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,

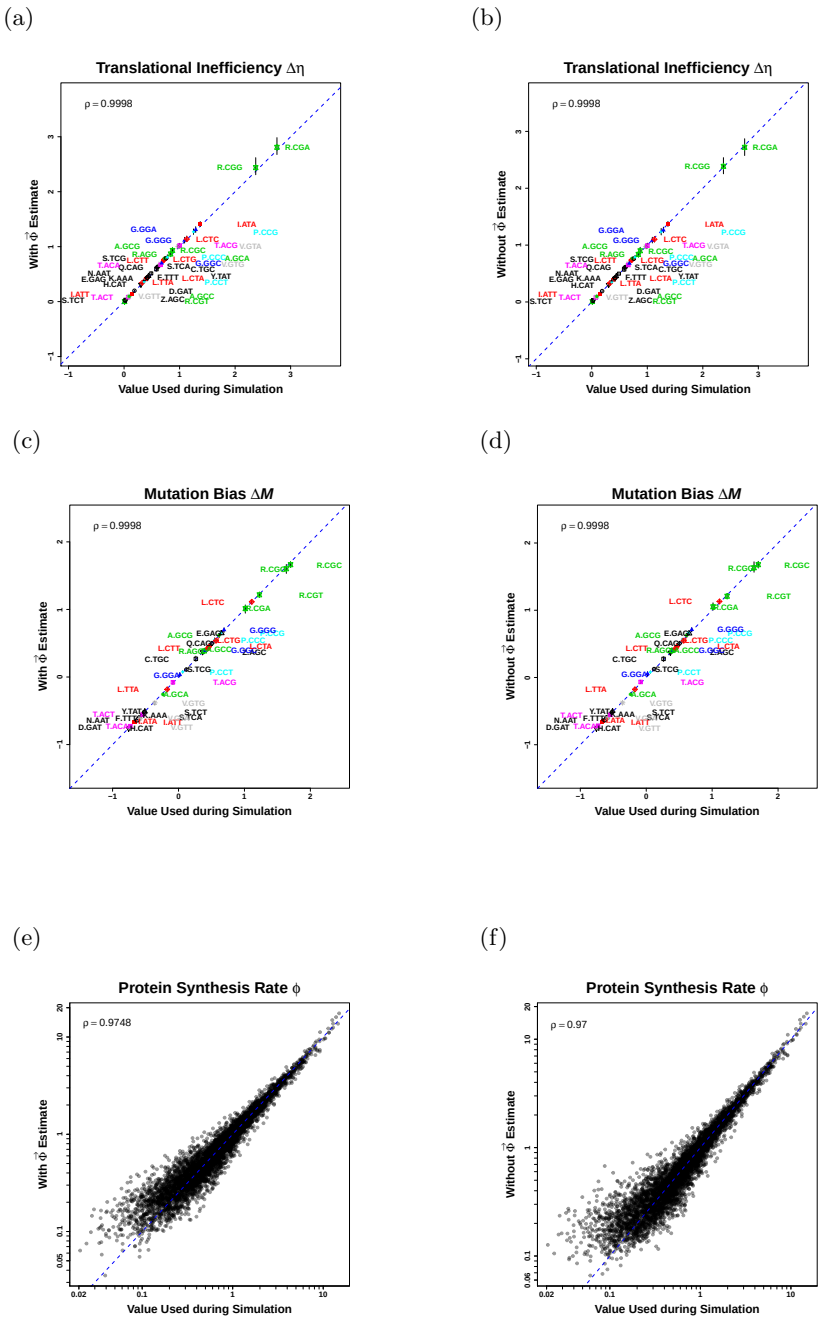

Figure S1: Comparison of estimated parameters versus actual parameters used to simulate data under the model ROC SEMPPR. Here $\phi \sim$ LogN as assumed when fitting ROC SEMPPR. (a) Comparison of *with* $\vec{\Phi}$ ROC SEMPPR parameter estimates $\Delta\eta$ vs. actual data generating parameters $\Delta\eta$. (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.

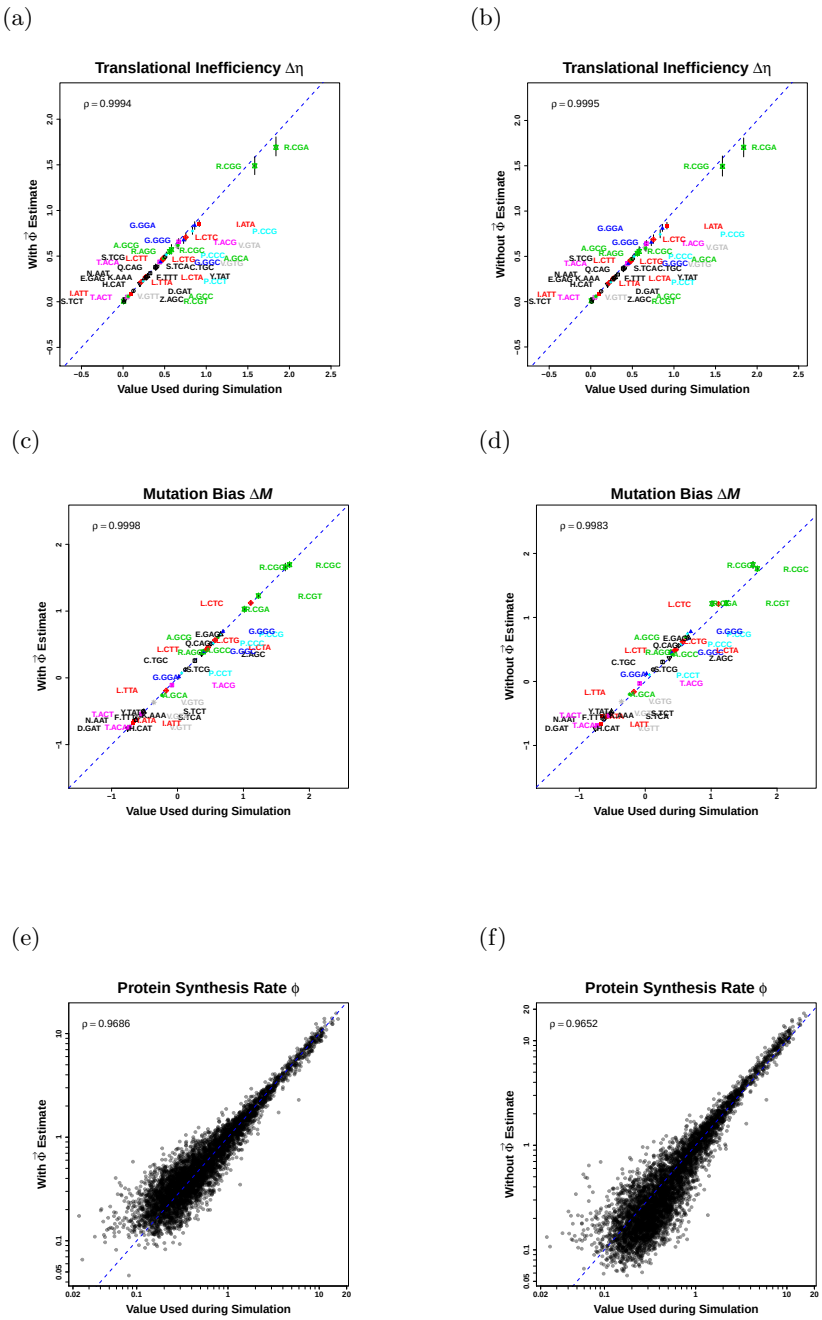

Figure S2: Comparison of estimated parameters versus actual parameters used to simulate data under the model ROC SEMPPR. Here ϕ values used in the simulation were based on the *with* $\vec{\Phi}$ fit of the *S. cerevisiae* S288c genome dataset and, as a result, do not follow a log-normal distribution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

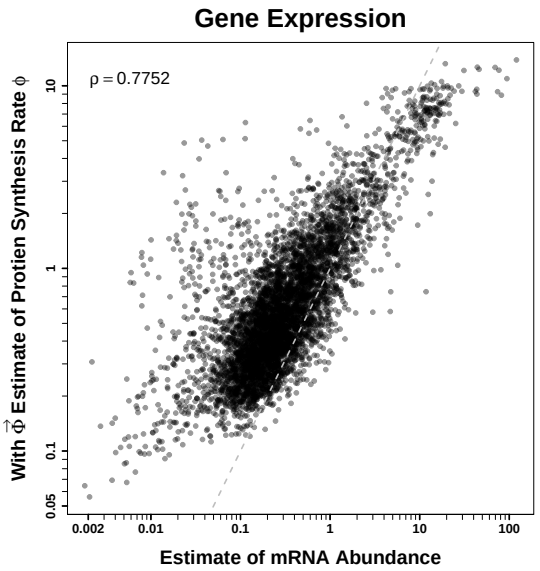

Figure S3: Comparison between posterior mean estimates of ϕ for the *with* $\vec{\Phi}$ model fit and $\vec{\Phi}$ data consisting of mRNA abundance measurements from Yassour *et al.* (2009).

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.

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

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

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

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

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

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

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

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

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

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

(a) Step $i = 0$ only.

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

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

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

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

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

Step 2. Update hyperparameters conditional on all other parameters:

(a) If using the fitting *with* $\vec{\Phi}$ model: update $\left(s_\varepsilon^{(i+1)}\right)^2 \sim \text{Inv-Gamma}((n_g - 1)/2, \left(S^{(i)}\right)^2/2)$ 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.

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.

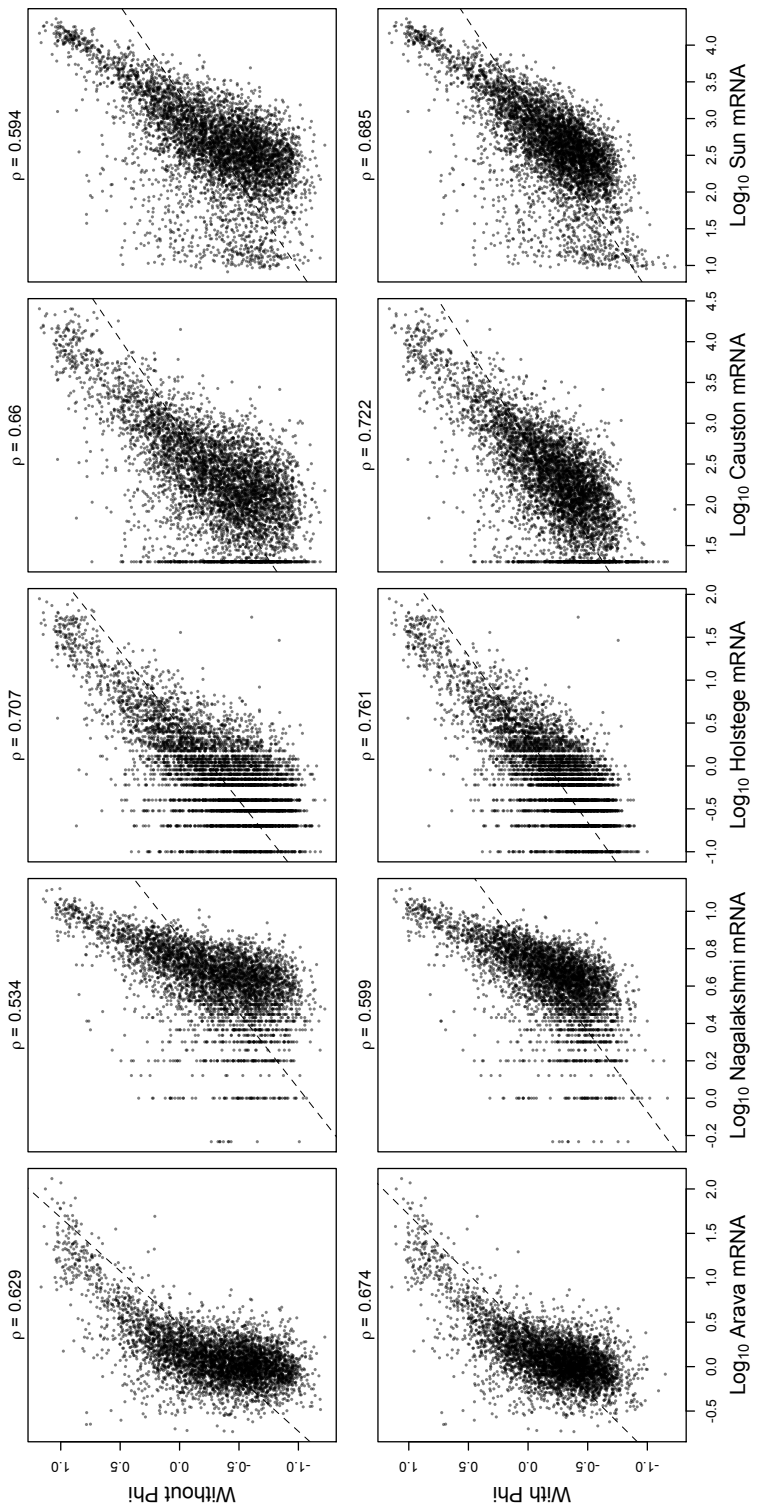

Figure S4: Scatter plot comparisons of *with* (Yassour measurements) and *without* $\vec{\Phi}$ posterior mean estimates to empirical measurements from four additional laboratories. The units for ϕ are protein/t and time is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. The empirical mRNA 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.

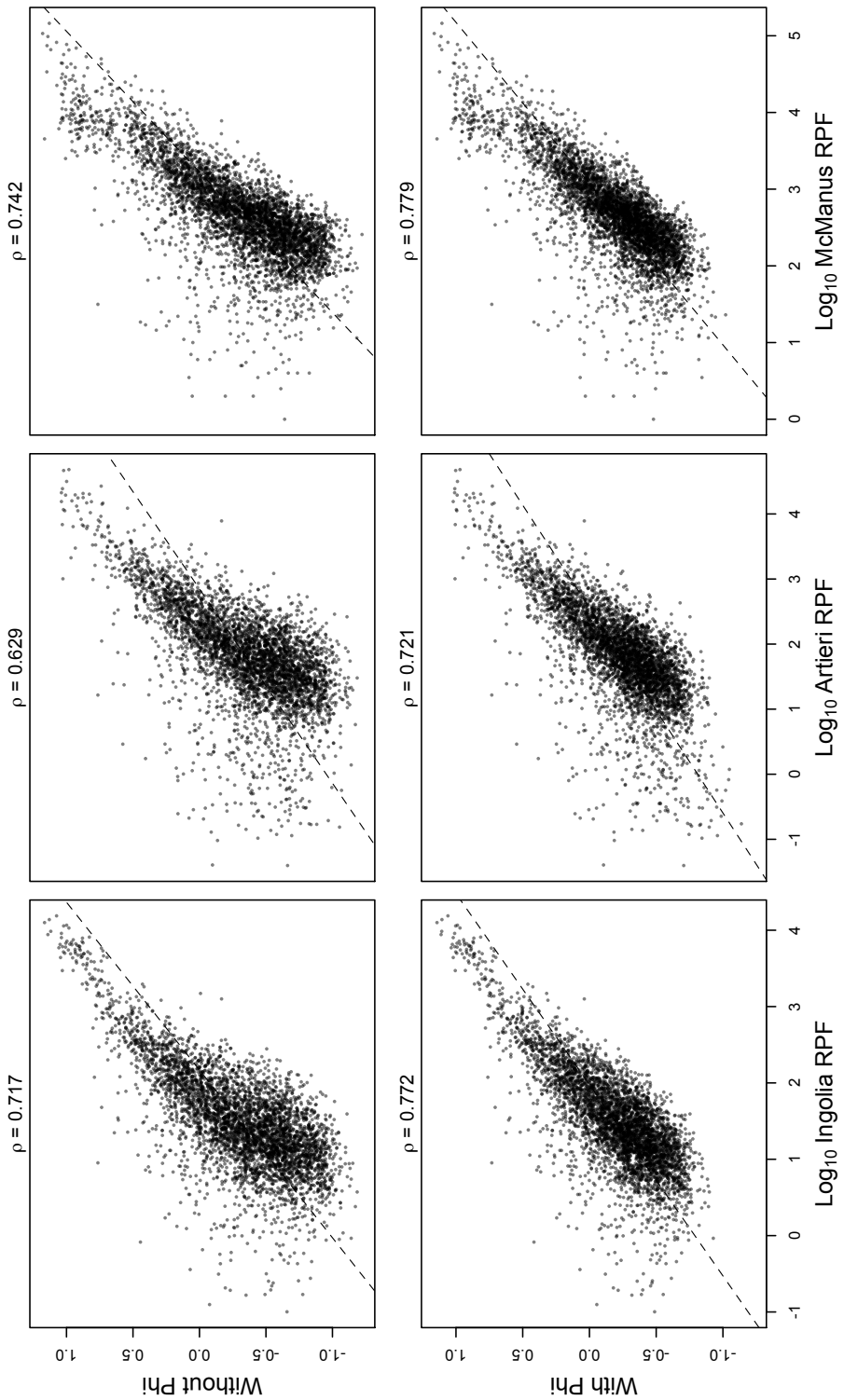

Figure S5: Scatter plot comparisons of *with* $\vec{\Phi}$ (Yassour) and *without* $\vec{\Phi}$ posterior mean estimates to empirical measurements from three ribosome profiling datasets from three different laboratories. The units for ϕ are protein/t and time is scaled such that the prior for ϕ satisfies $E(\phi) = 1$. 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.

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).