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

Premal Shah\textsuperscript{a,b,1} and Michael A. Gilchrist\textsuperscript{a,b}\textsuperscript{2}

\textsuperscript{a}Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville TN 37996; and \textsuperscript{b}National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville TN 37996

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 \textit{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 ribosomal 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 as a protein translation.

ribosome overhead cost \mid 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 \textit{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 \textit{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 \textit{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

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

where $x_i$ is the number of codons of type $i$ among the 61 sense codons used within a given coding sequence $x = \{x_1, x_2, \ldots, x_{61}\}$, $t_i$ 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 $\eta$, and hence are expected to be selected over their coding synonyms. Based on the work of Gilchrist (5), we assume an exponential fitness function $w(x|\phi) \propto e^{-\phi \eta(x)}$, where

Author contributions: P.S. and M.A.G. designed research; P.S. performed research; P.S. and M.A.G. contributed new reagent/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.

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.

June 21, 2011 | vol. 108 | no. 25 | 10231–10236

www.pnas.org/cgi/doi/10.1073/pnas.1016719108
q is the scaling constant seconds per ATP determining the relationship between the rate of ATP usage and fitness $w$ and $\phi$ 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 $\phi$ 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 $\phi$ but, instead, affects the cost of meeting the target $\phi$. Using the cost of producing a protein $\eta$ as the phenotype, we calculate the probability of observing a particular coding sequence given its expression level, $P(x|\phi)$. $P(x|\phi)$ is defined for each coding sequence in the synonymous codon genotype space $S$, for a given protein. Under the Fisher–Wright process (27–29), this probability is

$$P(x|\phi) \propto w(x|\phi)^N \prod_{i=1}^{61} \mu_i^{x_i},$$

where $N_e$ is the effective population size and $\mu_i$ is the sum of mutation rates to codon $i$ from its synonymous codons (29). Simply put, $P(x|\phi)$, the probability of observing a particular synonymous codon genotype for a given protein, is a combined function of mutation bias $\prod_{i=1}^{61} \mu_i$, natural selection for translational efficiency $w$, and genetic drift $N_e$. Given an expression level $\phi$, 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

$$E[f_i|\phi, aa_k] = \frac{\mu_i e^{-\frac{F}{mC\phi_i}}}{\sum_{j=1}^{n_k} \mu_j e^{-\frac{F}{mC\phi_j}}},$$

Eq. 3 describes how the expected frequency of a given codon changes with gene expression $\phi$ 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 $\phi$ 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 $\Delta_t = t_i - t_j$ as well as their relative mutation rate $\mu_i/\mu_j$. 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 $qC$.
codon-specific differences in elongation time $\Delta t_{ij}$, and relative mutation rate $\mu_i/\mu_j$ 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 $\phi$ for all the amino acids.

![Codon Frequency Graphs](image)

**Fig. 2.** Observed and predicted changes in codon frequencies with gene expression, specifically protein production rate $\phi$. 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 $\pm$ 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. $\rho_m$ represents the Pearson correlation between the mean of observed codon frequencies within a bin and predicted codon frequencies at mean $\phi$ value. $\rho_c$ represents the Pearson correlation between observed codon counts and predicted codon counts of all genes at their individual $\phi$ value.

Shah and Gilchrist PNAS | June 21, 2011 | vol. 108 | no. 25 | 10233
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_1 and Ser_2, 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 $Δ_{ij}$ 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 $ρ_k = 0.959$ (Fig. 3), explaining ~92% of observed variation in codon counts. Even at the level of individual amino acids, the correlation coefficients $ρ_k$ ranged from 0.81 to 0.99. All but two amino acids had $ρ_k > 0.9$, 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 $Δ_{ij} \sim 0$ 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.

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 $Δ_{ij}$ 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 $Δ_{ij}$ 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 $Δ_{ij}$ 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 ($ρ > 0.99$ for both $Δ_{ij}$ 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 $Δ_{ij}$ 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 ($ρ > 0.97$; 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.
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 $\Delta t_{fi}$. The high correlation between estimates of $\Delta t_{fi}$ from independent sources of genomic information (Fig. 4) suggests that our interpretation of the term $\Delta t_{fi}$ 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 $\Delta t_{fi}$ is akin to the additive fitness component used by Sella and Hirsh (29), scaled by $\phi$. 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 $\Delta t_{fi}$ 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 $\Delta t_{fi}$ 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 $t_{TGT} - t_{TGC} < 0$ 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 $\Delta t_{fi}$ can potentially be interpreted many ways, the high correlation between our predicted $\Delta t_{fi}$ and estimates of $\Delta t_{fi}$ based simply on tRNA gene copy numbers and wobble parameters (Fig. 4) indicates a mechanistic link between our estimates of $\Delta t$ 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 mutually 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 shown 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 $\Delta t_{fi}$ and $\mu_{fi}$ 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

$$\frac{\mu_{\theta1}e^{-N_{\theta1}C_{\theta1}}}{\mu_{\theta2}e^{-N_{\theta2}C_{\theta2}}} = \frac{1}{2} \frac{\mu_{\theta2}e^{-N_{\theta2}C_{\theta2}}}{\mu_{\theta1}e^{-N_{\theta1}C_{\theta1}}}$$

Codon usage in genes with low-expression $\phi$ is thought to be determined primarily by mutation biases (i.e., $N_{\theta1}C_{\theta1} = 0$). Because absolute mutation rates to each codon cannot be estimated directly, as it is only their ratios that affect codon usage, we estimated $\mu_{fi}$ by setting the mutation rate of an arbitrarily chosen codon to 1. Codon counts in low-expression genes can
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

\[
\frac{P(x_1)}{P(x_2)} \approx \frac{x_1}{x_2} \quad [5]
\]

Similarly, elongation times of codons affect 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 \( \Delta t_{ij} \) and \( \mu_{ij} \) for an amino acid with \( k \) codons and \( q_C \), by maximizing the following likelihood function (additional details are provided in SI Text, Analytical Solutions of the Model).

\[
\text{Lik}(\{i, \mu_{ij}, \Delta t_{ij}\}) = \prod_{i=1}^{k} \frac{\mu_{ij}^{x_{ij}} e^{-\mu_{ij}}}{x_{ij}!} \left[ x_{ij} + \Delta t_{ij} \right]^{\frac{\mu_{ij}}{\mu_{ij} + \Delta t_{ij}}}. \quad [6]
\]

In addition, we estimated the maximum likelihood value of \( q_C = 9.12 \times 10^{-7} \).

**Estimation of \( \Delta t_{ij} \) from tRNA Gene Copy Numbers.** To compare our estimates of \( \Delta t_{ij} \) with an independent source of genomic information, we estimated \( \Delta t_{ij} \) 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

\[
\frac{[\text{tRNA}]_i}{[\text{tRNA}] \times \text{wob}} \quad [7]
\]

where \( \text{wob} \) 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 \( \text{wob} = 1 \). 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 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.

38. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH (2004) Coding-sequence determinants of elongation times, we estimated the maximum likelihood value of \( q_C = 9.12 \times 10^{-7} \).
39. Curran and Lim (48)
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

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

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

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

where \( \phi \) 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

\[
P(\hat{x} | \phi) = \frac{w(\hat{x} | \phi)^{N_e}}{\sum_{y_i \in \Sigma} w(y | \phi)^{N_e}}, \tag{S4}
\]

where \( N_e \) is the effective population size and \( \Sigma \) is the entire synonymous codon genotype space, which has \( 2^3 \) 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 \( \eta \). 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

\[
P(\hat{x} | \phi) = \frac{\binom{n}{x_1} e^{-N_q C (x_1 t_1 + x_2 t_2)}}{\sum_{y_1=0}^{n} \binom{n}{y_1} e^{-N_q C (y_1 t_1 + y_2 t_2)}}. \tag{S5}
\]

Let \( \mu_1 \) and \( \mu_2 \) 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

\[
P(\hat{x} | \phi) \propto w(\hat{x} | \phi)^{N_e} \prod_{i=1}^{2} \mu_i^y, \tag{S6}
\]

\[
P(\hat{x} | \phi) = \frac{\binom{n}{x_1} e^{-N_q C (x_1 t_1 + x_2 t_2)}}{\sum_{y_1=0}^{n} \binom{n}{y_1} e^{-N_q C (y_1 t_1 + y_2 t_2)}} \prod_{i=1}^{2} \mu_i^y, \tag{S7}
\]

where \( \hat{x} = \{x_1, x_2\} \).

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

\[
E[x_k | \phi] = \frac{n}{x_k} x_k P(\hat{x} | \phi), \tag{S8}
\]

\[
= \frac{n}{x_k} \frac{\binom{n}{x_k} e^{-N_q C (x_k t_1 + x_{k-1} t_2)}}{\sum_{y_k=0}^{n} \binom{n}{y_k} e^{-N_q C (y_k t_1 + y_{k-1} t_2)}} \prod_{i=1}^{k} \mu_i^y. \tag{S9}
\]

and by symmetry

\[
E[x_2 | \phi] = \frac{n \mu_1 e^{-N_q C t_1} + \mu_2 e^{-N_q C t_2}}{N_e}, \tag{S10}
\]

\[
= n - E[x_1 | \phi]. \tag{S11}
\]

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 \( i^{th} \) codon is given as

\[
E[x_i | \phi] = \frac{n \mu_1 e^{-N_q C t_1} + \mu_2 e^{-N_q C t_2}}{\sum_{j=1}^{k} \mu_j e^{-N_q C t_j}}. \tag{S12}
\]

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

\[
E[f_i | \phi] = \frac{\mu_i e^{-N_q C t_i}}{\sum_{j=1}^{k} \mu_j e^{-N_q C t_j}}. \tag{S13}
\]

Variance around the expected value \( E[x_i | \phi] \) can also be calculated as

\[
\text{Var}[x_i | \phi] = \frac{n}{x_i} \left( x_i - E[x_i | \phi] \right)^2 P(\{x_1, x_2, \ldots, x_k\}), \tag{S14}
\]

\[
= \frac{n}{(\sum_{j=1}^{k} \mu_j e^{-N_q C t_j})^2} \left( \sum_{j=1}^{k} \mu_j e^{-N_q C t_j} \right). \tag{S15}
\]

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

\[
\eta(\hat{x}) = C(x_1 t_{11} + x_2 t_{12} + x_1 t_{21} + x_2 t_{22}), \tag{S16}
\]

where \( x_0 \) 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
\[ P(\bar{x}|\phi) = \frac{\left(\frac{n_1}{x_{11}}\right) \left(\frac{n_2}{x_{21}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(x_{1j}C\psi_{11} + x_{2j}C\psi_{12} + x_{3j}C\psi_{13} + x_{4j}C\psi_{14})}}{\sum_{\phi_{j1}=0}^{n_{11}} \sum_{\phi_{j2}=0}^{n_{21}} \left(\frac{n_1}{y_{11}}\right) \left(\frac{n_2}{y_{21}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(y_{1j}C\psi_{11} + y_{2j}C\psi_{12} + y_{3j}C\psi_{13} + y_{4j}C\psi_{14})}} \right], \quad \text{[S18]} \]

\[ = \frac{\left(\frac{n_1}{x_{11}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(x_{1j}C\psi_{11} + x_{2j}C\psi_{12} + x_{3j}C\psi_{13} + x_{4j}C\psi_{14})}}{\sum_{\phi_{j1}=0}^{n_{11}} \sum_{\phi_{j2}=0}^{n_{21}} \left(\frac{n_1}{y_{11}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(y_{1j}C\psi_{11} + y_{2j}C\psi_{12} + y_{3j}C\psi_{13} + y_{4j}C\psi_{14})}} \times \]

\[ = \frac{\sum_{\phi_{j1}=0}^{n_{11}} \left(\frac{n_1}{x_{11}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(x_{1j}C\psi_{11} + x_{2j}C\psi_{12} + x_{3j}C\psi_{13} + x_{4j}C\psi_{14})}}{\sum_{\phi_{j1}=0}^{n_{11}} \left(\frac{n_1}{y_{11}}\right) \prod_{j=1}^{n_1} \Pi_{j}^{e_{jk}} e^{-N_{ij}(y_{1j}C\psi_{11} + y_{2j}C\psi_{12} + y_{3j}C\psi_{13} + y_{4j}C\psi_{14})}} = P(\bar{x}_{1}|aa_{1})P(\bar{x}_{2}|aa_{2}). \quad \text{[S20]} \]

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

\[ \sum_{\phi_{j1}=0}^{n_{11}} P(\bar{x}_{2}|aa_{2}) = 1, \quad \text{[S21]} \]

\[ P(\bar{x}_{1}|aa_{1}) = \sum_{\phi_{j1}=0}^{n_{11}} P(\{\bar{x}_{1},\bar{x}_{2}\}). \quad \text{[S22]} \]

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

\[ \mathbb{E}[x_{11}|\phi] = \sum_{\phi_{j1}=0}^{n_{11}} \frac{x_{11}}{x_{11}} \sum_{x_{21}=0}^{n_{21}} P(\{\bar{x}_{1},\bar{x}_{2}\}). \quad \text{[S23]} \]


The above Eq. 26 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 number 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.

**Fig. S51.** Correlation between estimates of Δt and μ/μ0 using a random subset of 2,337 genes (half of the genome) and using the entire genome. We find a strong correlation (ρ > 0.99, P < 10−15) for both Δt and μ/μ0.
Observed and predicted changes in codon frequencies with gene expression for the second half of the genome using parameters $\Delta t$ and $\mu_i/\mu_j$ 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 ± 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. $\rho_m$ represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean $\phi$ value. $\rho_c$ represents the correlation between observed codon counts and predicted codon counts of all genes at their specific $\phi$ value.
**Fig. S3.** Correlation between observed codon counts and predicted codon counts of individual genes in the second half of the genome using parameters $\Delta t$ and $\mu_i/\mu_j$ 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. $\rho_c$ represents the correlation between observed codon counts and predicted codon counts of all genes at their specific $\phi$ value.

**Fig. S4.** Correlation between estimates of $\Delta t$ and $\mu_i/\mu_j$ using protein production rate $\phi$ for each gene and using mRNA abundances. We find a strong correlation ($\rho > 0.97$, $P < 10^{-15}$) for both $\Delta t$ and $\mu_i/\mu_j$. 

Using mRNA abundance

Using protein production rate

Estimated $\mu_i/\mu_j$

Estimated $\Delta t$

$\rho = 0.9983$

$\rho = 0.9776$
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 $\pm$ 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. $\rho_m$ represents the correlation between the mean of observed codon frequencies in a bin and predicted codon frequencies at mean mRNA abundance of the bin. $\rho_c$ represents the correlation between observed codon counts and predicted codon counts of all genes at their specific $\phi$ value.
### Table S1. Estimates of relative mutation rate ($\mu_i/\mu_j$)

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Codons</th>
<th>$\mu_i/\mu_j$</th>
<th>Amino acids</th>
<th>Codons</th>
<th>$\mu_i/\mu_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>tGCC/tGCA</td>
<td>0.6541</td>
<td>Pro</td>
<td>tGCC/tGCA</td>
<td>0.4460</td>
</tr>
<tr>
<td></td>
<td>tGCG/tGCA</td>
<td>0.4016</td>
<td></td>
<td>tGCG/tGCA</td>
<td>0.3630</td>
</tr>
<tr>
<td></td>
<td>tGCC/tGCA</td>
<td>1.0605</td>
<td></td>
<td>tGCC/tGCA</td>
<td>0.8008</td>
</tr>
<tr>
<td>Cys</td>
<td>tGGT/tGRC</td>
<td>1.6581</td>
<td>Gln</td>
<td>tGGT/tGRC</td>
<td>0.5026</td>
</tr>
<tr>
<td>Asp</td>
<td>tGAT/tGAC</td>
<td>1.9496</td>
<td>Arg</td>
<td>tGAT/tGAC</td>
<td>0.5325</td>
</tr>
<tr>
<td>Glu</td>
<td>tGAG/tGAA</td>
<td>0.4536</td>
<td></td>
<td>tGAG/tGAA</td>
<td>0.2012</td>
</tr>
<tr>
<td>Phe</td>
<td>tTTT/tTTC</td>
<td>1.5262</td>
<td></td>
<td>tTTT/tTTC</td>
<td>0.1376</td>
</tr>
<tr>
<td>Gly</td>
<td>tGAG/tGAA</td>
<td>0.7779</td>
<td></td>
<td>tGAG/tGAA</td>
<td>0.1104</td>
</tr>
<tr>
<td></td>
<td>tGAG/tGAA</td>
<td>0.5310</td>
<td></td>
<td>tGAG/tGAA</td>
<td>0.2946</td>
</tr>
<tr>
<td></td>
<td>tGAT/tGAA</td>
<td>1.6471</td>
<td>Ser</td>
<td>tGAT/tGAA</td>
<td>0.6861</td>
</tr>
<tr>
<td>His</td>
<td>tCAT/tCAC</td>
<td>1.8943</td>
<td></td>
<td>tCAT/tCAC</td>
<td>0.4736</td>
</tr>
<tr>
<td>Ile</td>
<td>tATC/tATA</td>
<td>0.7647</td>
<td></td>
<td>tATC/tATA</td>
<td>1.1472</td>
</tr>
<tr>
<td></td>
<td>tATC/tATA</td>
<td>1.4006</td>
<td></td>
<td>tATC/tATA</td>
<td>1.4752</td>
</tr>
<tr>
<td>Lys</td>
<td>tACG/tAAA</td>
<td>0.6811</td>
<td>Thr</td>
<td>tACG/tAAA</td>
<td>0.6185</td>
</tr>
<tr>
<td>Leu</td>
<td>tUCG/tUCA</td>
<td>0.4319</td>
<td></td>
<td>tUCG/tUCA</td>
<td>0.4740</td>
</tr>
<tr>
<td></td>
<td>tUCG/tUCA</td>
<td>0.8441</td>
<td></td>
<td>tUCG/tUCA</td>
<td>1.0249</td>
</tr>
<tr>
<td></td>
<td>tUCG/tUCA</td>
<td>0.9404</td>
<td>Val</td>
<td>tUCG/tUCA</td>
<td>0.7811</td>
</tr>
<tr>
<td></td>
<td>tUCG/tUCA</td>
<td>1.9598</td>
<td></td>
<td>tUCG/tUCA</td>
<td>0.8533</td>
</tr>
<tr>
<td></td>
<td>tUCG/tUCA</td>
<td>1.5350</td>
<td></td>
<td>tUCG/tUCA</td>
<td>1.5350</td>
</tr>
<tr>
<td>Asn</td>
<td>tAAAT/tAAC</td>
<td>1.5897</td>
<td>Tyr</td>
<td>tAAAT/tAAC</td>
<td>1.4217</td>
</tr>
</tbody>
</table>

### Table S2. Estimates of differences in elongation time ($\Delta \tau$)

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Codons</th>
<th>$\Delta \tau$</th>
<th>Amino acids</th>
<th>Codons</th>
<th>$\Delta \tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>tGCC−tGCA</td>
<td>−0.1108</td>
<td>Pro</td>
<td>tGCC−tGCA</td>
<td>0.1394</td>
</tr>
<tr>
<td></td>
<td>tGCC−tGCA</td>
<td>0.0551</td>
<td></td>
<td>tGCC−tGCA</td>
<td>0.2514</td>
</tr>
<tr>
<td></td>
<td>tGCC−tGCA</td>
<td>−0.1168</td>
<td></td>
<td>tGCC−tGCA</td>
<td>0.0396</td>
</tr>
<tr>
<td>Cys</td>
<td>tGGT−tGRC</td>
<td>−0.0289</td>
<td>Gln</td>
<td>tGGT−tGRC</td>
<td>0.1024</td>
</tr>
<tr>
<td>Asp</td>
<td>tGAT−tGAC</td>
<td>0.0125</td>
<td>Arg</td>
<td>tGAT−tGAC</td>
<td>0.1813</td>
</tr>
<tr>
<td>Glu</td>
<td>tGAC−tGAA</td>
<td>0.0585</td>
<td></td>
<td>tGAC−tGAA</td>
<td>0.6795</td>
</tr>
<tr>
<td>Phe</td>
<td>tTTT−tTTC</td>
<td>0.0419</td>
<td></td>
<td>tTTT−tTTC</td>
<td>0.1586</td>
</tr>
<tr>
<td>Gly</td>
<td>tGCG−tGGA</td>
<td>−0.1452</td>
<td></td>
<td>tGCG−tGGA</td>
<td>0.4932</td>
</tr>
<tr>
<td></td>
<td>tGCG−tGGA</td>
<td>−0.0593</td>
<td></td>
<td>tGCG−tGGA</td>
<td>0.0039</td>
</tr>
<tr>
<td></td>
<td>tGCG−tGGA</td>
<td>−0.2126</td>
<td>Ser</td>
<td>tGCG−tGGA</td>
<td>−0.0887</td>
</tr>
<tr>
<td>His</td>
<td>tCAT−tCAC</td>
<td>0.0281</td>
<td></td>
<td>tCAT−tCAC</td>
<td>0.0400</td>
</tr>
<tr>
<td>Ile</td>
<td>tATC−tATA</td>
<td>−0.2671</td>
<td></td>
<td>tATC−tATA</td>
<td>−0.0876</td>
</tr>
<tr>
<td></td>
<td>tATC−tATA</td>
<td>−0.2586</td>
<td></td>
<td>tATC−tATA</td>
<td>0.0054</td>
</tr>
<tr>
<td>Lys</td>
<td>tAAA−tAAA</td>
<td>−0.0443</td>
<td>Thr</td>
<td>tAAA−tAAA</td>
<td>−0.0950</td>
</tr>
<tr>
<td>Leu</td>
<td>tCTC−tCTA</td>
<td>0.1349</td>
<td></td>
<td>tCTC−tCTA</td>
<td>0.0600</td>
</tr>
<tr>
<td></td>
<td>tCTC−tCTA</td>
<td>0.0733</td>
<td></td>
<td>tCTC−tCTA</td>
<td>−0.0902</td>
</tr>
<tr>
<td></td>
<td>tCTC−tCTA</td>
<td>0.0674</td>
<td>Val</td>
<td>tCTC−tCTA</td>
<td>−0.1736</td>
</tr>
<tr>
<td></td>
<td>tCTC−tCTA</td>
<td>−0.0266</td>
<td></td>
<td>tCTC−tCTA</td>
<td>−0.0863</td>
</tr>
<tr>
<td></td>
<td>tCTC−tCTA</td>
<td>−0.0002</td>
<td></td>
<td>tCTC−tCTA</td>
<td>−0.1688</td>
</tr>
<tr>
<td>Asn</td>
<td>tAAAT−tAAC</td>
<td>0.0664</td>
<td>Tyr</td>
<td>tAAAT−tAAC</td>
<td>0.0683</td>
</tr>
</tbody>
</table>

Estimates of differences in elongation time ($\Delta \tau$) are given in seconds.

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

**Dataset S2.** Gene-specific observed and predicted codon counts