bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

Cedric L. Landerer

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

Russell Zaretzki

Department of Business Analytics and Statistics, University of Tennessee, Knoxville TN 37996

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

April 15, 2015

*Copy of manuscript and supporting materials archived at bioRχiv with **doi:** http://dx.doi.org/10.1101/009670
†Corresponding author: mikeg@utk.edu
‡Former address
§Current address

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Extracting biologically meaningful information from the continuing flood of genomic data is a major challenge in the life sciences. Codon usage bias (CUB) is a general feature of most genomes and is thought to reflect the effects of both natural selection for efficient translation and mutation bias. Here we present a mechanistically interpretable, Bayesian model (ROC SEMPPR) to extract biologically meaningful information from patterns of CUB within a genome. ROC SEMPPR, is grounded in population genetics and allows us to separate the contributions of mutational biases and natural selection against translational inefficiency on a gene by gene and codon by codon basis. Until now, the primary disadvantage of similar approaches was the need for genome scale measurements of gene expression. Here we demonstrate that it is possible to both extract accurate estimates of codon specific mutation biases and translational efficiencies while simultaneously generating accurate estimates of gene expression, rather than requiring such information. We demonstrate the utility of ROC SEMPPR using the *S. cerevisiae* S288c genome. When we compare our model fits with previous approaches we observe an exceptionally high agreement between estimates of both codon specific parameters and gene expression levels ($\rho > 0.99$ in all cases). We also observe strong agreement between our parameter estimates and those derived from alternative datasets. For example, our estimates of mutation bias and those from mutational accumulation experiments are highly correlated ($\rho = 0.95$). Our estimates of codon specific translational inefficiencies are tRNA copy number based estimates of ribosome pausing time ($\rho = 0.64$), and mRNA and ribosome profiling footprint based estimates of gene expression ($\rho = 0.53 - 0.74$) are also highly correlated, thus supporting the hypothesis that selection against translational inefficiency is an important force driving the evolution of CUB. Surprisingly, we find that for particular amino acids, codon usage in highly expressed genes can still be largely driven by mutation bias and 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, accessing this information does not require gene expression measurements, but instead carefully formulated biologically interpretable models.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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; Dunham *et al.*, 2012; Kundaje *et al.*, 2015), much of this information remains untapped. To address one aspect of this challenge, we present a method to estimate the 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 hypothesized selective forces affect the efficiency and efficacy of ORF translational through factors such as ribosome pausing times (Andersson and Kurland, 1990b; Bulmer, 1991; Sørensen and Pedersen, 1991; Plotkin and Kudla, 2011; Shah and Gilchrist, 2011a), 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), equalizing tRNA availability (Qian *et al.*, 2012), and the stability and/or accessibility of mRNA secondary structures (Kudla *et al.*, 2009; Tuller *et al.*, 2010; Gu *et al.*, 2012; 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) (see Methods).

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;

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Plotkin and Kudla, 2011). These differences, in turn, are predicted to lead to differences in ribosome pausing times and error rates between codons. Specifically, codons with higher abundances of cognate and near-cognate tRNAs are thought to have both shorter pausing times and lower error rates than codons with lower abundances of cognate and near-cognate tRNA (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 protein synthesis rate: 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, given sufficiently large N_e, such that high expression genes contain some signal of adaptation, the patterns of CUB observed within a genome should contain a significant amount of information about the average protein synthesis rate ϕ for a given gene. 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. 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 quantify or describe selection on synonymous codon usage.

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, because this view lacks an explicit population genetics framework it is likely overly simplistic and may lead to the misidentification of 'optimal' codons.

Phylogeny based models of protein evolution, some of which are derived from population genetics models, have also been used to generate estimates of codon-specific selection coefficients and mutation biases (Tamuri *et al.*, 2012; Rodrigue *et al.*, 2010; Yang and Nielsen, 2008). Other approaches have relied on intra-specific variation to make similar types of inferences (Keightley and Eyre-Walker, 2007; Lawrie *et al.*, 2013) or a combination of inter-specific divergence and intra-specific variation Akashi (1995). However, all of these approaches fail to disentangle how the contributions of mutation bias and natural selection change with gene expression. Furthermore, these models are either fitted independently across genes and thus estimate a large number of gene specific parameters from a relatively small amount of data or assume that the magnitude of selection is uniform across genes.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, 2011a; 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.*, 2009a; Li *et al.*, 2014, e.g). In contrast, the method proposed here does away with the necessity of having protein synthesis rate estimates (or their proxy) and *provides* estimates of the average protein synthesis rate for each gene, ϕ. Importantly, our method also provides estimates of codon specific mutation biases and translational inefficiencies, which is the additive contribution of a codon to the cost of protein synthesis.

Furthermore, we can combine our gene specific estimates of protein synthesis rates and codon specific translational inefficiencies to produce estimates of the strength of natural selection on synonymous substitutions on a gene by gene and codon by codon 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 estimates of 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, the commonly made assumption that all synonymous changes within a gene are neutral can bias values of dN/dS towards over-estimating the number of genes evolving under positive selection (Spielman and Wilke, 2015). 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 genome wide patterns of CUB using our Stochastic Evolutionary Model of Protein Production Rate (SEMPPR) (Gilchrist, 2007; Shah and Gilchrist, 2011a), we build on the Bayesian statistical advances of Wallace *et al.* (2013). Because the costs in our model can be interpreted as proportional differences in ribosome overhead costs (ROC) due to ribosome pausing, for simplicity we refer to the model formulated here as ROC SEMPPR (see Methods).

Using the *Saccharomyces cerevisiae* S288c genome as an example, we demonstrate that ROC SEMPPR can be used to accurately estimate differences in codon specific mutation biases and contributions to the protein synthesis cost-benefit function η *without* the need for gene expression data. ROC SEMPPR's codon specific estimates of mutation biases and translational inefficiencies generated without gene expression data match almost exactly those generated with gene expression data (Pearson correlation coefficient

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\rho > 0.99$ 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) (which was identified as the most reliable dataset out of five different mRNA abundance datasets by Wallace *et al.* (2013)).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). Further, our predictions show strong and significant correlations with measurements of mRNA abundance from four other labs and estimates of protein synthesis rates based on ribosome profiling data from three other labs (Supporting Figures S4 and S5).

By releasing our work as a stand alone package in R (see Chen *et al.* (2014)), researchers can potentially 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 of ROC SEMPPR 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 attribute this selection to the interaction of two underlying biological traits: difference between synonymous codons in their contribution to the cost-benefit ratio η for protein synthesis 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, $1/N_e$,

$$\Delta\eta_{i,j} = 2\, N_e\, q\, (\eta_i - \eta_j) \tag{1}$$

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 relative fitness cost of expending an extra 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 codon which makes the smallest additive contribution to η and is most favored by selection. Thus, $\Delta\eta_{i,1} = 0$ for $i = 1$ and $\Delta\eta_{i,1} > 0$ for $i > 1$. For notational simplicity, we will only include the subscripts when needed for clarity.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

At equilibrium, under the weak-mutation regime (Sella and Hirsh, 2005b; Shah and Gilchrist, 2011b; McCandlish and Stoltzfus, 2014), the expected frequency of observing a synonymous codon i (p_i) of an 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 unique codons

$$p_i = \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\,\phi\right]}{\sum_{j=1}^{n_a}\exp\left[-\Delta M_{j,1} - \Delta\eta_{j,1}\,\phi\right]}, \tag{2}$$

where $\Delta M_{i,1}$ is a unitless measure of codon specific mutation bias. Note that, as with $\Delta\eta$, $\Delta M_{i,1} = 0$ for $i = 1$ but, unlike with $\Delta\eta$, $\Delta M_{i,1}$ can be positive or negative. Further, because it relies on the stationary probability of observing a synonymous codon, ROC SEMPPR can only detect variation in mutation bias, not variation in absolute mutation rates. Additional model details can be found in the Methods and Materials.

The utility of Equation (2) is that it allows us to probabilistically link ROC SEMPPR's 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 gene specific protein synthesis rates, $\vec{\phi} = \{\phi_1, \phi_2, \ldots, \phi_{n_g}\}$, to the CUB patterns observed within and between ORF 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 the synonymous codon groups (TCA, TCC, TCG, TCT) and (AGC, AGT) for Ser requires at least one non-synonymous nucleotide substitution, we treated these two groups as if they were different amino acids, Ser_4 and Ser_2, respectively. So while strictly speaking, 18 of the canonical 20 amino acids use more than one codon, because we treat Ser as two separate amino acids, Ser_2 and Ser_4, for our purposes $n_{\text{aa}} = 19$. Assuming a log-normal 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 ROC SEMPPR model *without* $\vec{\Phi}$ data is the major advance of our work over Wallace *et al.* (2013). Tables with estimates of gene specific protein synthesis rates, $\vec{\phi}$, mutation biases, ΔM, and translational inefficiencies, $\Delta\eta$, based on ROC SEMPPR's posterior sampling for the *S. cerevisiae* genome can be found in the Supporting Materials.

Evaluating Model Parameter Estimates

Briefly, when fitted to the *S. cerevisiae* S288c genome, we find nearly perfect agreement between ROC SEMPPR's *with* and *without* $\vec{\Phi}$ estimates for codon specific protein synthesis translational inefficiencies, $\Delta\eta$, and mutation bias, ΔM (Pearson correlation $\rho > .99$ for both sets of parameters, see Figures 1) and

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2). We note that, with the exception Arginine's $\Delta\eta_{\mathrm{CGT,AGA}}$, the central 95% Credibility Intervals (CIs) for ROC SEMPPR's $\Delta\eta$ and ΔM parameters do not overlap with zero (see Supplemental Tables S1-S4). These results indicate that information on the genome scale parameters, $\Delta\vec{\eta}$ and $\Delta\vec{M}$ are robustly encoded and estimable from CUB patterns and that $\vec{\Phi}$, provides little additional information.

(Approximate Location of Figure 1)

(Approximate Location of Figure 2)

Instead of simply comparing our ROC SEMPPR model's *without* $\vec{\Phi}$ estimates of ΔM and $\Delta\eta$ to its *with* $\vec{\Phi}$ estimates, we can also compare these parameters to other data. Due to the detailed balance requirement of the stationary distribution of our population genetics model (Sella and Hirsh, 2005a), differences in ΔM values between codons that can directly mutate to one another will equal the log of the ratio of their mutation rates. Thus, our estimates of ΔM provide testable hypotheses about the ratio 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). These experimental estimates of mutation bias, ΔM^e, are calculated as

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

where $\frac{n_{i\to j}}{n_i}$ is the number of $i \to j$ mutations observed per n_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* ($\rho = 0.95$, Figure 3).

(Approximate Location of Figure 3)

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 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).

$$\Delta t_{i,j} = \frac{1}{\mathrm{tRNA}_i w_i} - \frac{1}{\mathrm{tRNA}_j w_j}. \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Specifically, tRNA_i is the gene copy number of the tRNA that recognize codon i and w_i is the wobble penalty between the anti-codon of tRNA_i and codon i. When a codon is recognized by its canonical tRNA, we set $w_i = 1$. 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) and Lim and Curran (2001). We find that our genome-wide estimates of Δt are positively correlated with empirical estimates of Δt in *S. cerevisiae* ($\rho = 0.64$, Figure 4).

(Approximate Location of Figure 4)

Predicting Protein Synthesis Rates

Given the strong correlation between ROC SEMPPR's *with* and *without* $\vec{\Phi}$ estimates of the codon specific mutation biases $\Delta\vec{M}$ and translational inefficiencies $\Delta\vec{\eta}$, it is not surprising that *with* and *without* $\vec{\Phi}$ estimates of ϕ from ROC SEMPPR are highly correlated ($\rho = 0.99$, Figure 5(a)). More importantly, the *without* $\vec{\Phi}$ based estimates of ϕ show substantial correlation with the mRNA abundance based estimates of $\vec{\Phi}$ values from Yassour *et al.* (2009) ($\rho = 0.72$, Figure 5 (b)). To be clear, these $\vec{\Phi}$ values are the same values used as inputs to the *with* $\vec{\Phi}$ model fits.

Supporting Figures S4 and S5 explore this issue further by plotting ROC SEMPPR's posterior mean estimates of ϕ produced *with* and *without* $\vec{\Phi}$ against eight sets of experimental data. This data includes three genome wide estimates based on ribosome-profiling (RPF) measurements (Ingolia *et al.*, 2009b; Artieri and Fraser, 2014; McManus *et al.*, 2014) and five other genome wide estimates of mRNA abundances (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* $\vec{\Phi}$ ROC SEMPPR fits and measured mRNA abundances range from 0.534 to 0.707, and measured RPF reads range from 0.629 to 0.742. 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* $\vec{\Phi}$ ROC SEMPPR predictions of ϕ. Similarly, correlation between ϕ estimates for ROC SEMPPR's *with* $\vec{\Phi}$ fits and RPF reads provide a 6% to 12% increase in explanatory power over its *without* $\vec{\Phi}$ predictions of ϕ.

(Approximate Location of Figure 5)

Changes in CUB with Protein Synthesis Rate

As first shown in Shah and Gilchrist (2011a), the relationship between codon usage and protein synthesis rate ϕ can range from simple and monotonic to complex. Figure 6 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Both ROC SEMPPR's *with* and *without* $\vec{\Phi}$ model fits accurately predict how CUB changes with protein synthesis rates (Figure 6). Indeed, the predicted changes in CUB between the *with* and *without* $\vec{\Phi}$ ROC SEMPPR 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.

Changes in codon frequency with ϕ are the result of a subtle interplay between natural selection for reducing η and mutation bias (Figure 6). 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 preferred 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.

(Approximate Location of Figure 6)

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 if $\bar{w} > 1$, 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 the selective advantage of codon i over the other codons is positive. For the reference codon 1 where, by definition, $\Delta\eta_{i,1} \geq 0$, we see that this inequality *always* holds. This criteria can only be met by the non-reference codon in amino acids with more than two synonyms and when there are other non-reference codons with lower fitnesses at appreciable probabilities. In the *S. cerevisiae* S288c genome, these conditions can occur 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 can increase with production rate has not been widely recognized by biologists.

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 (2011a). 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 6 where the changes in CUB predicted from the standard logistic

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (2011a) predicted codon AGC would be favored by selection over AGT, both of ROC SEMPPR's *with* and *without* $\vec{\Phi}$ fits predict the opposite. Although, this switch in order is 'significant' in that the 95% CI for $\Delta\eta_{AGT,AGC}$ is < 0, the amino acid Ser_2 is used at very low frequency in high expression genes and its 97.5% CI boundary lies very close to 0. (The upper boundary lies at 0.00387 and 0.000634 for the *with* and *without* $\vec{\Phi}$ ROC SEMPPR fits, respectively.) As a result, this discrepancy is not strongly supported and warrants further investigation.

In summary, for 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, and Arg) selection for reducing η in high expression genes is substantially tempered by the force of mutation bias.

Estimating Selection on Synonymous Codon Usage

The assumptions of the ROC SEMPPR 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 on synonymous codon j in comparison to codon i in a gene with an average protein synthesis rate ϕ is,

$$S\left(\Delta\eta_{i,j}, \phi\right) = -\Delta\eta_{i,j}\,\phi. \tag{5}$$

We remind the reader that $\Delta\eta$ includes the effective population size, N_e, in its definition. As a result, our selection coefficients S are measured relative to the strength of genetic drift, $1/N_e$, as is commonly done. The distribution of S across all genes for each alternative to an amino acid's reference codon are illustrated in Figure 7 and summarized in Table 1. Tables with genome wide gene and codon specific estimates of S can found in the Supporting Materials. Recall that S is scaled by ϕ and that the distribution of ϕ values across genes appears to follow a heavy tailed distribution. As a result even though, by definition, the average value of ϕ is 1, the large majority of genes have ϕ values less than 1. As a result, although purifying selection on synonymous codons is universal, its selection coefficients are usually quite small (i.e. > -0.5). Nevertheless, because our framework utilizes information on CUB held across genes, we

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

can clearly detect the signature of selection at the genome level, specifically in the form of $\Delta\eta$ values whose posterior credibility intervals differ from 0, while other approaches might fail.

(Approximate Location of Figure 7)

(Approximate Location of Table 1)

Figure 8 compares our *without* $\vec{\Phi}$ ROC SEMPPR based estimates of S to those estimated using the FMutSel phylogenetic model of Yang and Nielsen (2008) using PAML (Yang, 2007) for the 106 genes in the Rokas *et al.* (2003) dataset. Overall we observe reasonable qualitative agreement between the two models with the majority of codon specific predictions having correlation coefficients $\rho > 0.3$. Unfortunately, while PAML provides maximum likelihood point estimates of parameters, it does not provide any confidence intervals for these parameters. Given the large number of parameters (> 60) estimated from each coding sequence by FMutSel, the confidence intervals for each parameter is likely to be large and, hence, could explain much of the variation we observe between ROC SEMPPR and FMutSel parameter estimates. Nonetheless, for 85% of the codons examined (34/40), we observe is a significant ($p < 0.05$) and positive linear relationship between the ROC SEMPPR and the FMutSel estimates of S (see Table S11). Of the remaining 6 codons, half exhibit a positive, but non-significant relationship between ROC SEMPPR and FMutSel's estimates of S, while the other half exhibit a negative, but again non-significant, relationship between estimates of S. Thus for 92% of the codons, both the ROC SEMPPR and FMutSel estimates of S agree qualitatively.

The three exceptions to this qualitative agreement are codons CGT (Arg), TCT (Ser_4), and ACT (Thr) and it is worth noting two points. First, the central 95% CI for CGT (Arg) overlaps with 0 in both the *with* and *without* $\vec{\Phi}$ ROC SEMPPR model fits. Second, the Ser_4 codon TCT and Thr codon ACT are two of the four codons that ROC SEMPPR indicates have been misidentified as 'optimal' codons in the past. Relative to the ROC SEMPPR reference codons, TCT and ACT have small $\Delta\eta$ values, ~ 0.01, and ~ 0.05 respectively, and large ΔM values, ~ -0.5 for both. Thus, it appears in these last two cases the FMutSel model is misattributing the CUB towards these codons to selection rather than mutation (see Figure 6).

(Approximate Location of Figure 8)

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. 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

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

et al., 2012; Rodrigue *et al.*, 2010; Yang and Nielsen, 2008). While important advances, these models fail to leverage information on CUB encoded across genes. In contrast, ROC SEMPPR estimates selection coefficients and other key parameters by assuming a common directionality of selection on CUB, but where the strength of selection varies with protein synthesis rate.

As a result, ROC SEMPPR provides a modeling framework which can quickly extract information on codon specific translational inefficiencies $\Delta\eta$, mutation biases ΔM, and gene specific estimates of protein synthesis rates $\vec{\phi}$, using only genome wide patterns of CUB. This ability stems from the hypotheses that the intergenic variation in patterns of CUB observed within a genome reflect a lineage's evolutionary responses to selection against inefficient protein translation as well as 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 for *S. cerevisiae* ROC SEMPPR's *without* $\vec{\Phi}$ 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 for the near future.

Of course ROC SEMPPR may not work for all organisms. For example, some organisms may evolve under N_e values too small for adaptation in CUB to occur. Under these conditions, our method should fail to confidently identify the selectively preferred codon (i.e. our credibility intervals for our $\Delta\eta$ parameters will overlap with 0). However, because our estimates of $\Delta\eta$ are based on the analysis of the entire genome simultaneously rather than the combination of independent assessments of individual genes, our method may be able to detect the signature of selection on CUB in organisms where it previously went undetected. Alternatively, there may be organisms where N_e is so large that, as a result, there is not enough variation in CUB to reliably estimate our parameters. Assuming we retain our flat priors on $\Delta\eta$, in these cases we expect our estimates of $\Delta\eta$ for the most selectively favored codons to continually increase in magnitude rather than eventually stabilizing. Such behavior reflects a lack of information in the data rather than a flaw in our model and has been observed in other approaches, such as those using inter- and intra-specific variation (Yang and Nielsen (2008) and Lawrie *et al.* (2013), respectively). ROC SEMPPR may also fail to work with organisms whose adaptation in CUB are driven by more complex or less consistent selective forces. If these forces are uncorrelated across amino acids within a gene or

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

varied greatly with position within a gene, then our method should not be able to confidently identify the selectively preferred codons, similar to the case with of organisms with small N_e.

While direct, codon specific estimates of ΔM and $\Delta\eta$ do not exist, data from mutation accumulation lines and tRNA copy number can be used as proxies. Reassuringly, we observe strong and consistent agreement between ROC SEMPPR's parameter estimates and these proxies. In addition, when comparing ROC SEMPPR's estimates of S to the FMutSel model we observe general agreement between our estimates with the three key exceptions likely due to relatively small differences in translational inefficiencies between these synonymous codons and their most efficient alternative and strong mutation bias against the most efficient, which is misinterpreted by FMutSel as selection. In contrast, the selective values S on synonymous codon usage we estimate using ROC SEMPPR are substantially smaller than those estimated by Lawrie *et al.* (2013) based on intra-specific variation in four fold degenerate sites for *Drosophila melanogaster*. While we have no immediate explaination for these differences we do note that Lawrie *et al.* (2013) acknowledge that the high S values they estimate are the exception rather than the rule for population genetic studies, including those looking at non-synonymous substitutions.

Because of ROC SEMPPR's derivation from population genetics, it should be possible to take any observed intra-specific variation into account by expanding our codon counts likelihood function in equation (8) to be calculated across the polymorphic alleles in proportion to their frequencies. Further, given that the directionality of selection in ROC SEMPPR is estimated using information from across the genome, our ability to detect site specific violations of the model should be much greater than when analyzing the CUB of each gene separately. Expanding ROC SEMPPR to utilize inter-specific variation, however, is more complex and will require expanding the model to include the effects of non-synonymous substitutions and phylogenetic history.

For organisms that can be cultured in the laboratory, researchers can utilize experimental techniques to measure mRNA, ribosome profiles, 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 ($\bar{s_\varepsilon} = 0.929$) indicates that the error in a given measurement ranges over an order of magnitude. In terms of protein abundance measurements, most proteomic studies have difficulty quantifying membrane bound proteins [Durr *et al.* (2004); Babu *et al.* (2012); Chen *et al.* (2013)]. Furthermore, both transcriptomic and proteomic measurements are, by their very nature, restricted to the specific growth conditions used. Unfortunately, the frequency with which organisms outside of the lab encounter such conditions is generally unknown. This is particularly important for understanding a pathogenic organism, where expression of genes involved in its persistence

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and spread are highly dependent on their hosts and are difficult to mimic *in vitro*.

The predictions of protein synthesis rates $\vec{\phi}$ generated by ROC SEMPPR 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 incongruences between mRNA abundance measurements and ϕ estimates may indicate genes undergoing extensive post-transcriptional regulation, a hypothesis that can be evaluated experimentally.

The fact that the additional information provided by the $\vec{\Phi}$ data from Yassour *et al.* (2009) leads to a relatively small increase in the quality of our predictions of $\vec{\Phi}$ data from other labs may seem surprising. However, we believe this behavior indicates that the information in $\vec{\Phi}$ about gene specific protein synthesis rates is largely redundant with the information held within the CUB patterns within a gene and across a genome.

Of course a skeptic might proffer a different interpretation, i.e. the model is somehow ignoring or insensitive to the information in $\vec{\Phi}$. We, however, believe this is not the case for the following reasons. First, ROC SEMPPR was carefully formulated to combine the information from independent $\vec{\Phi}$ measurements and the CUB of each gene in a straightforward and logical manner (See Supporting Materials: Fitting of Model to Genomic Data and Noisy Measurements and Equation (S1) in particular). Instead of *a priori* assuming one source of information is better than the other, ROC SEMPPR actually evaluates the relative quality of each source of information in explaining the observed bias in codon usage for a gene across the n_{aa} amino acids. Next, given the fact that the 95% Posterior Credibility Intervals for $\Delta\eta$ differ for at least one pair of codons for each amino acid indicates that the information held within the CUB patterns is reliable. In contrast, ROC SEMPPR's estimate of the error in $\vec{\Phi}$ indicates that the empirical measurements $\vec{\Phi}$ are noisy, consistent with the findings from other studies. For example, Wallace *et al.* (2013) looked at the the correlation in $\vec{\Phi}$ measurements between independent labs and found non-trivial disagreement in their values. Finally, and perhaps most convincingly, the *without* $\vec{\Phi}$ version of ROC SEMPPR treats the ϕ values as missing values and is able to predict their values to a similar level of accuracy observed between empirical measurements from different laboratories and using different platforms (Supporting Figures S4 and S5).

Accessing information on $\vec{\Phi}$ using a mechanistic, model based approach as developed here has addi-

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tional, 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. In contrast, our estimates of codon-specific translation inefficiencies $\Delta\eta$ and expression levels $\vec{\phi}$ 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) or polymorphism data (McDonald and Kreitman, 1991). These tests generally assume that synonymous changes are neutral. However, (Spielman and Wilke, 2015) have recently shown, ignoring selection on synonymous changes can lead to a false positive signal of a gene evolving in response to diversifying 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 (Spielman and Wilke, 2015).

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_4, 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, if our inferences are correct, has led to misidentification of the 'optimal' codon.

In addition to codon-specific translation inefficiencies $\Delta\eta$, 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 (2011a), 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 evolution 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 (Gilchrist *et al.*, 2009). 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. Finally, the pausing time at a codon is also influenced by several factors such as downstream mRNA folding (Yang *et al.*, 2014), presence of polybasic stretches (Brandman *et al.*, 2012) as well as co-translational folding of the growing polypeptide (Thanaraj and Argos, 1996; Pechmann and Frydman, 2013). While the contributions of these factors to ribosomal pausing times are often idiosyncratic and vary widely between genes, they can all influence the cost-benefit ratio η. The situation becomes even more complex and non-linear when considering how nonsense and missense errors along with various factors influencing pausing time costs 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 of 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods and Materials

Modeling Natural Selection on Synonymous Codons

Following the notation and framework introduced in Gilchrist (2007) and Shah and Gilchrist (2011a), 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 or expected protein synthesis rate across the genome is one, i.e. $E(\phi) = 1$. The cost-benefit ratio $\eta(\vec{c})$ represents the expected cost, in ATPs, to produce one functional protein from the coding sequence $\vec{c} = \{c_1, c_2, \ldots c_n\}$ where c_i represents the codon used at position i in a protein of length n. In its most general form, $\eta(\vec{c}) = E(\text{Cost}|\vec{c})/E(\text{Benefit}|\vec{c})$, where $E(\text{Cost})$ is the expected direct and indirect energetic costs incurred by a cell when a ribosome initiates translation of a transcript containing $\vec{c}$. Similarly, $E(\text{Benefit}|\vec{c})$ is the expected benefit, relative to a complete and error free protein, received by a cell when a ribosome initiates translation of a transcript containing $\vec{c}$. By definition, in the absence of translation errors, ribosomes will only produce complete and error free proteins, i.e. for ROC SEMPPER $E(\text{Benefit}) = 1$. Thus any differences in η are the result of differences in $E(\text{Cost})$ between alternative $\vec{c}$'s and $E(Cost)$ simplifies to $a_1 + \sum_{i=1}^{n}(a_2 + v\, t(c_i))$ where a_1 is the direct and indirect cost of translation initiation, a_2 is the direct cost of peptide elongation (4 ATPs/amino acid), $t(c_i)$ is the average pausing time a ribosome takes to translate codon c_i, and v scales this indirect cost of ribosome pausing from units of time to ATPs. Based on these definitions, $\eta(\vec{c})\phi$ 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,

$$W(\vec{c}) \propto \exp\left[-q\,\eta(\vec{c})\,\phi\right]. \tag{6}$$

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 $\vec{c}$ is independent between positions. As a result, the information held within $\vec{c}$ can be summarized by the number of times each synonymous codon is used within $\vec{c}$. 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 within $\vec{c}$ will follow a multinomial distribution with the probability vector $\vec{p} = \{p_1, \ldots, p_{n_a}\}$. Here, for $i = 1, \ldots, n_a$,

$$p_i\left(\Delta\vec{M}, \Delta\vec{\eta}, \phi\right) = \frac{\exp\left[-\Delta M_{i,1} - \Delta\eta_{i,1}\phi\right]}{\sum_{j=1}^{n_a}\exp\left[-\Delta M_{j,1} - \Delta\eta_{j,1}\phi\right]} \tag{7}$$

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, 2005a; Shah and Gilchrist, 2011a; Wallace *et al.*, 2013). Following Sella and Hirsh (2005a), for $N_e \gg 1$, 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. 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 ROC SEMPPR's *with* and *without* $\vec{\Phi}$ model fits.

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

$$\prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\Delta\vec{M}_i, \Delta\vec{\eta}_i, \phi_j, s_\phi \middle| \vec{k}_{i,j}\right) \propto \prod_{i=1}^{n_{\text{aa}}}\prod_{j=1}^{n_g} f\left(\vec{k}_{i,j} \middle| \vec{p}_{i,j}\left(\Delta\vec{M}, \Delta\vec{\eta}, \phi\right), n_{i,j}\right) f\left(\phi_j|s_\phi\right) f\left(s_\phi\right) \tag{8}$$

where the likelihood of 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 (7), $\vec{p}_{i,j}$ 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

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

(Approximate Location of Figure 9)

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 log-normal 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 ROC SEMPPR's *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 $\Delta\vec{\eta}$ 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).

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

In order to generate reasonable starting values for $\vec{\phi}$ in the *without* $\vec{\Phi}$ version of ROC SEMPPR, 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 $\text{LogN}(m = -\left(s_{\phi}^{(0)}\right)^2/2, s = s_{\phi}^{(0)})$ 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* $\vec{\Phi}$ ROC SEMPPR model fit, 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 (7) 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 between 20 and 35%. 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`.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Estimating Selection Coefficients using FMutSel

In order to evaluate the consistency of our estimates of $S = -\Delta\eta\,\phi$ with other approaches, we used the dataset from Rokas *et al.* (2003) which consisted of 106 aligned genes from 8 yeast species. Details of the model fitting can be found in Kubatko *et al.* (view)(available at `bioRxivdoi:http://dx.doi.org/10.1101/007849`), but briefly, we used the maximum likelihood tree found by Rokas *et al.* (2003) and then generated MLEs of of the stationary probability of a given codon under the FMutSel model from Yang and Nielsen (2008) using CODONML in PAML 4.4 (Yang, 2007). Using the same notation as in Yang and Nielsen (2008) we have,

$$\pi_J = \pi_{j_1}\pi_{j_2}\pi_{j_3}\exp[F]$$

where, for a given gene, π_J represents the stationary probability of observing codon J given nucleotide specific mutational bias terms π_{j_1}, π_{j_2}, and π_{j_3} and where $F = \ln(\text{Fitness})2N_e$. It follows that the comparable selection coefficients on synonymous codon usage relative to our reference codon 1 is,

$$S_{\text{YN}} = \Delta F_{i,1} = F_i - F_1 = \ln\left(\pi_I/\pi_1\right) + \ln\left(\pi_{j_1}\pi_{j_2}\pi_{j_3}/\pi_{1_1}\pi_{1_2}\pi_{1_3}\right) \tag{9}$$

A list of these parameter estimates can be found in the Supporting Materials.

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 additional support from the University of Tennessee). We are grateful to the 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. We would also like to thank W. Preston Hewgley, Brian O'Meara, Ivan Erill, and Patrick O'Neill for their helpful discussions and suggestions and Laura Kubatko for providing the FMutSel output. Finally, we like to thank our two anonymous reviewers whose comments and suggestions greatly improved the quality of this article.

References

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

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in drosophila DNA. *Genetics*, 139: 1067–1076 ER.

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., *et al.* 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.

Artieri, C. G. and Fraser, H. B. 2014. Evolution at two levels of gene expression in yeast. *Genome Res*, 24(3): 411–421.

Babu, M., Vlasblom, J., Pu, S., *et al.* 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.

Brandman, O., Stewart-Ornstein, J., Wong, D., *et al.* 2012. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell*, 151(5): 1042–1054.

Bulmer, M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation 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., *et al.* 2013. High-mass matrix-assisted laser desorption ionization-mass spectrometry of integral membrane proteins and their complexes. *Anal. Chem.*, 85(7): 3483–3488.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chen, W.-C., Zaretzki, R., Howell, W., *et al.* 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.

Dunham, I., Kundaje, A., Aldred, S. F., *et al.* 2012. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414): 57–74.

Durr, E., Yu, J., Krasinska, K. M., *et al.* 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., *et al.* 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gu, W. J., Wang, X. F., Zhai, C. Y., Xie, X. Y., and Zhou, T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. *Mol. Biol. Evol.*, 29: 3037–3044.

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., *et al.* 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. 2009a. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(5924): 218–223.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. 2009b. 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.

Keightley, P. D. and Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177: 2251–2261.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., *et al.* 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811): 525–528.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Kubatko, L. S., Shah, P., Herbei, R., and Gilchrist, M. In Review. A codon model of nucleotide substitution with selection on synonymous codon usage. *Mol. Phylogenet. Evol.*

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.

Kundaje, A., Meuleman, W., Ernst, J., *et al.* 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539).

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.

Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. 2013. Strong purifying selection at synonymous sites in d. melanogaster. *PLoS Genet.*, 9.

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.

McCandlish, D. M. and Stoltzfus, A. 2014. Modeling evolution using the probability of fixation: History and implications. *Q. Rev. Biol.*, 89(3): 225–252.

McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328): 652–654.

McManus, C. J., May, G. E., Spealman, P., and Shteyman, A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res*, 24(3): 422–430.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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., *et al.* 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., *et al.* 2012. The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic 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.

Qian, W., Yang, J.-R., Pearson, N. M., Maclean, C., and Zhang, J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet*, 8(3): e1002603.

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.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425: 798–804.

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

Sella, G. and Hirsh, A. E. 2005b. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA*, 102(27): 9541–9546.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. 2011a. 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. and Gilchrist, M. A. 2011b. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci USA*, 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.

Spielman, S. J. and Wilke, C. O. 2015. The relationship between dn/ds and scaled selection coefficients. *Mol. Biol. Evol.*

Sun, M., Schwalb, B., Schulz, D., *et al.* 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., *et al.* 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, J.-R., Chen, X., and Zhang, J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol*, 12(7): e1001910.

Yang, Z. H. 2007. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24: 1586–1591.

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., *et al.* 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

				Quantile						
AA	Codon	$\bar{S}$	var(S)	1%	5%	25%	50%	75%	95%	99%
A	GCA	-0.4743	0.6171	-4.3994	-1.7817	-0.4715	-0.2096	-0.1135	-0.0620	-0.0466
A	GCC	-0.0648	0.0115	-0.6014	-0.2436	-0.0645	-0.0287	-0.0155	-0.0085	-0.0064
A	GCG	-0.6420	1.1306	-5.9551	-2.4117	-0.6382	-0.2837	-0.1536	-0.0839	-0.0631
C	TGC	-0.3777	0.3913	-3.5032	-1.4187	-0.3754	-0.1669	-0.0903	-0.0493	-0.0371
D	GAT	-0.1201	0.0396	-1.1142	-0.4512	-0.1194	-0.0531	-0.0287	-0.0157	-0.0118
E	GAG	-0.2527	0.1752	-2.3440	-0.9493	-0.2512	-0.1117	-0.0605	-0.0330	-0.0248
F	TTT	-0.2741	0.2061	-2.5425	-1.0297	-0.2725	-0.1211	-0.0656	-0.0358	-0.0269
G	GGA	-0.8286	1.8834	-7.6861	-3.1127	-0.8237	-0.3662	-0.1982	-0.1083	-0.0815
G	GGC	-0.4436	0.5398	-4.1149	-1.6664	-0.4410	-0.1960	-0.1061	-0.0580	-0.0436
G	GGG	-0.7052	1.3641	-6.5412	-2.6490	-0.7010	-0.3116	-0.1687	-0.0921	-0.0693
H	CAT	-0.1966	0.1060	-1.8233	-0.7384	-0.1954	-0.0869	-0.0470	-0.0257	-0.0193
I	ATA	-0.8874	2.1604	-8.2318	-3.3337	-0.8822	-0.3922	-0.2123	-0.1159	-0.0872
I	ATT	-0.0906	0.0225	-0.8403	-0.3403	-0.0901	-0.0400	-0.0217	-0.0118	-0.0089
K	AAA	-0.2661	0.1943	-2.4686	-0.9997	-0.2646	-0.1176	-0.0637	-0.0348	-0.0262
L	CTA	-0.2636	0.1906	-2.4452	-0.9902	-0.2620	-0.1165	-0.0631	-0.0344	-0.0259
L	CTC	-0.7185	1.4162	-6.6650	-2.6991	-0.7143	-0.3175	-0.1719	-0.0939	-0.0706
L	CTG	-0.4662	0.5961	-4.3242	-1.7512	-0.4634	-0.2060	-0.1115	-0.0609	-0.0458
L	CTT	-0.4601	0.5807	-4.2679	-1.7284	-0.4574	-0.2033	-0.1101	-0.0601	-0.0452
L	TTA	-0.2128	0.1242	-1.9741	-0.7995	-0.2116	-0.0941	-0.0509	-0.0278	-0.0209
N	AAT	-0.3164	0.2746	-2.9347	-1.1885	-0.3145	-0.1398	-0.0757	-0.0413	-0.0311
P	CCC	-0.5061	0.7027	-4.6948	-1.9013	-0.5031	-0.2237	-0.1211	-0.0661	-0.0498
P	CCG	-0.8002	1.7567	-7.4230	-3.0061	-0.7955	-0.3537	-0.1914	-0.1046	-0.0787
P	CCT	-0.2319	0.1476	-2.1513	-0.8712	-0.2306	-0.1025	-0.0555	-0.0303	-0.0228
Q	CAG	-0.3840	0.4046	-3.5624	-1.4427	-0.3818	-0.1697	-0.0919	-0.0502	-0.0378
R	AGG	-0.5378	0.7935	-4.9890	-2.0204	-0.5347	-0.2377	-0.1287	-0.0703	-0.0529
R	CGA	-1.7330	8.2391	-16.0757	-6.5103	-1.7228	-0.7659	-0.4146	-0.2264	-0.1704
R	CGC	-0.5568	0.8504	-5.1646	-2.0915	-0.5535	-0.2461	-0.1332	-0.0727	-0.0547
R	CGG	-1.5092	6.2486	-13.9999	-5.6696	-1.5003	-0.6670	-0.3611	-0.1972	-0.1484
R	CGT	-0.0080	0.0002	-0.0744	-0.0301	-0.0080	-0.0035	-0.0019	-0.0010	-0.0008
S	TCA	-0.3942	0.4264	-3.6571	-1.4810	-0.3919	-0.1742	-0.0943	-0.0515	-0.0388
S	TCG	-0.4927	0.6660	-4.5705	-1.8509	-0.4898	-0.2178	-0.1179	-0.0644	-0.0484
S	TCT	-0.0121	0.0004	-0.1120	-0.0454	-0.0120	-0.0053	-0.0029	-0.0016	-0.0012
T	ACA	-0.4278	0.5020	-3.9679	-1.6069	-0.4252	-0.1890	-0.1023	-0.0559	-0.0421
T	ACG	-0.6503	1.1603	-6.0327	-2.4431	-0.6465	-0.2874	-0.1556	-0.0850	-0.0639
T	ACT	-0.0482	0.0064	-0.4468	-0.1810	-0.0479	-0.0213	-0.0115	-0.0063	-0.0047
V	GTA	-0.6310	1.0922	-5.8529	-2.3703	-0.6272	-0.2789	-0.1509	-0.0824	-0.0620
V	GTG	-0.4308	0.5092	-3.9966	-1.6185	-0.4283	-0.1904	-0.1031	-0.0563	-0.0424
V	GTT	-0.0570	0.0089	-0.5288	-0.2142	-0.0567	-0.0252	-0.0136	-0.0074	-0.0056
Y	TAT	-0.2857	0.2239	-2.6501	-1.0732	-0.2840	-0.1263	-0.0683	-0.0373	-0.0281
Z	AGC	-0.0248	0.0017	-0.2297	-0.0930	-0.0246	-0.0109	-0.0059	-0.0032	-0.0024

Table 1: Summary statistics for gene specific selection coefficients on synonymous codon usage $S = -\Delta\eta\,\phi$ from the *without* $\vec{\Phi}$ ROC SEMPPR model fit to the *S. cerevisiae* genome. The selection coefficient S was calculated relative to the most translationally efficient codon for a given amino acid on a gene by gene basis.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

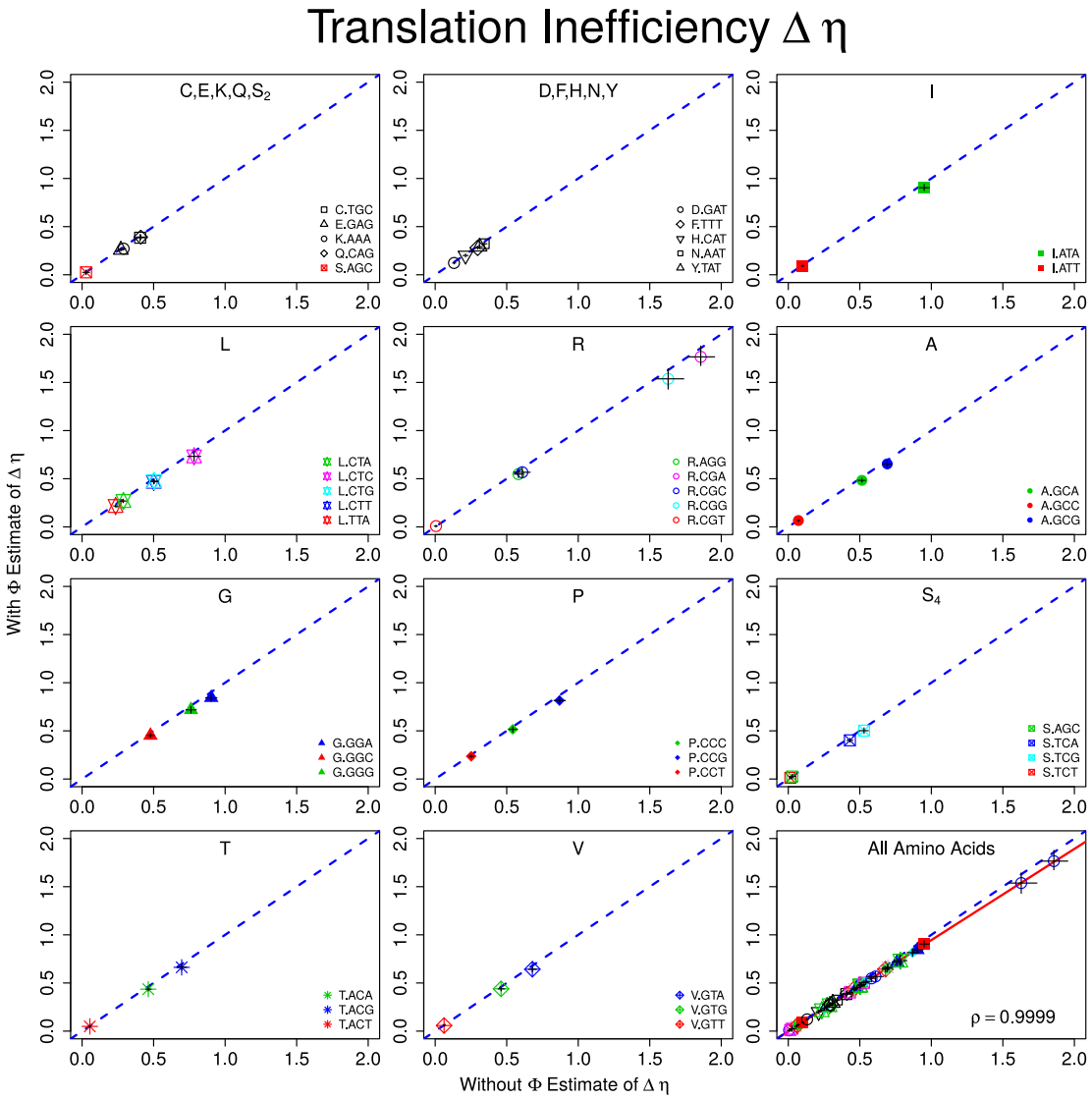


Figure 1: Comparison of *with* and *without* $\vec{\Phi}$ ROC SEMPPR estimates for codon specific differences in translational efficiencies $\Delta\eta$ which have units $1/(t$ protein) where the units of time are set such that the average protein synthesis rate across the genome, $\bar{\phi}$, equals 1. To improve legibility of the plots the two codon amino acids have been combined into two plots and all of the amino acids with > 2 codons into separate plots. The dashed blue line represents the 1:1 line between axes and error bars indicate the 95% posterior credibility intervals (CIs) for each parameter. For both the *with* and *without* $\vec{\Phi}$ fits of ROC SEMPPR, all codons but one, Arg codon CGT, have CIs that do not overlap with 0. As illustrated in the last plot, a linear regression between estimates of $\Delta\eta$ for all codons produces a correlation coefficient $\rho > 0.999$.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

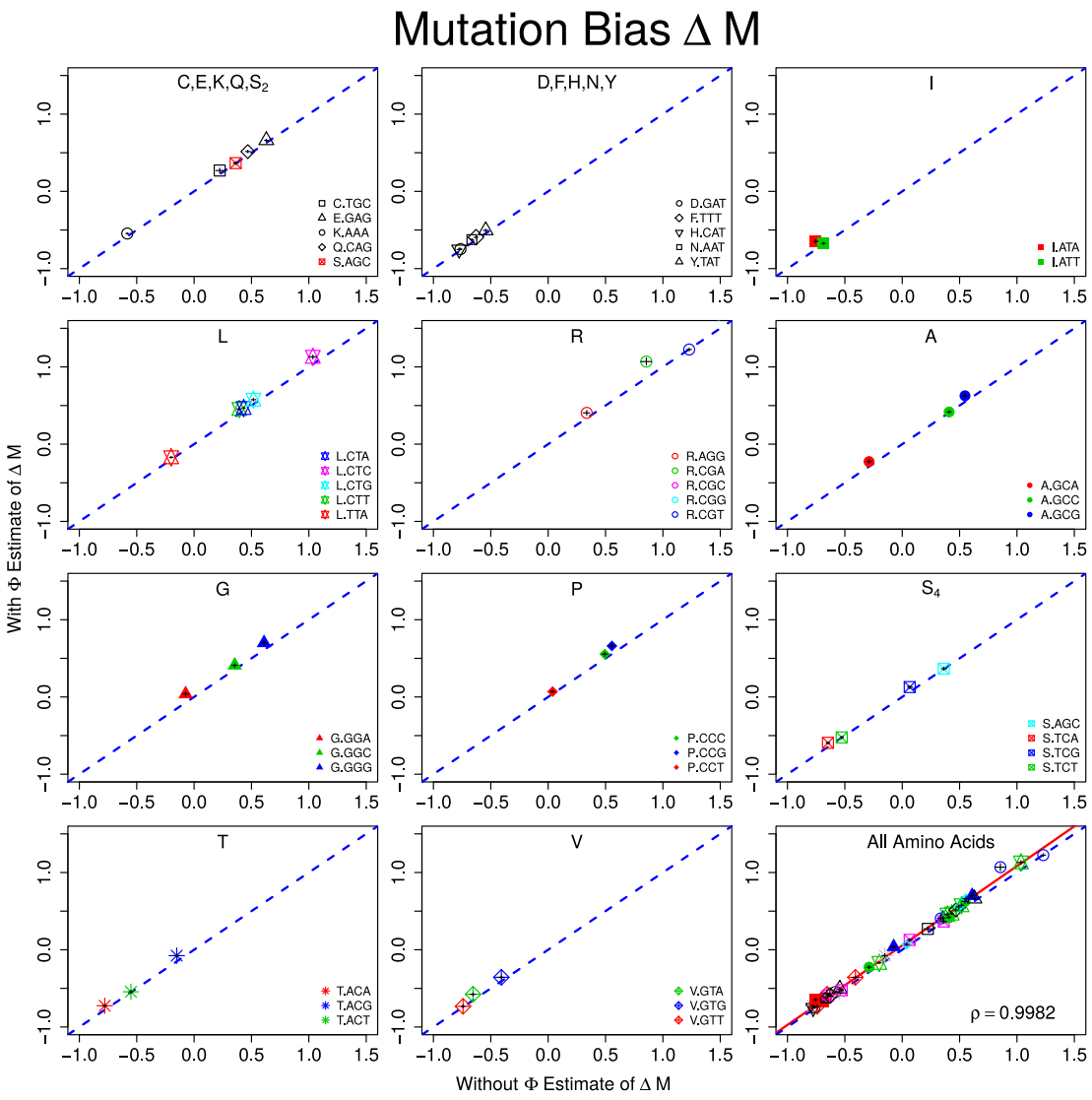


Figure 2: Comparison of *with* and *without* $\vec{\Phi}$ ROC SEMPPR estimates for codon specific differences in mutation biases terms ΔM which are unitless. Specifically, $\Delta M_{i,1}$ equals the natural logarithm of the ratio of the frequencies of synonymous codon 1 to i in the absence of natural selection. To improve legibility of the plots the two codon amino acids have been combined into two plots and all of the amino acids with > 2 codons into separate plots. The dashed blue line represents the 1:1 line between axes and error bars indicate the 95% posterior credibility intervals (CIs) for each parameter. For both the *with* and *without* fits of ROC SEMPPR, all codons have CIs that do not overlap with 0. As illustrated in the last plot, a linear regression between estimates of ΔM for all codons produces a correlation coefficient $\rho > 0.998$.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

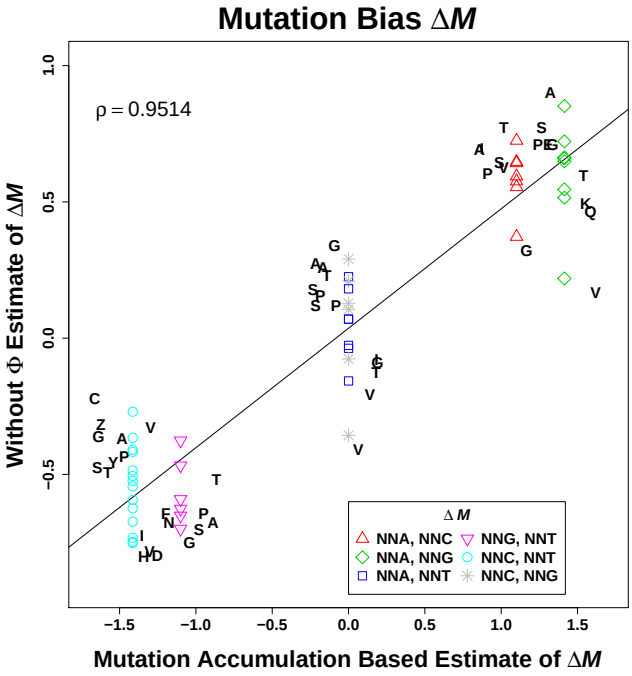


Figure 3: Comparison of *without* $\vec{\Phi}$ estimates of codon specific mutation biases ΔM and estimates generated from mutation accumulation experiments (Zhu *et al.*, 2014). For each amino acid the codon with the shortest pausing time is used as a reference and are not shown because, by definition their ΔM values are 0. Pearson correlation coefficient ρ for all of the codons is given. The solid line represents the best fit linear regression.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

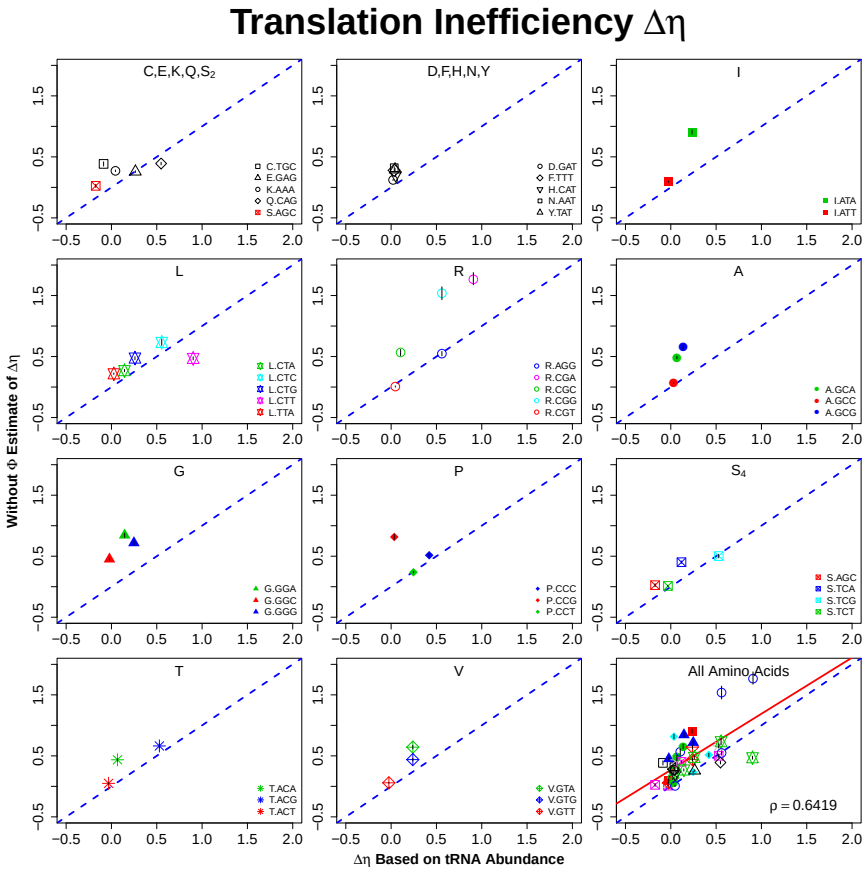
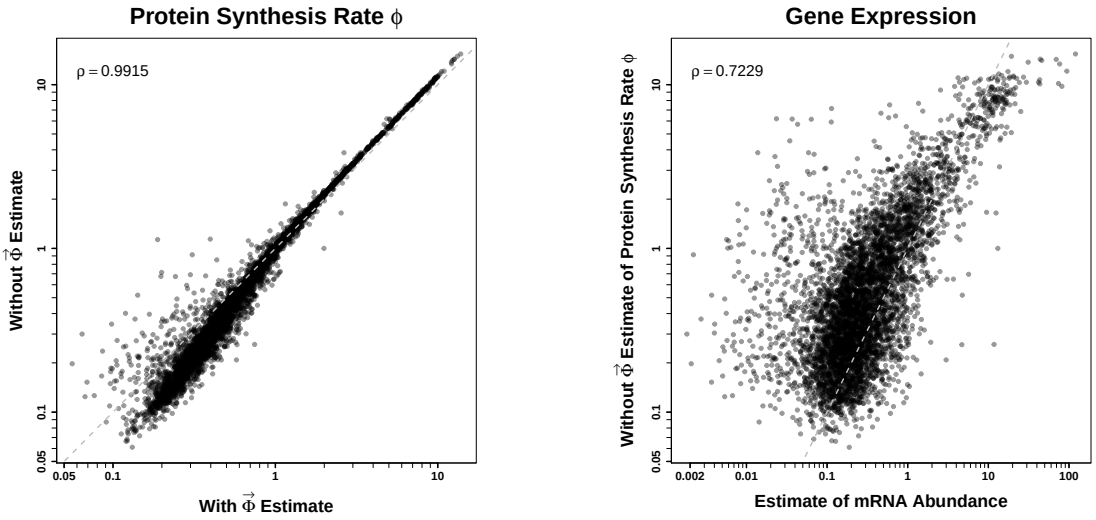


Figure 4: Comparison of *without* $\vec{\Phi}$ estimates of codon specific translational inefficiencies $\Delta\eta$ and estimates of differences in ribosome pausing times, Δt based on tRNA gene copy number and wobble inefficiencies. For each amino acid the codon with the shortest pausing time is used as a reference and are not shown because, by definition their $\Delta\eta$ values are 0. Pearson correlation coefficient ρ for all of the codons is given. The dashed blue line represents the 1:1 line and the red line represents the best fit linear regression line.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a) (b)

Figure 5: Evaluation of predicted gene expression levels between models and empirical measurements from Yassour *et al.* (2009). (a) Comparison of *with* and *without* $\vec{\Phi}$ ROC SEMPPR estimates of protein synthesis rates, $\hat{\phi}$. The units for ϕ are protein/t and time t is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. Note the very strong correlation between the *with* and *without* $\vec{\Phi}$ estimates of ϕ for the high expression genes. (b) Comparison of *without* $\vec{\Phi}$ estimates of ϕ and empirical measurements of mRNA abundances, $\vec{\Phi}$. The empirical mRNA abundance measurements, [mRNA], are being used here as a proxy for protein synthesis rates, i.e. [mRNA] $\propto$ protein/t. The measurements are scaled such that the mean [mRNA] value is 1. Pearson correlation coefficients ρ are given and the dashed gray line indicates 1:1 line.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

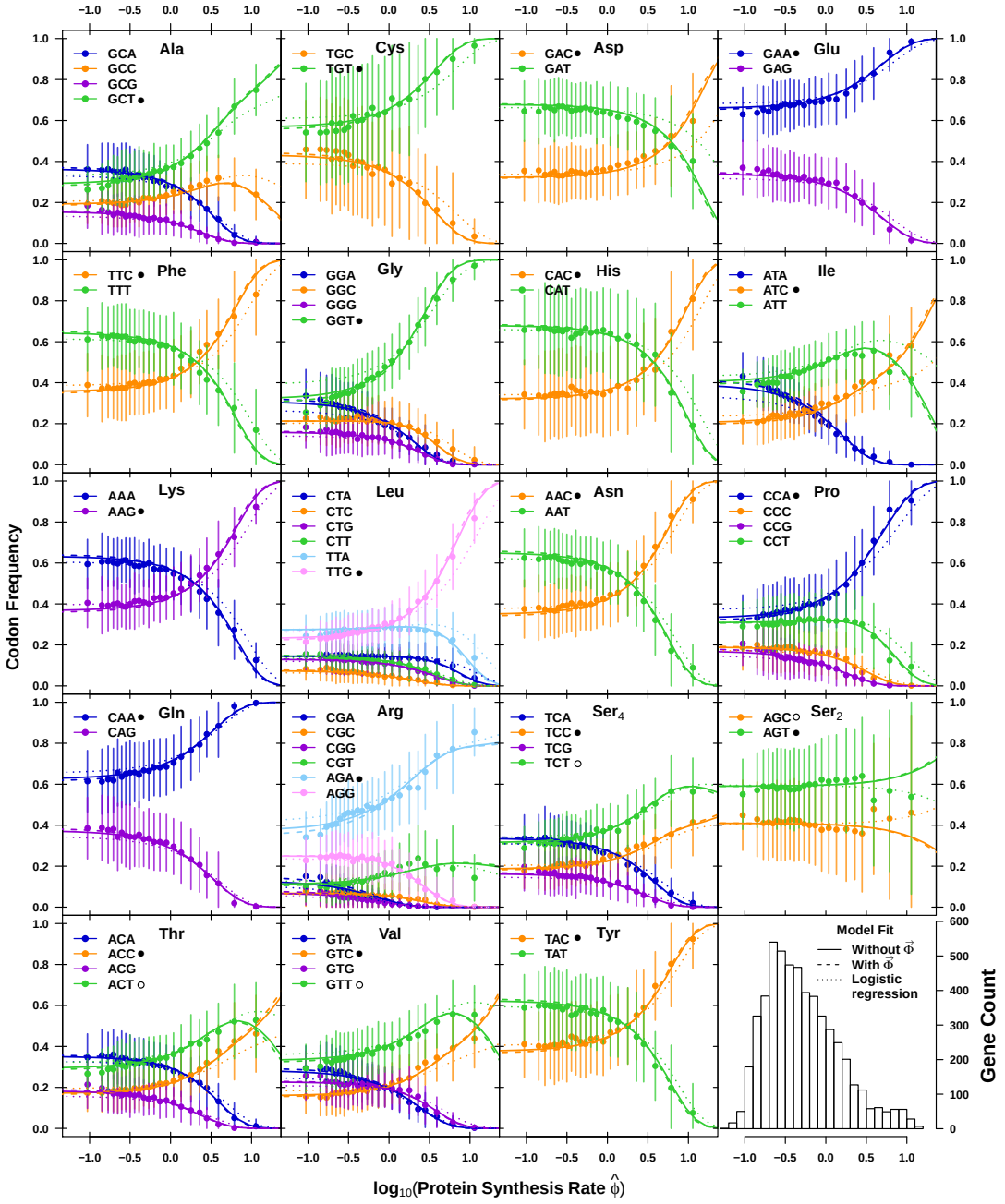


Figure 6: Model predictions and observed codon usage frequencies as a function of estimated protein synthesis rate ϕ for the *S. cerevisiae* S288c genome. The units for ϕ are protein/t and time t is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. Each amino acid is represented by a separate subplot. Solid, dashed, and dotted lines represent the *without* $\vec{\Phi}$, *with* $\vec{\Phi}$ ROC SEMPPR model fits, and a simple logistic regression approach where the estimation error in $\vec{\Phi}$ is ignored, respectively. None of the parameter estimates' 95% Credibility Intervals overlap with 0 except $\Delta\eta_{CGT,AGA}$. 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 ROC SEMPPR 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 $\hat{\phi}$ are based on the *with* $\vec{\Phi}$ ROC SEMPPR model fits, thus representing our best estimate of their values.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

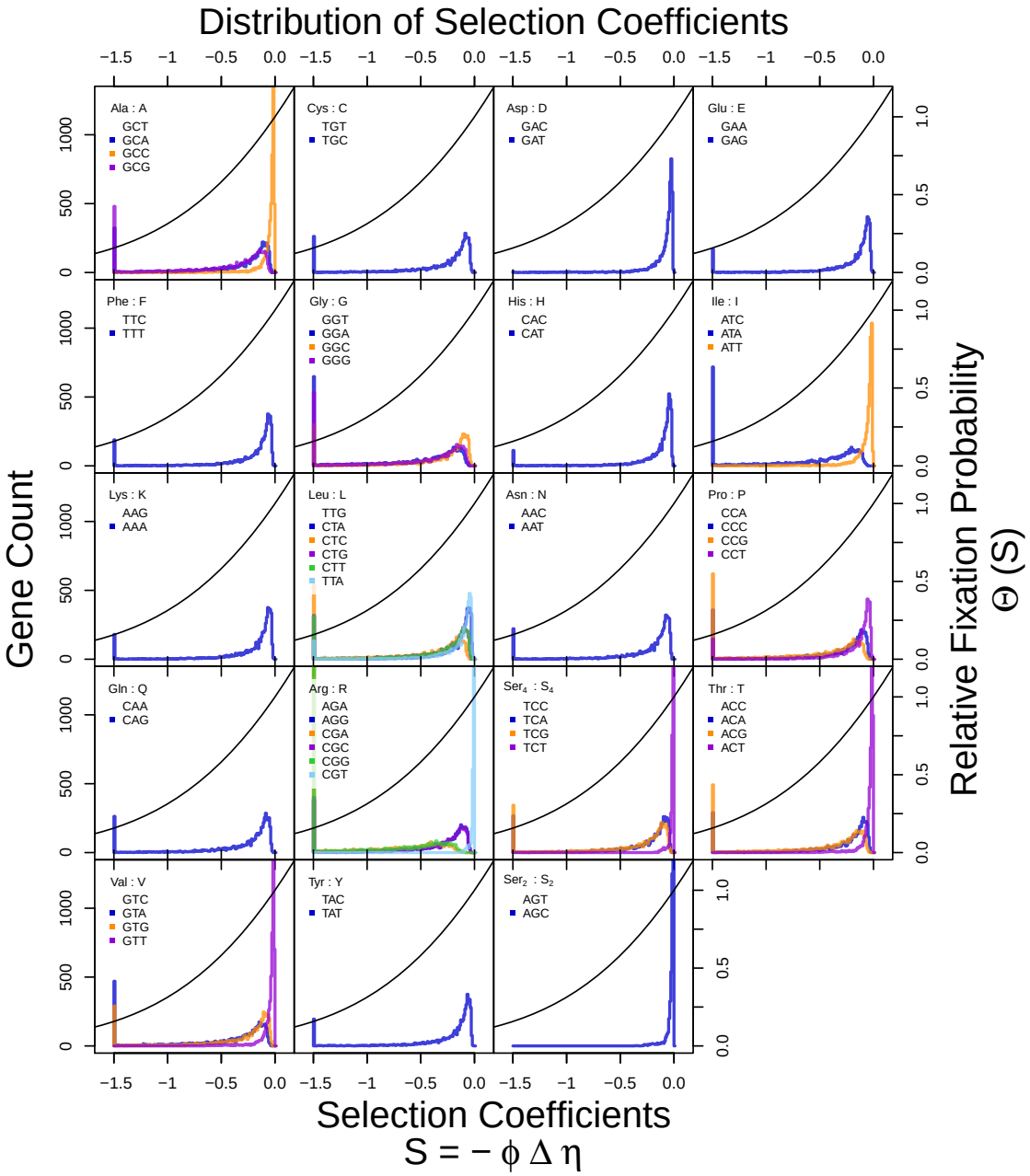


Figure 7: Distribution of gene specific selection coefficients on synonymous codon usage $S = -\Delta\eta\,\phi$ from the *without* $\vec{\Phi}$ model fit to the *S. cerevisiae* genome. Selection coefficient S were calculated on a gene by gene basis and relative to the most translationally efficient codon for a given amino acid (which is the codon listed first in the legend). The reference codon, which is most favored by selection and for which, by definition, $S = 0$, is listed first within the legend of each panel. Genes with $S \leq -2$ were combined together into a single bin. For reference, the fixation probability of a codon relative to a pure drift process, $\Theta(S) = 2S/(1 - \exp[-2S])$, are also plotted ($-$ line). Summary statistics can be found in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

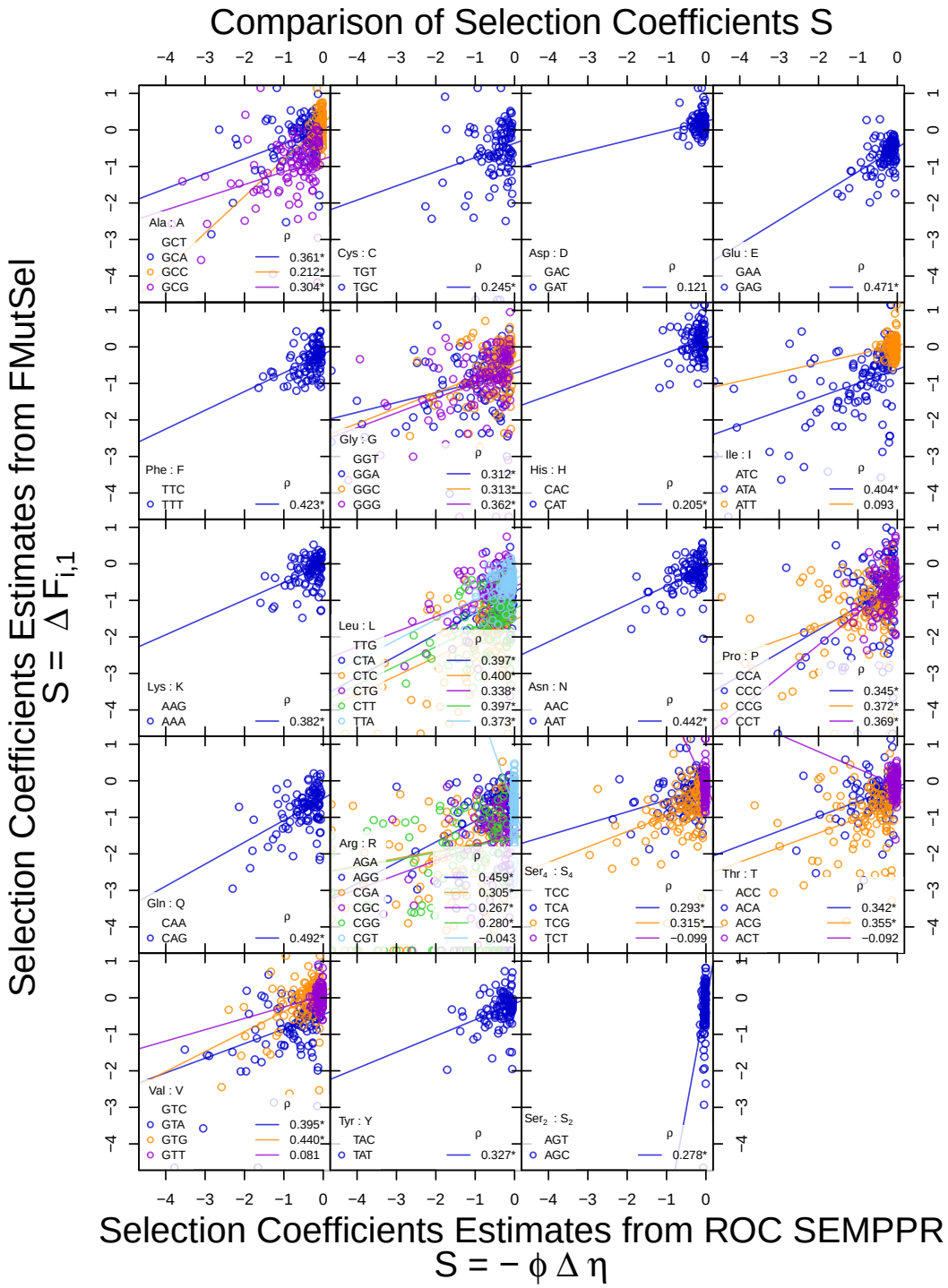


Figure 8: Comparison of gene specific selection coefficients on synonymous codon usage $S = -\Delta\eta\,\phi$ from the *without* $\vec{\Phi}$ model fit to the *S. cerevisiae* genome and those from fitting the FMutSel model from Yang and Nielsen (2008) for 106 yeast genes used in Rokas *et al.* (2003) as estimated by Kubatko *et al.* (view) For more details see the main text. Selection coefficient S were calculated on a gene by gene basis and relative to the most translationally efficient codon for a given amino acid (which is the codon listed first in the legend). Lines indicate linear regression line best fit and the corresponding correlation coefficients are listed as well with a $*$ indicating model fits with $p < 0.05$. Under the FMutSel model, monomorphic sites across species can lead to estimates of $S = -\infty$, these observations are plotted on the x-axis.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

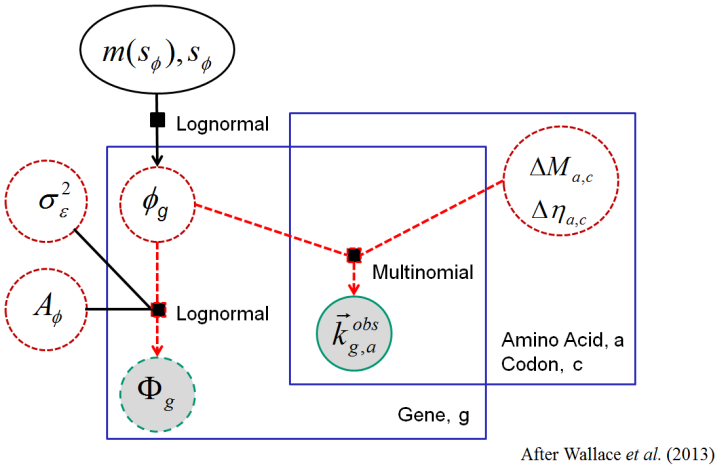


Figure 9: Dependence graph of *with* and *without* ROC SEMPPR methods. Shaded circles $\vec{\Phi}_j$ and $k_{i,j}$ represent observed data. Dashed circles represent key random model parameters while the solid oval represents a random hierarchical parameter. Solid black squares provide information on the distributional relationships between quantities. Large rectangular boxes represent replication of each model component across both amino acids and genes, e.g. pausing, and mutation parameters differ across amino acids but are common across genes, while counts $k_{i,j}$ differ across both amino acids and genes.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supporting Materials

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

Model Validation using Simulated Data

In order to verify the reliability of the *with* and *without* $\vec{\Phi}$ ROC SEMPPR 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* $\vec{\Phi}$ ROC SEMPPR 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* $\vec{\Phi}$ ROC SEMPPR 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 ROC SEMPPR. 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* $\vec{\Phi}$ ROC SEMPPR 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 ~ 0.96 (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,

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

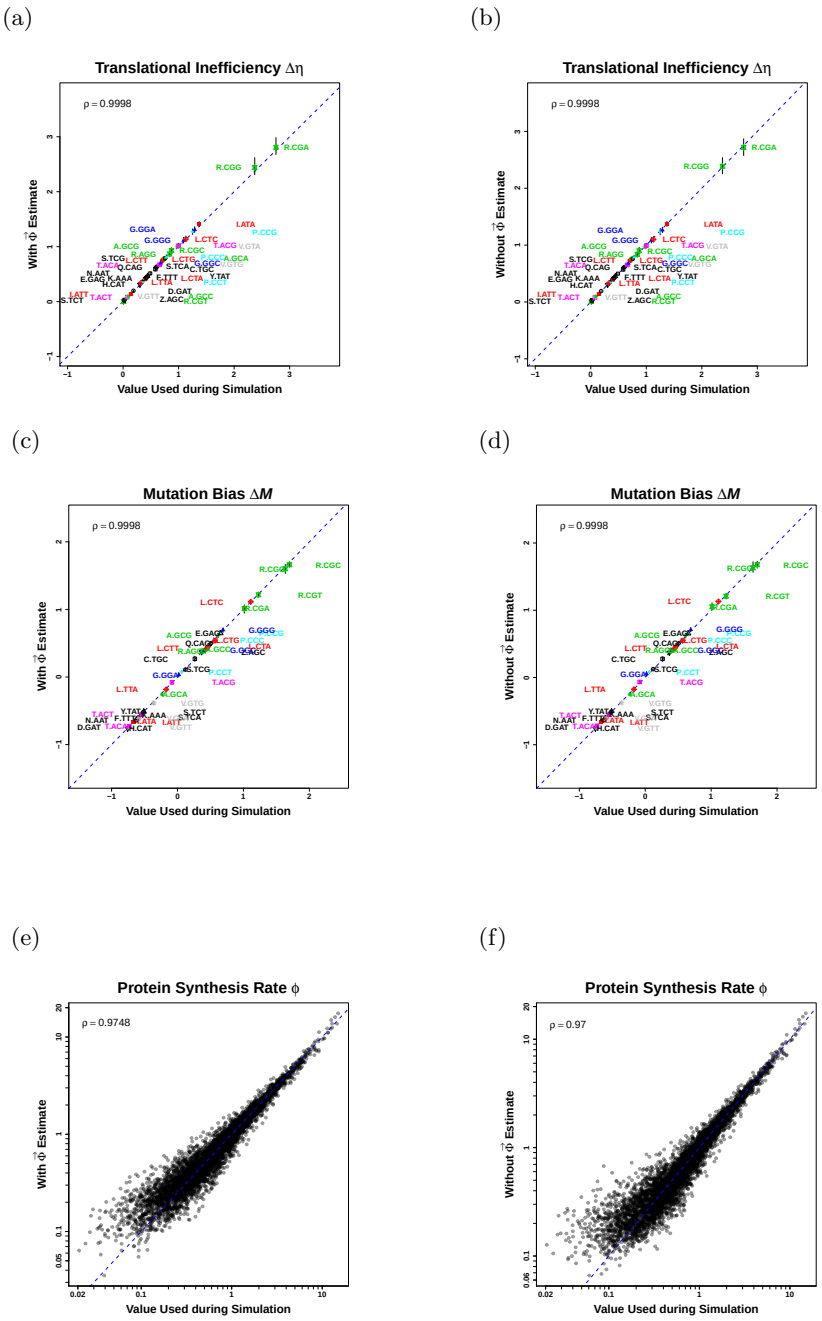


Figure S1: Comparison of estimated parameters versus actual parameters used to simulate data under the model ROC SEMPPR. Here $\phi \sim$ LogN as assumed when fitting ROC SEMPPR. (a) Comparison of *with* $\vec{\Phi}$ ROC SEMPPR parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$.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

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

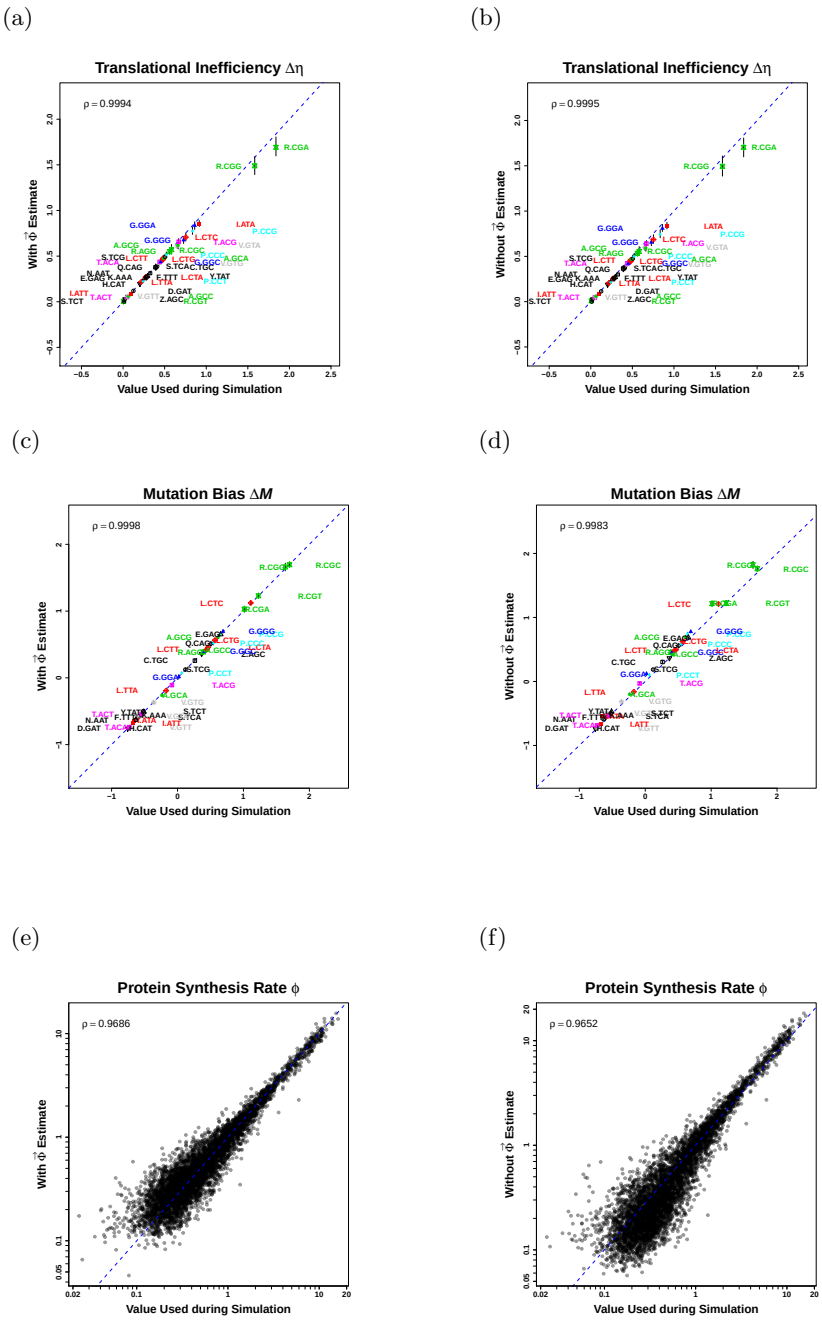


Figure S2: Comparison of estimated parameters versus actual parameters used to simulate data under the model ROC SEMPPR. Here ϕ values used in the simulation were based on the *with* $\vec{\Phi}$ fit of the *S. cerevisiae* S288c genome dataset and, as a result, do not follow a log-normal distribution as assumed when fitting ROC SEMPPR: (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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

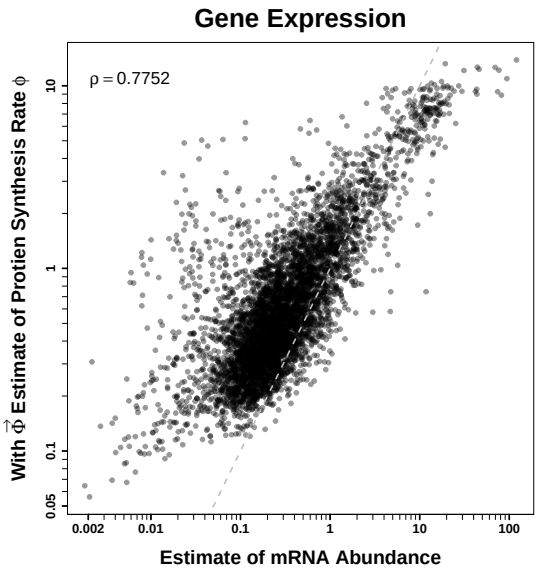


Figure S3: Comparison between posterior mean estimates of ϕ for the *with* $\vec{\Phi}$ model fit and $\vec{\Phi}$ data consisting of mRNA abundance measurements from Yassour *et al.* (2009).

if the additional measurement noise is not taken into account a subtle biases on ϕ and $\vec{\Delta\eta}$ is introduced. 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, $\Delta\vec{\eta}$. 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 ROC SEMPPR 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 $\Delta\vec{\eta}$ in pre- and post-MCMC data processing step, we include the estimation of the scaling term A_Φ discussed in the last section.

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

where, as before, $\Delta\vec{M}_i$ and $\Delta\vec{\eta}_i$ 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 (2), $n_{i,j}$ 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 $\Delta\vec{M}_i$, $\Delta\vec{\eta}_i$, 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 $\left(s_\varepsilon^{(i+1)}\right)^2 \sim \text{Inv-Gamma}((n_g - 1)/2, \left(S^{(i)}\right)^2/2)$ where $\left(S^{(i)}\right)^2 = \sum_{j=1}^{n_g}(\log\vec{\Phi}_j - A_\Phi^{(i)} - \log\phi_j^{(i)})^2$.

(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 $m^{(i+1)} = -\left(s^{(i+1)}\right)^2/2$.

(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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 20-35% 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.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

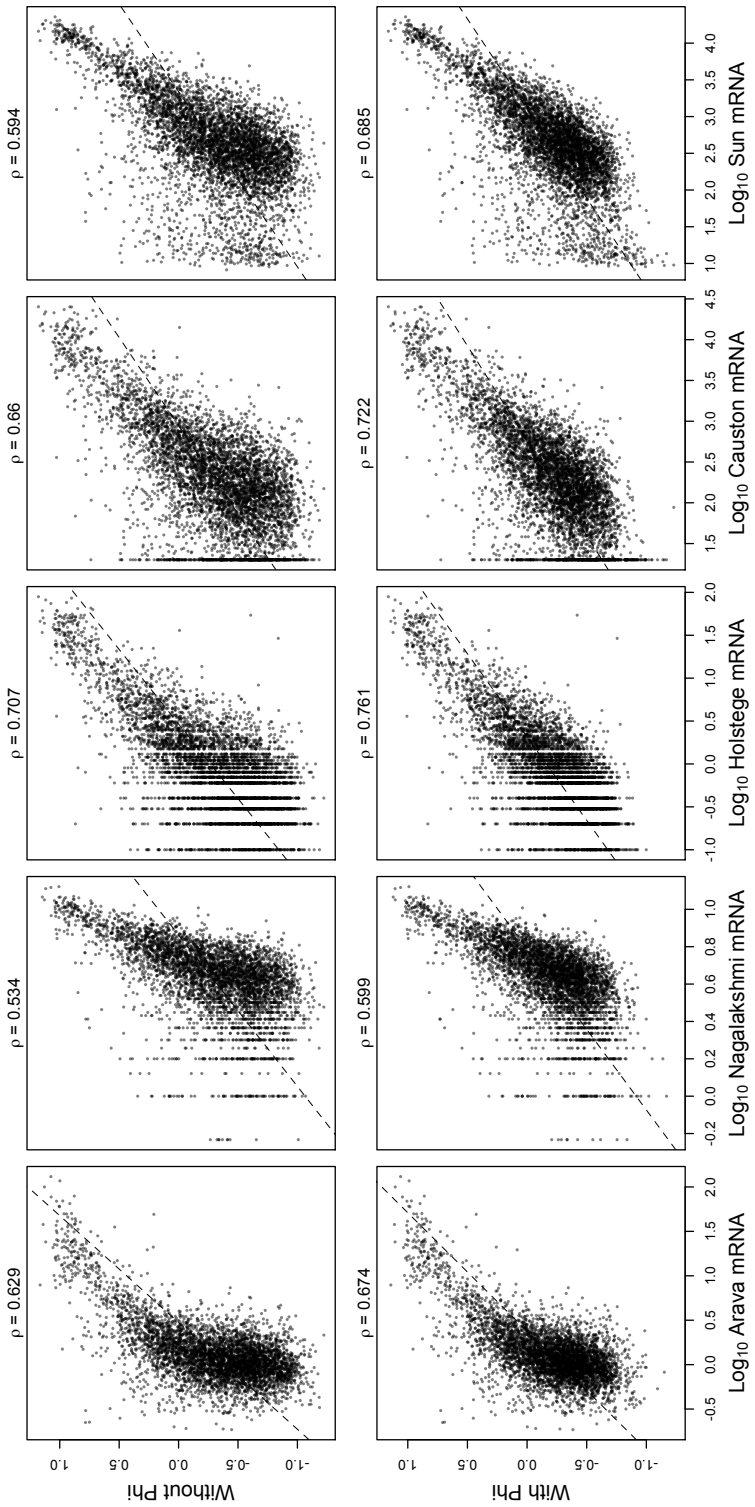


Figure S4: Scatter plot comparisons of *with* (Yassour measurements) and *without* $\vec{\Phi}$ posterior mean estimates to empirical measurements from four additional laboratories. The units for ϕ are protein/t and time is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. The empirical mRNA abundance measurements, [mRNA], are being used here as a proxy for protein synthesis rates, i.e. [mRNA] $\propto$ protein/t. The measurements are scaled such that the mean [mRNA] value is 1. Pearson correlation coefficients ρ are given and the dashed black line represents the fit of a linear regression model.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

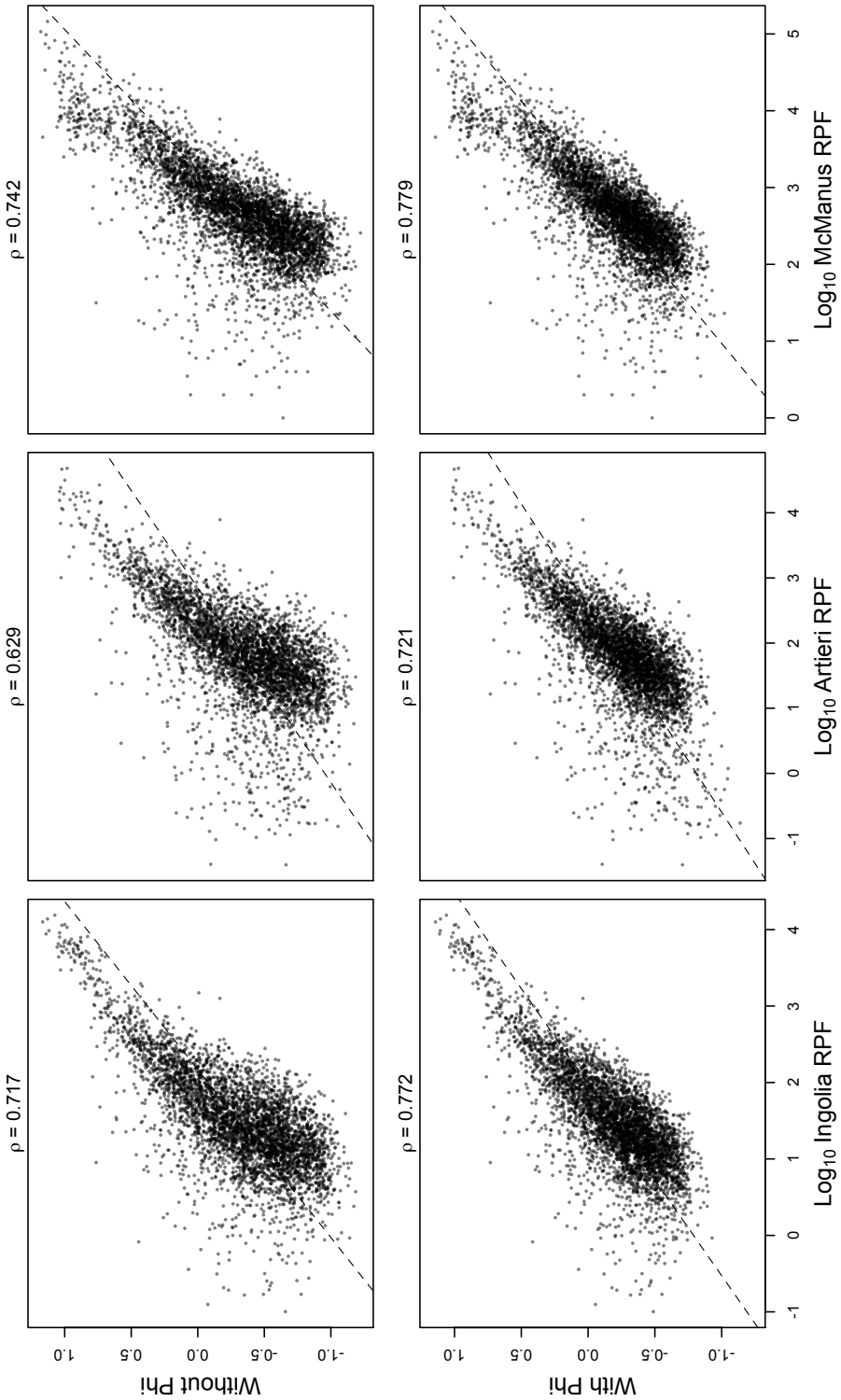


Figure S5: Scatter plot comparisons of *with* $\vec{\Phi}$ (Yassour) and *without* $\vec{\Phi}$ posterior mean estimates to empirical measurements from three ribosome profiling datasets from three different laboratories. The units for ϕ are protein/t and time is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. The empirical ribosome profiling measurements were originally in units of reads per kilobase of transcript per million mapped (rpkm) corrected for mRNA length. These measurements are scaled such that the mean rpkm value is 1. Pearson correlation coefficients ρ are given and the dashed black line represents the fit of a linear regression model.

bioRxiv preprint doi: https://doi.org/10.1101/009670; this version posted April 17, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Tables

Data in supplemental tables can be downloaded from **doi:** `http://dx.doi.org/10.1101/009670`

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).

S10. Additional measurements of protein synthesis rates from ribosome profiling experiments from multiple laboratories of *S. cerevisiae* Genome (s.cerevisiae.rpf.measurements.tsv).

S11. Results from linear regression of FMutSel estimates of S vs. *without* $\vec{\Phi}$ ROC SEMPPR estimates of S for the 106 genes in the Rokas *et al.* (2003) dataset (FMutSel_S_vs_ROC_wo_phi_S_regressions.txt).