Estimating gene expression and codon specific translational efficiencies, mutation biases, and selection coefficients from genomic data

Michael A. Gilchrist*

Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

Wei-Chen Chen

Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville TN 37996†

Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD 20993‡

Premal Shah

Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6313

Russell Zaretzki

Department of Statistics, Operations & Management Science, University of Tennessee, Knoxville TN 37996

National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996

September 26, 2014

*Corresponding author: mikeg@utk.edu
†Former address
‡Current address

Abstract

The time and cost of generating a genomic dataset is expected to continue to decline dramatically in the upcoming years. As a result, extracting biologically meaningful information from this continuing flood of data is a major challenge in biology. In response, we present a powerful Bayesian MCMC method based on a nested model of protein synthesis and population genetics. Analyzing the patterns of codon usage observed within a genome, our algorithm extracts and decouples information on codon specific translational efficiencies and mutation biases as well as gene specific expression levels for all coding sequences. This information can be combined to generate gene and codon specific estimates of selection on synonymous substitutions. One major advance over previous work is that our method can be used without independent measurements of gene expression. Using the *S. cerevisiae* S288c genome, we compare our model fits *with* and *without* independent gene expression measurements and observe an exceptionally high correlation between our codon specific parameters and gene specific expression levels ($\rho > 0.99$ in all cases). We also observe robust correlations between our predictions generated *without* independent expression measurements and previously published estimates of mutation bias, ribosome pausing time, and empirical estimates of mRNA abundance ($\rho = 0.53 - 0.72$). Our results indicate that failing to take mutation bias into account can lead to the misidentification of an amino acid's 'optimal' codon. In conclusion, our method demonstrates that an enormous amount of biologically important information is encoded within genome scale patterns of codon usage and this information can be accessed through carefully formulated, biologically based models.

Introduction

Genomic sequences encode a trove of biologically important information. Over 49,600 genomes are currently available from the Genomes OnLine Database (Pagani *et al.*, 2012) alone and the flow of newly sequenced genomes is expected to continue far into the future. As a result, developing ways to turn this data into useful information is one of the major challenges in the life sciences today. Although great strides have been made in extracting this information, ranging from the simple, e.g. identification of protein coding regions, to the more difficult, e.g. identification of regulatory elements (Hughes *et al.*, 2000; Wasserman and Sandelin, 2004), much of this information remains untapped. To address one aspect of this challenge, we present a method to estimate expression levels of every gene, codon specific selection coefficients, and mutation biases *solely* from patterns of codon usage bias (CUB) in protein coding sequences within a genome.

One of the earliest arguments against neutrality between synonymous codon usage was given by Clarke (1970). Since then, evidence for selection acting on CUB has been repeatedly observed. CUB clearly varies systematically within and between open reading frames (ORFs) within a species as well as across species (Grantham *et al.*, 1980; Ikemura, 1981a, 1985; Bennetzen and Hall, 1982; Sharp and Li, 1987; Andersson and Kurland, 1990b; Qin *et al.*, 2004; Gilchrist and Wagner, 2006; Chamary *et al.*, 2006; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). These patterns in CUB are driven by two evolutionary forces: mutation bias and natural selection (Ikemura, 1981a; Bulmer, 1988, 1991). Current evidence supports multiple selective forces contributing to the evolution of CUB. Most of these selective forces affect the translational efficiency of an ORF through factors such as ribosome pausing times (Andersson and Kurland, 1990b; Bulmer, 1991; Sørensen and Pedersen, 1991; Kudla *et al.*, 2009; Shah and Gilchrist, 2011), missense and nonsense errors (Kurland, 1987, 1992; Akashi, 1994; Gilchrist, 2007; Drummond and Wilke, 2008, 2009), co-translational protein folding (Thanaraj and Argos, 1996; Kimchi-Sarfaty *et al.*, 2007; Tsai *et al.*, 2008; Pechmann and Frydman, 2013), and the stability of mRNA secondary structures (Kudla *et al.*, 2009; Tuller *et al.*, 2010; Bentele *et al.*, 2013). The relative importance of each of these selective forces is expected to vary both within and between genes. The effects of these forces can be unified within a single framework by considering how the codon usage of a given ORF alters the ratio of the expected cost of protein synthesis over the expected benefit of protein synthesis, or the cost-benefit ratio η for short (Gilchrist *et al.*, 2009).

One likely way different synonymous codons lead to changes in a gene's cost-benefit ratio η results from differences in the abundances of cognate and near cognate tRNAs and the stability of the Watson-Crick base pairing between a given codon and tRNA anticodons (Ikemura, 1981a; Zaher and Green, 2009; Plotkin and Kudla, 2011). These differences, in turn, lead to differences in ribosome pausing times

and error rates between codons; codons with higher abundances of cognate and near-cognate tRNAs are thought to be have both shorter pausing times and lower error rates (Ikemura, 1981a; Kurland, 1992, though see Shah and Gilchrist (2010) for a more nuanced view).

The assumption that natural selection favors codon usage which reduces the protein synthesis cost-benefit ratio η implies that the strength of this selection should scale with the gene's expression level: highly expressed genes should show the strongest bias for codons with shorter pausing times and error rates (Ikemura, 1981a, 1985; Sharp and Li, 1986, 1987). As a result, the patterns of CUB observed within a genome should contain a significant amount of information about a gene's expression level, specifically the average protein synthesis rate ϕ for a given ORF. Further, because low expression genes are under very weak selection to reduce η, their patterns of CUB should provide information on the mutational biases experienced within a genome.

Accessing this information held within CUB patterns of an organism's genome has been the focus of several decades of research in molecular evolution. However, most approaches examine mutation bias and selection in isolation and, ignore their possible interactions (Shah and Gilchrist, 2011). The strength of mutation bias has typically been investigated by comparing the differences in GC content of synonymous sites of codons to the rest of the gene (Galtier *et al.*, 2001; Knight *et al.*, 2001; Palidwor *et al.*, 2010). Numerous methods have been used to estimate codon specific selection coefficients. For example, Sharp and Li (1987) relied on the codon usage in a set of highly expressed genes to identify the 'optimal' codon for a given amino acid as these genes are under stronger selection to be translated efficiently and accurately. Approaches that focus on a subset of high expression genes in this way implicitly assume the contribution of mutation bias to CUB is overwhelmed by natural selection and, therefore, can be ignored. As our results show, this view is overly simplistic and has likely led to the misidentification of 'optimal' codons in some situations. Phylogenetic models of protein evolution have also been used to estimate codon-specific selection coefficients and mutation biases (Tamuri *et al.*, 2012; Rodrigue *et al.*, 2010; Yang and Nielsen, 2008), however, these approaches have only been applied on a gene by gene basis, and lack the cohesive framework we use here.

We, along with others, have previously worked to link gene expression levels to patterns of CUB by nesting a mechanistic model of protein translation into a population genetics model of allele fixation in order to estimate codon specific mutation and selection parameters (Gilchrist and Wagner, 2006; Gilchrist, 2007; Shah and Gilchrist, 2011; Wallace *et al.*, 2013). Although these methods represent significant advances in estimating codon specific mutation biases and selection coefficients from genomic data, they are limited to genomes with independent measurements of gene specific protein synthesis rates or a close proxy. Historically, mRNA abundances have been used as such a proxy due to the fact that generating reliable genome scale measurements of protein synthesis is an expensive undertaking (Arava

et al., 2005; Ingolia *et al.*, 2009; Li *et al.*, 2014, e.g). In contrast, the method proposed here not only does away with the necessity of having protein synthesis rate estimates (or their proxy), but actually *provides* estimates of the average protein synthesis rate for each gene. This is in addition to our method also providing estimates codon specific mutation biases and translational inefficiencies. For this study, the translational inefficiency of a codon is specifically defined as the additive contributions of that codon to the cost-benefit ratio of protein synthesis, η.

Furthermore, combining our estimates of protein synthesis rates and codon specific translational inefficiencies allows us to generate estimates of the strength of natural selection on synonymous substitutions on a gene by gene basis. Estimating gene-specific selection coefficients on synonymous codons is critical to determining whether a gene is evolving under purifying or positive selection. Current models to identify the selection regime under which a gene evolves rely on estimated the rates of non-synonymous changes to rates of synonymous changes (dN/dS) (Li *et al.*, 1985; Nei and Gojobori, 1986; Yang and Nielsen, 2000). However, these models assume that all synonymous changes within a gene are neutral. As a result, values of dN/dS are likely biased leading to over-estimates of the number of genes evolving under positive selection ($dN/dS > 0$). By accurately estimating strength of selection on synonymous changes, researchers can begin to explicitly incorporate these effects into methods for identifying purifying and positive selection.

In order to extract information from the CUB patterns within a genome, we fit our model using a Bayesian approach which builds on our recent work (Shah and Gilchrist, 2011) and advances by Wallace *et al.* (2013). Using the *Saccharomyces cerevisiae* S288c genome as an example, we demonstrate that our model can be used to accurately estimate differences in codon specific mutation biases and contributions to η *without* the need for gene expression data. Our codon specific estimates of mutation biases and translational inefficiencies match almost exactly with the parameter estimates generated when the model is fit with empirical gene expression data (Pearson correlation coefficient $\rho > 0.99$ for both sets of parameters). In the end, we observe a Pearson correlation coefficient of $\rho = 0.72$ between our predicted protein synthesis rates and the mRNA abundances from Yassour *et al.* (2009). The variation between our predictions and Yassour *et al.* (2009)'s measurements is on par with the variation observed between mRNA abundance measurements from different laboratories (Wallace *et al.*, 2013).

By releasing our work as a stand alone package in R (see Chen *et al.* (2014)), researchers can now take the genome of any microorganism and obtain accurate, quantitative information on the effect of synonymous substitutions on protein translation costs, gene expression levels, and the strength of selection on codon usage bias.

Results

The posterior means estimated from our Bayesian MCMC simulation demonstrate two key facts: 1) we are able to estimate the strength of selection on synonymous codon usage bias from the patterns of codon usage observed within a genome and, 2) we can simultaneously, attribute this selection to the interaction of two underlying biological traits: differences in their contribution to the cost-benefit ratio η for protein synthesis between synonymous codons and the protein synthesis rate of the ORF ϕ averaged across its various environments and lifestages.

For this study, we scale our codon specific translational inefficiencies relative to the strength of genetic drift, $1/N_e$, such that $\Delta\eta_{i,j} = 2\,N_e\,q\,(\eta_i - \eta_j)$ where q described the proportional decline in fitness per ATP wasted per unit time. More specifically, $\Delta\eta_{i,j}$ describes the difference in the contribution of synonymous codons i and j to the protein synthesis cost benefit-ratios of an ORF, $(\eta_i - \eta_j)$, scaled by effective population size $N_e \gg 1$ and the value of a single ATP per unit time q. The greater the contribution of a codon to η, the greater its inefficiency. For a set of synonymous codons, by convention, we define codon 1 as the codon with the lowest inefficiency, i.e. the smallest additive contribution to η.

Based on our model assumptions, it follows that the stationary probability p_i of observing codon i being used for its cognate amino acid in a gene with an average protein synthesis rate ϕ follows a multinomial logistic distribution. Specifically, for a given amino acid a with n_a codons

$$p_i = \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\,\phi\right]}{\sum_{j=1}^{n_a}\exp\left[\Delta M_{j,1} - \Delta\eta_{j,1}\,\phi\right]}, \tag{1}$$

where $\Delta M_{1,i}$ is a measure of codon specific mutation bias. Additional model details can be found in the Methods and Materials.

The utility of Equation (1) is that it allows us to probabilistically link the parameters of interest, i.e. codon specific differences in mutation biases, $\Delta\vec{M} = \{\Delta M_{j,1} | j = 2,\ldots,n_a; a = 1,2,\ldots,n_{\text{aa}}\}$, translational inefficiencies $\Delta\vec{\eta} = \{\Delta\eta_{j,1} | j = 2,\ldots,n_a; a = 1,2,\ldots,n_{\text{aa}}\}$, and the set of gene specific protein synthesis rates, $\vec{\phi} = \{\phi_1, \phi_2, \ldots, \phi_{n_g}\}$, to the CUB patterns observed within and between the genes of a given genome. The terms n_{aa} represents the number of amino acids that use multiple codons while n_g represents the number of genes in the genome, respectively. Because moving between codon groups Ser_2 and Ser_4 requires more than one nucleotide substitution, we split the set of six synonymous codons for Ser into two groups of four and two codons, Ser_2 and Ser_4, respectively. Because of this splitting of the Ser codons and the fact that amino acids Met and Trp are encoded by only one codon, $n_{\text{aa}} = 19$. Assuming a lognormal distribution (LogN) with a mean of 1 as the prior for ϕ allows us to employ a random walk Metropolis chain to estimate the posteriors for $\Delta\vec{\eta}$, $\Delta\vec{M}$, and $\vec{\phi}$ without the need for any laboratory measurements of gene expression, $\vec{\Phi}$. This ability to fit our model *without* $\vec{\Phi}$ data is the main

advance of our work over Wallace *et al.* (2013).

The assumptions of our model imply that the codon specific translational inefficiencies are independent of codon position within a sequence. As a result, the relative strength of purifying selection *against* synonymous codon j relative to a given codon i in a gene with an average protein synthesis rate ϕ is,

$$s\left(\Delta\eta_{i,j},\phi\right) = \Delta\eta_{i,j}\phi. \tag{2}$$

Tables with these estimates of gene and codon specific selection coefficients as well as summary statistics for the posterior distributions for ΔM, $\Delta\eta$, $\vec{phi}$ can be found in the Supporting Materials.

Evaluating Model Parameter Estimates

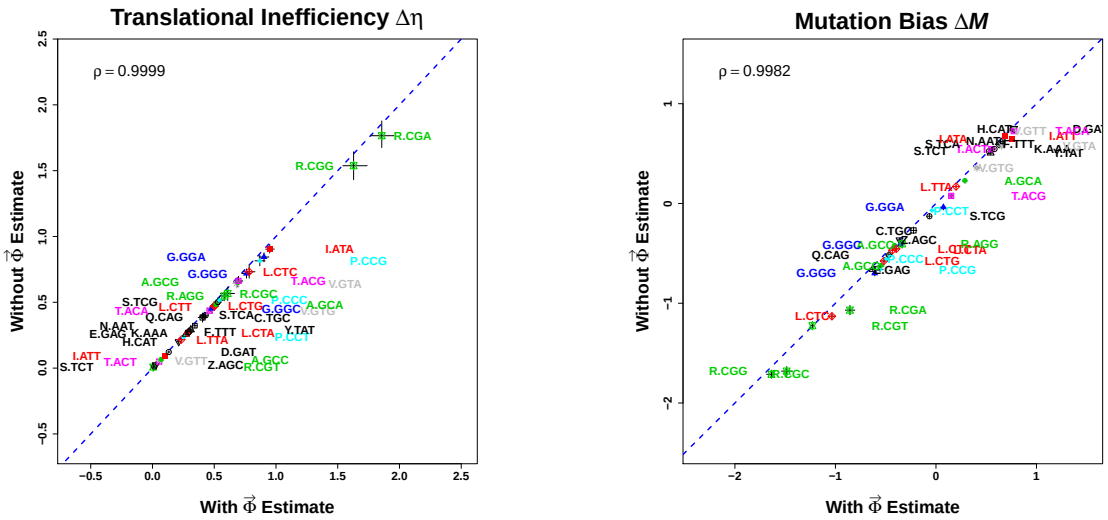
Briefly, when fitting to the *S. cerevisiae* S288c genome, we find nearly perfect agreement between *with* and *without* $\vec{\Phi}$ parameter estimates for codon specific mutation bias and protein synthesis translational inefficiencies, ΔM and $\Delta\eta$ (Pearson correlation $\rho > .99$ for both sets of parameters, see Figure 1). These results indicate that information on the genome scale parameters, $\Delta\vec{M}$ and $\Delta\vec{\eta}$ are robustly encoded and accessible within CUB patterns and that $\vec{\Phi}$, provides little additional information.

Instead of simply comparing our *without* $\vec{\Phi}$ estimates of ΔM and $\Delta\eta$ to our *with* $\vec{\Phi}$ estimates, we can also compare these parameters to other data. Because differences in ΔM values between codons that can directly mutate to one another should equal the log of the ratio of their mutation rates, our estimates of ΔM provide a testable hypothesis to compare against empirical estimates of mutation rates in *S. cerevisiae*. We use estimates of per base-pair mutation rates from a recent high-throughput mutation accumulation experiment in *S. cerevisiae* (Zhu *et al.*, 2014). Therefore empirical estimates of ΔM^e can be represented as

$$\Delta M^e_{NNN_i,NNN_j} = \ln\left[\frac{\frac{n_{i\to j}}{n_i}}{\frac{n_{j\to i}}{n_j}}\right] \tag{3}$$

where $\frac{n_{i\to j}}{n_i}$ is the number of $i \to j$ mutations observed per i bases in the genome. Since mutations in mutation accumulation experiments are strand agnostic, i.e. they do not distinguish between the coding and template strand nucleotides, we cannot distinguish between the mutations NNC→NNG and NNG→NNC nor NNA→NNT and NNT→NNA. As a result, our empirical estimates of $\Delta M^e_{C,G}$ and $\Delta M^e_{A,T}$ are set to 0. We find that our estimates of codon specific mutation rates correlate highly with empirical mutation rates in *S. cerevisiae* ($\rho = 0.95$).

Unlike mutation bias parameters, empirical estimates of the codon specific differences in translational efficiencies do not exist. However, one of the simplest ways of linking a codon to η is based on the indirect



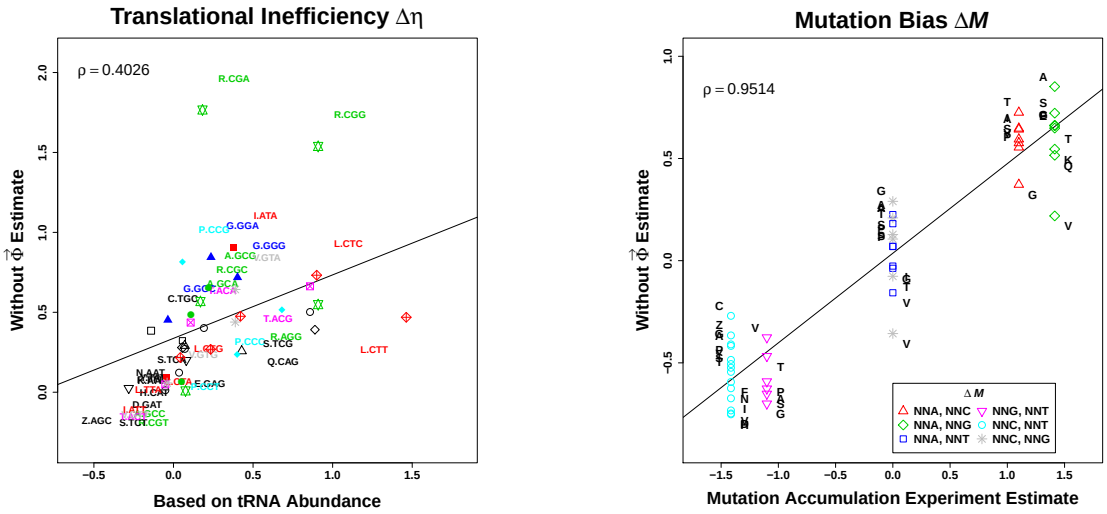
(a) (b)

Figure 1: Comparison of *with* and *without* $\vec{\Phi}$ estimates for codon specific differences in (a) translational efficiencies $\Delta\eta$ and (b) mutation biases ΔM. Error bars, which are especially difficult to see in (b), illustrate 95% posterior probability intervals for each parameter.

cost of codon specific ribosome pausing during translation. That is, $\eta_i - \eta_j \propto t_i - t_j$ where t_i is the average time a ribosome pauses when translating codon i. We calculate empirical estimates of pausing times based on a simple model of translation where pausing times at a codon depend only on its cognate tRNA abundances and associated wobble parameters (Ikemura, 1981b; Andersson and Kurland, 1990a; Sørensen and Pedersen, 1991; Kanaya *et al.*, 1999; Gilchrist and Wagner, 2006; Zaher and Green, 2009; Shah *et al.*, 2013).

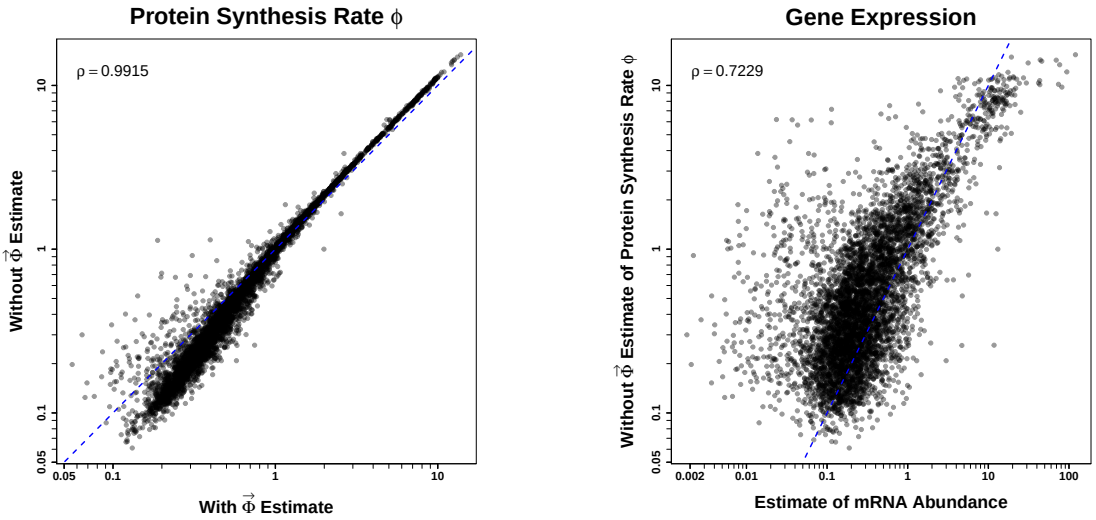
$$\Delta t_{i,j} = \frac{1}{\sum_{k \in \mathrm{T}_{aa_x}} \mathrm{tRNA}_k w_{k,i}} - \frac{1}{\sum_{k \in T_{aa_x}} \mathrm{tRNA}_k w_{k,j}} \tag{4}$$

where T_{aa_x} is the set of all tRNAs that recognize the codons for amino acid x. Specifically, tRNA_k is the gene copy number of tRNAs of type k and $w_{k,i}$ is the wobble penalty between the anti-codon of tRNA_k and codon i. When a codon is recognized by its canonical tRNA, we set $w_{k,i} = 1$. We assume a purine-purine (RR) or pyrimidine-pyrimidine (YY) wobble penalty to be 39% and a purine-pyrimidine (RY/YR) wobble penalty to be 36% based on Curran and Yarus (1989); Lim and Curran (2001). If there are two mismatches between the anti-codon of a tRNA and a codon i, as in the case of 6-codon amino acids, we set the wobble parameter $w_{k,i} = 0$. We find that our genome-wide estimates of Δt are positively correlated with empirical estimates of Δt in *S. cerevisiae* ($\rho = 0.4$).



(a) (b)

Figure 2: Comparison of *with* and *without* $\vec{\Phi}$ estimates for codon specific parameters to other types of data. (a) Comparison of *without* $\vec{\Phi}$ estimates of codon specific translational inefficiencies $\Delta\eta$ and estimates of differences in ribosome pausing times, Δt based on tRNA gene copy number and wobble inefficiencies. See (Shah and Gilchrist, 2011; Shah *et al.*, 2013) for more details. (b) Comparison of *without* $\vec{\Phi}$ estimates of codon specific mutation biases ΔM and estimates of mutation rates from mutation accumulation experiments Zhu:2014cp. For each amino acid, the codon with the shortest pausing time is used as reference codons and are not shown because, by definition, their $\Delta\eta$ and ΔM values are 0. Pearson correlation coefficients ρ are given and the black line represents the best fit regression line.



(a) (b)

Figure 3: Evaluation of predicted gene expression levels between models and empirical data. (a) Comparison of *with* and *without* estimates of protein synthesis rates, ϕ. (b) Comparison of *without* $\vec{\Phi}$ estimates of ϕ and empirical measurements of mRNA abundances, $\vec{\Phi}$, from Yassour *et al.* (2009). Pearson correlation coefficients ρ are given and blue dashed line indicates 1:1 line. Note the very strong correlation between the *with* and *without* $\vec{\Phi}$ estimates of ϕ for the high expression genes.

Predicting Protein Synthesis Rates

Given the strong correlation between the *with* and *without* $\vec{\Phi}$ estimates of the codon specific mutation biases $\Delta\vec{M}$ and translational inefficiencies $\Delta\vec{\eta}$, not surprisingly, we find that *with* and *without* $\vec{\Phi}$ estimates of ϕ are highly correlated ($\rho = 0.99$, Figure 3(a)). More importantly, the *without* $\vec{\Phi}$ based estimates of ϕ show substantial correlation with the mRNA abundance measurements $\vec{\Phi}$ values from Yassour *et al.* (2009) ($\rho = 0.72$, Figure 3 (b)). To be clear, these $\vec{\Phi}$ values are the same values used as inputs to the *with* $\vec{\Phi}$ model fits.

Figure S4 explores this issue further by plotting posterior mean estimates of ϕ produced *with* and *without* $\vec{\Phi}$ against five other experimental measurements from Arava *et al.* (2003); Nagalakshmi *et al.* (2008); Holstege *et al.* (1998); Sun *et al.* (2012). The *with* $\vec{\Phi}$ posterior estimates are generated using mRNA abundance measurements from Yassour *et al.* (2009) and are, therefore, independent of the measurements from other labs. Correlation between ϕ estimates for the *without* $\vec{\Phi}$ fits and measured mRNA abundances range from 0.534 to 0.707. The correlation between ϕ estimates for the *with* $\vec{\Phi}$ fits and mRNA provide only a 7% to 15% increase in explanatory power over the *without* predictions of ϕ. As we saw above, these high correlations between the *without* $\vec{\Phi}$ estimates of ϕ and mRNA abundance measurements are on par with the correlations typically observed when comparing mRNA measurements from different labs and platforms ($\rho = 0.6$ to 0.9 as reported in Wallace *et al.* (2013)).

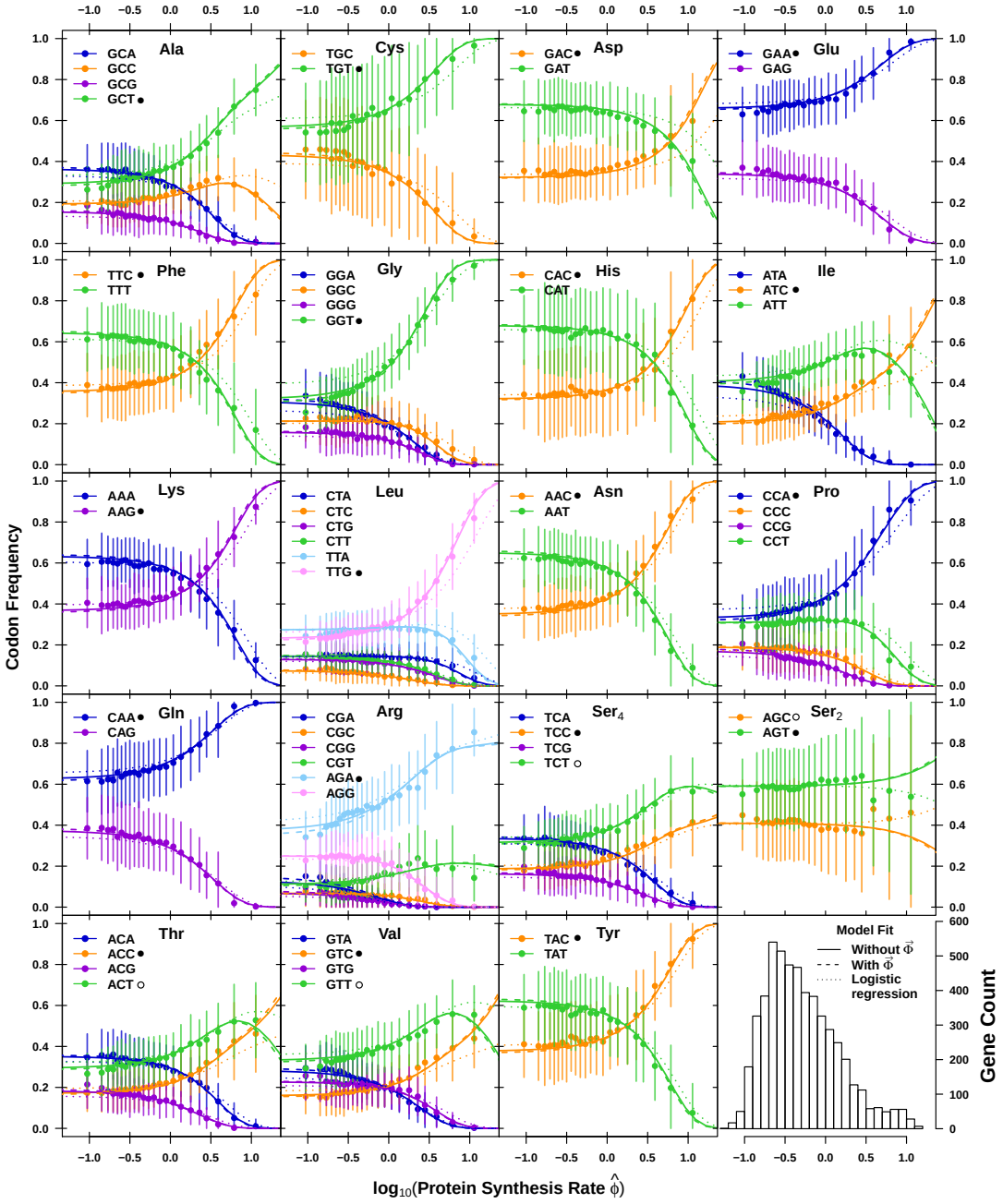
Changes in CUB with Protein Synthesis Rate

As first shown in Shah and Gilchrist (2011), the relationship in codon usage with protein synthesis rate ϕ can range from simple and monotonic to complex. Figure 4 illustrates how codon usage changes across approximately 2 orders of magnitudes of $\hat{\phi}$ for each of the $n_{\text{aa}} = 19$ multicodon amino acids. Both the *with* and *without* $\vec{\Phi}$ model fits accurately predict how CUB changes with protein synthesis rates (Figure 4). Indeed, the predicted changes in CUB between the *with* and *without* $\vec{\Phi}$ model fits are almost indistinguishable from one another, reflecting the strong agreement between their estimates of ΔM and $\Delta\eta$ across models as discussed above.

Independent of amino acid, in genes with protein synthesis rates substantially lower than the average, i.e. $\log_{10}(\hat{\phi}) \lesssim -0.5$, codon usage is largely determined by mutation bias terms ΔM. For about half of the amino acids (e.g. Cys, Lys, and Pro), in genes with protein synthesis rates 10 or more times greater than average, i.e. $\log_{10}(\hat{\phi}) \geq 1$, codon usage is largely determined by selection for the codon with the smallest translational inefficiency $\Delta\eta$. This result is largely consistent with the frequent assumption that in the set of genes with the highest expression levels the most translationally efficient codon dominates. However, for the amino acids (e.g. Ala, Ile, Ar, and Ser_2) selection for reducing η in high expression genes is substantially tempered by the effects of mutation bias.

Changes in codon frequency with ϕ are the result of a subtle interplay between natural selection for reducing η and mutation bias. The simplest cases involve two codon amino acids where the same codon is favored both by selection and mutation bias, i.e. Cys, Glu, and Ser_2. In these three cases, the selectively and mutationally favored codon 1 is used preferentially across all protein synthesis rates and the frequency of the prefered codon increases monotonically with ϕ. The next simplest cases involve two codon amino acids where codon 1 is favored by selection and codon 2 is favored by mutation bias, e.g. Asp, Asn, and Phe. In these cases, the mutationally favored codon 2 is used preferentially at low ϕ values and the selectively favored codon 1 is used preferentially used in genes at high ϕ values. Nevertheless, as before the codon frequency changes monotonically with ϕ.

More complex, non-monotonic changes in codon frequencies can occur in amino acids that use three or more codons. For example, the Ile codon ATC has the lowest translational inefficiency $\Delta\eta$ and, therefore, is the most favored codon by natural selection while ATT has the second lowest translational inefficiency. As a result, both codons initially increase in frequency with increasing ϕ at the expense of the most inefficient codon ATA. However, once the frequency of ATA approaches 0, selection for ATC begins driving the frequency of ATT down. These non-monotonic changes in codon frequency is most notable in Ala, Ile, Thr, and Val. Examining the derivative of $\vec{p}$ with respect to ϕ indicates that a given codon i will increase in frequency with ϕ, if $\sum_{j \neq i} p_j(\phi)\Delta\eta_{ij} > 0$ i.e. if the sum of the derivatives of



(a)

Figure 4: Model predictions and observed codon usage frequencies as a function of estimated protein synthesis rate $\hat{\phi}$ for the *S. cerevisiae* S288c genome. Each amino acid is represented by a separate subplot. Solid, dashed, and dotted lines represent the *without* $\vec{\Phi}$, *with* $\vec{\Phi}$, and a simple logistic regression approach where the estimation error in $\vec{\Phi}$ is ignored, respectively. Genes are binned by their expression levels with solid dots indicating the mean codon frequency of the genes in the respective bin. Error bars indicate the standard deviation in codon frequency across genes within a bin. For each amino acid, the codon favored by natural selection for reducing translational inefficiency is indicated by a •. The four ○ indicate codons that have been previously identified as 'optimal' but our model fits indicate these codons actually are the second most efficient codons. A histogram of the $\hat{\phi}$ values is presented in the lower right corner. Estimates of protein synthesis rates ϕ are based on the *with* $\vec{\Phi}$ model fits, thus representing our best estimate of their values.

the selective advantage of codon i over the other codons is positive. For the reference codon 1 where, by definition, $\Delta\eta_{1,j} \geq 0$, we see that this inequality *always* holds. This criteria can only be met by the non-reference codon if there are other codons at appreciable probabilities with lower fitness. For non-reference codons this criteria can only be met in amino acids with more than two synonyms. In the *S. cerevisiae* S288c genome, this occurs when the codon most favored by natural selection is strongly disfavored by mutation. Although this non-linear quality of multinomial logistic regression is well known among statisticians, the fact that non-optimal codons other than the choice most favored by selection one can increase with production rate has not been widely recognized.

If we ignore the noise in the $\vec{\Phi}$ data, our *with* $\vec{\Phi}$ model fitting simplifies to the standard logistic regression model applied in Shah and Gilchrist (2011). This simplification results in a slight distortion of ΔM estimates and a general attenuation of our estimates of $\Delta\eta$ (Wallace *et al.*, 2013). The effect of this attenuation can be seen in Figure 4 where the changes in CUB predicted from the standard logistic regression model fit lag behind the predicted changes when either the error in $\vec{\Phi}$ is accounted for or the $\vec{\Phi}$ data is not used. In the case of Ser_2 controlling for error leads to a change in the codon identified as being favored by natural selection. While Shah and Gilchrist (2011) predicted codon AGC would be favored by selection over AGT, both *with* and *without* $\vec{\Phi}$ fits predict the opposite. Although, this switch in order is 'significant' in that the 95% posterior interval for $\Delta\eta_{AGT,AGC} < 0$, the amino acid Ser_2 is used at very low frequency in high expression genes and its 97.5% boundary posterior probability boundary lies very close to 0. (The upper boundary lies at -0.00387 and -0.000634 for the *with* and *without* $\vec{\Phi}$ fits, respectively.) As a result, this discrepancy is not strongly supported and warrants further investigation.

Discussion

Recent advances in technology have led a remarkable and continuing decrease in the cost of genome sequencing. What is now needed are robust models and computational tools that allow researchers to access the information encoded within these genomes. The methods developed here address this need by providing a modeling framework and computational methods which can quickly extract information on codon specific mutation biases ΔM, translational inefficiencies $\Delta\eta$, and gene specific estimates of protein synthesis rates $\vec{\phi}$, using only genome wide patterns of CUB. This ability stems from the fact that the intergenic variation in patterns of CUB observed within a genome reflect a lineage's evolutionary responses to selection for efficient protein translation and mutation bias. Our results clearly show that these CUB patterns contain remarkably large amounts of useful quantitative information and the use of carefully constructed, mechanistically driven mathematical models can greatly improve our ability to access and interpret this information. Indeed, we find that our *without* $\vec{\Phi}$ gene expression measurements estimates of ΔM, $\Delta\eta$, and $\vec{\phi}$ values match almost exactly with the *with* $\vec{\Phi}$ estimates of these parameters.

By removing the need for gene expression data $\vec{\Phi}$ and, instead, providing reliable predictions of their average protein synthesis rates $\vec{\phi}$, the methods developed here should be especially helpful for molecular-, systems-, and micro-biologists for whom genomic sequence data are both abundant and inexpensive to obtain. For example, the protein translation rates we estimate $\vec{\phi}$ should contain useful information about the physiology and ecology of the organism. Indeed, for the large number of sequenced micro-organisms that cannot be easily cultured in the laboratory, their genome sequence may become the primary source of information about their biology in the near future.

For organisms that can be cultured in the laboratory, researchers can utilize experimental techniques to measure mRNA and protein abundances. Even though impressive gains have been made in our ability to measure these quantities at a genome scale, abundance data still have limitations. For example, mRNA abundance measurements have been shown to vary substantially between labs using the same strain and the same general conditions (Wallace *et al.*, 2013). Indeed, our posterior mean estimate of the error in mRNA abundance measurement indicates that the 95% credible interval for a given ranges over an order of magnitude. In terms of protein abundance measurements, most proteomic studies have difficulty quantifying membrane bound proteins [but see (Durr *et al.*, 2004; Babu *et al.*, 2012; Chen *et al.*, 2013) for recent advances]. Furthermore, for both transcriptomic and proteomic work, the measurements are, by their very nature, restricted to the specific growth conditions used. The importance of those conditions to the distribution and abundance of the organism outside of the lab is often unclear. This is particularly important for understanding a pathogenic organism, where expression of genes involved in its persistence and spread are highly dependent on their hosts and are difficult to obtain in cell culture.

The predictions of protein synthesis rates $\vec{\phi}$ we generate here contain independent and complementary information to that found in mRNA or protein abundance measurements. As a result, this information can be used on its own or in combination with other measures of gene expression. For example, our work provides estimates of protein production based on the average environment that an organism's lineage has experienced. These estimates of average gene expression can be used to further contextualize gene expression measurements in different environments. For example, comparing the ϕ values for proteins involved in different, environment specific pathways should give researchers an understanding of the relative importance these environments in the lineage's evolutionary history. At a finer scale, gene specific incongruencies between mRNA abundance measurements and ϕ estimates may indicate genes undergoing extensive post-transcriptional regulation, a hypothesis that can be evaluated experimentally. As a final example, based on ribosome footprinting data generated from *E. coli* growing in the lab, Li *et al.* (2014), show that protein synthesis rates can vary several fold between genes on the same polycistronic transcript. Our method provides the same type of information, but, instead of involving a large experimental effort, only requires looking at CUB patterns in a sequenced genome. Therefore, our methods should help advance the understanding of the molecular mechanisms responsible for intracistronic variation in translation initiation rates.

The fact that the additional information provided by the $\vec{\Phi}$ data from Yassour *et al.* (2009) leads to a relatively small increase in our predictions of $\vec{\Phi}$ data from other labs may be surprising at first. However, note that $\vec{\Phi}$ in the form of mRNA abundance measurements are (a) only a proxy for protein synthesis rates $\vec{\phi}$, (b) imprecise, and (c) that the CUB patterns themselves contain much more information than was previously recognized.

Accessing information on $\vec{\Phi}$ using a mechanistic, model based approach as developed here has additional, distinct advantages over more ad-hoc approaches frequently used by other researchers. Quantifying selection on synonymous codons is important for phylogenetic inference. Classical codon substitution models of protein evolution typically assume that synonymous codons of an amino acid are selectively neutral. Recently, several models have been proposed that estimate selection coefficients of all 61 sense codons either on a whole gene basis or on a site-by-site basis (Tamuri *et al.*, 2012; Rodrigue *et al.*, 2010; Yang and Nielsen, 2008). However, these estimates of codon-specific selection coefficients are a measure of selection on both non-synonymous and synonymous changes. Moreover, these estimates vary from gene to gene and as a result, it is unclear how to interpret these values in the context of an entire genome. Our estimates of codon-specific translation inefficiencies and expression levels provide an independent measure of selection on synonymous codons from a single genome. By incorporating these measures in codon substitution models, researchers would be able to measure selection on non-synonymous changes either within a gene or on a site-by-site basis. In addition, current measures to identify the selective regime

in which a gene evolves, e.g. positive, negative or nearly-neutral, are based on estimating the number of non-synonymous to synonymous changes (dN/dS) (Li *et al.*, 1985; Nei and Gojobori, 1986; Yang and Nielsen, 2000). However, for the sake of simplicity and due to absence of reliable estimates of selection on synonymous changes, dN/dS assumes that synonymous changes within a sequence are neutral. Ignoring selection on synonymous changes will lead to higher expected dS ratios, leading to misclassification of genes as evolving under positive selection when they might be evolving neutrally or even under purifying selection. By using our codon-specific estimates of translation inefficiencies, researchers will now be able to explicitly account for biases in estimates of dS due to selection on synonymous changes.

Estimates of codon-specific translation inefficiencies are also important for practical applications such as codon-optimization algorithms that are used to increase heterologous gene expression, for e.g. insulin expression in *E. coli*. When heterologous genes are expressed in a particular model organism such as *E. coli* or *S. cerevisiae*, their codon usage is 'optimized' by assuming that the most frequently used codon in a set of highly expressed genes is the optimal one. This approach implicitly assumes that natural selection against translational inefficiencies overwhelms any mutation bias. In several amino acids that use more than two synonymous codons, e.g. Ser, Thr and Val, genes with highest expression are more often encoded by the mutationally favored, second-best codon rather than the mutationally disfavored 'optimal' codon. As a result, relying on the codon usage of highly expressed genes appears to be overly simplistic in the case of the *S. cerevisiae* genome and can lead to misclassification of the optimal codon within a genome.

In addition to codon-specific translation efficiencies, we also estimate codon-specific mutation biases ΔM. We find that the direction of mutation biases between synonymous codons is consistent across all amino acids and in the same direction as genomic AT content. However, as we documented in Shah and Gilchrist (2011), ΔM for similar sets of nucleotides differ significantly between amino acids. For instance, in the case of two-codon amino acids with C-T wobble, we find that $\Delta M_{NNC,NNT}$ ranges from 0.27 to 0.75. For genes with low expression levels (i.e. $\phi < 1$), this corresponds to ratios of T-ending codons to C-ending codons between amino acids ranging from 1.3 to 2.1. One possible explanation for this wider than expected range of mutation biases could be context-dependence of mutation rates. Recent high-throughput mutation accumulation experiments in yeast support this idea, estimating that the mutation rate at a particular nucleotide depends on the context of surrounding nucleotides: the **C** nucleotide in the context of C**C**G has several fold higher mutation rate than in the context of C**C**T (Zhu *et al.*, 2014).

Despite the numerous advances outlined above, our work is not without its limitations. One important limitation stems from our assumption that codons contribute to the cost-benefit ratio of protein translation in an additive manner. While this assumption is consistent with certain costs of protein translation, such as ribosome pausing, it ignores many others selective forces potentially shaping the evo-

lution of CUB. For example, the cost of nonsense errors, i.e. premature termination events, are generally expected to increase with codon position along an ORF and, thus, lead to a non-additive contribution of a given codon to the cost-benefit ratio η. Similarly, if one assumes that the main effect of missense errors is to reduce the functionality of the protein produced, then the cost of these errors is expected to depend greatly on specific details such as the structural and functional role of the amino acid at which the error occurs and the physiochemical differences between the correct and the erroneously incorporated amino acids. The situation becomes even more complex and non-linear when considering how pausing time costs, nonsense errors, and/or missense errors combine to affect η. In all of these situations, the nonlinear mapping between a codon sequence and η makes direct evaluation of the likelihood function difficult. In such situations alternative, approximate methods and simulation techniques, such as those developed by (Murray *et al.*, 2006), will become necessary. Expanding our approach to include these additional selective forces should allow us to quantitatively evaluate the separate contributions ribosome pausing time, nonsense errors, and missense errors have made to the evolution of CUB for a given species. Doing so will allow us to address the long held goal in molecular and evolutionary biology of accurately quantifying the factors contributing to the evolution of CUB within a coding sequence and across a genome.

Methods and Materials

Modeling Natural Selection on Synonymous Codons

Following the notation and framework introduced in Gilchrist (2007) and Shah and Gilchrist (2011), we assume that for each gene, the organism has a target, average protein synthesis rate ϕ. Protein synthesis rates have units of 1/time; for convenience and ease of interpretation, we define our time units such that the average protein synthesis rate across the genome is one, i.e. $E(\phi) = 1$. The cost-benefit ratio $\eta(\vec{k})$ represents the expected cost, in ATPs, to produce one functional protein from the coding sequence $\vec{k} = \{c_1, c_2, \ldots c_n\}$ where c_i represents the codon used at position i in a protein of length n. As a result, $\eta(\vec{k})\phi$ represents the average energy flux an organism must expend to meet its target production rate for a given protein. If we assume that every ATP/time spent leads to a small, proportional reduction in genotype fitness q, then the fitness of a given genotype is,

$$W(\vec{k}) \propto \exp\left[-q\,\eta(\vec{k})\,\phi\right]. \tag{5}$$

In the simplest scenarios, such as when there is selection to minimize ribosome pausing during protein synthesis, a synonymous codon i makes an additive, position independent contribution to η. In this scenario, the evolution of the codons in $\vec{k}$ is independent between positions. Given these assumptions, within the ORF of a given gene the stationary probability of observing a set of codon counts $\vec{k} = \{k_1, \ldots, k_{n_a}\}$ for a given amino acid with n_a synonymous codons will follow a multinomial distribution with the probability vector $\vec{p} = \{p_1, \ldots, p_{n_a}\}$. Here, for $i = 1, \ldots, n_a$,

$$p_i\left(\Delta\vec{M}, \Delta\vec{\eta}, \phi\right) = \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\phi\right]}{\sum_{j=1}^{n_a}\exp\left[-\Delta M_{j,1} - \Delta\eta_{j,1}\phi\right]} \tag{6}$$

where $\Delta M_{i,1}$ is a measure of codon specific mutation bias and $\Delta\eta_{i,1}$ is a measure of translational inefficiency. Specifically, $\Delta M_{i,1} = \ln\left(p_1/p_i\right)|_{\phi=0}$, that is the natural logarithm of the ratio of the frequencies of synonymous codon 1 to i in the absence of natural selection. Following the detailed balance assumptions in our population genetics model, in the specific cases where codons i and 1 can mutate directly between each other, $\Delta M_{i,1}$ is also equal to the log of the ratio of the mutation rates between the two codons (Sella and Hirsh, 2005; Shah and Gilchrist, 2011; Wallace *et al.*, 2013). Following Sella and Hirsh (2005), so long as $N_e \gg 1$, for both a haploid and diploid Fisher-Wright populations, we scale the differences in the contribution two synonymous codons make to η relative to genetic drift, i.e. $\Delta\eta_{i,j} = 2N_e\left(\eta_i - \eta_j\right)$. Because the reference codon 1 is determined by pausing time values, $\Delta M_{i,1}$ values can be both negative and positive, unlike $\Delta\eta_{1,i}$.

Fitting the Model to Genomic Data

Our main goal is to estimate codon specific differences in mutation bias, $\Delta\vec{M}$, translational inefficiencies, $\Delta\vec{\eta}$, and protein synthesis rates for all genes, $\vec{\phi} = \{\phi_1, \phi_2, \ldots \phi_n\}$ from the information encoded in the codon usage patterns found across a genome. To test our approach we used the *S. cerevisiae* S288c genome file `orf_coding.fasta.gz` which was posted on 03 February 2011 by Saccharomyces Genome Database `http://www.yeastgenome.org/` (Engel *et al.*, 2014)). This data contains 5,887 genes and consists of the ORFs for all "Verified" and "Uncharacterized" genes as well as any transposable elements. To fit the *with* $\vec{\Phi}$ model we used RNA-seq derived mRNA abundance measurements from Yassour *et al.* (2009). We combined the abundance measures from the four samples, YPD0.1, YPD0.2, YPD15.1, and YPD15.2, taken during log growth phase and used the geometric mean of these values as a proxy for relative protein synthesis rates ϕ'. As is commonly done by empiricists, we rescaled our ϕ' values such that they summed to 15,000. Because our *with* $\vec{\Phi}$ model fits estimate the scaling term, $\exp(A_\Phi)$, the only effect of this rescaling is on our estimate of A_Φ. To reduce noise in the $\vec{\Phi}$ data, we only used genes with at least three non-zero measurements. After combining, the intersection of 5,887 DNA ORF sequences and 6,303 mRNA abundance measurements, produced 5,346 ORF's in common to both datasets. These 5,346 genes made up the final dataset used for the *with* and *without* $\vec{\Phi}$ model fits.

Using an MCMC approach we sample from the posterior distribution, according to the equation

$$\prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\Delta\vec{M}_i, \Delta\vec{\eta}_i, \phi_j, s_\phi \middle| \vec{k}_{i,j}\right) \propto \prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\vec{k}_{i,j} \middle| \vec{p}_{i,j}, n_{i,j}\right) f\left(\phi_j | s_\phi\right) f\left(s_\phi\right) \tag{7}$$

where the codon counts, $\vec{k}_{i,j}$, are naturally modeled as a multinomial distribution (Multinom) for the amino acid i in the ORF of gene j as defined in Equation (6), $\vec{p}_{i,j}$ is an inverse multinomial logit function (mlogit^{-1}) of $\Delta\vec{M}_i$, $\Delta\vec{\eta}_i$, and ϕ_j, and $f\left(\phi_j | s_\phi\right)$ is the prior for the protein synthesis rate $\phi_j \sim \text{LogN}(m_\phi, s_\phi)$. In order to enforce the restriction that $E[\phi_j] = 1$ for all genes we include the constraint that $m_\phi = -s_\phi^2/2$. As a result there is only one free parameter for the distribution $f(\phi_j|s_\phi)$. Further, we propose a flat prior for s_ϕ, i.e. $f(s_\phi) = 1$ for $s_\phi > 0$.

Figure 5 presents an overview of the structure of our approach, but to summarize,

$$\begin{aligned}
\vec{k}_{i,j} &\sim \text{Multinom}(n_{i,j}, \vec{p}_{i,j}), \\
\vec{p}_{i,j} &= \text{mlogit}^{-1}(-\Delta\vec{M}_i - \Delta\vec{\eta}_i\phi_j), \\
\phi_j &\sim \text{LogN}(-s_\phi^2/2, s_\phi), \text{ and} \\
\Delta\vec{M}_i,\, \Delta\vec{\eta}_i,\, s_\phi &\propto 1.
\end{aligned}$$

Our MCMC routine provides posterior samples of the genome wide parameters $\Delta\vec{\eta}$, $\Delta\vec{M}$, and s_ϕ and

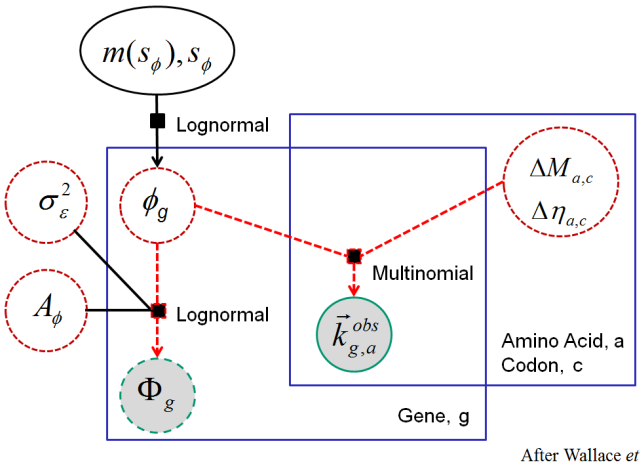


Figure 5: Dependence graph of *with* and *without* methods. Shaded circles $\vec{\Phi}_j$ and $k_{i,j}$ represent observed data. Dashed circles represent key random model parameters while the solid oval represents a random hierarchical parameter. Solid black squares provide information on the distributional relationships between quantities. Large rectangular boxes represent replication of each model component across both amino acids and genes, e.g. inefficiency and mutation parameters differ across amino acids but are common across genes, while counts $k_{i,j}$ differ across both amino acids and genes.

the gene specific, protein synthesis parameters $\vec{\phi}$. We refer to this model as our *without* $\vec{\Phi}$ model.

We refer to the more general model which incorporates information on ϕ_j from noisy protein synthesis measurements or their proxy, such as mRNA abundances, as the *with* $\vec{\Phi}$ model. This model differs from that of Wallace *et al.* (2013) in that (a) we assume ϕ_j is drawn from a lognormal distribution rather than an asymmetric Laplace distribution, (b) we include and estimate an explicit empirical scaling term A_Φ for the $\vec{\Phi}$ data, and (c) as in the *without* $\vec{\Phi}$ approach, we force the prior for ϕ_j, $f(\phi_j|s_\phi)$, to have $E[\phi_j] = 1$ instead of rescaling estimates of ϕ_j as a post-processing step. This prevents the introduction of additional biases in our parameter estimates. See the Supporting Materials for more details.

Model Fitting Details: We briefly describe the model fitting procedure here; full details can be found in Chen *et al.* (Prep). The code was originally based on a script published by Wallace *et al.* (2013), which was modified extensively and expanded greatly. Unless otherwise mentioned, all model fits were carried out using R version 3.0.2 (R Core Team, 2013) using standard routines, specifically developed routines, and custom scripts. All code was run on a multicore workstation with AMD Opteron 6378 processors. For both the *with* and *without* model fits, it takes <30 min and less than 3GB of memory to run 10,000 iterations of a chain when using 5,346 genes of *S. cerevisiae* S288c genome. Each MCMC sampling iteration was divided into three parts:

(1) conditional on a new set of parameters, propose new $\Delta\vec{M}$ and $\vec{\Delta t}$ values independently for each amino acid,

(2) conditional on the updates of (1), propose a new s_ϕ value for the prior distribution of $\vec{\phi}$, and

(3) conditional on the updates of (2), propose new $\vec{\phi}$ values independently for each gene. Update the new set of parameters and return to (1).

In all three phases, proposals were based on a random walk with step sizes normally distributed around the current state of the chain.

In order to generate reasonable starting values for $\vec{\phi}$ in the *without* $\vec{\Phi}$ model, we first calculated the SCUO value for each gene (Wan *et al.*, 2006) and then ordered the genes according to these corresponding values. We then simulated a random vector of equal dimension to $\vec{\phi}$ from a LogN($m = -\left(s_\phi^{(0)}\right)^2/2, s = s_\phi^{(0)}$) distribution where $s_\phi^{(0)}$ represents the initial value of s_ϕ and controls the standard deviation of ϕ. Next, these random $\vec{\phi}$ variates were rank ordered and assigned to the corresponding gene of the same SCUO rank. As a result, the rank order of a gene's initial ϕ_j value, $\phi_{j,0}$, was the same as the rank order of its SCUO value. We tried a variety of $s_\phi^{(0)}$ values and they all converged to similar parameter values. For the *with* $\vec{\Phi}$ model, we tried both the SCUO based approach and using the $\vec{\Phi}$ data to initialize our values of ϕ. In this second scenario, we set $\phi_j^{(0)} = \bar{X}_j^g / \sum_{i=1}^n \bar{X}_i^g$ where $\bar{X}_j^g$ represents the geometric mean of the observed mRNA abundances for gene j. As in the *without* $\vec{\Phi}$ case, we found the *with* $\vec{\Phi}$ chains consistently converged to the same region of parameter space independent of the initial ϕ values. It is worth noting that the structure of the probability function defined in Equation (6) is such that if the rank order of ϕ_i^0 were reversed from their true order, the model would converge to a similar quality of model fit and the signs of the parameters would change. Thus it is recommended that model fits be checked to ensure that the final estimates of ϕ for housekeeping genes, such as and ribosomal proteins, are much greater than 1.

Treating our initial protein synthesis rates ϕ for the entire genome as explanatory variables, the initial values for $\Delta\vec{M}$ and $\vec{\Delta\eta}$ were generated via multinomial logistic regression using the `vglm()` function of the **VGAM** package (Yee, 2013). We also used the covariance matrix returned by `vglm()` as the proposal covariance matrix for $\Delta\vec{M}$ and $\vec{\Delta\eta}$ for each amino acid. In order to make our random walk more efficient, we used an adaptive proposal function for all parameters in order to reach a target range of acceptance rates. For example, the covariance matrix of the step sizes was multiplied by a scalar value that was then increased or decreased by 20% every 100 steps when the acceptance rate of a parameter set was greater than 35% or less than 20%, respectively. The variance terms of the random walks for the $\vec{\phi}$ and the global parameter s_ϕ were also adjusted in a similar manner.

The results presented here were generated by running the MCMC algorithm for 10,000 iterations and, after examining the traces of the samples for evidence of convergence, selecting the last 5,000 iterations as our posterior samples. The arithmetic means of the posterior samples were used as point estimates based on the mean of our posterior samples. Posterior credibility intervals (CI) are generated by excluding the

lower and upper 2.5% of samples. Additional details on the model fit can be found in the Supporting Materials and in (Chen *et al.*, Prep). The code is implemented in an R package **cubfits** (Chen *et al.*, 2014) which is freely available for download at `http://cran.r-project.org/package=cubfits`.

Acknowledgments

We wish to acknowledge financial support for this project from NSF grants MCB-1120370 (M.A.G. and R.Z.) and EOB (Brian O'Meara, M.A.G., and R.Z.). Additional support was also provided by the National Institute for Mathematical and Biological Synthesis (NSF:DBI-1300426 with addtionall support from the University of Tennessee). We would also like to thank researchers in the RDAV group at the National Institute for Computational Sciences: George Ostrouchov, Drew Schmidt, and Pragnesh Patel who contributed to an earlier attempt to address this problem. Finally, we would like to thank W. Preston Hewgley, Cedric Landerer, Brian O'Meara, Ivan Erill, and Patrick O'Neill for their helpful discussions and suggestions.

References

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3): 927–935.

Andersson, S. G. and Kurland, C. G. 1990a. Codon preferences in free-living microorganisms. *Microbiol. Rev.*, 54(2): 198–210.

Andersson, S. G. E. and Kurland, C. G. 1990b. Codon preferences in free-living microorganisms. *Microbiological Reviews*, 54: 198–210.

Arava, Y. 2003. Genome-wide analysis of mRNA translation profiles in Saccharomycescerevisiae. *Proc Natl Acad Sci USA*, 100(7): 3889–3894.

Arava, Y., Wang, Y. L., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, 100: 3889–3894.

Arava, Y., Boas, F. E., Brown, P. O., and Herschlag, D. 2005. Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res.*, 33: 2421–2432.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D. M., Burston, H. E., Vizeacoumar, F. J., Snider, J., Phanse, S., Fong, V., Tam, Y. Y. C., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S.,

Punna, T., Christopolous, C., Wong, V., Yu, A., Zhong, G., Li, J., Stagljar, I., Conibear, E., Wodak, S. J., Emili, A., and Greenblatt, J. F. 2012. Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, 489(7417): 585–589.

Bennetzen, J. L. and Hall, B. D. 1982. Codon selection in yeast. *J Biol Chem*, 257(6): 3026–3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol*, 9: 675.

Bulmer, M. 1988. Are codon usage patterns in unicellular organisms determined by selectionmutation balance? *J Evol Biol*, 1(1): 15–26.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3): 897–907.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7(2): 98–108.

Chen, F., Gerber, S., Heuser, K., Korkhov, V. M., Lizak, C., Mireku, S., Locher, K. P., and Zenobi, R. 2013. High-mass matrix-assisted laser desorption ionization-mass spectrometry of integral membrane proteins and their complexes. *Anal. Chem.*, 85(7): 3483–3488.

Chen, W.-C., Zaretzki, R., Howell, W., Landerer, C., Schmidt, D., and Gilchrist, M. A. 2014. cubfits: Codon usage bias fits. R Package, `http://cran.r-project.org/package=cubfits`.

Chen, W.-C., Zaretzki, R., and Gilchrist, M. A. *In Prep.* cubfits: an R package for codon usage bias fits. *Bioinform.*

Clarke, B. 1970. Darwinian evolution of proteins. *Science*, 168: 1009–1011.

Curran, J. F. and Yarus, M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol*, 209(1): 65–77.

Drummond, D. A. and Wilke, C. O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2): 341–352.

Drummond, D. A. and Wilke, C. O. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*, 10(10): 715–724.

Durr, E., Yu, J., Krasinska, K. M., Carver, L. A., Yates, J. R., Testa, J. E., Oh, P., and Schnitzer, J. E. 2004. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nature Biotechnology*, 22(8): 985–992.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., and Cherry, J. M. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: Then and now. *G3: Genes—Genomes—Genetics*, 4(3): 389–398.

Fuller, W. A. 1987. *Measurement Error Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2): 907–911.

Gilchrist, M., Shah, P., and Zaretzki, R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183: 1493–1505.

Gilchrist, M. A. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol. Biol. Evol.*, 24: 2362–2373.

Gilchrist, M. A. and Wagner, A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J. theor. Biol.*, 239: 417–434.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, 8: R49–R62 ER.

Hershberg, R. and Petrov, D. A. 2008. Selection on codon bias. *Annu. Rev. Genet.*, 42: 287–299.

Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5): 717–728.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5): 1205–1214.

Ikemura, T. 1981a. 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 *Escherichia-coli* translational system. *J. Mol. Biol.*, 151: 389–409.

Ikemura, T. 1981b. 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(3): 389–409.

Ikemura, T. 1985. Codon usage and transfer-rna content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2: 13 – 34.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(5924): 218–223.

Kanaya, S., Yamada, Y., Kudo, Y., and 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(1): 143–155.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811): 525–528.

Knight, R. D., Freeland, S. J., and Landweber, L. F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*, 2(4): RESEARCH0010.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. 2009. Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science*, 324(5924): 255–258.

Kurland, C. G. 1987. Strategies for efficiency and accuracy in gene expression. *Trends Biochem Sci*, 12: 126–128.

Kurland, C. G. 1992. Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.*, 26: 29–50.

Li, G. W., Burkhardt, D., Gross, C., and Weissman, J. S. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157: 624–35.

Li, W.-H., Wu, C. I., and Luo, C. C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, 2(2): 150–174.

Lim, V. I. and Curran, J. F. 2001. Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. *RNA*, 7(7): 942–957.

Marin, J. and Robert, C. 2007. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. 2006. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881): 1344–1349.

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5): 418–426.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., and Kyrpides, N. C. 2012. The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata. *Nuc. Acids Res.*, 40(D1): D571–D579.

Palidwor, G. A., Perkins, T. J., and Xia, X. 2010. A general model of codon bias due to GC mutational bias. *PLoS ONE*, 5(10): e13431.

Pechmann, S. and Frydman, J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol*, 20(2): 237–243.

Plotkin, J. B. and Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12(1): 32–42.

Qin, H., Wu, W. B., Comeron, J. M., Kreitman, M., and Li, W. H. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168: 2245–2260.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci USA*, 107(10): 4629–4634.

Sella, G. and Hirsh, A. E. 2005. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U.S.A.*, 102: 9541–9546.

Shah, P. and Gilchrist, M. A. 2010. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet*, 6(9).

Shah, P. and Gilchrist, M. A. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. U.S.A.*, 108(25): 10231–10236.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. B. 2013. Rate-limiting steps in yeast protein translation. *Cell*, 153(7): 1589–1601.

Sharp, P. M. and Li, W. H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, 24: 28–38.

Sharp, P. M. and Li, W. H. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15: 1281 – 1295.

Sørensen, M. A. and Pedersen, S. 1991. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol*, 222(2): 265–280.

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K. C., Seizl, M., Tresch, A., and Cramer, P. 2012. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res*, 22(7): 1350–1359.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3): 1101–1115.

Thanaraj, T. A. and Argos, P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci*, 5(8): 1594–1612.

Tsai, C.-J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., and Nussinov, R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol*, 383(2): 281–291.

Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA*, 107(8): 3645–3650.

Wallace, E. W. J., Airoldi, E. M., and Drummond, D. A. 2013. Estimating selection on synonymous codon usage from noisy experimental data. *Mol. Biol. Evol.*, 30: 1438–1453.

Wan, X. F., Zhou, J., and Xu, D. 2006. Codono: a new informatics method for measuring synonymous codon usage bias within and across genomes. *Int. J. Gen. Syst.*, 35: 109–125.

Wasserman, W. W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4): 276–287.

Yang, Z. H. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17: 32–43.

Yang, Z. H. and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, 25: 568–579.

Yassour, M., Kapian, T., Fraser, H. B., Levin, J. Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S. J., Khrebtukova, I., Gnirke, A., Nusbaum, C., Thompson, D. A., Friedman, N., and Regev, A. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 106: 3264–3269.

Yee, T. 2013. VGAM: Vector generalized linear and additive models. R Package version 0.9-3.

Zaher, H. S. and Green, R. 2009. Fidelity at the molecular level: Lessons from protein synthesis. *Cell*, 136: 746–762.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. A. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci USA*, 111(22): E2310–8.

Supporting Materials

Supporting materials for *Estimating gene expression and codon specific translational efficiencies, mutation biases, and selection coefficients from genomic data* by Gilchrist *et al.* (In Review).

Model Validation using Simulated Data

In order to verify the reliability of the *with* and *without* $\vec{\Phi}$ model fits we apply both methods to simulated data. Data set $\mathbb{S}_1$, is generated from a model with ϕ values following a LogN distribution while $\mathbb{S}_2$ uses the estimates of ϕ obtained from our analysis of the S288c genome with $\vec{\Phi}$ data.

Analysis of both simulated datasets show that both the *with* and *without* $\vec{\Phi}$ methods produce accurate and unbiased estimates of the mutation bias parameters $\Delta\vec{M}$ under all circumstances ($\rho > 0.99$, Figures S1 & S2, panels c & d). We also obtained accurate estimates of differences in ribosome pausing times $\Delta\vec{\eta}$. Both *with* and *without* $\vec{\Phi}$ model fits produced near perfect recovery of $\Delta\vec{\eta}$ parameters when applied to simulated dataset $\mathbb{S}_1$ ($\rho > 0.99$,Figure S1, panels a & b).

When applied to simulated dataset $\mathbb{S}_2$, both *with* and *without* $\vec{\Phi}$ estimates of $\Delta\vec{\eta}$ showed strong agreement with parameter values ($\rho > 0.99$, Figure S2, panels a & b). We did, however, observe a small downward bias in their absolute values ($\sim 7\%$). This is a special case of attenuation bias (Fuller, 1987) which results from the ϕ values in $\mathbb{S}_2$ being distributed with a heavier right tail than the corresponding LogN distribution with the same mean and variance.

Comparing the *with* and *without* $\vec{\Phi}$ estimates of protein synthesis rates, e.g. the posterior means, $\bar{\phi}$, and the ϕ values used in our simulations illustrates the predictive power of our model. For example, analysis of the simulated dataset $\mathbb{S}_1$ indicates that under ideal conditions we observe correlation coefficients between the log of our protein synthesis estimates, $\log(\bar{\phi})$, and the log of their true values, $\log(\phi)$ of ~ 0.96 for both the *with* and *without* $\vec{\Phi}$ model fits (Figure S1). Even when the true distribution of ϕ values violates the LogN assumption as in $\mathbb{S}_2$, we still observe correlation coefficients between $\log(\bar{\phi})$ and $\log(\phi)$ of ~ 0.95 (Figure S2).

Scaling Bias due to Noise and Inherent Uncertainty

Because measurements of mRNA abundances, whether via microarray florescence or sequencing data, are usually not scaled to any particular unit, researchers often use either the sum of all the measurements or their mean value as a means of scaling their results. While it is intuitive to scale the data in this way, if the additional measurement noise is not taken into account a subtle biases on ϕ and $\Delta\vec{\eta}$ is introduced.

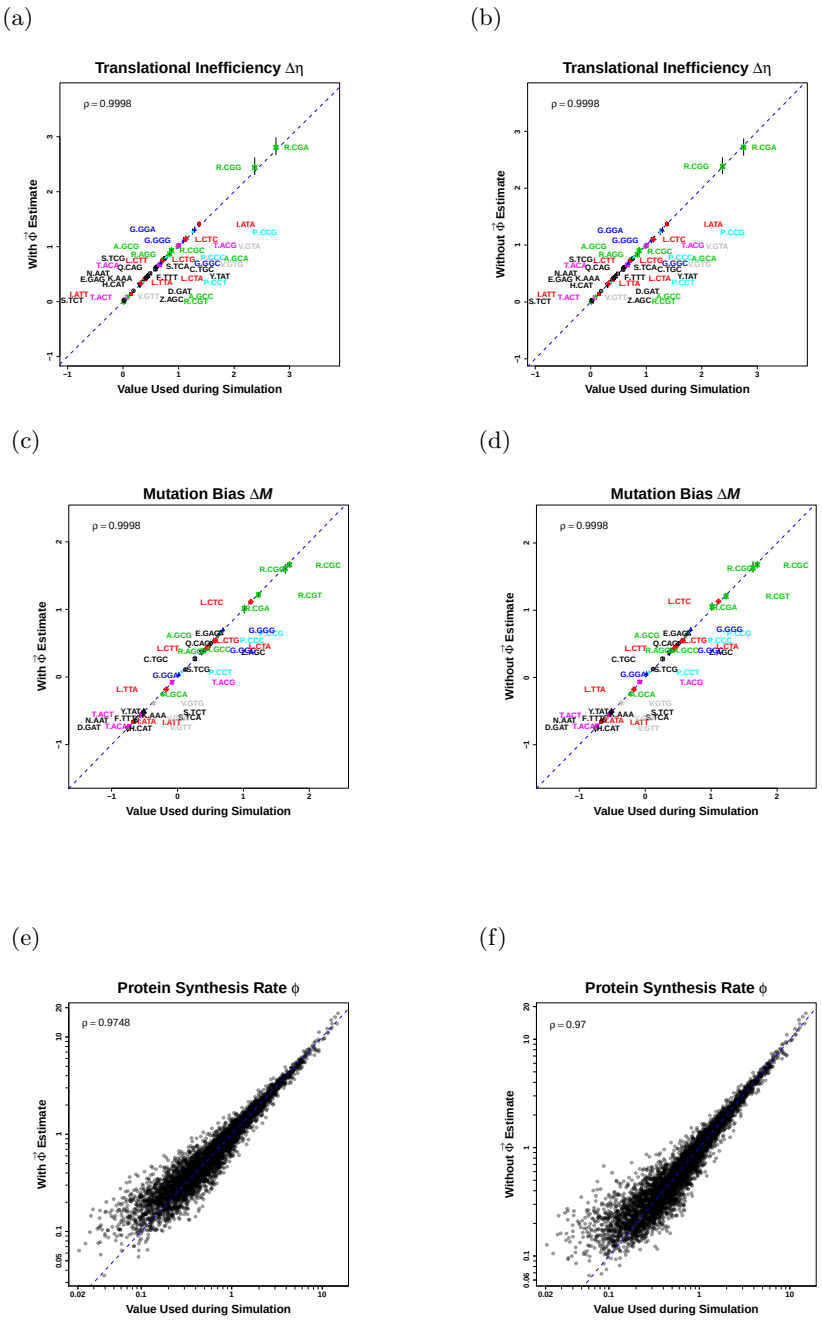


Figure S1: Comparison of estimated parameters versus actual parameters used to simulate data under the model. Here $\phi \sim$ LogN as assumed when fitting our model. (a) Comparison of *with* $\vec{\Phi}$ parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$. (b) Comparison of *without* $\vec{\Phi}$ parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$. (c) Comparison of *with* $\vec{\Phi}$ parameter estimates ΔM vs. actual data generating parameters ΔM. (d) Comparison of *without* $\vec{\Phi}$ parameter estimates ΔM vs. actual data generating parameters ΔM. (e) Comparison of *with* $\vec{\Phi}$ parameter estimates ϕ vs. actual data generating parameters ϕ. (f) Comparison of *without* $\vec{\Phi}$ parameter estimates ϕ vs. actual data generating parameters ϕ.

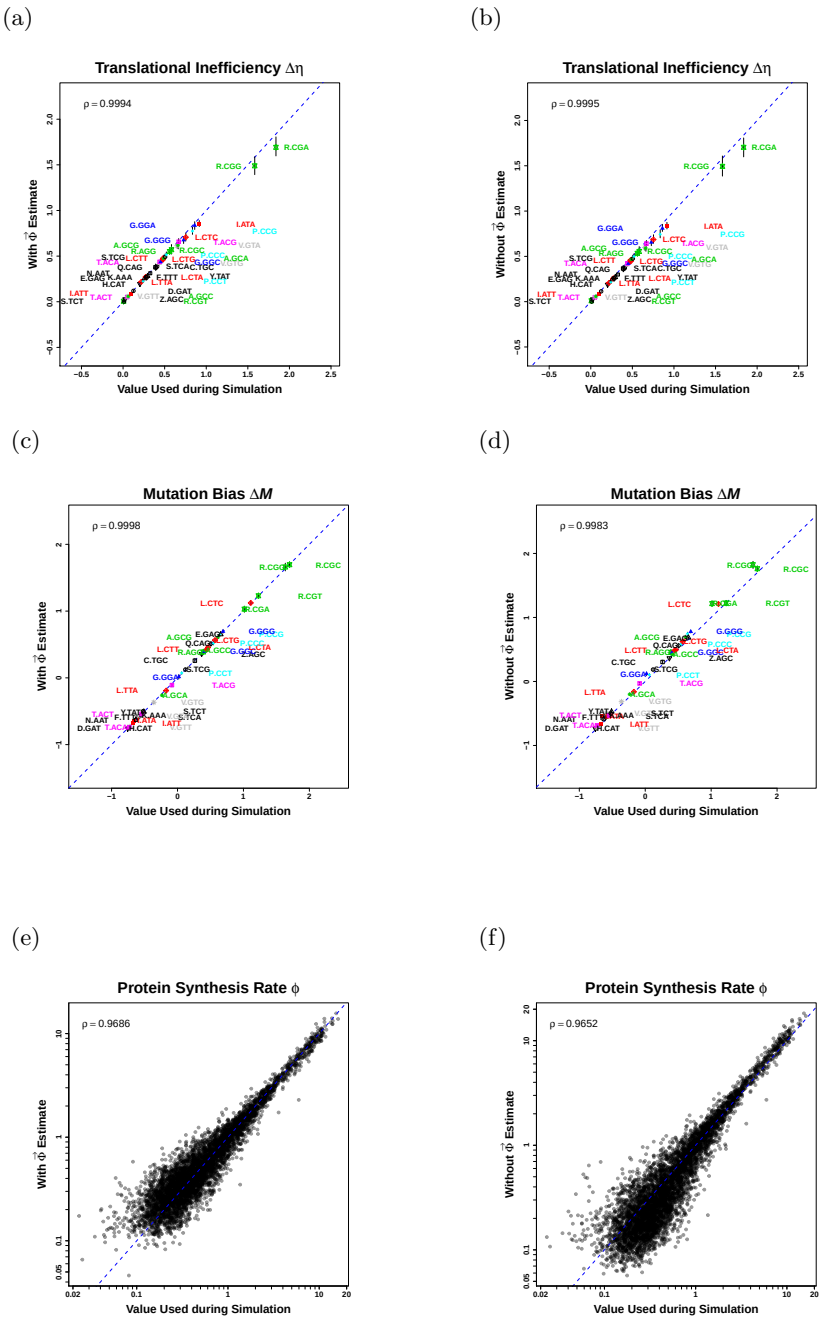


Figure S2: Comparison of estimated parameters versus actual parameters used to simulate data under the model. Here ϕ values used in the simulation were based on the *with* $\vec{\Phi}$ fit of the *S. cerevisiae* S288c genome dataset and, as a result, do not follow a LogNormal distribution as assumed when fitting our model: (a) Comparison of *with* $\vec{\Phi}$ parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$. (b) Comparison of *without* $\vec{\Phi}$ parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$. (c) Comparison of *with* $\vec{\Phi}$ parameter estimates ΔM vs. actual data generating parameters ΔM. (d) Comparison of *without* $\vec{\Phi}$ parameter estimates ΔM vs. actual data generating parameters ΔM. (e) Comparison of *with* $\vec{\Phi}$ parameter estimates ϕ vs. actual data generating parameters ϕ. (f) Comparison of *without* $\vec{\Phi}$ parameter estimates ϕ vs. actual data generating parameters ϕ.

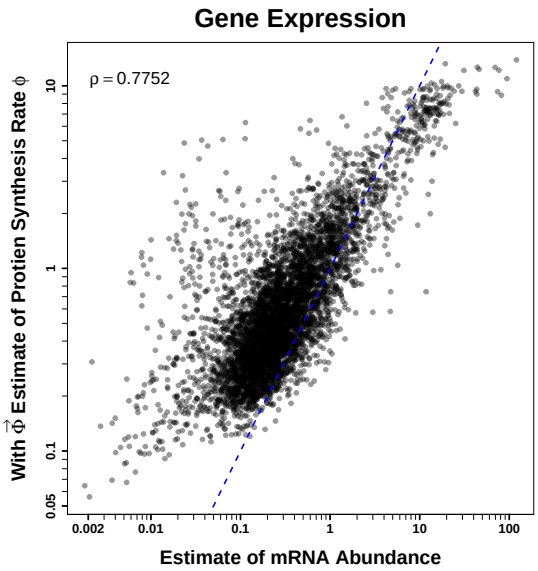


Figure S3: Comparison between posterior mean estimates of ϕ for the *with* $\vec{\Phi}$ model fit and $\vec{\Phi}$ data consisting of mRNA abundance measurements from Yassour *et al.* (2009).

The nature of the bias can be most easily illustrated when we assume that both the signal and the noise follow log-normal distributions, however, the effects should be present as long as the noise is not symmetrically distributed around the underlying signal values.

For example, let ϕ'_i represent the true, unscaled protein synthesis rate of gene i, i.e. $\ln(\phi'_i) = \ln(\phi_i) + A_\Phi$ and assume that, across the genome, $\phi' \sim \text{LogN}(m_{\phi'}, s_{\phi'})$, such that $E(\phi') = \exp\left[m_{\phi'} + s^2_{\phi'}/2\right]$. Let $\Phi_{i,j}$ represent a given noisy observation or estimate of ϕ'_i, i.e. $\Phi_{i,j}$ is part of our $\vec{\Phi}$ data set. Also let $\Phi_{i,j} = \phi'_i \varepsilon_j$ where $\varepsilon_j \sim \text{LogN}(0, s_\varepsilon)$ and implies that the observation $\Phi_{i,j}$ is log normally distributed around the true values ϕ'_i. Even though the noise is centered around the true value, because the log-normal distribution is asymmetric, $E(\Phi_i|\phi'_i) = \phi'_i \exp[s^2_\varepsilon/2] > \phi'_i$ and when considering the entire distribution $E(\Phi) = \exp\left[m_{\phi'} + s^2_{\phi'}/2 + s^2_\varepsilon/2\right] = E(\phi')\exp[s^2_\varepsilon/2]$. Thus we see that the mean of our observed values is actually greater than the mean of the true signals underlying them and, as a result, if one scales by the sum or the mean of these observed values the resulting values will be biased downward by a factor of $\exp[s^2_\varepsilon/2]$. To remove this bias, we introduce an additional scaling term A_Φ such that $m_{\phi'} = A_\Phi - s^2_{\phi'}/2$ and, as a result, $E(\phi') = \exp\left[A_\Phi\right]$ and $E(\Phi) = \exp\left[A_\Phi + s^2_\varepsilon/2\right]$. Our empirical data provides an estimate of $E(\Phi)$ and the inconsistency between the degrees of adaptation in CUB observed across genes and their expression levels greater than that expected due to genetic drift allows us to estimate s_ε while, simultaneously estimating A_Φ.

Finally, we note that simply scaling one's estimates of x by the mean of these estimates during the

MCMC run also introduces bias. This is because our estimates of ϕ_i' during the MCMC, Φ_{MCMC} are imprecise and, as a result, their mean value will be overestimated. Assuming our uncertainty in x is log-normally distributed $\text{LogN}(m = 0, s = s_{\text{MCMC}})$, $E(\Phi_{MCMC}) = E(\phi')E(s^2_{\text{MCMC}}/2)$. As a consequence, the scaled protein synthesis rates, ϕ, are biased downward leading to an overestimation in the absolute differences in pausing times between codons, $\vec{\Delta\eta}$. The effects of this bias are actually evident in Wallace *et al.* (2013) Figure 5A where the estimates of the coefficients differ from the values used during their simulations. Including the parameter A_Φ, which explicitly models this scaling terms, provides a simple way to avoid these issues.

Fitting of Model to Genomic Data and Noisy Measurements of Protein Synthesis

We generalize our model to include the extraction of information from noisy, unscaled measurements of protein synthesis for each gene, i.e. $\vec{\Phi}_j$. This is essentially the same model as Wallace *et al.* (2013) except instead of rescaling estimates of $\vec{\phi}$ and $\vec{\Delta\eta}$ in pre- and post-MCMC data processing step, we include the estimation of the scaling term A_Φ discussed in the last section.

$$\prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\vec{\Delta M}_i, \vec{\Delta\eta}_i, \phi_j, s_\phi, A_\Phi, s_\varepsilon^2 \middle| \vec{k}_{i,j}, n_{i,j}, \vec{\Phi}_j\right) \propto$$
$$\prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\vec{k}_{i,j}\middle|\vec{p}_{i,j}, n_{i,j}\right) f\left(\vec{\Phi}_j\middle|\phi_j, A_\Phi, s_\varepsilon^2\right) f\left(\phi_j|s_\phi\right) f\left(s_\phi\right) f\left(A_\Phi\right) f\left(s_\varepsilon^2\right) \tag{S1}$$

where, as before, $\vec{\Delta M}_i$ and $\vec{\Delta\eta}_i$ are the mutation and selection coefficients respectively for amino acid i, $\vec{k}_{i,j}$ are the codon counts following a multinomial distribution for the amino acid i in the ORF of gene j as defined in Equation (1), $n_{i,j}$ is the sum of all codon counts related to a particular amino acid i in the gene j, $\vec{p}_{i,j}$ is an inverse multinomial logit function of $\vec{\Delta M}_i$, $\vec{\Delta\eta}_i$, and ϕ_j, $f\left(\phi_j|s_\phi\right)$ is the prior for the protein synthesis rate $\phi_j \sim \text{LogN}(-s_\phi^2/2, s_\phi)$, and $f(s_\phi) = 1$.

Additionally, we assume that $\log\left(\vec{\Phi}_j\right) \sim \text{N}\left(\log\left(\phi_j\right) + A_\Phi, s_\varepsilon^2\right)$, i.e. the log transformed measurements $\log\left(\vec{\Phi}_j\right)$ are offset by a constant A_Φ and normally distributed around $\log(\phi) + A_\Phi$ with variance s_ε^2. We also assume $f(A_\Phi) = 1$ and $f(s_\varepsilon^2) \propto 1/s_\varepsilon^2$. Both A_Φ and s_ε^2 are genome scale parameters and are estimated in the *with* $\vec{\Phi}$ model. In the future, the assumption that s_ε^2 is the same across genes could be relaxed. In the absence of any $\vec{\Phi}$ data, the $f\left(\vec{\Phi}_j\middle|\phi_j, A_\Phi, s_\varepsilon^2\right)$, $f\left(A_\Phi\right)$, and $f\left(s_\varepsilon^2\right)$ terms are undefined and drop out.

The system below summarizes the expressions just given describing Equation (S1):

$$\begin{aligned}
\vec{k}_{i,j} &\sim \text{Multinom}(n_{i,j}, \vec{p}_{i,j}), \\
\vec{p}_{i,j} &= \text{mlogit}^{-1}(-\Delta\vec{M}_i - \Delta\vec{\eta}_i\phi_j), \\
\log\left(\vec{\Phi}_j\right) &\sim \text{N}(\log(\phi_j) + A_\Phi, s_\varepsilon^2), \\
\phi_j &\sim \text{LogN}(-s_\phi^2/2, s_\phi), \\
\Delta\vec{M}_i,\, \Delta\vec{\eta}_i,\, s_\phi,\, A_\Phi &\propto 1, \text{ and} \\
f(s_\varepsilon^2) &\propto 1/s_\varepsilon^2.
\end{aligned}$$

To fit the *without* and *with* $\vec{\Phi}$ models, we apply the following algorithm with a superscript (i) indicating the i^{th} iteration of an MCMC chain.

Step 1. Update $\Delta\vec{M}$ and $\Delta\vec{\eta}$ conditional on all other parameters in the i^{th} iteration through a random walk Metropolis-Hasting (MH) algorithm:

(a) Step $i = 0$ only.

i. Calculate SCUO value for each gene following Wan *et al.* (2006).

ii. Generate random ordered values $\phi^{(0)}$ by simulating from $\text{LogN}(m = -s_\phi^{2(0)}/2, s = s_\phi^{(0)})$, and sorting them in the same order as the SCUO values to maintain the rank order of production rates among genes.

iii. Given $\phi^{(0)}$, for each amino acid a estimate initial values $\Delta\vec{M}_a^{(0)}$, $\Delta\vec{\eta}_a^{(0)}$, and the covariance matrix of these estimates $\Sigma^{(0)}_{\Delta\vec{M}_a, \Delta\vec{\eta}_a}$ using multinomial logistic regression.

(b) For each amino acid, independently simulate a new proposal for $(\Delta\vec{M}_a, \Delta\vec{\eta}_a)$ jointly from a multivariate normal distribution which has mean $(\Delta\vec{M}_a^{(i)}, \Delta\vec{\eta}_a^{(i)})$ and covariance $c_a^{(i)}\Sigma^{(0)}_{(\Delta\vec{M}_a, \Delta\vec{\eta}_a)}$ with initial adaptive scaling factor $c_a^{(0)} = 1$. See Marin and Robert (2007, Chapter 2) for details on incorporating a covariance matrix in practice.

(c) Accept the proposal with the MH probability based on the acceptance ratio and set $\Delta\vec{M}_a^{(i+1)}$ and $\Delta\vec{\eta}_a^{(i+1)}$ accordingly for all amino acids.

Step 2. Update hyperparameters conditional on all other parameters:

(a) If using the fitting *with* $\vec{\Phi}$ model: update $s_\varepsilon^{2(i+1)} \sim \text{Inv-Gamma}((n_g - 1)/2, S^{2(i)}/2)$ where $S^{2(i)} = \sum_{j=1}^{n_g}(\log\vec{\Phi}_j - A_\Phi^{(i)} - \log\phi_j^{(i)})^2$.

(b) Update $s_\phi^{(i+1)}$ using a random walk MH with proposal distribution $\text{LogN}(\log s_\phi^{(i)}, \sigma_{s_\phi}^{(i)})$ with initial value $\sigma_{s_\phi}^{(0)} = 1$ for the adaptive scaling factor of MCMC. Also, set $m^{(i+1)} = -s^{2(i+1)}/2$.

(c) If fitting *with* $\vec{\Phi}$ model: update $A_\Phi^{(i+1)}$ using a random walk MH with proposal distribution $\text{N}(A_\Phi^{(i)}, \sigma_{A_\Phi}^{2(i)})$ with initial value $\sigma_{A_\Phi}^{(0)} = 0.1$ for the adaptive MCMC scaling factor.

Step 3. Update protein translation rates conditional on Steps 1 and 2 and all other parameters:

For each gene j, generate ϕ_j through a random walk MH step:

(a) Propose ϕ_j from $\text{LogN}(\phi_j^{(i)}, \sigma_{\phi_j}^{(i)})$ with initial value $\sigma_{\phi_j}^{(0)} = 1$ for the adaptive MCMC scaling factor.

(b) Accept the proposal with the MH probability based on the acceptance ratio and set $\phi_j^{(i+1)}$ accordingly.

Step 4. Update all adaptive scaling factors if the acceptance rate of each set of parameters falls outside the 20-30% acceptance rate in the above Steps 1, 2, and 3 in order to sample the posterior distribution efficiently.

Comparison of Predicted Protein Synthesis Rates ϕ to Independent mRNA Abundance Measurements

Figure S4 compares posterior mean estimates of ϕ produced *with* (using the mRNA abundance measurements of Yassour *et al.* (2009)) and *without* $\vec{\Phi}$ to four additional lab measurements of mRNA abundances reported by Arava (2003); Nagalakshmi *et al.* (2008); Holstege *et al.* (1998); Sun *et al.* (2012). These values can be found in Table S9. Correlation coefficients are provided for each figure and tend to be slightly higher for estimates generated using the *with* $\vec{\Phi}$ algorithm. Although this seems to indicate that *with* $\vec{\Phi}$ estimates are superior, it is worth noting that these data measure mRNA expression levels. Because the *without* $\vec{\Phi}$ algorithm estimates protein synthesis rates, fundamentally a different quantity, we would expect these estimates to differ. Because the *with* $\vec{\Phi}$ measurement algorithm shrinks the protein synthesis estimates toward the mRNA expression observations, it is natural that *with* $\vec{\Phi}$ estimates show higher correlation with measurements from other laboratories.

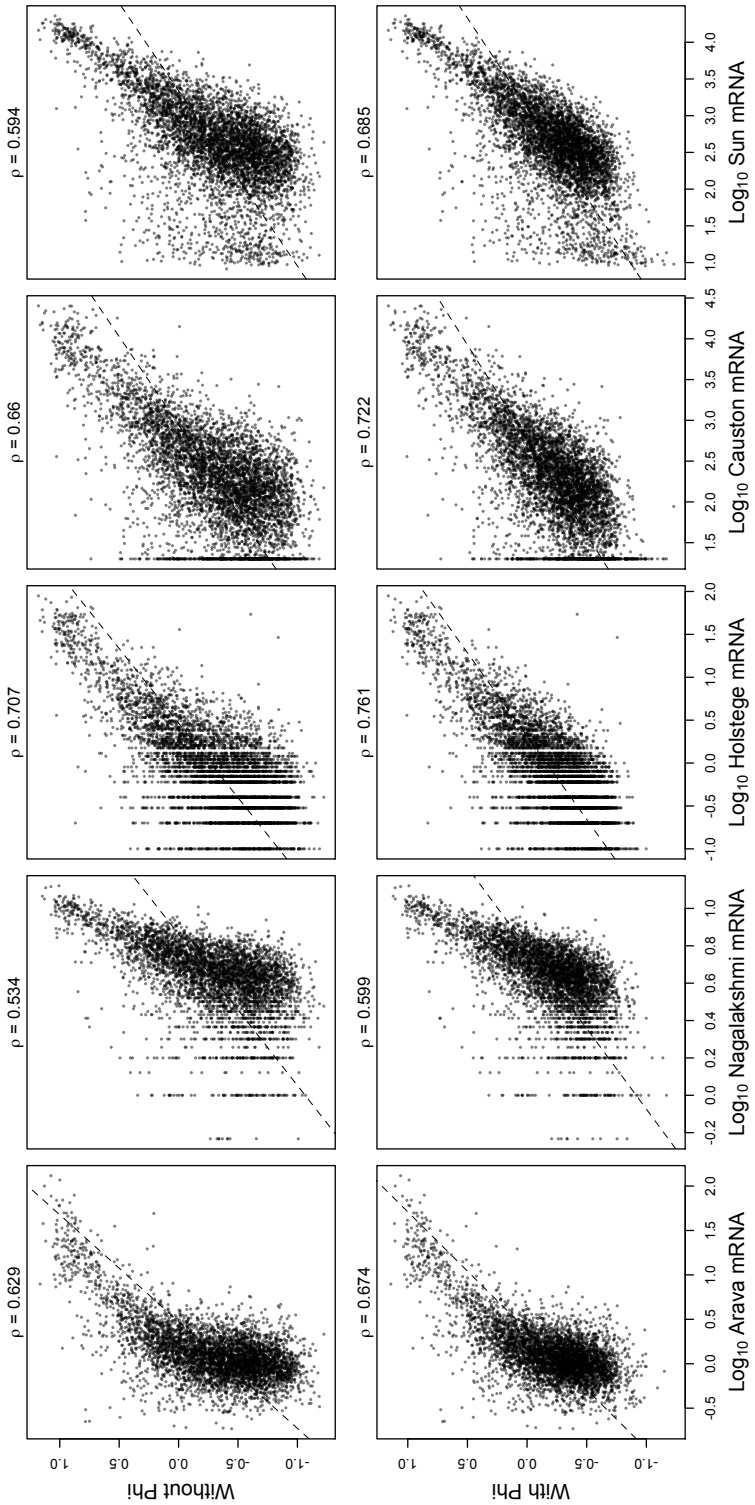


Figure S4: Scatter plot comparisons of *with* (Yassour measurements) and *without* $\vec{\Phi}$ posterior mean estimates to empirical measurements from four additional laboratories. Correlation coefficients between predictions and measurements are provided. *with* $\vec{\Phi}$ (Yassour) measurements consistently show slightly higher correlations.

Supplemental Tables

S1. Summary statistics of posterior estimates of ΔM for *S. cerevisiae* S288c genome estimated *with* $\vec{\Phi}$ (s288c_deltam_wphi.tsv).

S2. Summary statistics of posterior estimates of ΔM for *S. cerevisiae* S288c genome estimated *without* $\vec{\Phi}$ (s288c_deltam_wophi.tsv).

S3. Summary statistics of posterior estimates of $\Delta\eta$ for *S. cerevisiae* S288c genome estimated *with* $\vec{\Phi}$ (s288c_deltaeta_wphi.tsv).

S4. Summary statistics of posterior estimates of $\Delta\eta$ for *S. cerevisiae* S288c genome estimated *without* $\vec{\Phi}$ (s288c_deltaeta_wophi.tsv).

S5. Summary statistics of posterior estimates of ϕ for *S. cerevisiae* S288c genome estimated *with* $\vec{\Phi}$ (s288c_phi_wphi.tsv).

S6. Summary statistics of posterior estimates of ϕ for *S. cerevisiae* S288c genome estimated *without* $\vec{\Phi}$ (s288c_phi_wophi.tsv).

S7. Gene and codon specific selection coefficients for *S. cerevisiae* S288c genome estimated *with* $\vec{\Phi}$ (s288c_selection_coefficient_wphi.tsv).

S8. Gene and codon specific selection coefficients for *S. cerevisiae* S288c genome estimated *without* $\vec{\Phi}$ (s288c_selection_coefficient_wophi.tsv).

S9. Additional absolute mRNA measurements from multiple laboratories of *S. cerevisiae* Genome (s.cerevisiae.mRNA.measurements.tsv).