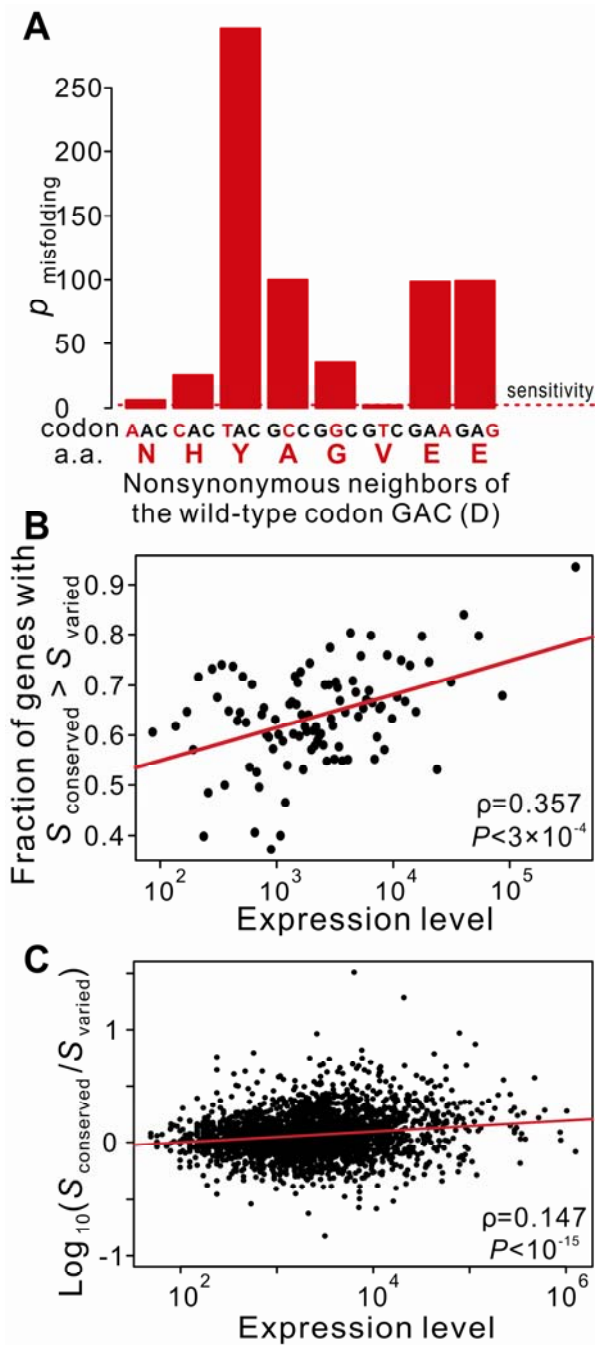


**Fig. S5.** Protein aggregation propensity decreases with the rise of gene expression level in yeast. Aggregation propensity is predicted from protein sequences using TANGO (A and C) or AGGRESCAN (B and D). In both algorithms, higher scores indicate higher aggregation propensities. (A) Relationship between gene expression level and protein aggregation propensity predicted by TANGO. (B) Relationship between gene expression level and protein aggregation propensity predicted by AGGRESCAN. (C) Protein aggregation propensity predicted by TANGO is significantly different between the 5% most expressed and the 5% least expressed genes ( $P < 10^{-6}$ , Mann-Whitney test). (D) Protein aggregation propensity predicted by AGGRESCAN is significantly different between the 5% most expressed and the 5% least expressed genes ( $P = 0.027$ , Mann-Whitney test). In (A) and (B), correlation coefficients and significance levels are determined by Spearman's rank correlation tests and the red lines are estimated using locally weighted scatterplot smoothing. Note that 40 genes with negative TANGO scores are not included in (A) because of the use of log-scaled Y-axis. In (C) and (D), means and standard errors are shown.

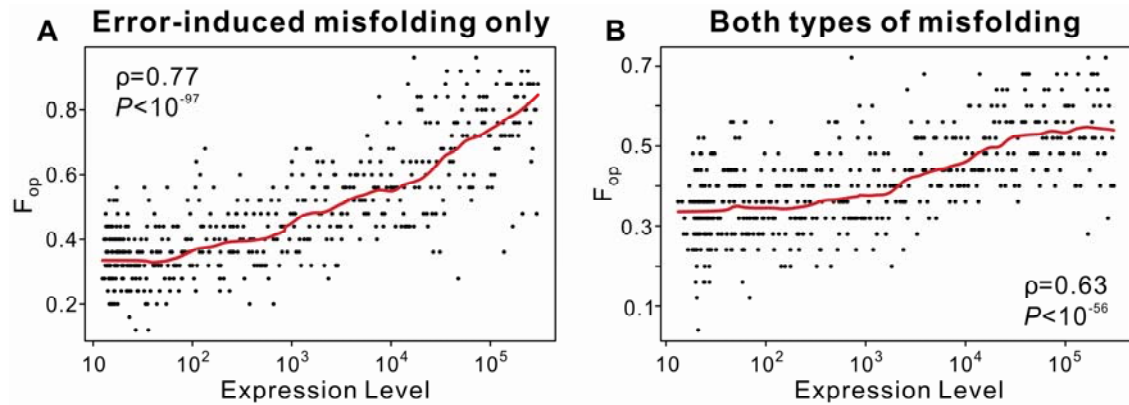


**Fig. S6.** Structure-based I-mutant predictions of  $\Delta\Delta G$  also show that codons minimizing the probability of protein misfolding are used more frequently in highly expressed yeast genes than in lowly expressed genes. **(A)** The fraction ( $f_{\text{matching-codon}}$ ) of wild-type codons in a gene that match the codons with the smallest  $p_{\text{misfold}}$  increases with the protein expression level. **(B)** The fraction ( $f_{\text{matching-aa}}$ ) of wild-type amino acids that match the amino acids encoded by the codons with the smallest  $p_{\text{misfold}}$  increases with the protein expression level. In both panels, genes are divided into 10 equal-size bins. Error bars indicate standard errors. Correlations and  $P$ -values are estimated from unbinned data, using Spearman's rank correlation tests.





**Fig. S7.** Evolutionary conservation of amino acid residues correlates with the mutational sensitivity to misfolding, when sensitivity is redefined by the minimal  $p_{\text{misfold}}$  of a codon's one-nonsynonymous-mutation neighbors. Proteins with at least 3 varied sites are considered. (A) An example (codon #58 of YAL001C, which encodes a subunit of the RNA polymerase III transcription initiation factor complex) showing the measurement of the mutational sensitivity ( $S$ ) of a codon, indicated by the dotted red line. Here,  $p_{\text{misfold}}$  is the protein misfolding probability of a mutant relative to that of the wild-type gene. The nucleotide differences from the wild-type as well as the altered amino acids are colored in red. (B) Fraction of genes with  $S_{\text{conserved}} > S_{\text{varied}}$  increases significantly with expression level. Here,  $S_{\text{conserved}}$  and  $S_{\text{varied}}$  are the mean  $S$  values for codons with conserved and varied amino acids between *S. cerevisiae* and *S. paradoxus* orthologs, respectively. The genes are grouped into 100 equal-size bins according to the yeast protein expression level. (C) The ratio of  $S_{\text{conserved}}$  and  $S_{\text{varied}}$  within a gene is positively correlated with its expression level. In (B) and (C), correlation coefficients and significance levels are determined by Spearman's rank correlation tests.



**Fig. S8.** The fraction of preferred synonymous codons ( $F_{op}$ ) in the gene sequences at the 100,000th generation of the simulation when the sequences are generated under (A) the translational robustness hypothesis (error-induced misfolding only) or (B) under the overarching misfolding-avoidance hypothesis (both types of misfolding). Correlation coefficients and significance levels are determined by Spearman's rank correlation tests. The red lines are estimated using locally weighted scatterplot smoothing.