Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift

Premal Shah[a,b,1] **and Michael A. Gilchrist**[a,b]

[a]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville TN 37996; and [b]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville TN 37996

Edited by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved May 4, 2011 (received for review November 6, 2010)

The genetic code is redundant with most amino acids using multiple codons. In many organisms, codon usage is biased toward particular codons. Understanding the adaptive and nonadaptive forces driving the evolution of codon usage bias (CUB) has been an area of intense focus and debate in the fields of molecular and evolutionary biology. However, their relative importance in shaping genomic patterns of CUB remains unsolved. Using a nested model of protein translation and population genetics, we show that observed gene level variation of CUB in *Saccharomyces cerevisiae* can be explained almost entirely by selection for efficient ribosomal usage, genetic drift, and biased mutation. The correlation between observed codon counts within individual genes and our model predictions is 0.96. Although a variety of factors shape patterns of CUB at the level of individual sites within genes, our results suggest that selection for efficient ribosome usage is a central force in shaping codon usage at the genomic scale. In addition, our model allows direct estimation of codon-specific mutation rates and elongation times and can be readily applied to any organism with high-throughput expression datasets. More generally, we have developed a natural framework for integrating models of molecular processes to population genetics models to quantitatively estimate parameters underlying fundamental biological processes, such a protein translation.

ribosome overhead cost | protein production rate

For many organisms, the preferential usage of certain codons, commonly referred to as codon usage bias (CUB), is strongly correlated with corresponding tRNA abundances and expression levels (1, 2). Explanations for these correlations abound; the most favored ones include selection against translational errors (3–5), selection for translational efficiency (6–8), effects on protein folding (9), and stability of mRNA secondary structures (10, 11). Because different combinations of these factors could lead to very similar patterns of codon usage, their relative importance in shaping the evolution of CUB is unknown (10, 12, 13). We believe that this uncertainty over their relative importance is, in large part, attributable to a lack of mechanistic models of processes hypothesized to give rise to these patterns (exceptions are found in refs. 5, 6, 13, and 14). Although most theories of codon usage predict that the degree of bias in codon usage should increase with gene expression (1, 4, 15), they lack any specific quantitative predictions about the rate and nature of these changes. This is because most commonly used indices of CUB, such as frequency of optimal codons (F_{op}) (1), codon adaptation index (*CAI*) (16), and codon bias index (*CBI*) (17), are both heuristic and aggregate measures of CUB and fail to define explicitly the factors responsible for the evolution of CUB. In contrast, we show that a mechanistic model of protein translation that explicitly includes the effects of biased mutation, genetic drift, and selection for efficient ribosome usage can explain the genome-wide codon usage patterns in *Saccharomyces cerevisiae*. Although ours is not the first attempt at using mechanistic models to explain CUB in a population genetics context (5, 6), it is unique in its ability to estimate codon-specific parameters and quantitatively predict how codon frequencies change with gene expression. We find that our model can explain ~92% of the observed variation in CUB across the *S. cerevisiae* genome.

Model

Protein synthesis is the most energetically expensive process within a cell (19). During the log-phase of growth in *S. cerevisiae*, about 60% of transcriptional machinery is devoted to making about 2,000 ribosomes every minute (20). Because ribosomes are large complexes with a finite life span and are expensive to manufacture, one would expect strong selection for their efficient usage during protein translation (6, 21–23). Coding sequences that use faster codons free up ribosomes from the mRNA, leading to smaller polysome sizes as well as an increase in the pool of free ribosomes. Given that protein translation is limited by the initiation rate, an increased pool of free ribosomes will lead to an overall increase in the translation initiation and protein production rate (6, 24). Thus, we explicitly define selection for translation efficiency as selection for an increased pool of free ribosomes (6, 22). In the absence of other factors, selection for translation efficiency should favor coding sequences that use codons with shorter elongation times and the strength of this selection should increase with gene expression (6, 7, 23, 25). If selection for translational efficiency is a major force driving the evolution of CUB in *S. cerevisiae*, we should be able to predict the CUB of a gene based on the differences in elongation times of synonymous codons, mutational bias, and its expression level.

Based on the work of Gilchrist (5) and Gilchrist et al. (15), we begin our model by first noting that in the absence of translation errors, the expected cost-benefit ratio (cost for short) for production of a single functional protein is simply

$$\eta(\vec{x}) = C\sum_{i=1}^{61} x_i\, t_i, \qquad [1]$$

where x_i is the number of codons of type i among the 61 sense codons used within a given coding sequence $\vec{x} = \{x_1, x_2, \ldots x_{61}\}$, t_i is the expected elongation time for codon i, and C is a scaling factor that represents the overhead cost of ribosome usage in ATPs per second. Codons that have shorter elongation times will lead to lower costs η, and hence are expected to be selected over their coding synonyms. Based on the work of Gilchrist (5), we assume an exponential fitness function $w(\vec{x}|\phi) \propto e^{-q\phi\eta(\vec{x})}$, where

Author contributions: P.S. and M.A.G. designed research; P.S. performed research; P.S. and M.A.G. contributed new reagents/analytic tools; P.S. analyzed data; and P.S. and M.A.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: pshah1@utk.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1016719108/-/DCSupplemental.

q is the scaling constant seconds per ATP determining the relationship between the rate of ATP usage and fitness w and ϕ is a measure of gene expression, specifically protein production rate (proteins per second). Modeling the cost of protein production in terms of ATP implicitly assumes that the organism is resource-limited and there exists selection for efficient ATP usage to maximize reproductive output. However, if the organism is not limited by resources, one would expect selection to maximize reproductive rate instead (26); in such a case, parameters C and q would be ATP-independent. This would, however, not affect the behavior of our model.

It is also important to note the distinction between the protein production rate and the translation rate of a ribosome across an mRNA. This lack of distinction has been the source of confusion over the role of gene expression in shaping patterns of codon usage in the past (6, 24). In addition, although the protein production rate of a gene changes during a single cell's lifetime, the ϕ value used here is the target time-averaged rate at which the protein will be produced. In this scenario, a change from an optimal codon to a suboptimal codon does not affect ϕ but, instead, affects the cost of meeting the target ϕ. Using the cost of producing a protein η as the phenotype, we calculate the probability of observing a particular coding sequence given its expression level, $P(\vec{x}|\phi)$. $P(\vec{x}|\phi)$ is defined for each coding sequence in the synonymous codon genotype space S_c for a given protein. Under the Fisher–Wright process (27–29), this probability is

$$P[\vec{x}|\phi] \propto w(\vec{x}|\phi)^{N_e} \prod_{i=1}^{61} \mu_i^{x_i}, \qquad [2]$$

where N_e is the effective population size and μ_i is the sum of mutation rates to codon i from its synonymous codons (29). Simply put, $P(\vec{x}|\phi)$, the probability of observing a particular synonymous codon genotype for a given protein, is a combined function of mutation bias $\prod_{i=1}^{61}\mu_i^{x_i}$, natural selection for translational efficiency w, and genetic drift N_e. Given an expression level ϕ, the probability of observing a set of codons for one amino acid is independent of the probability of observing a set of codons for another amino acid (*SI Text, Analytical Solutions of the Model*). This independence allows us to calculate the expected frequencies of codons within an amino acid independent of codon compositions of other amino acids. The resulting expected frequency of codon i of amino acid aa_k that has n_k synonymous codons is given by

$$\mathbb{E}[f_i|\phi, aa_k] = \frac{\mu_i e^{-N_e q C \phi t_i}}{\sum_{j \in n_k} \mu_j e^{-N_e q C \phi t_j}}. \qquad [3]$$

Eq. **3** describes how the expected frequency of a given codon changes with gene expression ϕ at its mutation-selection-drift equilibrium. To compare our model predictions with observed codon usage frequencies, we looked at the 4,674 verified nuclear genes that lack internal stops in *S. cerevisiae* (5) (Dataset S1). Because time-average target protein production rates of genes are not available for any organism, we use estimates of protein production rates during log growth as proxies. Empirical estimates of protein production rate ϕ were obtained from a study by Gilchrist (5), which combines mRNA abundance (30) and ribosome occupancy datasets (31, 32) (Dataset S1). The effective population size was set to $N_e = 1.36 \times 10^7$ based on the effective population size of its closely related species *Saccharomyces paradoxus* (19). Note that because N_e is scaled by qC in Eq. **3**, any error in our estimate of N_e will only affect our estimates of qC and not the behavior of our predictions.

Results

Model Behavior. The general behavior of our model is illustrated in Fig. 1, which shows the simple case of one amino acid with two codons. It demonstrates how expected frequencies of the codons change with gene expression with respect to differences in the elongation times of the codons $\Delta t_{ij} = t_i - t_j$ as well and their relative mutation rate μ_i/μ_j. As expected, codon usage in genes with low expression is primarily determined by their relative mutation rates, whereas codon usage in genes with high expression is determined by the differences in their elongation times. When both natural selection for translation efficiency and mutation biases favor the same codon, the lines representing expected frequencies of codons (Fig. 1, red lines) do not cross. However, when the direction of mutation bias is opposite to that of natural selection, the lines representing expected frequencies of codons cross (Fig. 1, blue lines).

Model Fit to *S. cerevisiae* Genome. Using Eq. **2**, we calculated the maximum likelihood estimates for the composite parameter qC,

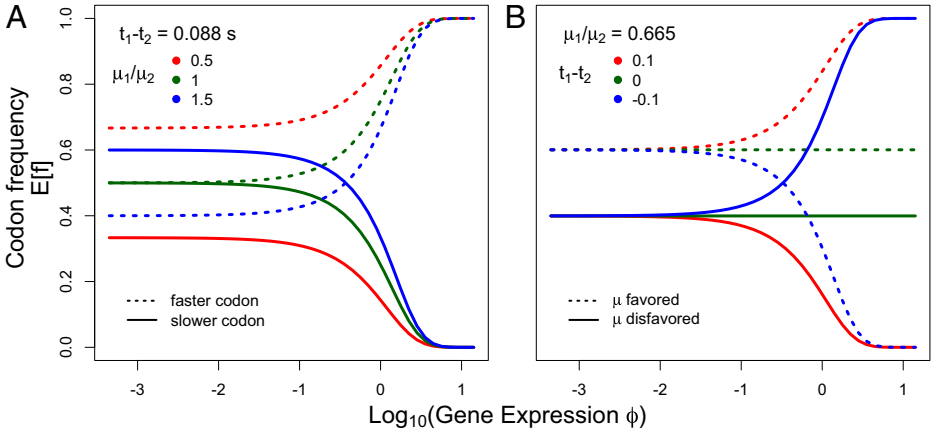

Fig. 1. Effect of varying relative mutation rate (μ_i/μ_j), elongation time (Δt_{ij}), and protein production rate (ϕ) on the expected codon frequencies (E[f]) in a hypothetical two-codon amino acid. (*A*) Effect of changing μ_i/μ_j on E[f] with ϕ. Solid lines represent the codon with a longer elongation time t_1, and dotted lines represent the codon with a shorter elongation time t_2. Mutation bias has a greater effect on E[f] at low ϕ, whereas at very high ϕ, the E[f] of codons converge to the same values, irrespective of μ_i/μ_j. (*B*) Effect of changing $t_i - t_j$ on their expected frequencies E[f] with respect to ϕ. Solid lines represent the codon with a lower relative mutation rate μ_1, and dotted lines represent the codon with a higher mutation rate μ_2. Differences in elongation times between the two codons $t_i - t_j$ has little effect on E[f] at low ϕ. However, at high ϕ, as $t_i - t_j$ changes, so does the difference in their expected frequencies E[f].

codon-specific differences in elongation time Δt_{ij}, and relative mutation rate μ_i/μ_j using 4,674 genes of the *S. cerevisiae* genome (more details are provided in *Materials and Methods* and Tables S1 and S2). Although our model uses $2(k-1)$ parameters for each amino acid with k codons, we show that it is far from being overparameterized because it uses genome-scale datasets (*SI Text, Argument Against Model Overparametrization*). The fit of our model predictions with observed data is illustrated in Fig. 2. Specifically, Fig. 2 shows how the observed and predicted codon frequencies change with gene expression ϕ for all the amino acids

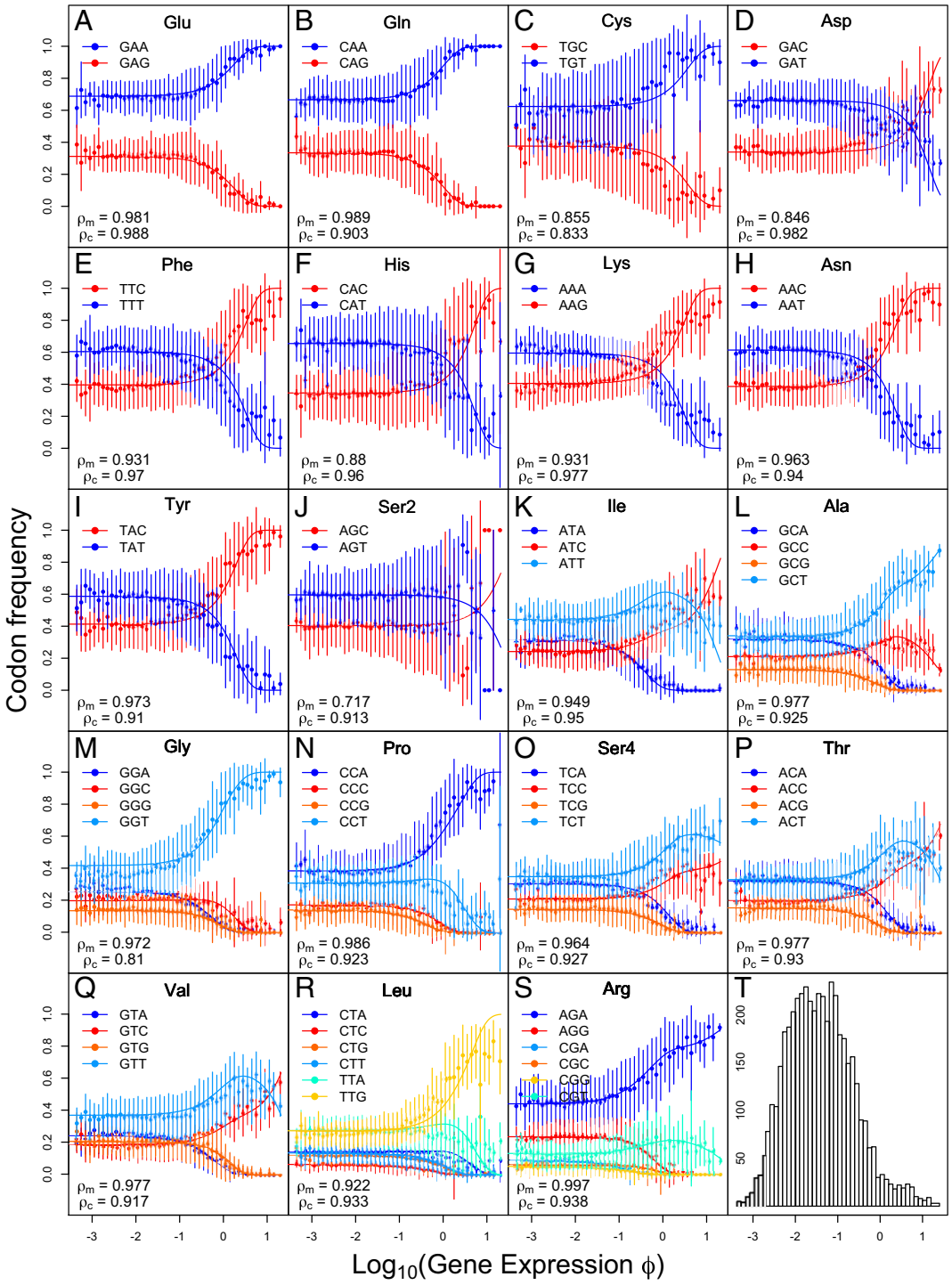

Fig. 2. Observed and predicted changes in codon frequencies with gene expression, specifically protein production rate ϕ. *A–S* correspond to specific amino acids, where codons ending in A or T are shown in shades of blue and codons ending in G or C are shown in shades of red. Solid dots and vertical bars represent mean ±1 SD of observed codon frequencies of genes in a given bin. The expected codon frequencies under our model are represented by solid lines. (*T*) Histogram of genes in each bin. We used $k-1$ codons of an amino acid with k codons in estimating correlation coefficients. ρ_M represents the Pearson correlation between the mean of observed codon frequencies within a bin and predicted codon frequencies at mean ϕ value. ρ_c represents the Pearson correlation between observed codon counts and predicted codon counts of all genes at their individual ϕ value.

that use multiple codons. Because the set of synonymous codons for Ser occurs in blocks of two and four codons separated by more than a single mutation step, we treat each of the blocks as separate amino acids, Ser_2 and Ser_4, respectively. The fit of our model can be quantified on a per-amino acid basis based on the Pearson correlation ρ_M between the mean of binned observed codon frequencies and predicted codon frequencies at the mean ϕ value. The ρ_M values ranged from 0.72 to 0.99, with a median value of 0.936.

Although many indices of adaptation have been proposed to estimate the degree of codon bias within a gene, there exists no method or index that makes predictions on codon counts of individual genes. For instance, if a particular gene has a protein production rate ϕ, what should the distribution of its codons be given its amino acid sequence? To address this question directly, we used our estimates of Δt_{ij} and μ_i/μ_j (Tables S1 and S2) to evaluate on a per-gene basis the expected codon frequencies for each amino acid using Eq. **3** (Dataset S2). We find that the correlation between observed and predicted codon counts is ρ_c = 0.959 (Fig. 3), explaining ~92% of observed variation in codon counts. Even at the level of individual amino acids, the correlation coefficients ρ_c ranged from 0.81 to 0.99. All but two amino acids had $\rho_c > 0.9$, indicating that the high correlation was consistent across all amino acids. In summary, we find that our model does an excellent job of predicting how the observed codon frequencies in *S. cerevisiae* change with gene expression ϕ.

One key insight from this work is that in *S. cerevisiae* for amino acids with more than two codons, the frequencies of preferred codons with similar elongation time $\Delta t_{ij} \sim 0$ can change in a nonmonotonic manner with gene expression ϕ. For instance, in the case of Thr, the frequency of codon ACT increases from low to moderate levels of gene expression log(ϕ) but decreases at high gene expression and is replaced by codon ACC. This nonmonotonic behavior is the result of complex interplay between mutation biases and translation selection. Specifically, although both codons ACC and ACT have shorter elongation times than their other coding synonyms ACG and ACA, codon ACC has the shortest elongation time. However, unlike codon ACC, ACT is favored by mutation bias; thus, its frequency initially increases with gene expression. We call this phenomenon "mutational inertia," whereby the frequency of a suboptimal codon transiently increases with gene expression attributable to mutation bias. This nonmonotonic behavior runs counter to traditional explanations, where the frequency of an optimal codon is expected to monotonically increase and that of a suboptimal codon is expected to monotonically decrease with gene expression (16, 33). We observed these effects of mutational inertia in most amino acids with more than two codons. Although nonmonotonic changes in codon frequencies with gene expression have been documented previously (34), the mechanisms responsible for this behavior have not been put forth. We believe this interesting and complex interplay between mutation biases and selection for efficient translation has been obscured because of an overemphasis on indices in studies of CUB. Our study illustrates the advantages of the model-based approach used here over heuristic approaches. In addition and as indicated by the crossing of lines representing codon frequencies, 7 of 10 amino acids with two codons in Fig. 2 *D–J* show mutation biases in a direction opposite to that of natural selection. In other words, codons with high frequencies in low-expression genes are not the same as the ones preferred in high-expression genes. Along with explaining these previously described patterns (35–37), we quantify the changes in codon frequencies with gene expression.

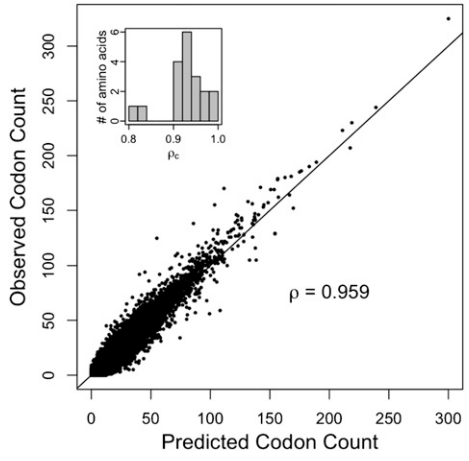

Fig. 3. Correlation between observed codon counts and predicted codon counts of individual genes. We used codon counts of $k - 1$ codons of an amino acid with k codons. Ignoring Met and Trp (one-codon amino acids) and splitting Ser into two blocks of four and two codons, there are 19 unique amino acid sets. Hence, the number of data points used is 4,674 × (59 − 19) = 186,960. We find a very high correlation (ρ = 0.959, $P < 10^{15}$) 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.

In addition to describing the genome-scale patterns of codon usage, our model allows for estimation of relative mutation rate μ_i/μ_j and differences in elongation times of these codons Δt_{ij} on a per-amino acid basis directly from the genome sequence and expression datasets. Interestingly, we find that estimates of relative mutation rates sometimes differed between amino acids. For instance, in the case of two-codon amino acids (Lys, Gln, and Glu), the NNA codons were always favored over NNG codons. However, the relative mutation rate μ_{NNG}/μ_{NNA} ranged from 0.45 to 0.68, with a mean of 0.546. These small but significant differences (*t* test, $P < 10^{-9}$ for every pair of amino acids) in the estimation of relative mutation rate may be attributable, in part, to the fact that our model does not allow for nonsynonymous substitutions, some of which may behave in a nearly neutral manner, especially in genes with low ϕ values.

We also compared our estimates of Δt_{ij} with estimates based on tRNA gene copy numbers as a proxy for tRNA abundances and wobble penalties (*Materials and Methods*). We find that these independently obtained estimates of Δt_{ij} are highly correlated (ρ = 0.801; Fig. 4).

Model Fit vs. Model Predictions. To demonstrate the predictive value of our model, we randomly partitioned the *S. cerevisiae* genome into two sets of 2,337 genes each with no significant bias in their distribution of gene expression levels ϕ (*t* test, $P > 0.4$). Parameters estimated using half of the genome were found to be highly correlated with our previous estimates based on the entire genome ($\rho > 0.99$ for both Δt_{ij} and μ_i/μ_j; Fig. S1). We then used the parameters estimated using the first set of genes to predict gene-specific codon counts in the second set of genes. The correlation coefficient between observed and predicted codon counts at the level of individual genes was 0.96 (Figs. S2 and S3). Because we do not have ribosome occupancy datasets to estimate protein production rates for most organisms, we estimated Δt_{ij} and μ_i/μ_j using mRNA abundances (5, 30) as proxies for protein production rate ϕ. We found a very high correlation between parameters estimated using mRNA abundances and protein production rates ($\rho >$ 0.97; Figs. S4 and S5). Because our model is based on mechanistic principles of protein translation, these parameters can be directly related to specific biological processes underlying protein translation. Our work demonstrates that, in principle, these parameters can be estimated directly from genomic and expression datasets,

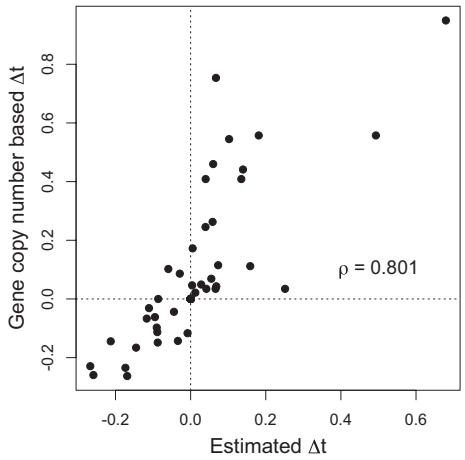

Fig. 4. Correlation between our model-based estimates of Δt_{ij} with Δt_{ij} estimated using tRNA gene copy numbers. We find a strong correlation (ρ = 0.801, $P < 10^{-9}$) between our model estimates and estimates of Δt_{ij} based on tRNA gene copy numbers, indicating that our estimates can be related to other biological estimates, such as tRNA abundances, directly.

as shown above. Estimation of these parameters can thus be easily extended to any sequenced organisms for which genome-scale mRNA expression datasets exist.

Discussion

Broader Interpretation of Δt_{ij}. The high correlation between estimates of Δt_{ij} from independent sources of genomic information (Fig. 4) suggests that our interpretation of the term Δt_{ij} is consistent with selection for translation efficiency as a major force in shaping patterns of codon usage. However, from a purely mathematical standpoint, the parameter Δt_{ij} is akin to the additive fitness component used by Sella and Hirsh (29), scaled by ϕ. Thus, its value can broadly be interpreted as an expression level-dependent selective coefficient associated with the specific codon pair. In the future, this broader interpretation should allow us to compare our genome-based estimates of Δt_{ij} with values expected under alternate hypotheses of the factors responsible for shaping codon usage patterns. For example, in the case of Cys, an interpretation of Δt_{ij} is difficult to justify based on a naive model of estimating elongation times from tRNA abundances. In *S. cerevisiae*, Cys is coded by a single tRNA, where the noncanonical codon TGT is recognized by wobble and assumed to be elongated at a slower rate than its synonym TGC (13, 38). Thus, our estimates of $t_{\mathrm{TGT}} - t_{\mathrm{TGC}} < 0$ cannot be explained on the basis of elongation times alone, because the sign of $\Delta t_{TGT,TGC}$ is opposite to that expected based on tRNA abundances and wobble. A variety of factors could potentially explain this discrepancy. First, because of its unique ability to form disulfide linkages, Cys might be under stronger selection to minimize missense errors than other amino acids. The fact that a codon with a slower elongation rate might be better at minimizing missense errors has also been predicted in a large number of other microorganisms (13). Second, as noted by Bennetzen and Hall (17), codons with side-by-side GC nucleotides may be selected against because of the high binding energies between codon-anticodon pairs. Despite the fact that Δt_{ij} can potentially be interpreted many ways, the high correlation between our predicted Δt_{ij} and estimates of Δt_{ij} based simply on tRNA gene copy numbers and wobble parameters (Fig. 4) indicates a mechanistic link between our estimates of Δt and differences in elongation times of codons.

In summary, our work shows that genome-scale patterns of codon usage can be largely explained by the effects of genetic drift, mutational biases, and natural selection for efficient usage of ribosomes (i.e., translational efficiency). Although a variety of indices have been proposed to estimate the degree of adaptation of a gene based on its CUB, our method makes predictions in the opposite direction as well (i.e., predicting codon counts of a gene, given its expression level). Our model of translation efficiency also allows us to estimate codon-specific elongation times (selection coefficients) as well as relative mutation rates. In addition, we make quantitative predictions on how individual codon frequencies should change with gene expression in yeast. Although selection for translational efficiency appears to be sufficient to explain most of the genome-scale patterns of codon usage, this does not preclude the effects of other selective forces on the evolution of CUB. For instance, selection for translation accuracy (minimizing translation missense errors) has long been argued to be a dominant force in driving the evolution of CUB (3, 39, 40). However, current data suggest that only ~10–50% of missense errors disrupt protein function (41, 42), and therefore cannot explain the high frequencies of ~100% of mutationally disfavored codons in Phe, Asn, and Tyr amino acids (Fig. 2). Moreover, the assumptions underlying Akashi's test (3) used to support the translation accuracy hypothesis are not always justified (13). Nevertheless, selection for translation accuracy can explain codon usage at functionally and/or structurally critical sites of a protein (40). Because codons that minimize missense errors may not necessarily be the ones that minimize elongation times (13), our model is likely insufficient to explain the codon usage at these sites. Similarly, adaptation against nonsense errors has been documented in *S. cerevisiae* (14, 15) and other organisms (43). In addition, factors indirectly related to protein translation, such as mRNA secondary structures at the 5′ region of a gene, have been shown to be under selection for efficient binding of ribosomes to mRNAs, and hence can affect the frequency of codon usage at these sites (10, 11).

Clearly, although a number of selective mechanisms have been proposed to explain and likely contribute to specific patterns of codon usage, the combined effects of these forces in shaping genomic patterns of codon usage are not well understood (4, 24). To decipher the relative importance of these forces on the evolution of CUB, mechanistic models that explicitly take into account tRNA competition and intraribosomal dynamics (13) as well as effects of amino acid substitutions on protein structure and function (42) need to be developed. As with our previous work (5), our model demonstrates the strength of such an approach and provides a natural framework for expansion to include other selective forces as well. More generally, this approach will allow us to estimate parameters underlying fundamental biological processes, such as protein translation, quantitatively and to improve our understanding of how evolutionary forces shape genomic patterns and processes.

Materials and Methods

Estimation of Δt_{ij} and μ_i/μ_j from Observed Data. In the case of an amino acid with k codons, the change in codon frequencies across the entire range of gene expression can be determined by $2(k-1)$ parameters for codon-specific mutation rates and elongation times. For instance, in the case of amino acids with two codons, the frequency of any one codon depends only on the difference in the elongation times of the two codons and the ratio of their mutation rates

$$\mathbb{E}[x_1|\phi] = \frac{n\mu_1 e^{-N_e qC\phi t_1}}{\mu_1 e^{-N_e qC\phi t_1} + \mu_2 e^{-N_e qC\phi t_2}} = \frac{1}{1 + \frac{\mu_2}{\mu_1} e^{-N_e qC\phi (t_2 - t_1)}}. \quad [4]$$

Codon usage in genes with low-expression ϕ is thought to be determined primarily by mutation biases (i.e., $N_e qC\phi \approx 0$). Because absolute mutation rates to each codon cannot be estimated directly, as it is only their ratios that affect codon usage, we estimated μ_i/μ_j by setting the mutation rate of an arbitrarily chosen codon to 1. Codon counts in low-expression genes can

EVOLUTION

then be assumed to follow a multinomial distribution with parameters determined by their mutation rates. Thus, in the case of an amino acid with two codons whose codon counts are x_1 and x_2, the maximum likelihood estimate of relative mutation rate is approximately

$$\frac{\mu_2}{\mu_1} \approx \frac{x_2}{x_1}. \quad [5]$$

Similarly, elongation times of codons affect codon usage only as their differences ($t_1 - t_2$). Thus, during parameter estimation of elongation times, we set the elongation time of an arbitrarily chosen codon within each amino acid to 1 and estimated the differences in elongation times of other codons with respect to that codon. We used the NEWUOA optimization algorithm (44), which is utilized in R to estimate Δt_{ij} and μ_i/μ_j for an amino acid with k codons and qC, by maximizing the following likelihood function (additional details are provided in *SI Text, Analytical Solutions of the Model*).

$$\text{Lik}\,(\vec{t}, \vec{\mu}|\phi, \vec{x}) = P(\vec{x}|\phi) = \prod_{i=1}^{k} \left(\frac{\mu_i e^{-N_e qC\phi t_i}}{\sum_{j=1}^{k} \mu_j e^{-N_e qC\phi t_j}} \right)^{x_i}. \quad [6]$$

In addition, we estimated the maximum likelihood value of $\widehat{qC} = 9.12 \times 10^{-7}$.

Estimation of Δt_{ij} from tRNA Gene Copy Numbers. To compare our estimates of Δt_{ij} with an independent source of genomic information, we estimated Δt_{ij} using tRNA gene copy numbers and wobble effects. Following the work of Dong et al. (2) and Kanaya et al. (45), we use tRNA gene copy numbers in yeast obtained from GtRNAdb (46) as proxies for tRNA abundances. We assume that the expected waiting time at a codon t_i is inversely proportional to its cognate tRNA abundances based on an exponential waiting process

$$[\text{tRNA}_i] \propto \text{Gene copy number of tRNA}_i, \quad [7]$$

$$t_i = \frac{a}{[\text{tRNA}_i] \times wob}, \quad [8]$$

where *wob* is the wobble penalty attributable to codon-anticodon mismatch and *a* is a scaling constant. When a codon is recognized by its canonical tRNA, we set *wob* = 1. Based on the work of Curran et al. (47) and Curran and Lim (48), we assume a purine-purine or pyrimidine-pyrimidine wobble penalty to be 39% and a purine-pyrimidine wobble penalty to be 36%. We set the scaling constant *a* such that the harmonic mean of elongation rates of all codons is 10 aa per second (5, 14).

ACKNOWLEDGMENTS. We thank J. Plotkin, B. O'Meara, F. Ubeda de Torres, I. Juric, and two anonymous reviewers for comments on the manuscript. M.A.G. thanks B. Burstein and S. Gardial for their indirect support of this work. Support for this project was provided by the Department of Ecology and Evolutionary Biology at the University of Tennessee, Knoxville; the National Institute for Mathematical and Biological Synthesis; and the Tennessee Science Alliance. P.S. received additional funding from a National Institute for Mathematical and Biological Synthesis Graduate Research Assistantship.

1. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 151:389–409.
2. Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* 260:649–663.
3. Akashi H (1994) Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. *Genetics* 136:927–935.
4. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10:715–724.
5. Gilchrist MA (2007) Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol* 24:2362–2372.
6. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.
7. Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. *Curr Opin Genet Dev* 8:688–693.
8. Coleman JR, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* 320:1784–1787.
9. Kimchi-Sarfaty C, et al. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315:525–528.
10. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258.
11. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* 107:3645–3650.
12. Arava Y, Boas FE, Brown PO, Herschlag D (2005) Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res* 33:2421–2432.
13. Shah P, Gilchrist MA (2010) Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet* 6:e1001128.
14. Gilchrist MA, Wagner A (2006) A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J Theor Biol* 239:417–434.
15. Gilchrist MA, Shah P, Zaretzki R (2009) Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* 183:1493–1505.
16. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 24:28–38.
17. Bennetzen JL, Hall BD (1982) Codon selection in yeast. *J Biol Chem* 257:3026–3031.
18. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res* 32:5036–5044.
19. Wagner A (2005) Energy constraints on the evolution of gene expression. *Mol Biol Evol* 22:1365–1374.
20. Warner JR (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* 24:437–440.
21. Kurland C (1987) Strategies for efficiency and accuracy in gene expression. *Trends Biochem Sci* 12:126–128.
22. Lovmar M, Ehrenberg M (2006) Rate, accuracy and cost of ribosomes in bacterial cells. *Biochimie* 88:951–961.
23. Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* 42:287–299.
24. Plotkin JB, Kudla G (2011) Synonymous but not the same: The causes and consequences of codon bias. *Nat Rev Genet* 12:32–42.
25. Akashi H (2003) Translational selection and yeast proteome evolution. *Genetics* 164: 1291–1303.
26. MacArthur R, Wilson E (1967) *The Theory of Island Biogeography* (Princeton Univ Press, Princeton).
27. Wright S (1969) *Evolution of the Genetics of Population. The Theory of Gene Frequencies* (Univ of Chicago Press, Chicago), Vol 2.
28. Gavrilets S (2004) *Fitness Landscapes and the Origin of Species: Monographs in Population Biology* (Princeton Univ Press, Princeton), Vol 41.
29. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.
30. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T (2004) Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics* 3:1083–1092.
31. Arava Y, et al. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 100:3889–3894.
32. MacKay VL, et al. (2004) Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: Response of yeast to mating pheromone. *Mol Cell Proteomics* 3:478–489.
33. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci USA* 96:4482–4487.
34. Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J Evol Biol* 1:15–26.
35. Sharp PM, Devine KM (1989) Codon usage and gene expression level in Dictyostelium discoideum: Highly expressed genes do 'prefer' optimal codons. *Nucleic Acids Res* 17: 5029–5039.
36. Musto H, Romero H, Zavala A (2003) Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum. *Microbiology* 149:855–863.
37. Peixoto L, Fernández V, Musto H (2004) The effect of expression levels on codon usage in Plasmodium falciparum. *Parasitology* 128:245–251.
38. Gromadski KB, Rodnina MV (2004) Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. *Mol Cell* 13:191–200.
39. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102:14338–14343.
40. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
41. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* 240:421–433.
42. Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. *Proc Natl Acad Sci USA* 101:9205–9210.
43. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH (2004) Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* 168:2245–2260.
44. Powell MJD (2006) The NEWUOA software for unconstrained optimization without derivatives. *Large-Scale Nonlinear Optimization*, eds Di Pillo G, Roma M (Springer, New York), pp 255–297.
45. Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238:143–155.
46. Chan PP, Lowe TM (2009) GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* 37(Database issue):D93–D97.
47. Curran JF, Yarus M (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol* 209:65–77.
48. Lim VI, Curran JF (2001) Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. *RNA* 7: 942–957.

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.

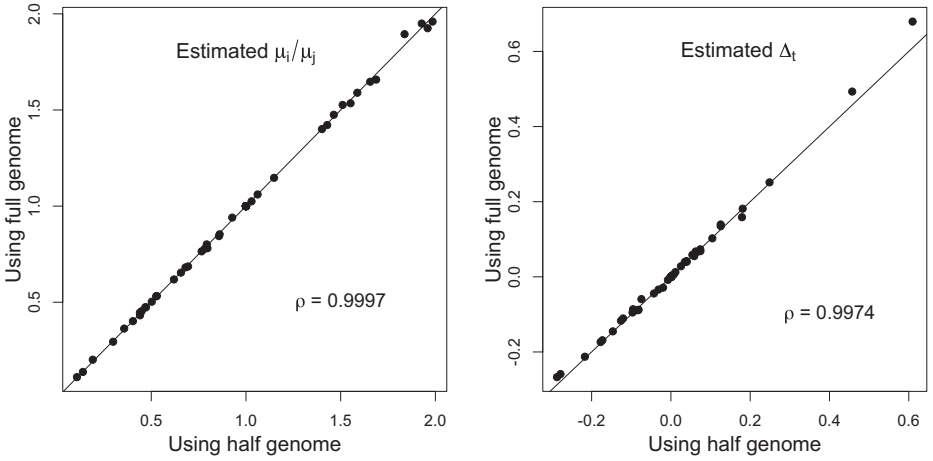

Fig. S1. Correlation between estimates of Δts and μ_i/μ_j using a random subset of 2,337 genes (half of the genome) and using the entire genome. We find a strong correlation ($\rho > 0.99$, $P < 10^{-15}$) for both Δt and μ_i/μ_j.

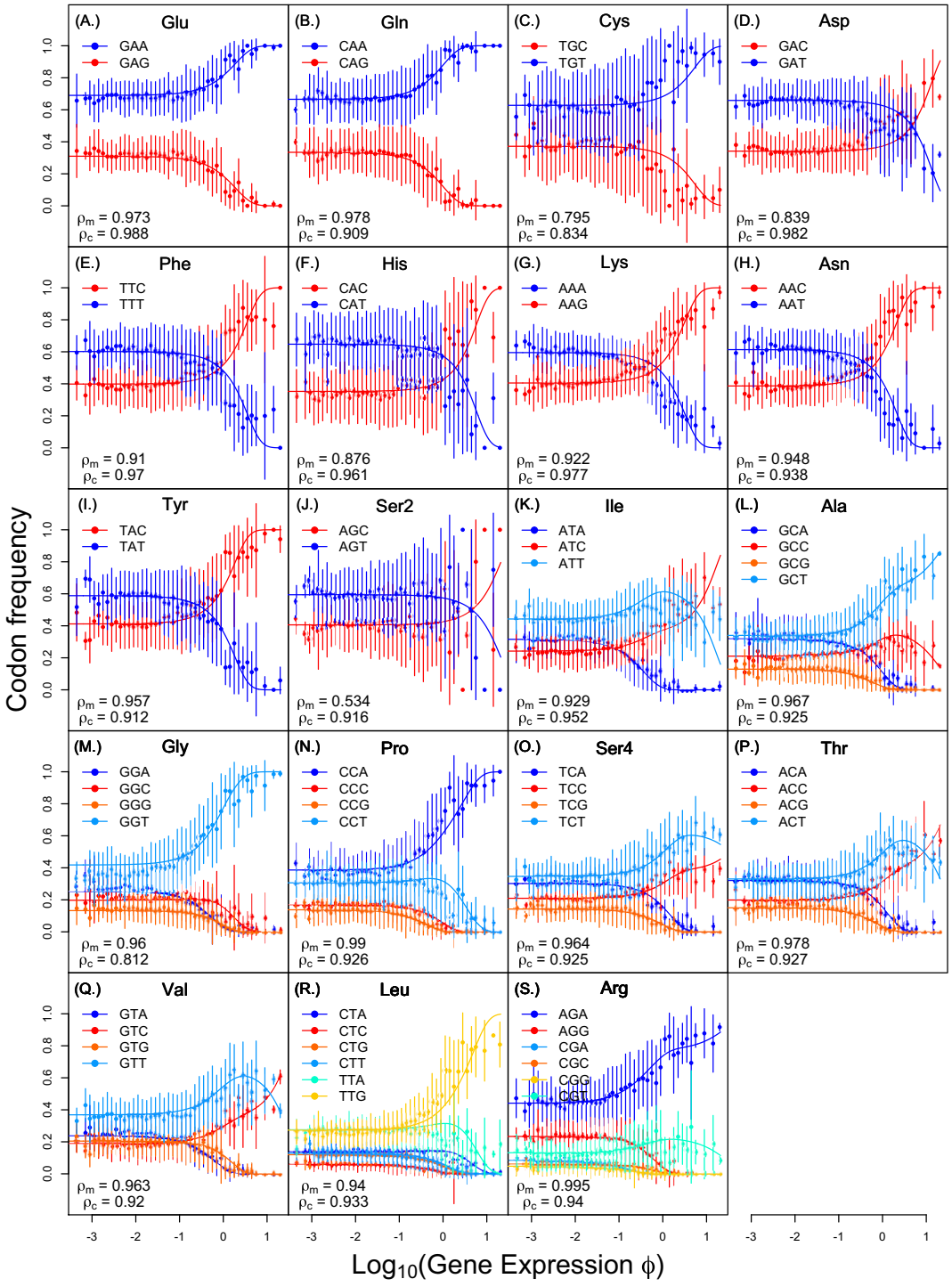

Fig. S2. Observed and predicted changes in codon frequencies with gene expression for the second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half. *A–S* correspond to a specific amino acid, where codons ending in A/T are shown in shades of blue and codons ending in G/C 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.

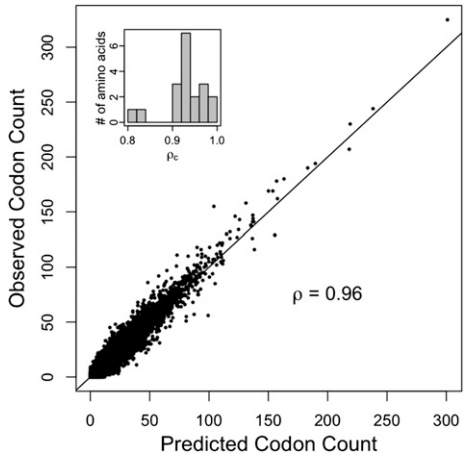

Fig. S3. Correlation between observed codon counts and predicted codon counts of individual genes in the second half of the genome using parameters Δt and μ_i/μ_j estimated using the first half. We find a very high correlation ($\rho = 0.96$, $P < 10^{-15}$) between our model predictions and observed 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.

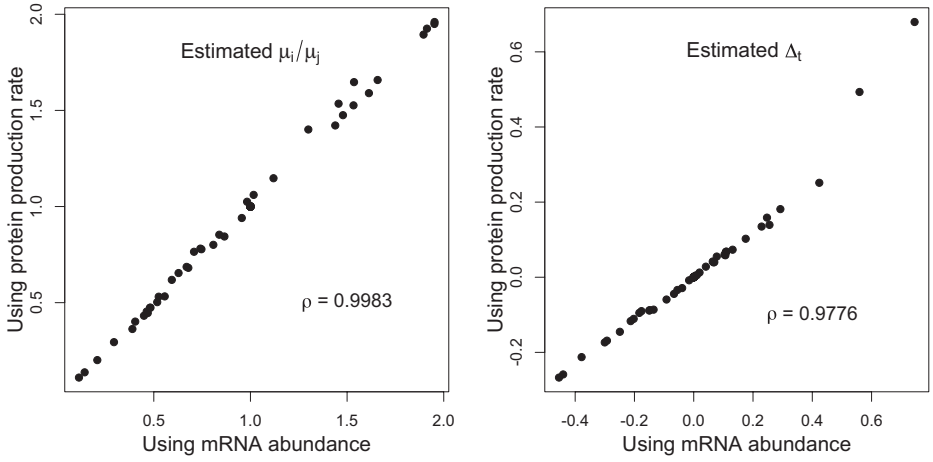

Fig. S4. Correlation between estimates of Δts and μ_i/μ_j using protein production rate ϕ for each gene and using mRNA abundances. We find a strong correlation ($\rho > 0.97$, $P < 10^{-15}$) for both Δt and μ_i/μ_j.

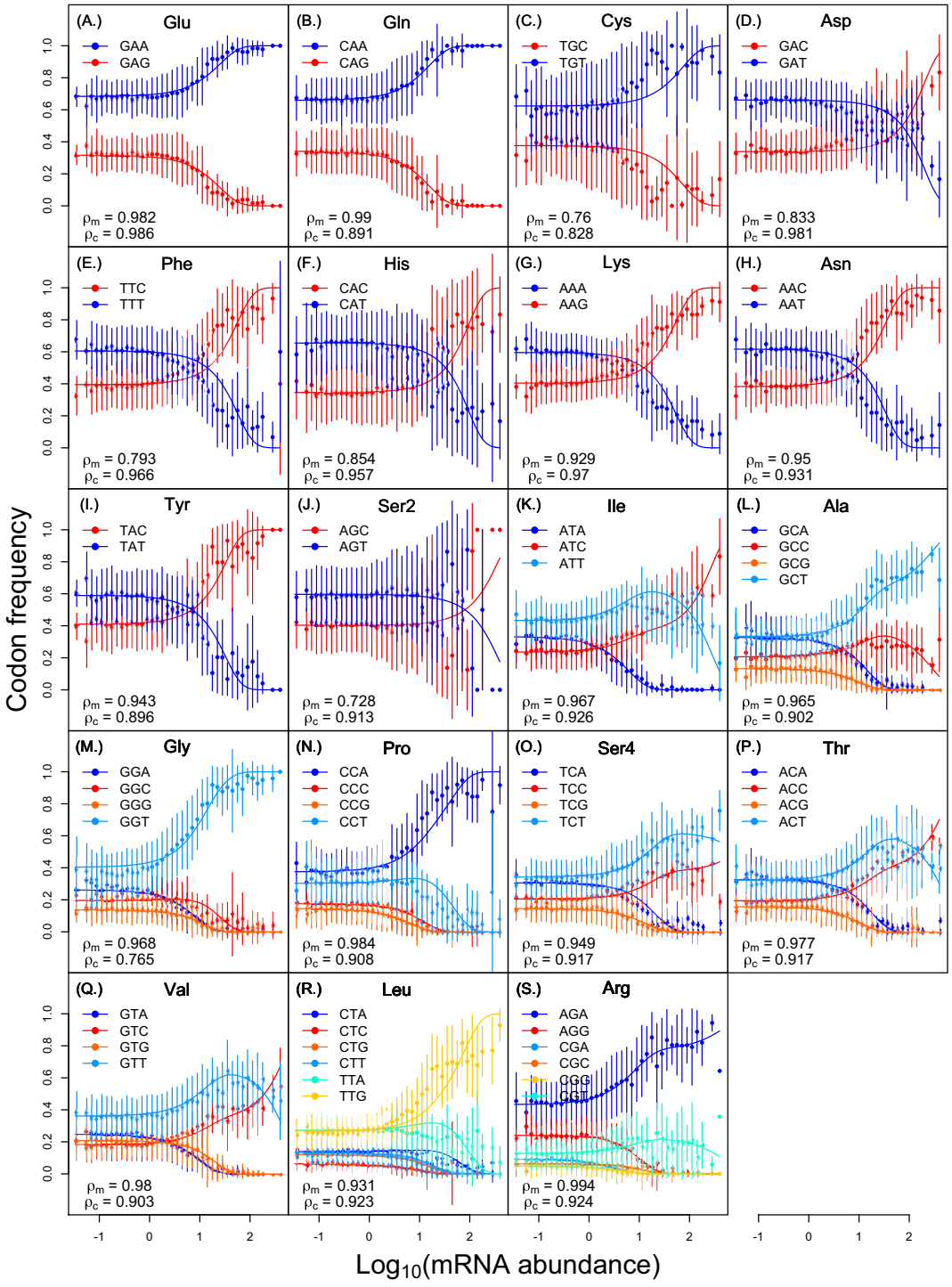

Fig. S5. Observed and predicted changes in codon frequencies with gene expression, specifically mRNA abundances. *A*–*S* correspond to a specific amino acid, where codons ending in A/T are shown in shades of blue and codons ending in G/C are shown in shades of red. Solid dots and vertical bars represent 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