Supporting Information

Shah and Gilchrist 10.1073/pnas.1016719108

SI Text

S1. Analytical Solutions of the Model. ***One amino acid with two codons.*** Consider a gene sequence of length n composed of a single two-codon amino acid, whose average elongation times are t_1 and t_2. Let x_1 and $x_2 = n - x_1$ be the respective codon counts. The expected cost of ribosome usage during protein production is then given as

$$\eta(\vec{x}) = C\sum_{i=1}^{2} x_i t_i, \qquad [S1]$$

$$= C(x_1 t_1 + x_2 t_2), \qquad [S2]$$

where C is the cost of ribosome usage in ATP per second. We assume an exponential fitness function w described as

$$w(\vec{x}|\phi) = e^{-q\phi\eta(\vec{x})} = e^{-q\phi C(x_1 t_1 + x_2 t_2)}, \qquad [S3]$$

where ϕ is the protein production rate, a measure of gene expression, and q is the scaling constant determining the relationship between cost of ATP usage to organismal fitness w.

Following the methods used in studies (1–4), the probability of observing an allele across the entire genotype space at equilibrium is given by

$$P(\vec{x}|\phi) = \frac{w(\vec{x}|\phi)^{N_e}}{\sum_{y\in S_c} w(\vec{y}|\phi)^{N_e}}, \qquad [S4]$$

where N_e is the effective population size and S_c is the entire synonymous codon genotype space, which has 2^n alleles in this simple case. Because the cost of protein production is independent of codon order within a gene, multiple synonymous alleles could give rise to the same cost η. In the case of two codons, the number of alleles with the same cost is represented by a binomial coefficient and for amino acids with more than two codons, the combinations will be represented by a multinomial coefficient

$$P(\vec{x}|\phi) = \frac{\binom{n}{x_1} e^{-N_e q\phi C(x_1 t_1 + x_2 t_2)}}{\sum_{y_1=0}^{n}\binom{n}{y_1} e^{-N_e q\phi C(y_1 t_1 + y_2 t_2)}}. \qquad [S5]$$

Let μ_1 and μ_2 represent the rate of mutations to the two codons, as described by Sella and Hirsh (4).

Taking mutational biases into account, the probability of observing a given allele is given as

$$P(\vec{x}|\phi) \propto w(\vec{x}|\phi)^{N_e}\prod_{i=1}^{2}\mu_i^{x_i}, \qquad [S6]$$

$$P(\vec{x}|\phi) = \frac{\binom{n}{x_1} e^{-N_e q\phi C(x_1 t_1 + x_2 t_2)}\prod_{i=1}^{2}\mu_i^{x_i}}{\sum_{y_1=0}^{n}\binom{n}{y_1} e^{-N_e q\phi C(y_1 t_1 + y_2 t_2)}\prod_{i=1}^{2}\mu_i^{y_i}}, \qquad [S7]$$

where $\vec{x} = \{x_1, x_2\}$.

Given the protein production rate ϕ (gene expression) of a gene and the elongation time t of codons, the expected count of each codon is given as

$$\mathbb{E}[x_1|\phi] = \sum_{x_1=0}^{n} x_1 P(\vec{x}|\phi), \qquad [S8]$$

$$= \sum_{x_1=0}^{n} x_1 \frac{\binom{n}{x_1} e^{-N_e q C\phi(x_1 t_1 + x_2 t_2)}\prod_{i=1}^{2}\mu_i^{x_i}}{\sum_{y_1=0}^{n}\binom{n}{y_1} e^{-N_e q C\phi(y_1 t_1 + y_2 t_2)}\prod_{i=1}^{2}\mu_i^{y_i}}, \qquad [S9]$$

$$= \frac{n\mu_1 e^{-N_e q C\phi t_1}}{\mu_1 e^{-N_e q C\phi t_1} + \mu_2 e^{-N_e q C\phi t_2}}, \qquad [S10]$$

and by symmetry

$$\mathbb{E}[x_2|\phi] = \frac{n\mu_2 e^{-N_e q C\phi t_1}}{\mu_1 e^{-N_e q C\phi t_1} + \mu_2 e^{-N_e q C\phi t_2}}, \qquad [S11]$$

$$= n - \mathbb{E}[x_1|\phi]. \qquad [S12]$$

One amino acid with k codons. Using the methods described above, it can be shown that for any amino acid with k codons, the expected count of the *ith* codon is given as

$$\mathbb{E}[x_i|\phi] = \frac{n\mu_i e^{-N_e q C\phi t_i}}{\sum_{j=1}^{k}\mu_j e^{-N_e q C\phi t_j}}. \qquad [S13]$$

Thus, the expected frequencies of each codon $f_i = x_i/n$ are given as

$$\mathbb{E}[f_i|\phi] = \frac{\mu_i e^{-N_e q C\phi t_i}}{\sum_{j=1}^{k}\mu_j e^{-N_e q C\phi t_j}}. \qquad [S14]$$

Variance around the expected value $\mathbb{E}[x_i|\phi]$ can also be calculated as

$$\mathrm{Var}[x_i|\phi] = \sum_{x_i=0}^{n}(x_i - \mathbb{E}[x_i|\phi])^2 P(\{x_1, x_2, \cdots, x_k\}), \qquad [S15]$$

$$= \frac{n\left(\prod_{j=1}^{k}\mu_j\right) e^{N_e q C\phi\sum_{j=1}^{k} t_j}}{\left(\sum_{j=1}^{k}\mu_j e^{N_e q C\phi t_j}\right)^2}. \qquad [S16]$$

Multiple amino acids with varying number of codons. In the case of real genes, which are composed of multiple amino acids, each with a varying number of codons, the expected counts and frequencies of codons can be estimated from the marginal distributions of each amino acid. For instance, consider the simple case of two amino acids with two codons each. The ribosomal overhead cost of protein production is given as

$$\eta(\vec{x}) = C(x_{11}t_{11} + x_{12}t_{12} + x_{21}t_{21} + x_{22}t_{22}), \qquad [S17]$$

where x_{ij} is the number of codons of type j of amino acid i in the gene. Let $n_1 = x_{11} + x_{12}$ and $n_2 = x_{21} + x_{22}$ be the counts of the two amino acids in the gene. As previously, the probability of observing an allele can be written as

$$P(\vec{x}|\phi) = \frac{\binom{n_1}{x_{11}}\binom{n_2}{x_{21}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12}+x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}}{\sum_{y_{11}=0}^{n_1}\sum_{y_{21}=0}^{n_2}\binom{n_1}{y_{11}}\binom{n_2}{y_{21}}\prod_{j=1}^{2}\mu_{1j}^{y_{1j}}\prod_{j=1}^{2}\mu_{2j}^{y_{2j}}e^{-N_e(y_{11}qC\phi t_{11}+y_{12}qC\phi t_{12}+y_{21}qC\phi t_{21}+y_{22}qC\phi t_{22})}}, \quad \text{[S18]}$$

$$= \frac{\binom{n_1}{x_{11}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12})}}{\sum_{y_{11}=0}^{n_1}\binom{n_1}{y_{11}}\prod_{j=1}^{2}\mu_{1j}^{x_{1j}}e^{-N_e(x_{11}qC\phi t_{11}+x_{12}qC\phi t_{12})}} \times \frac{\binom{n_2}{x_{21}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}}{\sum_{y_{21}=0}^{n_1}\binom{n_1}{y_{11}}\prod_{j=1}^{2}\mu_{2j}^{x_{2j}}e^{-N_e(x_{21}qC\phi t_{21}+x_{22}qC\phi t_{22})}}, \quad \text{[S19]}$$

$$= P(\vec{x}_1|aa_1)P(\vec{x}_2|aa_2). \quad \text{[S20]}$$

The marginal distribution of genotype space of a singe amino acid is given as

$$\sum_{x_{21}=0}^{n_2} P(\vec{x}_2|aa_2) = 1, \quad \text{[S21]}$$

$$P(\vec{x}_1|aa_1) = \sum_{x_{21}=0}^{n_2} P\left(\{\vec{x}_1,\vec{x}_2\}\right). \quad \text{[S22]}$$

Thus, the expected number of codons of a specific amino acid based on the marginal distribution of that amino acid can be calculated as

$$\mathbb{E}[x_{11}|\phi] = \sum_{x_{11}=0}^{n_1} x_{11}\sum_{x_{21}=0}^{n_2} P\left(\{\vec{x}_1,\vec{x}_2\}\right), \quad \text{[S23]}$$

$$= \sum_{x_{11}=0}^{n_1} x_{11}P(\vec{x}_1|aa_1)\sum_{x_{21}=0}^{n_2} P(\vec{x}_2|aa_2), \quad \text{[S24]}$$

$$= \sum_{x_{11}=0}^{n_1} x_{11}P(\vec{x}_1|aa_1), \quad \text{[S25]}$$

$$= \frac{n_1\mu_{11}e^{-N_eqC\phi t_{11}}}{\mu_{11}e^{-N_eqC\phi t_{11}} + \mu_{12}e^{-N_eqC\phi t_{12}}}. \quad \text{[S26]}$$

The above Eq. **S26** is equivalent to Eq. **S10**, which considers a gene sequence with only one amino acid and two codons.

S2. Argument Against Model Overparametrization. Although it may seem that the excellent fit between the observed and predicted values may be attributable to overfitting the data with a large numbers of parameters, this is not the case. For instance, in the case of an amino acid with k codons, there are $k - 1$ independent codon frequencies. Because the change in codon frequencies with gene expression can be thought of as a nonlinear regression, each codon should have a slope and an intercept. Thus, there are $2(k - 1)$ independent parameters for an amino acid with k codons. The relative mutation rates provide the estimates for intercepts, whereas differences in elongation times provide the estimates for their respective slopes. The beauty of our approach lies in the fact that our simple model, appropriately parameterized, leads to a correlation coefficient of 0.96.

1. Kimura M (1964) Diffusion models in population genetics. *J Appl Probab* 1:177–232.
2. Gavrilets S (2004) *Fitness Landscapes and the Origin of Species: Monographs in Population Biology* (Princeton Univ Press, Princeton), Vol 41.
3. Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.
4. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.

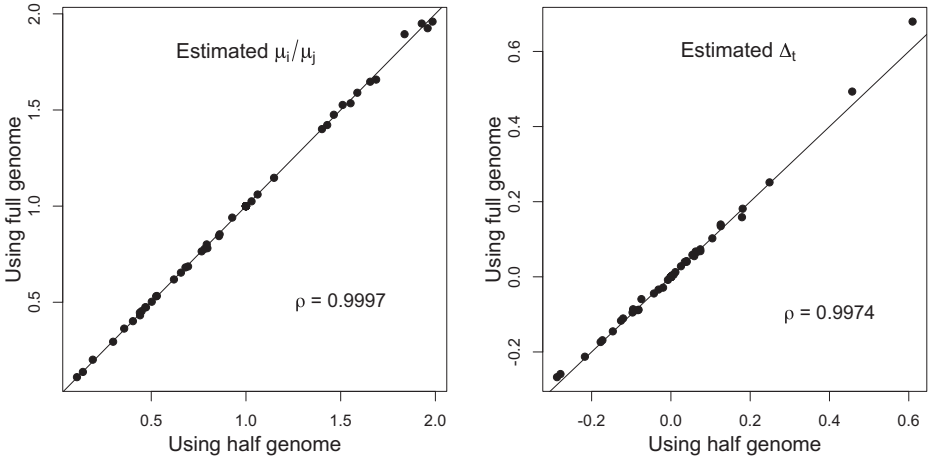

Fig. S1. Correlation between estimates of Δts and μ_i/μ_j using a random subset of 2,337 genes (half of the genome) and using the entire genome. We find a strong correlation ($\rho > 0.99$, $P < 10^{-15}$) for both Δt and μ_i/μ_j.

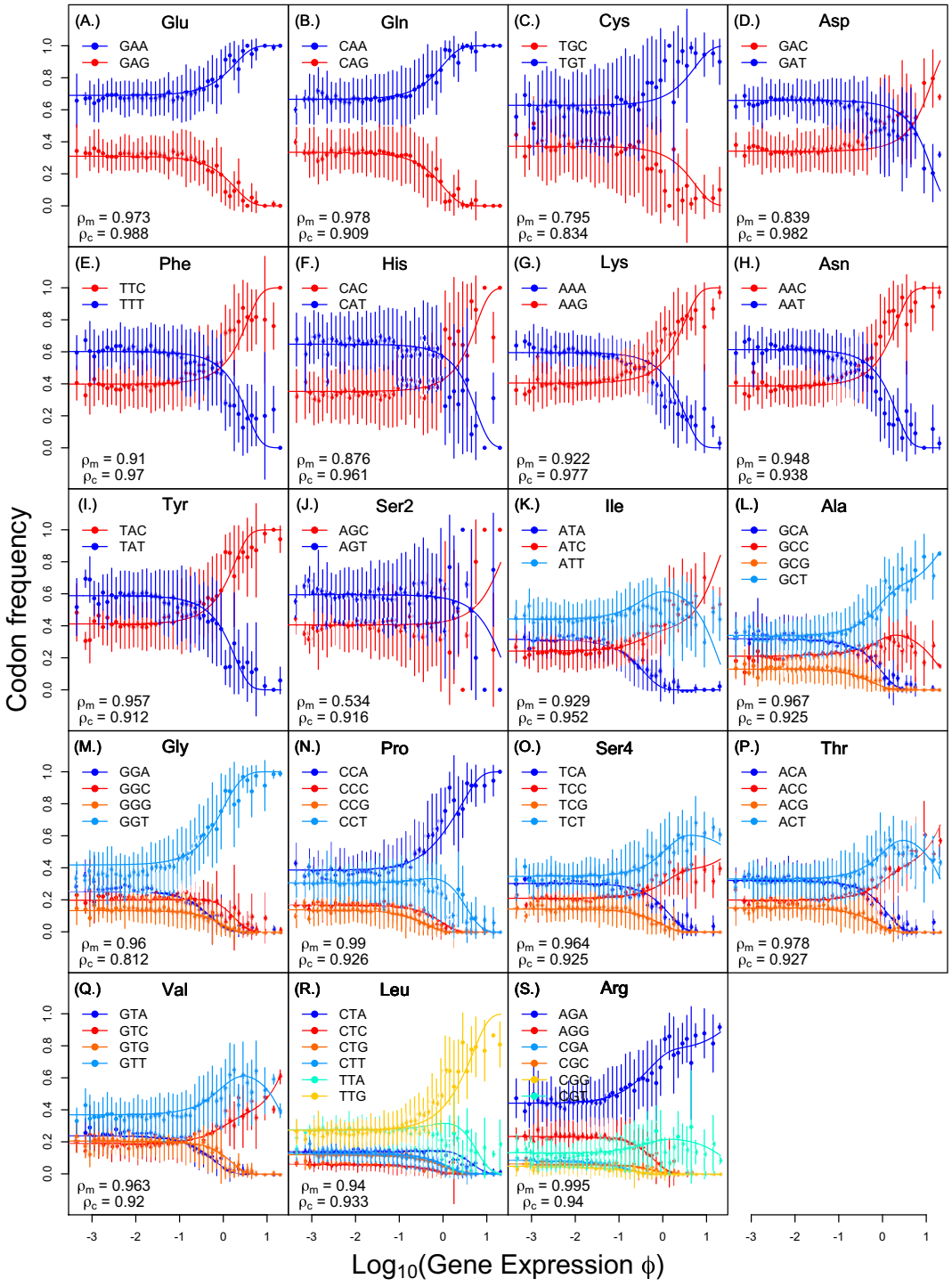

Fig. S2. Observed and predicted changes in codon frequencies with gene expression for the second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half. *A–S* correspond to a specific amino acid, where codons ending in A/T are shown in shades of blue and codons ending in G/C are shown in shades of red. Solid dots and vertical bars represent mean $\pm$ 1 SD of observed codon frequencies within genes, with protein production rates defined by the bin. The expected codon frequencies under our model are represented by solid lines. ρ_M represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean ϕ value. ρ_c represents the correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

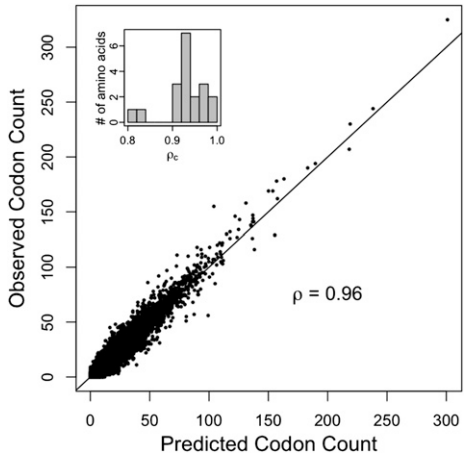

Fig. S3. Correlation between observed codon counts and predicted codon counts of individual genes in the second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half. We find a very high correlation ($\rho = 0.96$, $P < 10^{-15}$) between our model predictions and observed counts. (*Inset*) Distribution of correlation coefficients at the level of individual amino acids, indicating that our high correlation is not biased by specific amino acids and that we have a high correlation across all amino acids. ρ_c represents the correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

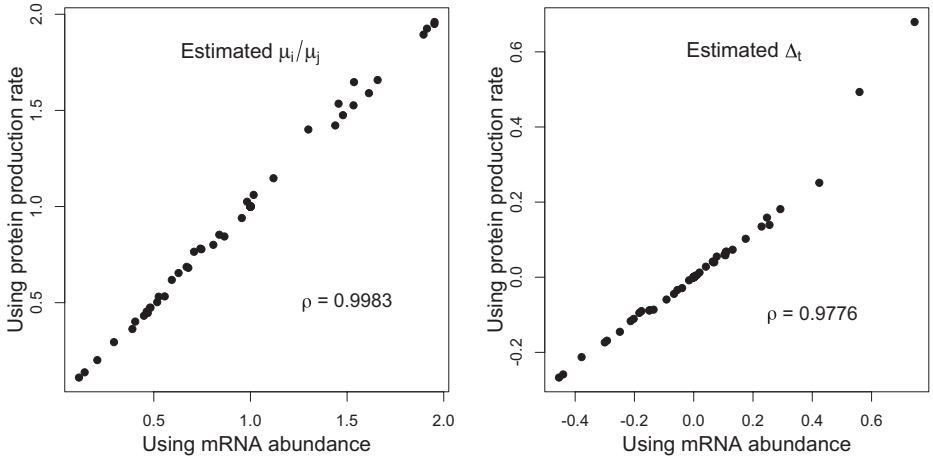

Fig. S4. Correlation between estimates of Δts and μ_i/μ_j using protein production rate ϕ for each gene and using mRNA abundances. We find a strong correlation ($\rho > 0.97$, $P < 10^{-15}$) for both Δt and μ_i/μ_j.

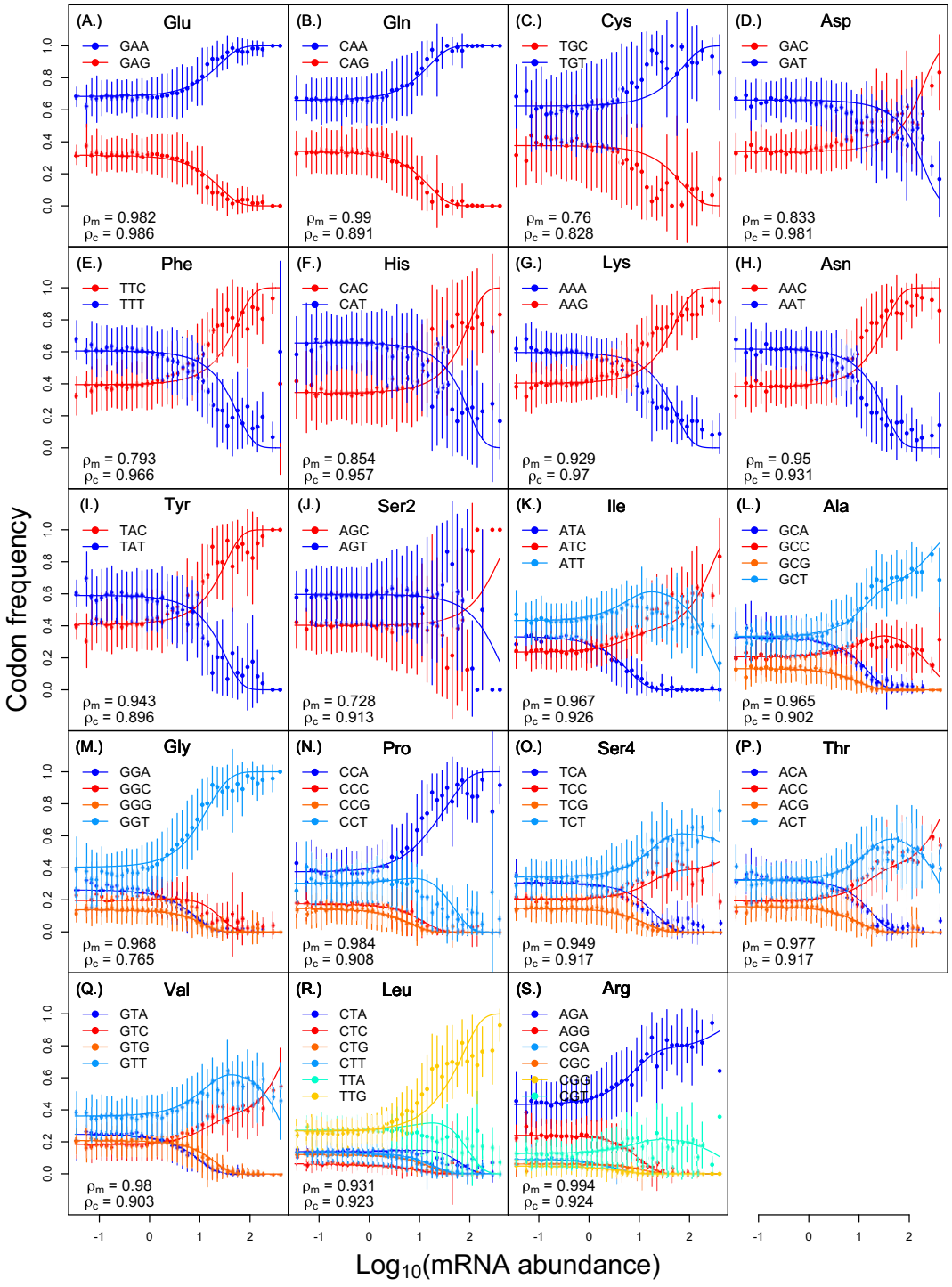

Fig. S5. Observed and predicted changes in codon frequencies with gene expression, specifically mRNA abundances. *A*–*S* correspond to a specific amino acid, where codons ending in A/T are shown in shades of blue and codons ending in G/C are shown in shades of red. Solid dots and vertical bars represent mean ± 1 SD of observed codon frequencies within genes, with mRNA abundances defined by the bin. The expected codon frequencies under our model are represented by solid lines. ρ_M represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean mRNA abundance of the bin. ρ_c represents the correlation between observed codon counts and predicted codon counts of all genes at their specific ϕ value.

Table S1. Estimates of relative mutation rate (μ_i/μ_j)

Amino acids	Codons	μ_i/μ_j	Amino acids	Codons	μ_i/μ_j
Ala	μ_{GCC}/μ_{GCA}	0.6541	Pro	μ_{CCC}/μ_{CCA}	0.4460
	μ_{GCG}/μ_{GCA}	0.4016		μ_{CCG}/μ_{CCA}	0.3630
	μ_{GCC}/μ_{GCA}	1.0605		μ_{CCT}/μ_{CCA}	0.8008
Cys	μ_{TGT}/μ_{TGC}	1.6581	Gln	μ_{CAG}/μ_{CAA}	0.5026
Asp	μ_{GAT}/μ_{GAC}	1.9496	Arg	μ_{AGG}/μ_{AGA}	0.5325
Glu	μ_{GAG}/μ_{GAA}	0.4536		μ_{CGA}/μ_{AGA}	0.2012
Phe	μ_{TTT}/μ_{TTC}	1.5262		μ_{CGC}/μ_{AGA}	0.1376
Gly	μ_{GGC}/μ_{GGA}	0.7779		μ_{GGG}/μ_{AGA}	0.1104
	μ_{GGG}/μ_{GGA}	0.5310		μ_{CGT}/μ_{AGA}	0.2946
	μ_{GGT}/μ_{GGA}	1.6471	Ser	μ_{TCC}/μ_{TCA}	0.6861
His	μ_{CAT}/μ_{CAC}	1.8943		μ_{TCG}/μ_{TCA}	0.4736
Ile	μ_{ATC}/μ_{ATA}	0.7647		μ_{TCT}/μ_{TCA}	1.1472
	μ_{ATT}/μ_{ATA}	1.4006		μ_{AGT}/μ_{AGC}	1.4752
Lys	μ_{AAG}/μ_{AAA}	0.6811	Thr	μ_{ACC}/μ_{ACA}	0.6185
Leu	μ_{CTC}/μ_{CTA}	0.4319		μ_{ACG}/μ_{ACA}	0.4740
	μ_{CTG}/μ_{CTA}	0.8441		μ_{ACT}/μ_{ACA}	1.0249
	μ_{CTT}/μ_{CTA}	0.9404	Val	μ_{GTC}/μ_{GTA}	0.7811
	μ_{TTA}/μ_{CTA}	1.9598		μ_{GTG}/μ_{GTA}	0.8533
	μ_{TTG}/μ_{CTA}	1.9253		μ_{GTT}/μ_{GTA}	1.5350
Asn	μ_{AAT}/μ_{AAC}	1.5897	Tyr	μ_{TAT}/μ_{TAC}	1.4217

Table S2. Estimates of differences in elongation time (Δt)

Amino acids	Codons	Δt	Amino acids	Codons	Δt
Ala	$t_{GCC}-t_{GCA}$	−0.1108	Pro	$t_{CCC}-t_{CCA}$	0.1394
	$t_{GCG}-t_{GCA}$	0.0551		$t_{CCG}-t_{CCA}$	0.2514
	$t_{GCC}-t_{GCA}$	−0.1168		$t_{CCT}-t_{CCA}$	0.0396
Cys	$t_{TGT}-t_{TGC}$	−0.0289	Gln	$t_{CAG}-t_{CAA}$	0.1024
Asp	$t_{GAT}-t_{GAC}$	0.0125	Arg	$t_{AGG}-t_{AGA}$	0.1813
Glu	$t_{GAC}-t_{GAA}$	0.0585		$t_{CGA}-t_{AGA}$	0.6795
Phe	$t_{TTT}-t_{TTC}$	0.0419		$t_{CGC}-t_{AGA}$	0.1586
Gly	$t_{GGC}-t_{GGA}$	−0.1452		$t_{CGG}-t_{AGA}$	0.4932
	$t_{GGG}-t_{GGA}$	−0.0593		$t_{CGT}-t_{AGA}$	0.0039
	$t_{GGT}-t_{GGA}$	−0.2126	Ser	$t_{TCC}-t_{TCA}$	−0.0887
His	$t_{CAT}-t_{CAC}$	0.0281		$t_{TCG}-t_{TCA}$	0.0400
Ile	$t_{ATC}-t_{ATA}$	−0.2671		$t_{TCT}-t_{TCA}$	−0.0876
	$t_{ATT}-t_{ATA}$	−0.2588		$t_{AGT}-t_{AGC}$	0.0054
Lys	$t_{AAG}-t_{AAA}$	−0.0443	Thr	$t_{ACC}-t_{ACA}$	−0.0950
Leu	$t_{CTC}-t_{CTA}$	0.1349		$t_{ACG}-t_{ACA}$	0.0600
	$t_{CTG}-t_{CTA}$	0.0733		$t_{ACT}-t_{ACA}$	−0.0902
	$t_{CTT}-t_{CTA}$	0.0674	Val	$t_{GTC}-t_{GTA}$	−0.1736
	$t_{TTA}-t_{CTA}$	−0.0266		$t_{GTG}-t_{GTA}$	−0.0863
	$t_{TTG}-t_{CTA}$	−0.0082		$t_{GTT}-t_{GTA}$	−0.1688
Asn	$t_{AAT}-t_{AAC}$	0.0664	Tyr	$t_{TAT}-t_{TAC}$	0.0683

Estimates of differences in elongation time (Δt) are given in seconds.

Dataset S1. List of *S. cerevisiae* genes used in the analyses and their protein production rates ϕ

Dataset S1

Dataset S2. Gene-specific observed and predicted codon counts

Dataset S2