Estimating Selection on Synonymous Codon Usage from Noisy Experimental Data

Edward W.J. Wallace,*[1,2] Edoardo M. Airoldi,[2,3] and D. Allan Drummond[1]

[1]Department of Biochemistry and Molecular Biology, University of Chicago

[2]Department of Statistics, and FAS Center for Systems Biology, Harvard University

[3]Broad Institute of MIT & Harvard, Cambridge, Massachusetts

***Corresponding author:** E-mail: ewallace@uchicago.edu.

Associate editor: John H. McDonald

Abstract

A key goal in molecular evolution is to extract mechanistic insights from signatures of selection. A case study is codon usage, where despite many recent advances and hypotheses, two longstanding problems remain: the relative contribution of selection and mutation in determining codon frequencies and the relative contribution of translational speed and accuracy to selection. The relevant targets of selection—the rate of translation and of mistranslation of a codon per unit time in the cell—can only be related to mechanistic properties of the translational apparatus if the number of transcripts per cell is known, requiring use of gene expression measurements. Perhaps surprisingly, different gene-expression data sets yield markedly different estimates of selection. We show that this is largely due to measurement noise, notably due to differences between studies rather than instrument error or biological variability. We develop an analytical framework that explicitly models noise in expression in the context of the population-genetic model. Estimates of mutation and selection strength in budding yeast produced by this method are robust to the expression data set used and are substantially higher than estimates using a noise-blind approach. We introduce per-gene selection estimates that correlate well with previous scoring systems, such as the codon adaptation index, while now carrying an evolutionary interpretation. On average, selection for codon usage in budding yeast is weak, yet our estimates show that genes range from virtually unselected to average per-codon selection coefficients above the inverse population size. Our analytical framework may be generally useful for distinguishing biological signals from measurement noise in other applications that depend upon measurements of gene expression.

Key words: **selection, codon usage, gene expression, noise.**

Article

Introduction

Codons encoding the same amino acid, although synonymous with respect to the protein sequence, appear in most genes at unequal frequencies. These frequencies differ markedly from those expected from genomic nucleotide composition alone, suggesting the action of selection in addition to mutational biases. The selective force at work has long been linked to protein synthesis—translation—by multiple convergent lines of evidence. Across taxa, genes with higher expression show stronger biases toward a subset of codons (Ikemura 1981, 1982). Further, the frequency of codons in a genome correlates with the abundance of their cognate transfer RNAs (tRNAs), and the correlation is higher in highly expressed genes (Ikemura 1981, 1982; Kanaya et al. 2001). Similar effects have been found in diverse organisms, from bacteria to mammals (Kanaya et al. 2001; Drummond and Wilke 2008; Plotkin and Kudla 2011), arise in all codon families, and span the full length of genes.

We aim to estimate the strength of translational selection on codon usage (SCU). Doing so in a meaningful way requires an understanding of the mechanistic basis of this selection. Recently, numerous observations and hypotheses about codon usage have emerged: Rare synonymous codons induce translational pausing to allow proteins to fold (Kimchi-Sarfaty et al. 2007); a stretch of codons ordered roughly by genomic frequency populate the 5′-end of coding sequences, putatively slowing ribosome transit early in the coding sequence (Tuller et al. 2010); in bacteria, codons which mimic the 16S ribosomal RNA-binding Shine–Dalgarno sequence are used at low frequency to avoid stalling ribosomes (Li et al. 2012); and codons at the 5′-end of mRNAs avoid forming stable RNA structures with their neighbors to facilitate ribosomal initiation (Kudla et al. 2009; Gu et al. 2010). Although these phenomena are of major interest, they address only a small subset of sites or codons, such that none of them can explain the core genome-wide signature of translational selection described earlier.

Although there is ongoing debate on the underlying causes of translational selection, the menu of viable mechanistic hypotheses remains short. Synonymous codons have been found to differ in the speed (Sørensen et al. 1989) and the accuracy with which they are translated (Dix and Thompson 1989; Kramer and Farabaugh 2007; Kramer et al. 2010). Both forms of selection play some role (Hershberg and Petrov 2008; Qian et al. 2012), though their relative contributions remain a matter of debate. Either of these causes is also compatible

with the observed correlation between codon preference, gene expression, and cognate tRNA abundance. Because the first step of tRNA binding on the ribosome is thought to be proportional to the tRNA abundance (Zaher and Green 2009), increasing the level of a codon's cognate tRNA will increase both the codon's speed and accuracy of translation. We argue below that selection on either speed or accuracy leads to SCU that scales with gene expression.

The speed hypothesis attributes selection on synonymous codons to differences in the rate at which ribosomes make complete proteins (Andersson and Kurland 1990; Bulmer 1991) (fig. 1). The rate of ribosomal initiation on an mRNA, considered rate limiting for protein synthesis (Mathews et al. 2007), is proportional to both the number of transcripts and the number of free ribosomes. Coding an mRNA with codons that are translated more rapidly allows ribosomes to complete translation sooner, freeing them to initiate translation of a new message and leading to a global increase in the rate of initiation. The global protein synthesis rate increase resulting from use of a faster codon is roughly proportional to the rate of ribosomal reads at that codon, which is in turn proportional to the mRNA expression level of the gene containing that codon. For organisms whose evolutionary success depends on their growth rate, and thus on the rate of protein synthesis, use of a faster codon leads to a selective advantage proportional to the level of gene expression. The speed hypothesis explains a constellation of correlations between bacterial growth rate, ribosomal RNA copy number, codon usage, tRNA copy number, and gene expression level (Vieira-Silva and Rocha 2010).

The accuracy hypothesis attributes selection on synonymous codons to differences in the accuracy with which they are translated (fig. 1). Tests specific for accuracy selection reveal its action in genomes from bacteria to mammals (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008). Approximately one in every 10^3–10^4 ribosomal reads of a codon results in a mistranslation in vivo (Ogle and Ramakrishnan 2005), such that very roughly, one in five average-length proteins will be translated with at least one error, an unknown but presumably small fraction of which will disrupt folding and function. Misfolded proteins possess intrinsic cytotoxicity (Drummond and Wilke 2009; Geiler-Samerotte et al. 2011), which reduces fitness in proportion to the number of misfolded proteins produced (Geiler-Samerotte et al. 2011). Consequently, the per-generation fitness cost or benefit of changing accuracy at a particular codon is the accuracy change at one codon multiplied by the number of codon reads per generation, which is again proportional to the expression level of the gene in question.

The key question is, how to estimate the per-ribosomal-read fitness cost of substituting one codon with its synonym from observed codon usage patterns? A recent major advance in quantifying mutation bias and SCU was made by Shah and Gilchrist (2011), who introduced a model that exploits an exact analogy between the codon frequencies predicted from population genetics (Bulmer 1991; Sella and Hirsh 2005) and a standard method in statistics called multinomial logistic regression (Gelman and Hill 2007). The dependence of selection on gene expression allows this model

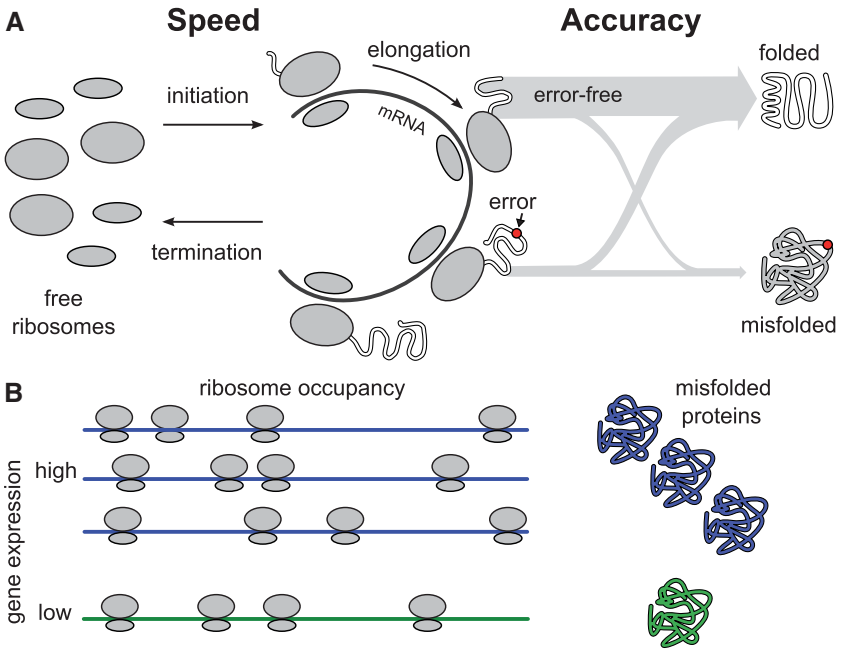

Fig. 1. Selection on speed and accuracy. (*A*) Free ribosomes initiate translation of an mRNA and begin elongation. Initiation is rate limiting for translation, and the availability of free ribosomes is affected by the rate of elongation, as the quicker a ribosome completes elongation and termination, the sooner it may initiate on another mRNA. Some proteins are produced with errors or mistranslation events. Though errors are rare, mistranslated proteins are more likely to misfold. (*B*) The number of ribosomes occupied, and the number of misfolded proteins produced, per gene, are each proportional to gene expression.

to disentangle the effects of mutation and selection, based on the assumption that selective pressures scale with gene expression, while mutational biases do not. Although the original model attributes selection to speed through the free-ribosome mechanism described above, it is equally compatible with accuracy selection, because it requires only that the selection pressure scale with gene expression level.

Gene expression levels vary in vivo and, similar to all empirical quantities, are measured with error. Following Raser and O'Shea (2005), we use the term "noise" in gene expression to refer to "the measured level of variation in gene expression among cells, regardless of source, within a supposedly identical population." Such noise in gene expression, if unaccounted for, distorts estimates of selection and mutation. The core problem can be understood by thinking about the inferences of selection and mutation expected if expression measurements were very noisy: Codon frequencies would necessarily correlate poorly with such measurements, leading to an apparently small effect of selection and inflating the effects of mutation. We show that selection and mutation coefficients predicted from different expression measurements differ systematically consistent with differing noise levels in these data. Moreover, we show that noise in gene expression leads to dramatic underestimation of selection and misestimation of mutational bias: If noise is substantial enough, selection is inferred to be absent.

We present an analytical framework that explicitly accounts for noise in measured gene expression, allowing unbiased estimation of the underlying selection and mutation coefficients in the presence of moderate-to-high amounts of noise. At the core of this method is a probabilistic model for observed gene expression that depends on the true but unobservable gene expression, which is encoded by a latent variable, as well as a model for the distribution of this latent gene expression. These models are combined with the population genetics model used in Shah and Gilchrist (2011) for codon usage conditional on latent gene expression. Using a combination of real and simulated data, we assess the extent of the bias expected from noise-blind methods, in a range of transcriptional noise settings. These confirm that noise-blind methods underestimate the strength of selection, whereas our framework gives unbiased estimates of selection and mutation, even with moderate to high noise in expression measurements.

Applying this framework to budding yeast, we infer selection coefficients which, although still compatible with weak selection on synonymous sites, are substantially higher than those inferred using noise-blind approaches. Estimates of selection coefficients derived from different data sets agree much more closely applying our framework than when noise is not modeled, confirming that our approach is more robust to measurement error. As an external check, the inferred mutation biases are highly correlated with direct experimental measurements. Because our model produces codon-specific estimates of translational selection, we then estimate the mean translational selection on a gene, which we term SCU. Our results confirm that genes in yeast range from apparently free of selection to strongly selected. This SCU correlates strongly with the codon adaptation index (CAI) (Sharp and Li 1987), the most common genewise measurement of synonymous codon usage, with the advantage that, unlike CAI, the numerical values of SCU have evolutionary meaning.

Results

Estimating Selection Coefficients and Mutation Bias from Genomic Data and Gene Expression Measurements

The frequency distribution of a codon in the genome results from mutation, selection, and genetic drift, which we model using population genetics theory (Bulmer 1991). We model selection on synonymous codons due to speed and accuracy as proportional to gene expression. Consider codon c, encoding amino acid a, in a gene g which has gene expression x_g measured in mRNA transcripts per cell. The population genetic parameters are S_c, the selection coefficient of translational speed, and accuracy per mRNA per cell, scaled by effective population size, and M_c is a mutation bias coefficient: The quantity $e^{M_c - M_{c'}}$ is the ratio of mutation rates between codons c and c'. Then the probability that codon c is found at a site coding for amino acid a is as follows:

$$\pi_{gc} = \frac{\exp(M_c + S_c x_g)}{\sum_{c' \mid a} \exp(M_{c'} + S_{c'} x_g)} \quad (1)$$

where the sum in the denominator is over all codons c', which code for amino acid a. This is the same model as in Shah and Gilchrist (2011); the derivation is summarized in supplementary text, Supplementary Material online. Given an organism with a sequenced genome, one may then use the codon counts across the genome, and measurements of gene expression, to estimate the mutation bias and selection coefficients.

Gene Expression Measurements Are Noisy

Once it is clear that gene expression must play a role in the inference of selection, the question becomes, which gene expression measurement should be used? Direct experimental measurements of gene expression by RNA sequencing in haploid budding yeast growing exponentially in rich medium have been made by multiple groups, but they vary widely. Figure 2*B* compares all replicate mRNA abundance measurements from three such published data sets (Ingolia et al. 2009; Lipson et al. 2009; Yassour et al. 2009) (cf. supplementary table S1, Supplementary Material online). Generally, two replicates from the same study have a correlation above 0.9, but using two different studies, the correlation varies between 0.6 and 0.9. The exception is that measurements from the same biological sample, measured with two different platforms, have a correlation below 0.9. It appears that the major source of variation is neither the much-studied phenomena of cell-to-cell variability nor technical variability in instrument runs. Rather, most variability occurs between platforms, and between

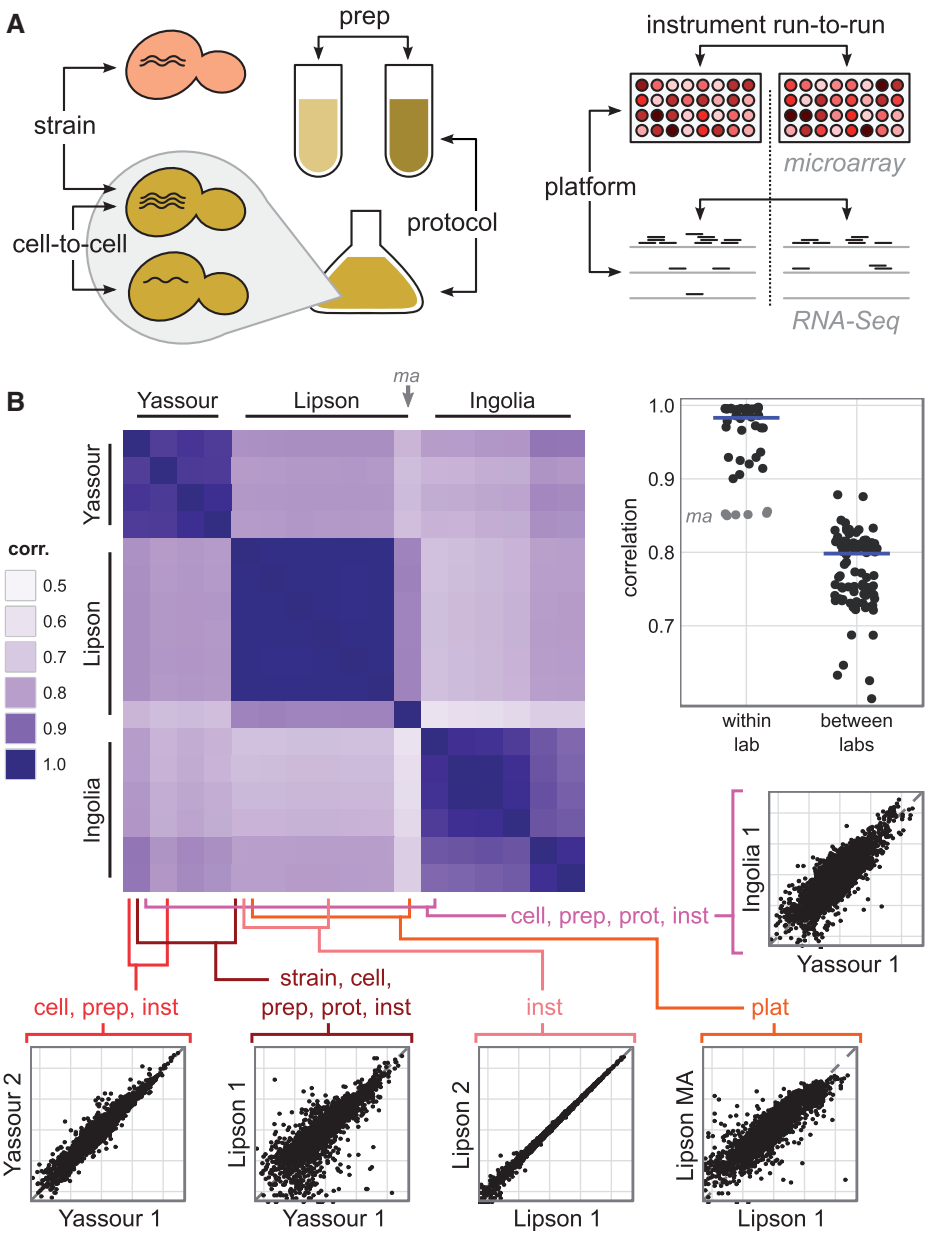

Fig. 2. Noise in gene expression in budding yeast. (*A*) Sources of variability in gene expression measurements. (*B*) Expression measurements compared from three recent studies of haploid budding yeast growing exponentially in rich medium (Ingolia et al. 2009; Lipson et al. 2009; Yassour et al. 2009). Main plot, pairwise correlation coefficients of log-transformed expression measurements from these studies. The gray label *ma* indicates the only replicate to use microarray rather than RNA-Seq measurements. Right, summary plot of all pairwise correlation coefficients, grouped according to whether the comparison is within or between laboratories; blue line represents median, and dots are dispersed horizontally for clarity. Replicates from the same study tend to agree more than replicates from two different studies. Below, selected pairs of expression measurements are compared on a log–log scale. In these expression plots, each point represents a single gene's measured mRNA abundance per cell.

laboratories, arising presumably from differences in the protocol used or in its implementation, sometimes termed "batch effects" (Leek et al. 2010).

These observations suggest that quantifying noise by examining replicate measurements performed by a single group, while placing bounds on noise produced by a clonal culture, sharply underestimates the actual experimental noise. Moreover, the error arising from use of average gene expression under laboratory conditions as a proxy for the true amounts of translation assumed to actually shape codon usage during the evolution of budding yeast—which we term "proxy error"—remains unestimated and is surely nontrivial.

The magnitude of noise in measured expression of a particular gene scales with that gene's expression (Bar-Even et al. 2006): This is confirmed in figure 2*B* where the plots of expression data sets on a log–log scale show that, equivalently, two measurements differ by a proportion that is roughly independent of expression level. Lognormally distributed noise, or equivalently normally distributed noise on a log-transformed scale, is the simplest way to model expression noise consistent with these observations.

Noise in Gene Expression Degrades the Inference of Selection and Mutation

Mutation and selection coefficients inferred from different measurements of gene expression differ systematically, by as much as 2-fold between data sets. As shown in figure 3*A*, selection coefficients inferred from one data set are roughly proportional to selection coefficients inferred from another, with a minimum R^2 of 0.91 across all pairs of data sets from figure 2*B* but with a slope far from unity. To demonstrate that adding noise suffices to produce such slope differences, figure 3*B* shows the effect of artificially adding lognormal noise to a single gene expression data set from Yassour et al. (2009); estimating selection coefficients from these again produces values which are proportional, with reduced slope, to those estimated from the original data set. The close resemblance of figure 3*A* and *B* demonstrates that, with respect to this inference, the differences between data sets are well described by simple differences in noise level.

As figure 3*C* shows, far from changing selection estimates randomly, noise pushes these selection estimates closer to zero. This is an example of attenuation error or regression dilution bias (Carroll et al. 2006). Intuitively, if the noise in measured expression were so large as to overwhelm the signal, then codon usage would have no dependence on measured expression, and the selection coefficient—the scaling of codon usage with expression—would be close to zero.

Our estimates of relative mutation rates will be biased (supplementary fig. S1, Supplementary Material online) to compensate for the misestimation of selection. The extent of bias in selection and mutation estimates depends on the expression noise, so it is necessary to estimate this noise to infer the parameters of biological interest.

Every data set has noise and suffers from proxy error. Moreover, expression measurements appear to be systematically noisier in some published studies than in others—figure 3*A* suggests that the Yassour et al. (2009) data sets produce the least-attenuated selection estimates. Least-noisy is not non–noisy, so fitting selection coefficients to any single expression data set in a noise-blind manner is certain to underestimate the magnitude of selection. Noise-induced attenuation of inferred selection is not uniform across codons, so that although it is easy to predict the approximate degree of underestimation of selection if its true value and the noise level are known, the inverse problem

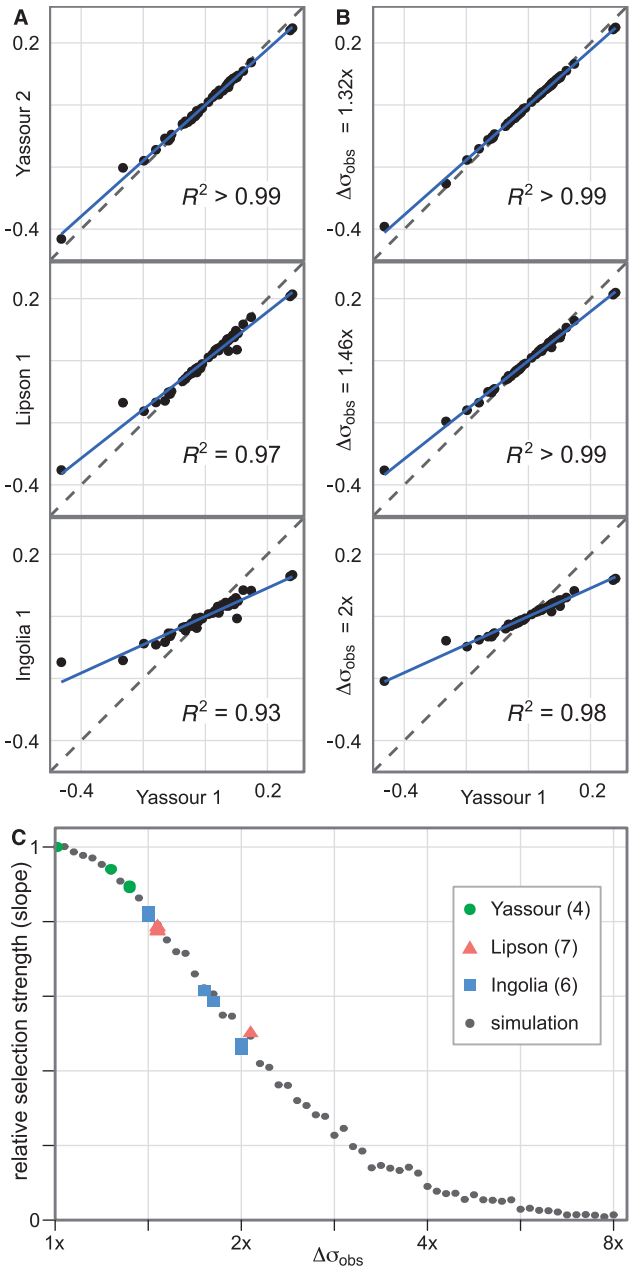

Fig. 3. Noise in gene expression affects selection estimates in budding yeast. (*A*) Codon selection estimates compared from a subset of the yeast expression data sets used in figure 2. Each dot represents a codon, its *x* co-ordinate the selection coefficient estimated from expression data in Yassour 1, and its *y* co-ordinate the selection coefficient estimated from an alternative data set. The solid line (blue, online) is the least-squares fit of the alternative selection coefficients to those from Yassour 1, with the R^2 values reported. The selection coefficients vary systematically between data sets, with those inferred from one data set resembling a scaled copy of those from another. (*B*) Codon selection estimates inferred from adding noise to a yeast expression data set. We artificially add lognormal noise to expression estimates from Yassour 1 and estimate the codon selection coefficients. The noise strength is given by its standard deviation on the log scale: $\Delta\sigma_{obs} = 1.32\times$ means that the root mean square difference between the logarithm of noise-free expression and the logarithm of noise-added expression is 1.32-fold. (*C*) Relative selection strength between data sets, summarized. Each small dot represents a simulated noisy expression data set, generated as in panel *B*, its *x* co-ordinate the noise strength, and its *y* co-ordinate the selection strength relative to Yassour 1, that is, the slope of the least squares fit (solid line in panel *B*). Estimated selection coefficients decline systematically as the noise strength increases. Each large shape represents a replicate published data set from one of Ingolia et al. (2009), Lipson et al. (2009), Yassour et al. (2009), its *y* co-ordinate the selection strength relative to Yassour 1 (slope of solid line in panel *A*), and its *x* co-ordinate the noise strength of the simulated data set with closest relative strength. There are four replicates from Yassour et al., seven replicates from Lipson et al., and six from Ingolia et al., although some of these are overplotted and visually indistinguishable.

of interest—inferring the true selection coefficient from given unknown noise levels—is hard. These statements hold for any estimate of a biological quantity dependent on gene expression: Failing to account for expression noise leads to an estimate that is biased and biased by an unknown amount.

Modeling Expression Noise Gives More Accurate Inference of Selection and Mutation

We present a method for inferring selection coefficients and mutation bias more accurately by directly modeling noise in gene expression and treating the unobserved underlying gene expression as a latent variable. There are three ingredients: first, the codon distribution conditional on latent gene expression and mutation and selection parameters described above; second, the observed gene expression conditional on the latent gene expression; and third, the distribution of latent gene expression. We model the expression observation with lognormal error as discussed earlier, introducing a parameter for the magnitude of observation noise: σ_{obs} is the standard deviation of measurement noise on a log scale. We model the distribution of latent gene expression as log-asymmetric Laplace, a long-tailed distribution that describes gene expression measurements better than the more standard lognormal distribution (Purdom and Holmes [2005] and supplementary fig. S2, Supplementary Material online). This third ingredient is necessary to ensure that the distribution of latent expression is constrained to plausible values. The model is described in Materials and Methods and supplementary text, Supplementary Material online; statistical considerations involved will be addressed in more detail in a future publication (Wallace et al., in preparation).

Briefly, our modeling assumptions imply a likelihood function. We fit the model to the data using an iterative Markov chain Monte Carlo (MCMC) sampling strategy, which provides not only point estimates of selection coefficients and mutation bias but also the full (posterior) distribution of the parameters given the data (Gelman et al. 2003). Because we do not have external knowledge of the parameters controlling latent gene expression nor measurement noise, we infer these parameters simultaneously with the selection and mutation coefficients.

A concern arises as to how to normalize the estimates in terms of the absolute mRNA abundance per cell. Gene expression in RNA-Seq is estimated in reads per kilobase, and this may be converted to units of mRNA count per cell by fixing the sum at an estimate of total mRNA abundance per cell derived independently. However, for a right-skewed error model such as the lognormal, fixing the sum of observed gene expression alone produces biased estimates, because the latent expression estimates have on average a lower sum. Accordingly, following the completion of the MCMC sampler, we normalize the estimates of latent gene expression by their sum at each iteration (see Materials and Methods).

Results from fitting to codon counts and mRNA abundance measurements from budding yeast, shown in figure 4, are as anticipated: For every abundance data set, estimates of selection coefficients are larger when noise is modeled than when using the noise-blind approach (fig. 4A). Because abundance data from Yassour et al. (2009) were suggested to be the least noisy in figure 3, we used the geometric mean of Yassour replicates as our primary data set (see Materials and Methods) and for comparison took a single replicate from each of Ingolia et al. (2009) and Lipson et al. (2009). For the Yassour data set, estimates of selection coefficients range from 0.7-fold to 4.7-fold higher than in the noise-blind approach, typically 1.7-fold higher. The rank order of selection coefficients changes slightly once we model expression noise (Yassour, Spearman rank correlation = 0.98). Summary statistics comparing noise-modeled with noise-blind estimates for each data set is shown in supplementary table S2, Supplementary Material online.

Similarly, for the Yassour data set, the rank order of mutation coefficients changes upon modeling noise (Spearman rank correlation = 0.82, see supplementary table S2, Supplementary Material online). Note that, when selection coefficients increase, mutation bias coefficients may increase or decrease, depending on whether the bias favors the selected codon.

Importantly, selection estimates from different abundance data sets agree more closely when noise is modeled (fig. 4C and supplementary fig. S4, Supplementary Material online). The results with noise modeled are very strongly correlated ($R^2 > 0.99$) for each pair of data sets and are much closer in magnitude than when noise is not modeled; the difference in magnitude may be partially explained by each data set detecting different total numbers of genes. This demonstrates that modeling noise makes inferences of selection more robust, in that the results depend less on the particular gene expression data set used.

When noise is ignored, selection-driven codon usage, which scales strongly with true gene expression, will be observed to scale poorly with measured gene expression and thus will be interpreted as mutational in origin. Because of this noise-induced cross-contamination, we expect that noise-blind estimates of selection and mutation will correlate; accounting for noise should reduce this effect. Indeed, in the Yassour data set, noise-blind estimates of mutation and selection have $R^2 = 0.21$, whereas when noise is modeled, mutation and selection coefficients are uncorrelated $R^2 = 0.001$.

For the Yassour data set, the estimated noise value is $\sigma_{obs} = 1.13 = \log(3.12)$, meaning that the average difference between the measured and true abundance value is 3.12-fold. Consistent with the observation that most variability occurs between platforms and between laboratories, the noise estimated between replicate measurements in the Yassour data set is considerably smaller, $\sigma_{obs} = 0.056 = \log(1.06)$.

We have argued that fitness costs of speed or accuracy at a codon in a particular mRNA scale with its protein synthesis rate, that is, the total rate of ribosomal reads per codon across all transcripts. We used mRNA abundance as a proxy for this protein synthesis rate, building upon considerable evidence that mRNA abundance strongly correlates with protein synthesis rate despite the layers of post-transcriptional regulation

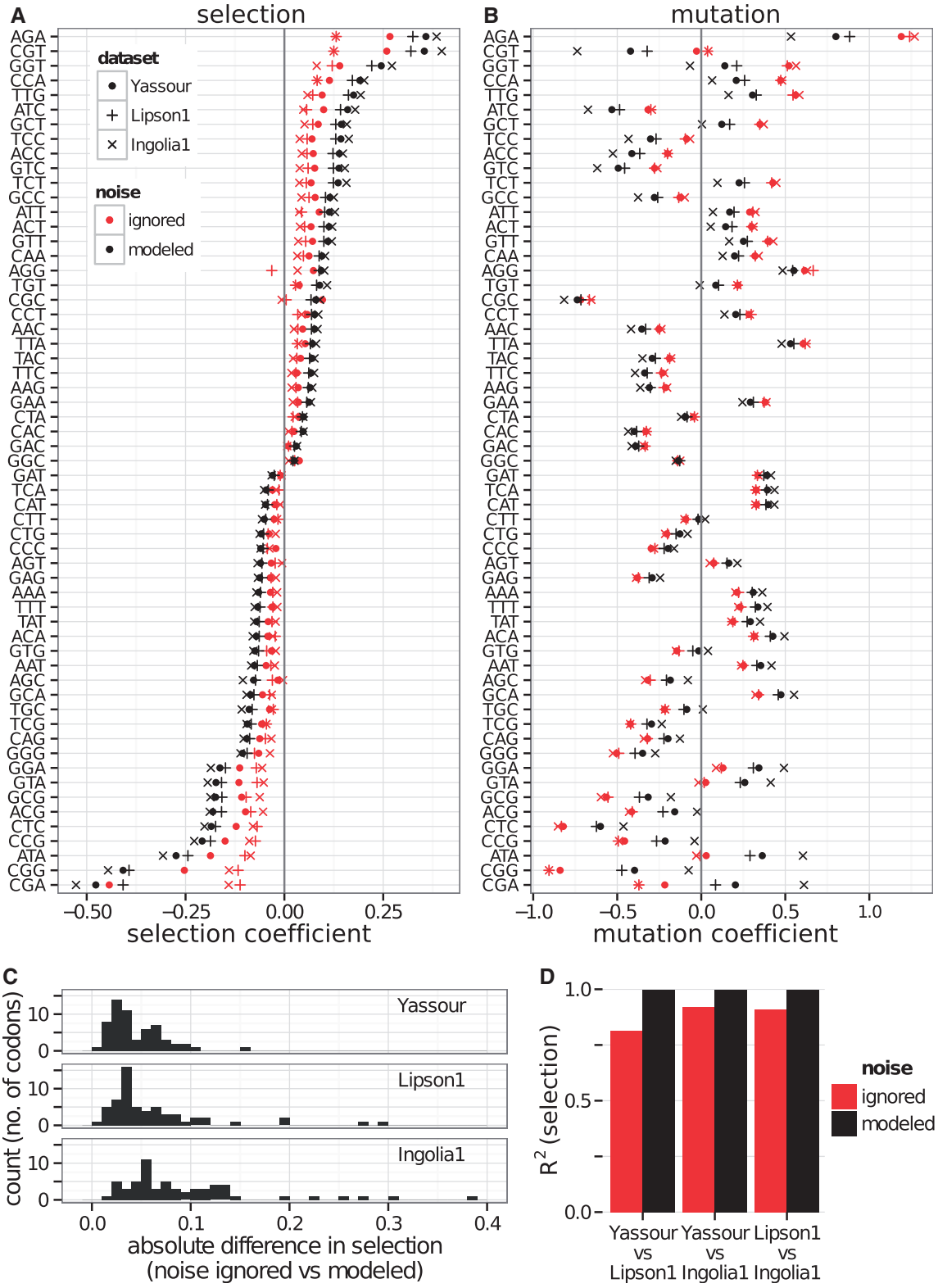

Fig. 4. Explicit modeling of noise reveals stronger synonymous-site selection in budding yeast. (*A*) Selection coefficients and (*B*) mutation coefficients: We compare estimates where we modeled noise in gene expression (black) or ignored noise in gene expression (grey; red, online), based on expression data taken from Yassour et al. (2009), Lipson et al. (2009), and Ingolia et al. (2009) (see Materials and Methods). These are displayed ranked in order of the noise-modeled selection coefficient in Yassour and labeled by codon and amino acid. The Yassour data, collected by amino acid, is shown in supplementary figure S3, Supplementary Material online. (*C*) Absolute difference in selection coefficient estimates for each data set with median values indicated. (*D*) R^2 (squared correlation) of each pair of selection estimates. Selection estimates are more closely correlated between data sets when noise is modeled than when noise is ignored.

(Arava et al. 2003; Mathews et al. 2007). Ribosome profiling, a recently developed technique that directly measures ribosome occupancy of messenger RNA, gives per-gene estimates of synthesis rate without employing any additional data, under the assumption that the average elongation rate does not vary systematically across codons or genes (Ingolia et al. 2009), a convenient notion which remains controversial.

We repeated the above analysis with the ribosome profiling data set of Ingolia et al. (2009) and find substantial agreement (supplementary fig. S6, Supplementary Material online). The selection coefficients inferred from the Yassour data set and the ribosome profiling data set, measured relative to transcripts per cell or estimated proteins produced per second per cell, respectively, are very strongly correlated ($R^2 > 0.99$). For the ribosome profiling data set, selection coefficients were underestimated by only 1.3-fold (median) in the noise-blind fit compared with the full model, yet expression observations had an estimated noise value of $\sigma_{obs} = \log(3.70)$, suggesting that the observation noise relative to underlying selective pressure was greater for this data set. We use estimates from the Yassour data set in subsequent analyses except where noted and provide the selection and mutation estimates from all data sets as a supplementary file, Supplementary Material online.

To test the model's ability to infer true values in the presence of noise, we generated data using known selection/mutation coefficients, added noise, and attempted to infer the coefficients from the noisy data. We used amino acid counts from the yeast genome, and estimated coefficients from the Yassour data set, to simulate data according to the model. As figure 5 shows, the re-estimated coefficients remain close to the coefficients used to generate the data across a range of measurement noise which contains that inferred for the Yassour data set. The simulation results shown in figure 5 verify that this approach produces approximately unbiased estimates given the model assumptions. In contrast, the noise-blind model systematically underestimates the magnitude of selection parameters and misestimates mutation parameters in a compensatory manner. At the noise level inferred for the Yassour data set, the noise-blind approach underestimates selection strength by roughly a factor of four.

The counts of codons per gene provide one measure of model fit (Shah and Gilchrist 2011). To what extent does the noise-corrected model fit codon counts better? We estimated the Bayes factor (Gelman et al. 2003), the difference between the posterior mean likelihood of codon counts of the full model and the noise-blind model. For the Yassour data set, the log of the Bayes factor was 2.343×10^4, which is statistically significant by any measure. The large size of the Bayes factor is due to the size of the data set and the improvement in fit: This means an average 0.8% increase in likelihood for each of the 2.7 million codon sites or an average a 61-fold increase in likelihood across the codon counts for each of the 5,709 genes.

In summary, the noise-blind approach substantially misestimates the selective and mutational forces acting on codon usage, and this misestimation is corrected, under some circumstances, by explicitly modeling noise.

Estimated Mutation Parameters Are Consistent with Experimental Measurements

The mutation bias coefficients in the model estimate the log difference in relative mutation rates between a codon and its

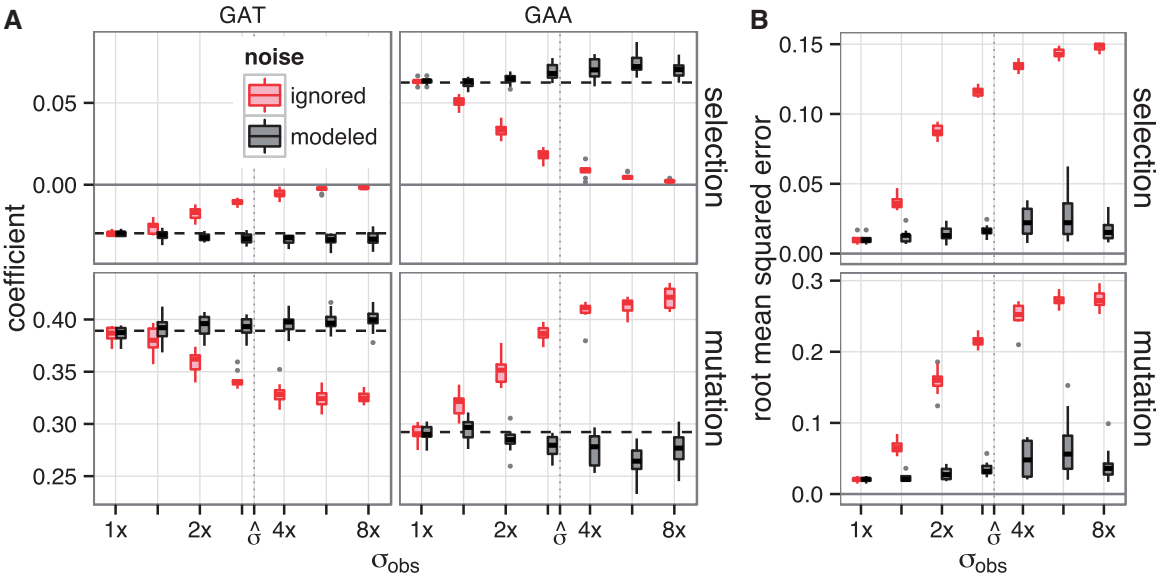

FIG. 5. Modeling recovers true coefficients from simulated noisy data. Mutation and selection parameters inferred from simulated data using amino acid counts from 1,000 yeast genes, with parameters taken from model fit to the yeast genome; 12 data sets were simulated for each value of the noise parameter σ_{obs}. (*A*) Parameters inferred for codons GAT (Asp) and GAA (Glu). Black boxplots indicate estimates from full model, and gray boxplots (red, online), those that do not model noise in expression; the vertical dashed line is the parameter value used to generate the data. (*B*) Root mean-squared error of selection and mutation parameters across all degenerate amino acids. To provide a reference, $\hat{\sigma}$ indicates the inferred noise parameter fitted to the Yassour data set.

synonym. These estimates represent strong predictions, because only steady-state sequence composition and mRNA levels are provided to the model. Moreover, they are testable predictions, as several groups have measured single-nucleotide mutation rates in budding yeast from which relative rates can be computed.

We used single-nucleotide mutation rates collated from Kunz et al. (1998), Lang and Murray (2008), and Ohnishi et al. (2004), as reported in Lynch et al. (2008), to estimate the relative difference in mutation rates between synonyms differing by a single nucleotide, generally at the third position. Experimentally measured rates are reported per base pair, and we do not distinguish lead and lagging strands, so for A → T the log ratio of counts $\log\left[N_{(A:T\rightarrow T:A)}/N_{(T:A\rightarrow A:T)}\right] = 0$, and similarly for C → G. These mutation parameters in the model correlate with experimental measurements of mutation bias, as shown in figure 6A ($R^2 = 0.77$). However, the model estimates smaller ratios than experimental measurements, with the linear regression coefficient of 0.38. This could be due to other unmodeled factors affecting mutation rates, such as sequence context.

In contrast, the inferred selection coefficients are poorly correlated with experimental measurements of mutation bias, with a regression coefficient of 0.05 ($R^2 = 0.02$) (fig. 6B), suggesting that the inference of selection is independent of mutational forces acting on the genome.

The question arises, are the selection and mutation estimates obtained so far compatible with a classical model where mutation is driven only by single-nucleotide changes? We fitted an alternative model where selection coefficients are fitted as previously, but mutation bias depends only on the nucleotide substitution, and so mutation coefficients are pooled across amino acids. This is analogous to the model of Yang and Nielsen (2008) but with gene expression as a coefficient of selection. Selection estimates when mutation is pooled, again directly modeling noise in gene expression, are in substantial agreement with those where mutation is modeled separately ($R^2 = 0.92$, supplementary fig. S7B, Supplementary Material online). The spread of values in figure 6A previously suggested that different amino acids show different mutation bias. Likewise, codon mutation coefficients derived from pooled mutation estimates are in slightly weaker agreement with those from unpooled estimates ($R^2 = 0.86$, supplementary fig. S7D, Supplementary Material online). Selection estimates necessarily change to compensate for these disagreements in mutation estimates. Interestingly, even when noise is ignored, selection estimates are larger when mutation is pooled than when mutation is fitted separately (supplementary fig. S7A, Supplementary Material online), especially for arginine codons containing a CpG dinucleotide. This additionally suggests that pooling mutation partially corrects for attenuation due to noise in gene expression.

Genewise Estimates of Translational Selection

We often wish to summarize selection at the gene level. To provide a measure of average codon selection for a polypeptide-encoding sequence, we introduce the measure SCU, the

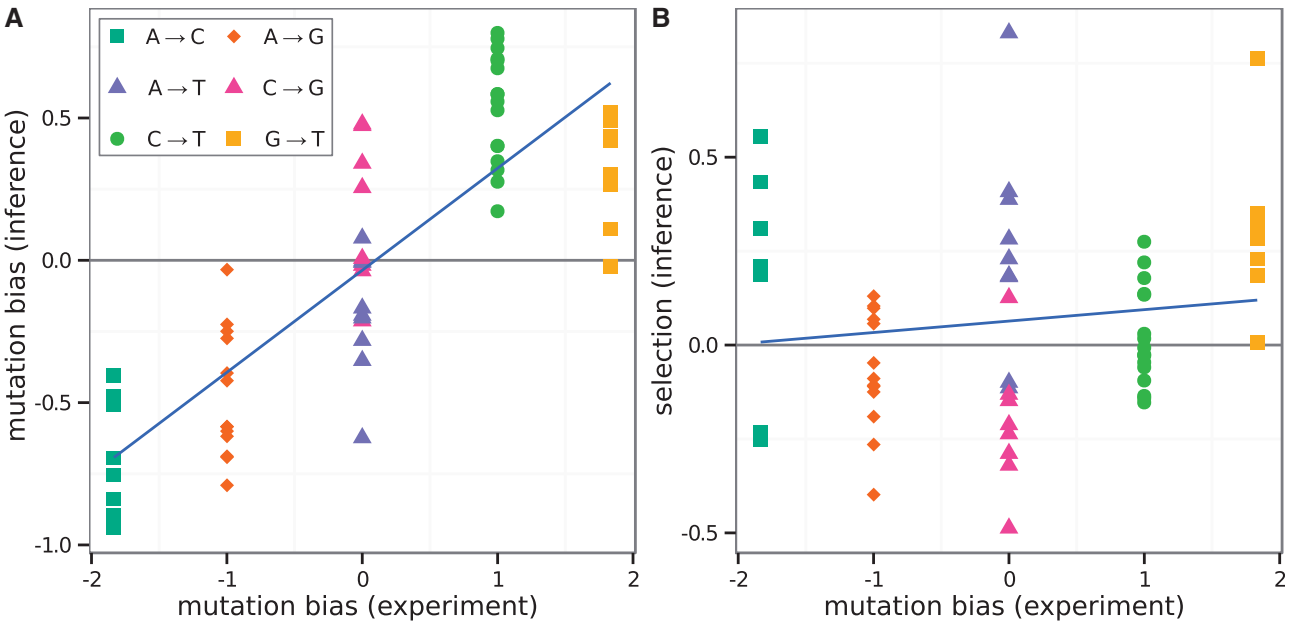

FIG. 6. Mutation parameters are consistent with experimental measurements in budding yeast. (A) Inferred mutation bias coefficients plotted against experimental estimates derived from single-nucleotide mutation rates in yeast. For example, the difference in model parameters $M_{GGC} - M_{GGA}$ is compared with the experimentally measured log ratio of counts $\log\left[(N_{(A:T\rightarrow C:G)}/N_{(C:G\rightarrow A:T)}\right]$ collated in (Lynch et al. 2008) (see Materials and Methods). The experimentally measured rates are reported per base pair, and we do not distinguish lead and lagging strands, so for A → T the value $\log\left[N_{(A:T\rightarrow T:A)}/N_{(T:A\rightarrow A:T)}\right] = 0$, and similarly for C → G. Blue line, linear regression fit of model parameters to direct measurements, has slope of 0.38 and $R^2 = 0.77$. (B) Selection estimates from model plotted against log ratio of directly measured single-nucleotide mutation rates. Linear regression fit has slope of 0.05 and $R^2 = 0.02$.

average per-codon selective advantage of a gene's observed encoding over that of a random gene encoding the same polypeptide. Our model provides a precise and useful meaning of "random" as a gene whose codons are chosen according to the inferred mutational biases alone, which we refer to as an unselected sequence. The codon selection coefficients inferred above are selective differences relative to an arbitrary point and gain evolutionary meaning only in relation to each other: $S_c - S_{c'}$ is the population-scaled per-transcript fitness advantage of replacing codon c' with c. We first define S'_c as the mean per-transcript selective advantage of codon c over its unselected synonymous alternatives (see Materials and Methods). Then the mean S'_c for a gene, mSCU, is the average of S'_c over all codons in that gene, and SCU is mSCU multiplied by the gene's transcript level. SCU is the mean fitness advantage, scaled by effective population size, that the organism gains from the gene's codon composition, relative to an unselected synonymous alternative.

Figure 7*A* shows the distribution of SCU values for all yeast genes compared with values for unselected synonymous alternatives generated from model-estimated mutation biases alone. These distributions partially overlap, suggesting that codon usage in some genes is driven primarily by selection and in others primarily by mutation bias.

Evolution acts efficiently on selective differences above $1/N_e$ and therefore efficiently selects an average codon in a gene if that gene's SCU > 1. Efficient selection means that lower-fitness alternatives are purged by natural selection, and in this context, it means that the most advantageous synonym will be used almost exclusively in the set of genes for which SCU > 1; this is true for 249 of 5,709 genes. Below SCU = 1, drift begins to dominate, allowing accumulation of lower-fitness codons.

The most popular of many codon usage metrics, the CAI (Sharp and Li 1987), provides a useful comparison for visualizing these effects. CAI assigns a score to each codon proportional to its frequency in a group of high-expression genes, with the highest score for each synonymous family equal to 1; the CAI for a gene is the geometric mean of these scores across its codons. Exclusive use of the highest scoring codons yields CAI = 1. Figure 7*B* compares SCU values to CAI scores for all genes in budding yeast, showing that they are strongly correlated. Genes with SCU > 1 have CAI values approaching 1.0.

The per-transcript average selective advantage mSCU (distributions in fig. 7*C*) provides a more direct comparison with CAI. Both log(CAI) and mSCU represent mean codon scores, and the relationship between mSCU and log(CAI) is

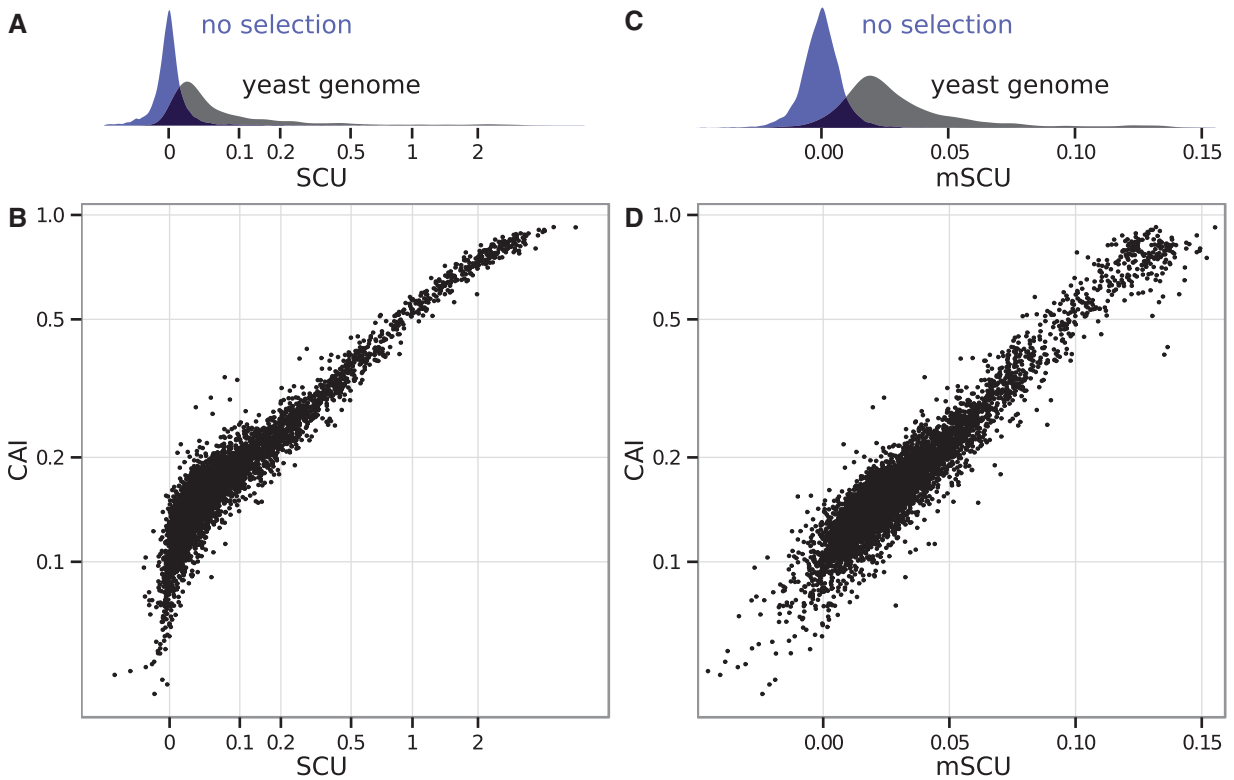

Fig. 7. Genewise estimates of translational selection in budding yeast. SCU measures the per-codon average selective advantage of a gene relative to an unselected gene (see text), and mean selection per codon mSCU measures the per-codon per-transcript selective advantage. (*A*) Distribution of SCU values for all budding yeast genes compared with unselected genes, generated with identical amino acid counts and model-estimated mutation biases absent selection. These distributions partially overlap, suggesting that codon usage in some genes is driven primarily by selection and in others primarily by mutation bias. Because SCU varies over orders of magnitude, yet has negative values, we have plotted it on a scale transformed by $\log(0.1 + x)$. (*B*) SCU values for all budding yeast genes compared with CAI, a popular metric of codon usage bias. Each point represents one gene. (*C*) Distribution of mSCU values across the yeast genome compared with those generated with identical amino acid counts and model-estimated mutation biases but without selection. (*D*) mSCU compared with CAI. The relationship between mSCU and log(CAI) is close to linear ($R^2 = 0.85$).

close to linear ($R^2 = 0.85$). Additionally, we compare CAI to latent gene expression in supplementary figure S8*B*, Supplementary Material online.

The key advantage of SCU and mSCU is their interpretability: We have estimated both mutation and selection across the genome in a manner that allows us to tease apart the influences of these forces. Our statistics directly estimate the strength of translational selection on the gene's coding sequence: One might view mSCU as CAI calibrated to reflect the genomic signature of selection and SCU as the total selective strength accounting for transcript level. Because virtually all metrics of codon bias correlate well, it is sensible to prefer one for which the numerical values have biological meaning. We report SCU and mSCU values for all verified and uncharacterized open reading frames (ORFs) in budding yeast in a supplementary file, Supplementary Material online.

Discussion

We estimated the strength of translational SCU in budding yeast. Our per-codon selection coefficients, although still compatible with weak selection on synonymous sites, are substantially higher than those previously inferred, because previous estimates did not account for the effects of noise in measured gene expression. We confirm that a substantial portion of the yeast genome is under translational selection. Reliable estimation of such selection is essential to understanding how natural selection shapes genomes.

The key challenge in codon usage analyses has been to separate the effects of mutation, selection, and drift (Bulmer 1991). Following Shah and Gilchrist (2011), we directly model mutation and translational selection, separating their effects by exploiting the scaling of translational selection with gene expression. Going further, we account for noise in gene expression, which we show confounds the effects of mutation with selection otherwise.

Previous estimates of selection on synonymous codons include that of Yang and Nielsen (2008), which models the background codon frequencies in coding genes as given by the genome's nucleotide composition. However, the nucleotide composition is itself affected by SCU: Simply counting nucleotide frequencies confounds the signatures of mutation and selection.

In contrast, Sharp et al. (2005) estimated SCU in bacteria by comparing a basket of highly expressed, highly biased genes, to others, using this internal control to limit the influence of mutation bias. They study a handful of amino acids, chosen for consistent mutation bias across bacteria. Our approach, although similar in spirit, applies to all codons, uses expression measurements for all genes, and takes a more rigorous position with respect to the kinds of selection which can be inferred, namely those which scale proportional to gene expression.

The explicit use of gene expression, we argue, is necessary to extract selective parameters with a mechanistic basis. The average speed of translation of a codon in a cell need not relate at all to the speed with which one ribosome translates one codon of that type: The average speed of translation is the number of translation events of that codon per unit time, and the number of translation events is primarily determined by the number of transcripts containing that codon. In contrast, the average codon translational speed per transcript has mechanistic meaning. By inferring the selection coefficient per transcript, our method converges on evolutionary parameters, which relate to the physical implementation of those parameters. This strategy is essential when evolutionary signals are used to learn about the operation of cells, one major goal of molecular evolutionary studies. In turn, this strategy's dependence on measurements, such as gene expression, makes accounting for measurement noise an essential part of making accurate inferences.

Another side effect of depending on gene-expression measurements is that the inferred selection coefficients depend on knowing transcript levels per cell. For example, a 4-fold higher estimate of transcript levels would mean 4-fold lower selection estimates, while preserving mutation estimates; estimates of genewise selection via SCU, in which selection estimates are multiplied by transcript levels, also remain the same. In budding yeast, total transcript levels are not without controversy: Some studies report the total number as approximately 15,000 mRNAs per cell (Hereford and Rosbash 1977), a number validated by a recent meta-analysis (von der Haar 2008). Recent data, applying single-molecule fluorescent detection techniques to a handful of low-abundance yeast genes, each showing a 3- to 6-fold higher expression than determined previously (Zenklusen et al. 2008). Although these differences are argued to demonstrate a higher total of approximately 60,000 mRNAs per cell, in view of the small and biased sample, the higher number cannot be regarded as definitive. This number surely also depends on growth state and environmental factors, making point estimates of limited utility. Whatever the correct absolute numbers, the underestimation of selection coefficients due to measurement noise remains.

Experimental evidence is sorely needed to test the predictions of our and others' models. High-quality studies are beginning to emerge. For example, CGA, the codon (encoding arginine) here found to be most strongly selected against in yeast, was recently found to have the strongest negative effect on gene expression in yeast (Letzring et al. 2010). Insertion of repeated CGA codons to a gene reduced that gene's expression, whereas insertion of other repeated codons, or of a single CGA codon, had a smaller effect. Although this is suggestive of a difference in translational speed, we cannot conclude that selection against CGA codons acts exclusively on speed or accuracy, as these are mutually correlated with cognate tRNA abundance.

High-resolution ribosome profiling data (Ingolia et al. 2009) in principle measure codon-specific elongation speeds. One study combined ribosome profiling data with a host of additional data—tRNA copy numbers, tRNA to ribosome ratios, assumptions about cognate and near-cognate binding, and models of tRNA diffusion—to estimate per-codon elongation rates, which correlated with codon usage patterns (Siwiak and Zielenkiewicz 2010); many biophysical assumptions employed in this study await experimental verification. Other studies, in

contrast, argue that ribosome profiling data show negligible speed variation between codons (Qian et al. 2012). Our selection estimates may be useful in providing a missing ingredient, namely, the minimum speed differences between codons that induce a selectable fitness difference.

There are many other causes of selection on coding sequences not modeled here. For example, translational pausing (Kimchi-Sarfaty et al. 2007), avoidance of particular motifs (Li et al. 2012), and mRNA secondary structure (Kudla et al. 2009; Gu et al. 2010). This is consistent with observed codon counts being more variable than is predicted here by mutation, translational selection, and drift alone. Despite this, averaging over the whole genome reveals the core signature of translational selection that codon usage varies with gene expression: The quantity reported here is the average selection coefficient on speed or accuracy per mRNA, across the genome. Future models might incorporate these other sources of selection either directly or indirectly via overdispersion (Gelman et al. 2003).

Although translational selection on synonymous codon usage due to either speed or accuracy is proportional to gene expression, substantial evidence demonstrates that codon location within genes affects selection. For example, codon usage is on average different in the 50 sites closest to the start codon (Tuller et al. 2010). Similarly, codon frequencies vary between sites within a gene according to protein structure (Zhou et al. 2009; Lee et al. 2010) and evolutionary conservation of amino acids (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008), providing evidence for selection specifically on accuracy. Although our current model cannot distinguish between the effects of speed and accuracy, extending the model to include site-specific effects may allow such a distinction.

As in Shah and Gilchrist (2011), mutations here are modeled with individual rates per synonymous family, a choice which simplifies the model yet may complicate the apparent biology. If mutational changes depended only on the nucleotide being mutated, the underlying mutational model would be considerably simpler, a model adopted in other studies (Yang and Nielsen 2008), which we additionally implemented. Here, mutation estimates for the two models, although highly correlated, are statistically significantly different, suggesting that codon mutation rates are not determined by nucleotide alone. Similarly, substantial prior evidence exists that mutation biases operate on more than just single nucleotides (Karlin and Mrázek 1996), with CpG mutation rates providing a canonical example (Bulmer 1986). Further, mutation rates—and presumably mutation biases—vary within genomes (Baer et al. 2007), for example, varying along a chromosome in budding yeast (Lang and Murray 2011). Our model's ability to reproduce experimentally measured mutational biases to a substantial extent, given only genome sequence and expression values as input, argues for its utility. Still, future work should address more mechanistically grounded mutational models.

The present work exemplifies a situation which is quite general. In standard regression models, unbiased estimates can be obtained given noise in the dependent variable but not the independent variable. Here, the independent variable, gene expression, is noisy; any regression-based estimates will be biased if noise is ignored. Indeed, we have shown that the bias in this particular case, measuring selection coefficients, is often several times the values of the parameters themselves. We expect that the strategies introduced here will be useful in contending with the unavoidable uncertainty that accompanies experimental measurements.

Materials and Methods

Data

We downloaded sequence information for verified and uncharacterized ORFs in *Saccharomyces cerevisiae* from the Saccharomyces Genome Database (http://www.yeastgenome.org/, last accessed April 2, 2013) (file: orf_coding.fasta) in February 2011. We downloaded expression data from Yassour et al. (2009), measured using RNA sequencing from *S. cerevisiae* strain BY4141, a derivative of strain S288C, grown to exponential phase in YPD medium. We restricted our analysis to ORFs whose expression was measured under these conditions in their study, a total of 5,709 genes. There were four replicate measurements (YPD0.1, YPD0.2, YPD15.1, and YPD15.2). Because errors were approximately lognormally distributed, we took as an expression summary the geometric mean of the reads per kilobase per million mapped reads (rpkm) from these measurements, equivalent to taking the log-scale mean, adding 1 to avoid taking log of zero. As in Yassour et al., we normalized this summary figure to a conservative total of 15,000 mRNAs per cell (Hereford and Rosbash 1977; von der Haar 2008), as addressed in the Discussion.

For comparison, we also downloaded mRNA abundance measurements from Lipson et al. (2009) and Ingolia et al. (2009); see supplementary table S1, Supplementary Material online, for details of labeling. We also normalized these read counts to a total of 15,000 mRNAs per cell. We downloaded ribosomal occupancy estimates from Ingolia et al. (2009), normalizing to a total of 13,000 translation events per second per cell (von der Haar 2008).

Model

The model has three components: a component relating codon counts to "latent" (i.e., true but unobserved) gene expression, an error component relating observed gene expression to latent gene expression, and a distribution of latent gene expression (fig. 8). The multinomial distribution of codon counts $\vec{y}_{ga}$ for amino acid a in gene g is specified by a vector of probabilities π_{ga} (whose elements π_{gac} specify the probability that codon c is used conditional on amino acid a in gene g), and the size parameter n_{ga} is given by the number of times amino acid a is observed in gene g. The vector of probabilities $\vec{\pi}_{ga}$ is then specified by a population-genetic/multivariate logistic regression model with two coefficients: $\vec{M}_a$, which captures the effect of mutation, and $\vec{S}_a$, which captures the effect of selection. The latter is modulated by the amount of "observed" gene expression x_g. The formula for the multinomial probability of a specific codon π_{gc} is as given

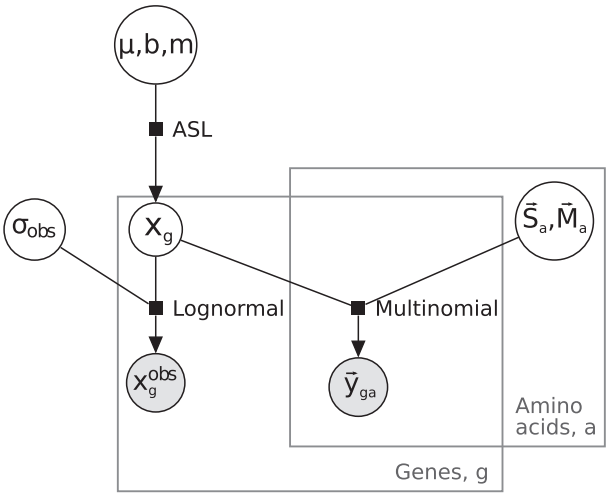

FIG. 8. Graphical representation of model. The representation as a probabilistic graphical model (Airoldi 2007) uses shaded circles to indicate observed quantities, and plain circles to indicate unobserved quantities; plates (gray rectangles) indicate that the likelihood factors both by gene and by amino acid. The three black squares represent the three components of the model discussed in the text and Materials and Methods.

in equation (1), which drops the amino acid subscript for clarity. The vectors $\vec{y}_{ga}$, $\vec{M}_a$, and $\vec{S}_a$ have one component for each codon coding for amino acid a. The error model for the observed gene expression x_g^{obs} is lognormal with variance σ^2_{obs}. The distribution of latent gene expression is log-asymmetric Laplace (ASL) with parameters μ_x, b_x, and m_x (Kotz et al. [2001] and supplementary text, Supplementary Material online, for details on the parametrization).

In equations, this is as follows:

$$\vec{y}_{ga} \sim \text{Multin}(n_{ga}, \vec{\pi}_{ga}) \quad (2a)$$

$$\vec{\pi}_{ga} = \text{mlogit}^{-1}(\vec{M}_a + \vec{S}_a x_g) \quad (2b)$$

$$\log(x_g^{\text{obs}}) \sim N(\log(x_g), \sigma^2_{\text{obs}}) \quad (2c)$$

$$\log(x_g) \sim \text{ASL}(\mu_x, b_x, m_x) \quad (2d)$$

where mlogit denotes the multinomial logistic function. Further details, including a discussion of the choices of distributions used, are in the supplementary text, Supplementary Material online.

We used (improper) uniform priors on the hyperparameters μ_x, σ_x^2, and σ^2_{obs}. We fit this using a Metropolis–Hastings within Gibbs sampler for MCMC inference, running four parallel chains for 10,000 iterations each following a burn-in of 2,000 iterations, thinning the results by keeping only every tenth iteration. The Gelman–Brooks–Rubin convergence diagnostic suggested that the chains had converged to a single distribution (Brooks and Gelman 1998), giving potential scale reduction values below 1.05 for all parameters. Further details are in the supplementary text, Supplementary Material online.

For the validation runs in figure 5, we ran each sample for 6,000 iterations following a burn-in of 2,000 iterations.

To calculate the Bayes factor, we drew 1,000 sets of parameter values uniformly at random from the post burn-in posterior draws from our MCMC sampler and took the mean of the likelihood for the codon portion of each of these; similarly for the noise-blind model, we drew 1,000 sets of parameter values at random from the normal approximation about its maximum likelihood and took the mean of these, the difference in log of these means is the log Bayes factor (Gelman et al. 2003).

All calculations were done in R (Ihaka and Gentleman 1996), using the VGAM package (Yee 2010) to fit the multinomial logistic regression model and the ggplot2 package (Wickham 2010) to produce the plots. An R package implementing the fits is included in supplementary text, Supplementary Material online.

Reporting Coefficients

For a right-skewed error model such as the lognormal, fixing the sum of observed gene expression alone produces biased estimates, because the (latent) true expression has on average a lower sum than noisy observed expression. Accordingly, following the completion of the MCMC sampler, we normalize the estimates of latent gene expression by their sum at each iteration. In other words, in iteration k, we have posterior draws $x_{g,k}$ for the latent gene expression, and we report

$$x'_{g,k} = x_{g,k} \frac{X}{\sum_g x_{g,k}} \quad (3)$$

where X is the total gene expression. Similarly, we multiply the parameter estimates for S at iteration k by $\frac{\sum_g x_{g,k}}{X}$. This strategy is permissible because estimates of absolute total gene expression are derived from experiments independent of the inference being performed; its success in producing unbiased estimates of the parameters is shown in figure 5.

We reported the median of the normalized MCMC draws as the inferred parameter. Confidence intervals given in supplementary files, Supplementary Material online, are the 95% intervals of the normalized MCMC draws, that is, the 2.5th and 97.5th percentiles.

Multinomial logistic regression predicts probabilities that do not change if a constant is added to all predictors: We chose the constant so that the sum of reported selection (mutation) coefficients across synonyms was zero, so that the reported coefficient represents the difference from the mean across synonyms. This differs from the reporting method in Shah and Gilchrist (2011), who chose a constant, so that, for each amino acid, the coefficients for a reference codon are zero, thus reporting the difference in selection (mutation) coefficients between a codon and the reference codon. In calculating SCU below, we report an alternative measure, where the expected selection if codon composition is given by mutation bias is zero. In all cases, if there were no selection (mutation bias) between synonyms, then all selection (mutation) coefficients for that amino acid would be zero.

To validate the fit of the model to simulated data, we used amino acid counts for 1,000 yeast genes and drew mutation and selection coefficients, as well as expression parameters,

from the reported fit to the yeast genome. For each of a range of observation noise parameters, we generated 12 simulated data sets according to the model, normalizing the observed gene expression to the original total, and then estimated the coefficients from each of these, including the sum normalization described here. The results are reported in figure 5.

Estimating Mutation Bias Directly

Direct measurements of single-nucleotide mutation rates in yeast were taken from Kunz et al. (1998), Lang and Murray (2008), and Ohnishi et al. (2004), reported in Lynch et al. (2008), as counts for every possible nucleotide substitution. The experimentally measured counts are reported per base pair and we do not distinguish lead and lagging strands, so for A → T the value $\log\left[N_{(\mathrm{A:T}\to\mathrm{T:A})}/N_{(\mathrm{T:A}\to\mathrm{A:T})}\right]=0$, and similarly for C → G. These were compared with the difference in mutation coefficients in our model between synonyms differing by the same nucleotide substitution.

Precisely, in the derivation of the model, we assume detailed balance, that is, that the mutation rate between synonymous codons c and d is given by

$$\mu_{c\to d}=\bar{\mu}e^{\frac{1}{2}(M_d-M_c)}, \tag{4}$$

where $\bar{\mu}$ has units of mutations per unit time, and M's are the mutation constants fit by the model (see supplementary material, Supplementary Material online, and Sella and Hirsh [2005]). Thus

$$M_d-M_c=\log\left[\frac{\mu_{c\to d}}{\mu_{d\to c}}\right]. \tag{5}$$

For example, the difference in model parameters for two glycine codons is compared with experimentally measured mutation counts as follows:

$$M_{\mathrm{GGG}}-M_{\mathrm{GGA}}=\log\left[\frac{\mu_{\mathrm{GGA}\to\mathrm{GGG}}}{\mu_{\mathrm{GGC}\to\mathrm{GGG}}}\right] \tag{6}$$

$$\approx\log\left[\frac{\mu_{\mathrm{A}\to\mathrm{G}}}{\mu_{\mathrm{G}\to\mathrm{A}}}\right] \tag{7}$$

$$\approx\log\left[\frac{N_{(\mathrm{A:T}\to\mathrm{G:C})}}{N_{(\mathrm{G:C}\to\mathrm{A:T})}}\right]. \tag{8}$$

Similarly, in fitting the mutation-pooled model in supplementary figure S7, Supplementary Material online, nucleotide-specific mutation rates were assigned to codons by the same equations. However, the mutation rate $\mu_{\mathrm{G}\to\mathrm{A}}$ was treated as independent of $\mu_{\mathrm{G}\to\mathrm{T}}$, allowing for mutation biases to differ in the transcribed and untranscribed strand.

Calculating Genewise Selection

To provide a measure of average codon selection for a sequence encoding a polypeptide, we estimate the average per-codon selective advantage conferred between the observed encoding compared with the selective advantage of a randomly encoded version of the same polypeptide. In the absence of selection, the frequency of codon c among synonymous sites would be given by

$$\pi_c^M=\frac{\exp(M_c)}{\sum_{c'\,|\,a}\exp(M_{c'})}. \tag{9}$$

Recall that the S_c selection coefficients represent selective differences relative to an arbitrary point chosen so that the mean S_c in a synonymous family is zero and are only evolutionarily meaningful as differences from each other. The average selective advantage of a codon over its synonyms, if those synonyms occur according to mutational biases alone, is

$$S_c'=S_c-\sum_{c'\,|\,a}\pi_{c'}^M S_{c'}. \tag{10}$$

If a gene's codon composition is produced by mutation bias alone, the expected fitness advantage conferred by its encoding relative to an unselected sequence is zero, and accordingly, the expected value of S_c' is zero.

A genewise statistic is obtained by taking the average across the codons in a gene and multiplying by gene expression in transcripts per cell

$$\mathrm{SCU}=x_g\frac{1}{L_g}\sum_i S_{c_i}' \tag{11}$$

where i indexes all sites in the gene coding for degenerate amino acids (equivalently, one can set $S_c=0$ for all Met and Trp codons), and L_g is the length of gene g counting only the degenerate amino acids. Because S_c is the selection coefficient per transcript per cell scaled by effective population size, so SCU has units of selection coefficient per site, scaled by effective population size. Note that SCU is independent of the figure used to normalize total gene expression, because it is a product of selection coefficients and estimated expression.

For direct comparison with CAI, we define mSCU as the average translational selection per transcript across all degenerate codons in a gene,

$$\mathrm{mSCU}=\frac{1}{L_g}\sum_i S_{c_i}'=\frac{1}{x_g}\mathrm{SCU}. \tag{12}$$

This is an arithmetic mean, whose calculation is directly analogous to that of log(CAI), because CAI is a geometric mean across codons in a gene (Sharp and Li 1987).

Supplementary Material

Supplementary text, file, figures S1–S8, and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Alexander Blocker and Eric Solís in the Airoldi group, and members of the Harvard FAS center for Systems Biology, for many helpful discussions. They thank

Sergey Kryazhimskiy, Arvind Subramaniam, and Claus Wilke for comments on the manuscript and Nicholas Ingolia and Jonathan Weissman for allowing them to use some of their unpublished data. The computations in this article were run on the Odyssey cluster supported by the FAS Science Division Research Computing Group at Harvard University. This work was supported by the Pew Foundation, by NIH NIGMS grants R01 GM088344 and P50 GM068763, by a Bauer Fellowship to D.A.D., and by NIH NIGMS grant R01 GM096193 and NSF grant IIS-1017967 to E.M.A. E.M.A. and D.A.D. are Alfred P. Sloan Research Fellows. D.A.D. is a Pew Scholar in the Biomedical Sciences.

References

Airoldi EM. 2007. Getting started in probabilistic graphical models. *PLoS Comput Biol.* 3:e252.

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

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev.* 54:198–210.

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* 100:3889–3894.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet.* 8: 619–631.

Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N. 2006. Noise in protein expression scales with natural protein abundance. *Nat Genet.* 38:636–643.

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat.* 7:434–455.

Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes. *Mol Biol Evol.* 3:322–329.

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

Carroll R, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. Measurement error in nonlinear models: a modern perspective. Boca Raton, Fla: Chapman & Hall/CRC.

Dix DB, Thompson RC. 1989. Codon choice and gene expression: synonymous codons differ in translational accuracy. *Proc Natl Acad Sci U S A.* 86:6888–6892.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 10:715–724.

Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci U S A.* 108:680–685.

Gelman A, Carlin J, Stern H, Rubin D. 2003. Bayesian data analysis. Boca Raton, Fla: Chapman & Hall/CRC.

Gelman A, Hill J. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge (UK): Cambridge University Press.

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol.* 6:e1000664.

Hereford LM, Rosbash M. 1977. Number and distribution of polyadenylated RNA sequences in yeast. *Cell* 10:453–462.

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

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *J Comput Graph Stat.* 5:299–314.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol.* 151: 389–409.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol.* 158:573–597.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol.* 53:290–298.

Karlin S, Mrázek J. 1996. What drives codon choices in human genes? *J Mol Biol.* 262:459–472.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315:525–528.

Kotz S, Kozubowski TJ, Podgórski K. 2001. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Boston: Birkhäuser.

Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* 13:87–96.

Kramer EB, Vallabhaneni H, Mayer LM, Farabaugh PJ. 2010. A comprehensive analysis of translational missense errors in the yeast *Saccharomyces cerevisiae*. *RNA* 16:1797–1808.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255–258.

Kunz BA, Ramachandran K, Vonarx EJ. 1998. DNA sequence analysis of spontaneous mutagenesis in *Saccharomyces cerevisiae*. *Genetics* 148: 1491–1505.

Lang GI, Murray AW. 2008. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* 178:67–82.

Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome vi are correlated with replication timing. *Genome Biol Evol.* 3:799–811.

Lee Y, Zhou T, Tartaglia GG, Vendruscolo M, Wilke CO. 2010. Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics* 10:4163–4171.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.* 11:733–739.

Letzring DP, Dean KM, Grayhack EJ. 2010. Control of translation efficiency in yeast by codon-anticodon interactions. *RNA* 16: 2516–2528.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538–541.

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. 2009. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol.* 27: 652–658.

Lynch M, Sung W, Morris K, et al. (11 co-authors). 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A.* 105:9272–9277.

Mathews M, Sonenberg N, Hershey JWB. 2007. Translational control in biology and medicine. Cold Spring Harbor: Cold Spring Harbor Laboratory press.

Ogle JM, Ramakrishnan V. 2005. Structural insights into translational fidelity. *Annu Rev Biochem.* 74:129–177.

Ohnishi G, Endo K, Doi A, Fujita A, Daigaku Y, Nunoshiba T, Yamamoto K. 2004. Spontaneous mutagenesis in haploid and diploid *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun.* 325: 928–933.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet.* 12:32–42.
Purdom E, Holmes SP. 2005. Error distribution for gene expression data. *Stat Appl Genet Mol Biol.* 4:Article 16.
Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* 8:e1002603.
Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences, and control. *Science* 309:2010–2013.
Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A.* 102:9541–9546.
Shah P, Gilchrist MA. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A.* 108:10231–10236.
Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33:1141–1153.
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–1295.
Siwiak M, Zielenkiewicz P. 2010. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Comput Biol.* 6: e1000865.
Sørensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol.* 207:365–377.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24:374–381.
Tuller T, Carmi A, Vetsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141:344–354.
Vieira-Silva S, Rocha EPC. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6:e1000808.
von der Haar T. 2008. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol.* 2:87.
Wickham H. 2010. A layered grammar of graphics. *J Comput Graph Stat.* 19:3–28.
Yang Z, 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, Kaplan T, Fraser HB, et al. (14 co-authors). 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. 2010. The VGAM package for categorical data analysis. *J Stat Software.* 32:1–34.
Zaher HS, Green R. 2009. Fidelity at the molecular level: lessons from protein synthesis. *Cell* 136:746–762.
Zenklusen D, Larson DR, Singer RH. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol.* 15:1263–1271.
Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 26:1571–1580.

Estimating selection on synonymous codon usage from noisy experimental data

Edward W.J. Wallace, Edoardo M. Airoldi, and D. Allan Drummond

Supplementary Text

Population Genetics Model

Here we present a derivation of the population genetics model used to describe the codon frequencies conditional on underlying gene expression. An alternative derivation and generalizations are found in Barton and Coe (2009).

Fixation probabilities.

Initially consider two alleles, denoted a and b, in a population. The mutation rate from a to b is $\mu_{a\to b}$, and that in the other direction is $\mu_{b\to a}$. Let the additive fitness difference be $s = s_{ab}$, meaning that the rate of increase of the number of individuals with allele a is e^s times the rate of increase for allele b. For small s, since $e^s \approx 1 + s$, our usage coincides with the selective advantage in classical population genetics. Given a single mutation to a in a population of b's, the probability of fixation of a is given by

$$\phi_{b\to a} \approx \frac{1 - e^s}{1 - e^{N_e s}} \tag{S1}$$

where N_e is the effective population size (see Gillespie (2004), p93, or any textbook in population genetics); for a diploid population, there would be additional constants in the exponents.. Similarly the fixation probability of a mutation to b in a population of a's is

$$\phi_{a\to b} = \frac{1 - e^{-s}}{1 - e^{-N_e s}}. \tag{S2}$$

Substitution matrix.

If the mutation rates are smaller than the fixation rates, then at a given time the population is likely to be fixed at a single allele. Suppose the population is fixed at b. Then, a mutation occurs to a at the rate $\mu_{b\to a}$, and this mutation is fixed with probability given by (S1). This means the substitution rate, i.e. the rate of replacement of allele b by allele a in the entire population, is given by

$$W_{a,b} = \mu_{b\to a}\phi_{b\to a}. \tag{S3}$$

In general, if we have alleles $j = 1, \ldots, n$, each of which has additive fitness s_j, and we define the difference in fitnesses as $s_{jk} = s_j - s_k$, then the analogous formulas hold:

$$\phi_{k\to j} = \frac{1 - e^{s_{jk}}}{1 - e^{N_e s_{jk}}} \approx e^{-N_e s_k} \frac{e^{s_k} - e^{s_j}}{e^{N_e s_k} - e^{N_e s_j}}, \tag{S4}$$

where we approximated $N_e \approx N_e - 1$; the substitution rate matrix is given by

$$W_{j,k} = \begin{cases} \mu_{k\to j}\phi_{k\to j} & \text{if } j \neq k, \\ 1 - \sum_{l\neq j} W_{j,l} & \text{if } j = k. \end{cases} \tag{S5}$$

Here $\mu_{k\to j}$ is the mutation rate from k to j, which may be zero for some pairs of alleles.

Equilibrium distribution.

Now we have the substitution matrix W, we may find its the equilibrium distribution. A fundamental result of Markov process theory is that, if the matrix W is irreducible, i.e., between any pair of alleles there is some series of substitutions between them all with nonzero rates, then the distribution of allele probabilities always approaches an equilibrium P^*. This distribution is a vector which does not change with the action of substitution described by W, so satisfies

$$WP^* = P^*. \tag{S6}$$

Under mild conditions on the mutation rates, then this process has the detailed balance property, meaning that the expected rate of flow between any two states is symmetric:

$$W_{jk}P^*_k = W_{kj}P^*_j. \tag{S7}$$

The necessary conditions on the mutation rates hold if there is some set of numbers $\bar{\mu}$ and M_j analogous to a "chemical potential", such that

$$\mu_{k\to j} = \bar{\mu}e^{\frac{1}{2}(M_j - M_k)}. \tag{S8}$$

This is true when, for example, we consider single-nucleotide mutations whose rates depend only on whether the mutation is $(AT) \to (GC)$, and whether it is purine to pyrimidine. These conditions are discussed in the supplementary information of Sella and Hirsh (2005), which notes that they incorporate "mutation schemes that are consistent with most of our knowledge about mutation in biology". We use this approximation, as (S7) considerably simplifies the estimation of the equilibrium distribution, implying that

$$\frac{P^*_j}{P^*_k} = \frac{W_{jk}}{W_{kj}} = \frac{\mu_{k\to j}e^{N_e s_j}}{\mu_{j\to k}e^{N_e s_k}} = \frac{e^{M_j + N_e s_j}}{e^{M_k + N_e s_k}} \tag{S9}$$

and so the equilibrium allele probabilities are given by

$$P^*_j = \frac{e^{M_j + N_e s_j}}{\sum_k e^{M_k + N_e s_k}}. \tag{S10}$$

This distribution is exactly that in a multinomial logistic regression (Gelman and Hill, 2007), meaning that the parameters may be fitted using maximum likelihood estimation, and any allele-specific predictors of s_j may be considered. Note also that adding a constant to every M_j, or to every s_j, does not change the predicted probabilities. This additive uncertainty is exploited in the course of fitting multinomial logistic regression by picking a reference allele k and setting $M_k = 0$ and $s_k = 0$. The choice of reference allele does not affect the predicted probabilities, only changing the parametrization.

Application to synonymous codon usage.

We employ the above models to study translational selection on synonymous codon usage, by considering as alleles different synonymous codings of the same gene. Suppose the additive fitness of codon c at site i in gene g is $s_{c|g,i}$, approximated as being independent of the codon context. This is clearly an inexact approximation as, for example, mRNA secondary structure depends on codon context as well as identity and is believed to affect translational selection. Similarly, we approximate the mutation parameter at site i in gene g as $M_{c|g,i}$ as being independent of context. Approximating the selective forces on codons as independent leads to the probability distribution of a gene sequence as

$$P(c_1 c_2 c_3 \dots c_n) = \prod_i \frac{e^{M_{c_i|g,i} + N_e s_{c_i|g,i}}}{\sum_{d|aa_{g,i}} e^{M_{d|g,i} + N_e s_{d|g,i}}} \tag{S11}$$

where codon c_i is at position i, the product is over all positions. We extend this to include all genes similarly by multiplying the probabilities for each gene sequence. In the sum in the denominator, $d|aa_{g,i}$ denotes summation only over all codons coding for amino acid found at g, i, since we are only considering synonymous codon substitutions. This effectively factorizes the fitting procedure so that the distribution for each amino acid is fit as a separate multinomial logistic regression. The model and fitting as generically described are identical to the protocol followed in Shah and Gilchrist (2011), and the population genetics model without expression as a predictor was previously presented in Bulmer (1991). One may choose any covariate measured at the site g, i to fit the probabilities, for example gene expression level, which depends only on the gene g, or protein structure data, which depends also on the position i. However, the fitting procedure by itself does not attribute fitted effects to selection or mutation; that requires interpretation after the model is fit.

In this paper, we approximate the mutation coefficient as independent of gene and position,

$$M_{c_i|g,i} = M_c \tag{S12}$$

and the selection coefficient depends on gene and position only through the expression level of the gene, x_g,

$$N_e s_{c_i|g,i} = S_c x_g. \tag{S13}$$

This leads to equations (1), and (2b), which includes a subscript to index the coefficients by amino acid.

This model, with no correction for noise, is an example of multinomial logistic regression; we fit it using a maximum likelihood procedure implemented in the statistical programming language R (Ihaka and Gentleman, 1996) by the VGAM package (Yee, 2010). Note that this derivation changes the interpretation of P^*, conceived in Sella and Hirsh (2005) as the time-average of a single population. Here the collection of all codon sites is an ensemble of realizations, sampled at a single point in time.

Inference via MCMC

As presented in the methods section, the model has three components: codon usage conditional on underlying gene expression, observed noisy gene expression conditional on underlying gene expression, and the distribution of underlying expression across genes. The multinomial distribution of codon counts $\vec{y}_{ga}$ for gene g and amino acid a is given by the population-genetic/multinomial logistic regression coefficients $\vec{M}_a$ for mutation and $\vec{S}_a$ for selection, the coefficient of gene expression x_g. The multinomial probabilities are as given in equation (1), which drops the amino acid subscript for clarity. The vectors $\vec{y}_{ga}$, $\vec{M}_a$ and $\vec{S}_a$ have one component for each codon coding for amino acid a. The error model is lognormal with variance σ^2_{obs}. The underlying distribution of gene expression is asymmetric Laplace with parameters μ_x, b_x, m_x. In equations, this is

$$\log(x_g) \sim ASL(\mu_x, b_x, m_x) \tag{S14}$$
$$\log(x_g^{obs}) \sim N(\log(x_g), \sigma^2_{obs}) \tag{S15}$$
$$\vec{\pi}_{ga} = \text{mlogit}^{-1}(\vec{M}_a + \vec{S}_a x_g) \tag{S16}$$
$$\vec{y}_{ga} \sim \text{Multin}(n_{ga}, \vec{\pi}_{ga}). \tag{S17}$$

We employed a representation of the asymmetric Laplace distribution parametrized as a continuous mixture of normal distributions, representing $l \sim ASL(\mu, b, m)$ as

$$z \sim N(0,1) \tag{S18}$$
$$w \sim \text{Expo}(1) \tag{S19}$$
$$l = \mu + mw + b\sqrt{2w}z \tag{S20}$$

where Expo denotes an exponential random variable. The conditional distribution of the augmented variable $w|l, \mu, m, b$ has a tractable cumulative distribution function (cdf) so that w may be drawn by generating a uniform random variable and numerically inverting the cdf via bisection, $u \sim F^{-1}_{w|l,\mu,m,b}[U(0,1)]$. Note also that the conditional distribution $(l|\mu, m, b, w)$ is Gaussian, so that the conditional distribution $(mu, m|l, w, b)$ is jointly Gaussian and $(b^2|l, w, \mu, m)$ is inverse-Gamma. In the full model this necessitates the use of a vector of augmented variables $\vec{w}$, one per gene.

We placed uniform prior distributions on the coefficients the coefficients M, S. We also placed (improper) uniform prior distributions on the hyperparameters μ_x, b_x, m_x, and σ^2_{obs}. The values of these are strongly constrained by the available data.

We fit this model using a Metropolis-Hastings-within Gibbs sampler (Gelman et al., 2003) for Markov chain Monte Carlo (MCMC) inference. Starting with dispersed initial estimates, at each iteration the algorithm proceeds:

1. For each amino acid a, update $M_a, S_a|x, y, n$ by independence chain Metropolis-Hastings step:
   (a) Estimate $\hat{M}_a, \hat{S}_a, \hat{\Sigma}_{M_a,S_a}|x, y, n$ via maximum-likelihood.
   (b) Propose M^*_a, S^*_a from $N(\hat{M}_a, \hat{S}_a, \hat{\Sigma}_{M_a,S_a})$ distribution.

   (c) Accept M_a^*, S_a^* with Metropolis-Hastings probability.

2. Update hyperparameters from exact conditional distributions:

   (a) Update $\sigma^2_{obs}|x, x^{obs}$, using Inverse-Gamma draw.

   (b) For each gene g, update $w_g|\log(x_g), \mu_x, m_x, b_x$ via numerical inversion of cdf from Uniform draw.

   (c) Update $(\mu_x, m_x)|x, w, b_x$, using Gaussian draw.

   (d) Update $b_x^2|\mu_x, m_x, x, w$, using Inverse-Gamma draw.

3. For each gene g, update $x_g|M, S, x_g^{obs}, y_g, n_g, \sigma^2_{obs}, \mu_x, \sigma_x^2$ by random walk Metropolis-Hastings step:

   (a) Propose $\log(x_g^*)$ from $N\left(\log(x_g^{(i)}), \tau_g\right)$, where $x_g^{(i)}$ is the value at the previous iteration and τ_g

   (b) Accept x_g^* with Metropolis-Hastings probability.

For the fit with mutation pooled by nucleotide, S_a are drawn as in step 1 above, but the nucleotide mutation bias parameters M are drawn by a Metropolis random walk step.
We implemented the algorithm in R, with some likelihood calculations in C++ using the Rcpp package, for speed. The reason for not using a packaged Gibbs sampler such as BUGS or JAGS is that these packages do not currently implement multinomial logistic regression, only binary logistic regression. We used the VGAM package for multinomial logistic regression, which calculates both the maximum likelihood parameter estimates $\hat{M}_a, \hat{S}_a$ and the covariance matrix about that estimate $\hat{\Sigma}_{M_a,S_a}$.
Fitting each dataset, we ran 4 parallel chains for 10,000 iterations each following a burn-in of 2,000 iterations. We reported the median of the MCMC draws as the inferred parameter. Confidence intervals give the central 95% intervals of the MCMC draws. The Gelman-Brooks-Rubin convergence diagnostic (Brooks and Gelman, 1998) suggested that the chains had converged to a single distribution, with a multivariate potential scale reduction factor below 1.05 for all parameters presented.

Discussion of modeling choices

Beyond the sources of selection and biological assumptions, specific technical choices made in our model are worth describing as a way to understand how the model may be improved or generalized. The model has three components: codon usage conditional on latent gene expression, the distribution of latent expression across genes, and observed gene expression conditional on latent gene expression. The Bayesian hierarchical framework of our model is modular, meaning that each component of the model may be adjusted independently of the others.
We modeled codon usage with a standard approximation in population genetics, applicable when mutation rates are small compared to selection-driven fixation, so that the population clusters at a single genotype at any given time (Bulmer, 1991; Shah and Gilchrist, 2011). Since translational selection is very weak in low-expression genes,

incorporating population heterogeneity at a single time might provide a better description of the data (Barton and Coe, 2009).

We modeled the distribution of expression across genes as log-asymmetric-Laplace, a phenomenological distribution that approximates observed gene expression (Purdom and Holmes (2005) and supplementary figure). The simplest distribution that covers the several orders of magnitude spanned by expression measurements is the lognormal; however, the true distribution of gene expression is more widely dispersed than a lognormal (Lu and King, 2009). The role of this part of the model is to regularize the estimates, i.e. to ensure that the distribution of latent expression is constrained to plausible values; while there are many choices of distribution, we chose the log-asymmetric-Laplace as a computationally tractable distribution which closely resembles the data.

We modeled noise in gene expression observations as lognormal, consistent with the standard deviation being constant on a log scale (figure 2), and as uncorrelated across genes, or intrinsic. This error distribution is an approximation to noise combined from many sources, including biological variation, measurement error from RNA sequencing, and proxy error, i.e. using exponential-growth mRNA abundance in the laboratory as a proxy for protein synthesis rate over evolutionary time. For example, the noise parameter, σ_{obs}, estimated in our model is much greater than that estimated in the four expression replicates from the Yassour, reflecting sources of noise not controlled for in replication.

A more detailed model might additionally take into account count variability arising in RNAseq, use multiple expression datasets, or allow for differing variability in expression across genes. More generally, gene expression varies over time and stochastically between cells, and our model accounts for only an average gene expression figure over time, essentially modeling codon usage as depending only on this average expression, and subsuming differences into the generic noise term. While it is likely that more detailed modeling of noise would lead to in-principle more accurate inferences, this would require further experimental measurements to estimate the error distribution arising at each step: in their absence, accounting for noise with a single phenomenological distribution is a clear improvement over leaving it unmodeled.

Additionally, extrinsic or correlated noise is thought to play a larger role than intrinsic noise in the variability of yeast gene expression (Raser and O'Shea, 2005). Our results would be affected if the codon composition of a group of genes were linked to the noise in their expression. We are not aware of any study linking codon composition to noisiness, beyond their common covariance with gene expression. However, in microarrays it is known that the measurement error for a gene may be dependent on sequence features of the probes used (Kerkhoven et al., 2008). It is possible that similar sequence-dependent measurement error may be present in RNA-Seq data. In general, systematic differences correlated with measured gene expression could bias our inference if not directly modeled, for example in translational robustness and protein stability, ribosomal loading, or protein length.

Since our model of noise in gene expression is agnostic to the source of this noise, it may apply to more general situations. For example, one could estimate translational selection in a related yeast species using codon counts in that species and expression data from *S. cerevisiae*; comparing between species would introduce an additional source of proxy error, but as long as the total noise was small enough, could lead to valuable biological insights.

Supplementary Figures

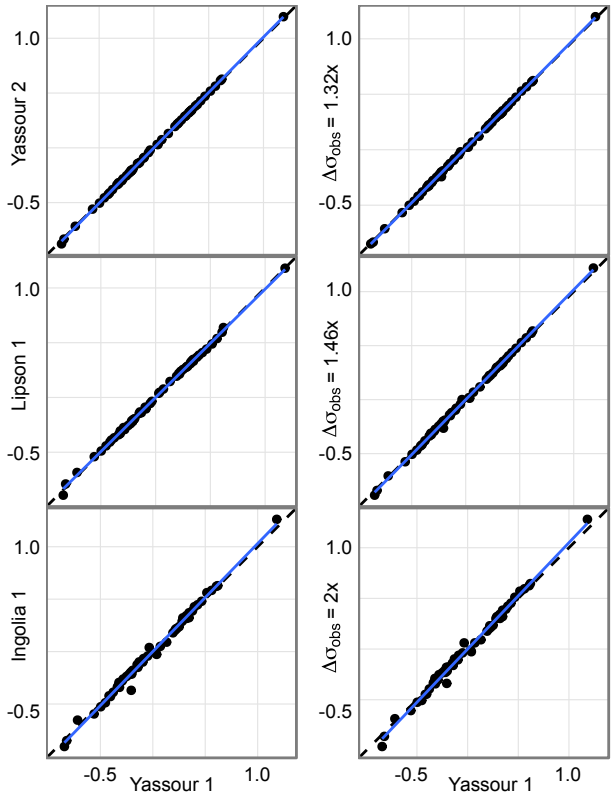

Figure S1: **Noise in gene expression and mutation estimates in budding yeast.** First column, mutation bias estimates compared from a subset of the expression datasets used in figure 2. Each dot represents a codon, its x-coordinate the mutation coefficient estimated from expression data in Yassour 1, and its y-coordinate the mutation coefficient estimated from an alternative dataset. The blue line is the least-squares fit of the alternative selection coefficients to those from Yassour 1. Second column, mutation bias estimates inferred from simulated noisy expression datasets. As in figure 3, we artificially add lognormal noise to expression estimates from Yassour 1, and estimate the codon mutation coefficients.

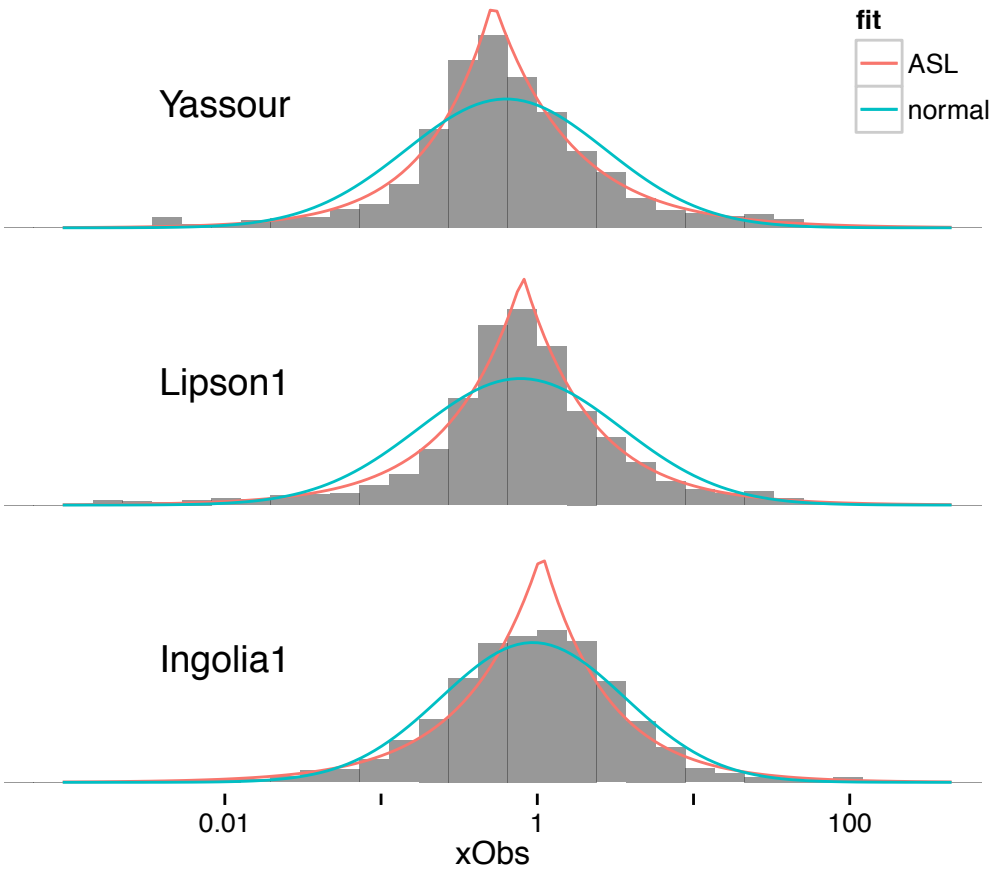

Figure S2: **Comparing distributions for underlying gene expression.** Asymmetric Laplace (ASL) and Normal fits to log-transformed x observations from three gene expression datasets. The ASL is a better fit for all: estimated by the Akaike information criterion, the improvement from using the ASL is $\Delta AIC = 1108,\ 817,\ 16$ for Yassour, Lipson1 and Ingolia1 respectively.

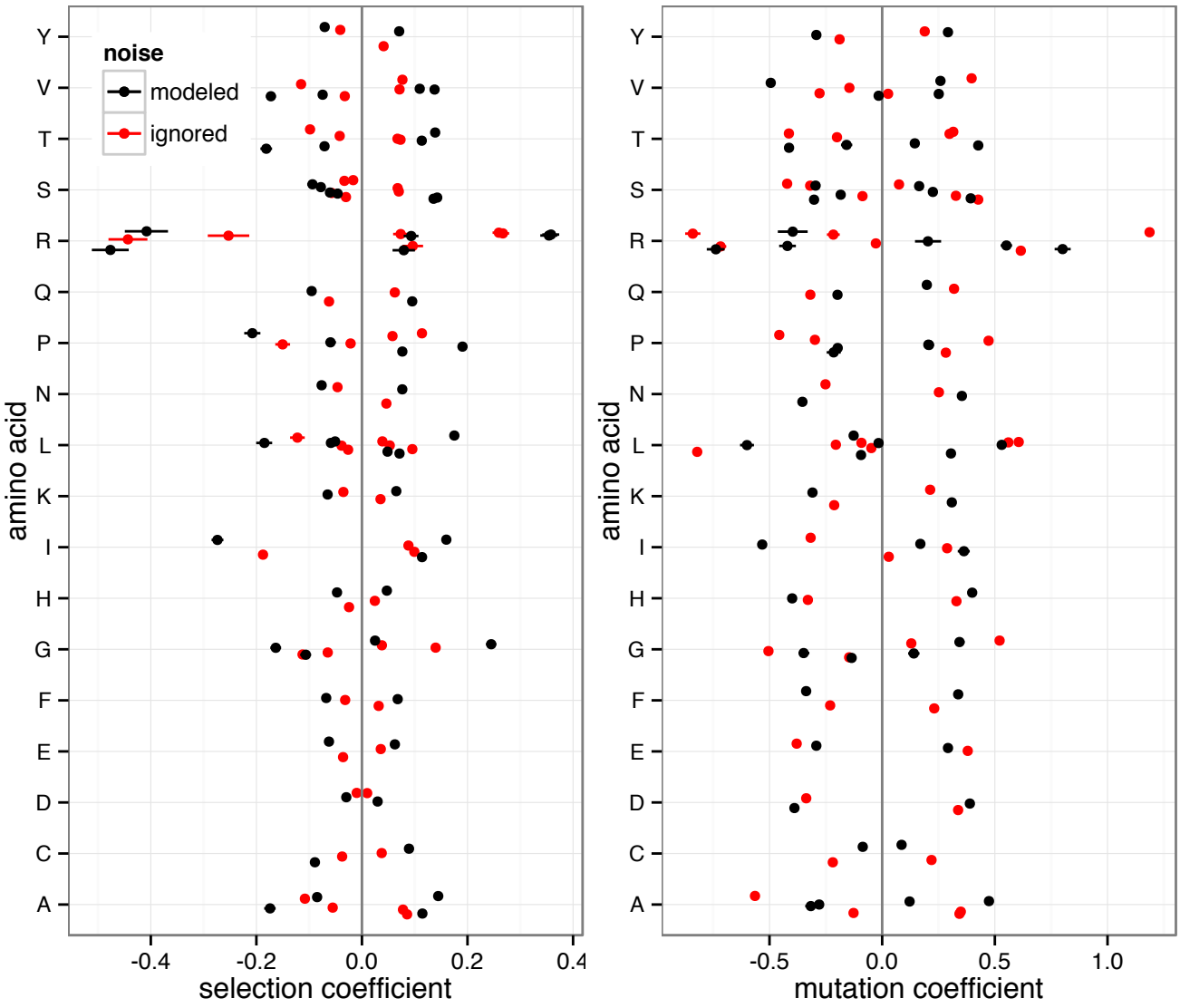

Figure S3: **Selection and mutation coefficients compared by amino acid.** This uses data from Yassour et al. (2009) plotted in figure 4, but with coefficients collected by amino acid; y-coordinates have been jittered to make all points visually distinguishable.

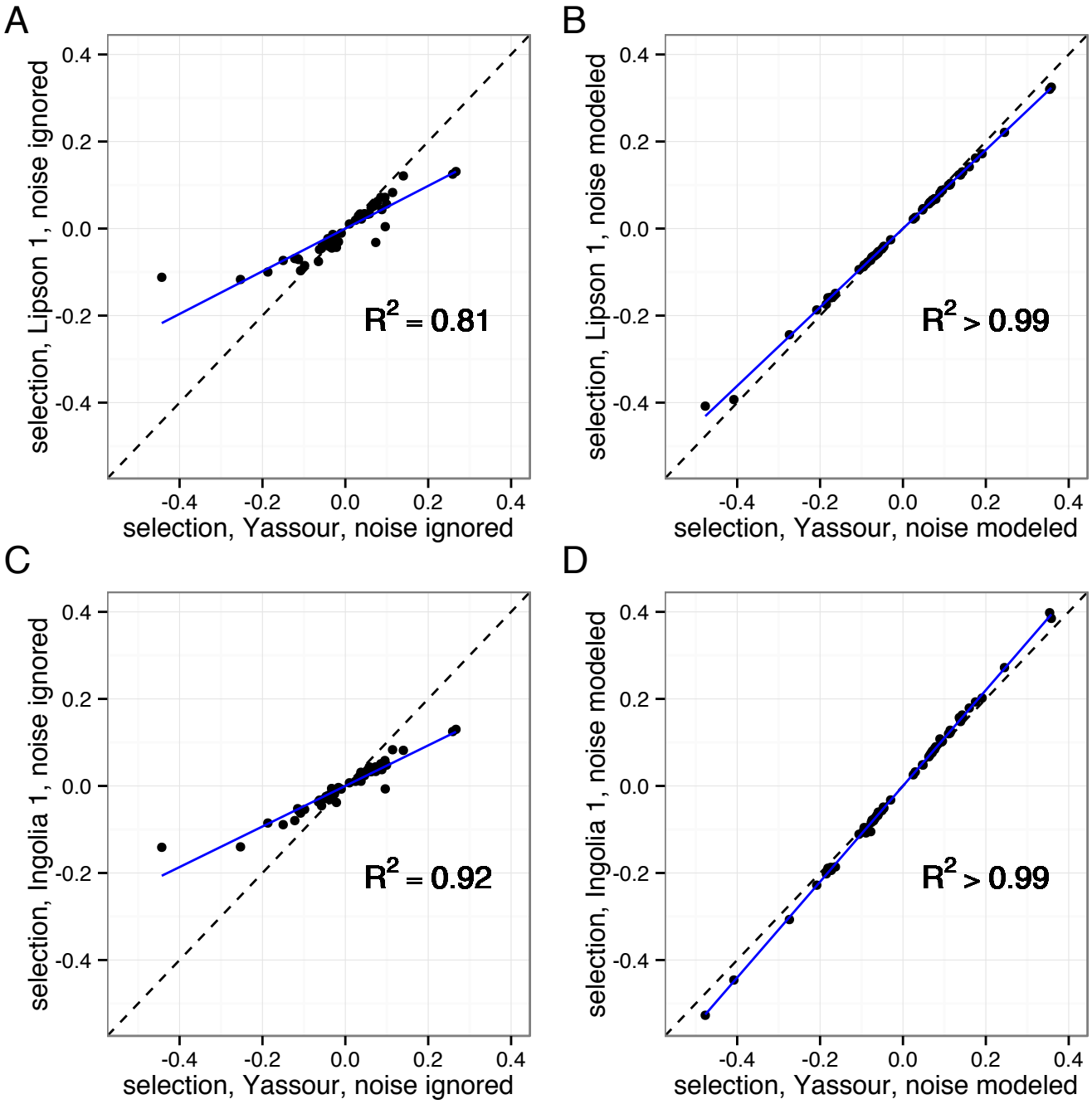

Figure S4: **Comparison of selection estimates from different mRNA abundance datasets.** Each dot represents a codon, its x-coordinate is the selection coefficient inferred from the geometric mean of Yassour et al's estimates of mRNA counts per cell (Yassour et al., 2009), and its y-coordinate is the selection coefficient inferred from alternative estimates of mRNA abundance, either Lipson 1 (Lipson et al., 2009) or Ingolia 1 (Ingolia et al., 2009).

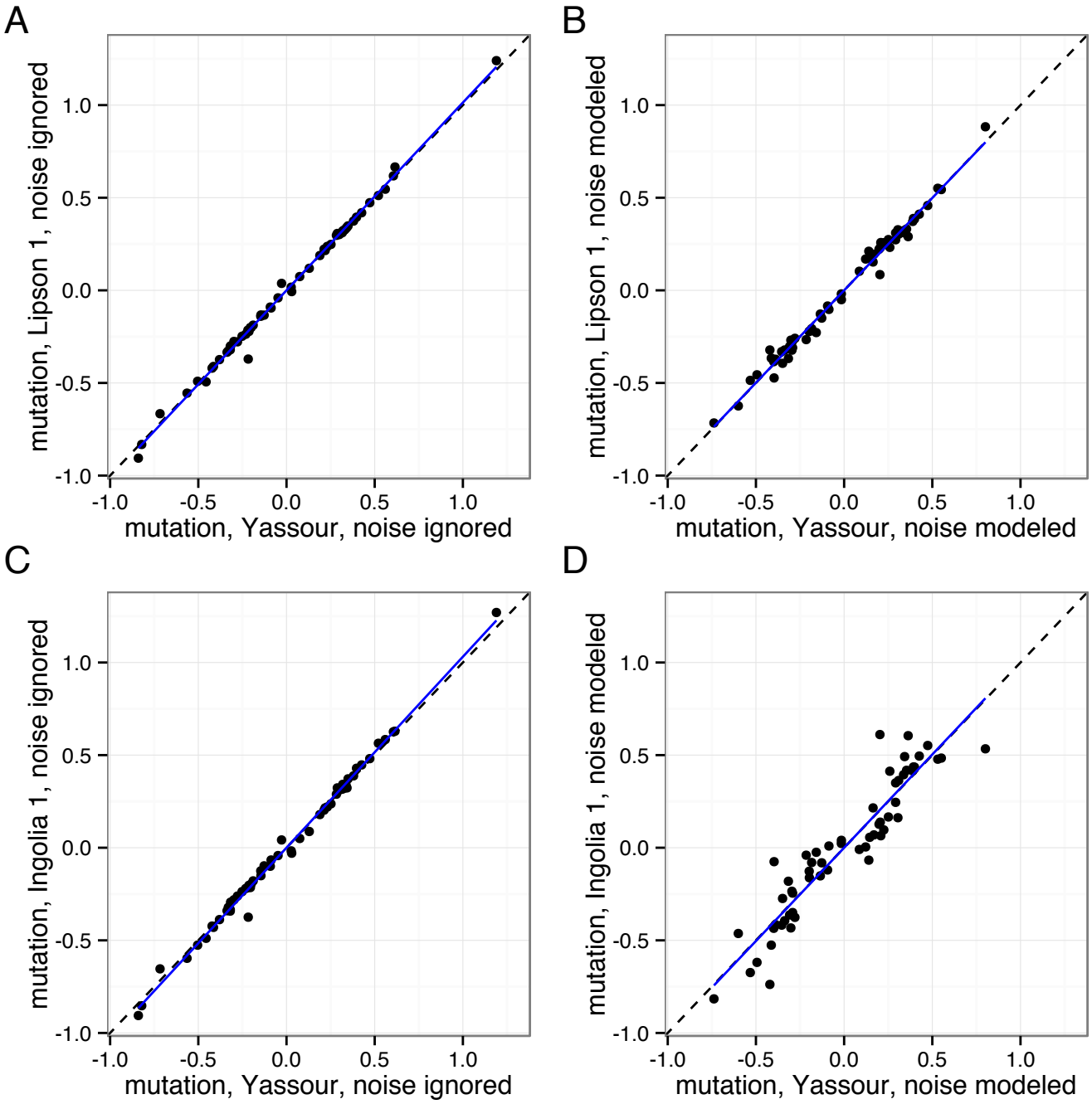

Figure S5: **Comparison of mutation estimates from different mRNA abundance datasets.** Each dot represents a codon, its x-coordinate is the mutation coefficient inferred from the geometric mean of Yassour et al's estimates of mRNA counts per cell (Yassour et al., 2009), and its y-coordinate is the mutation coefficient inferred from alternative estimates of mRNA abundance, either Lipson 1 (Lipson et al., 2009) or Ingolia 1 (Ingolia et al., 2009).

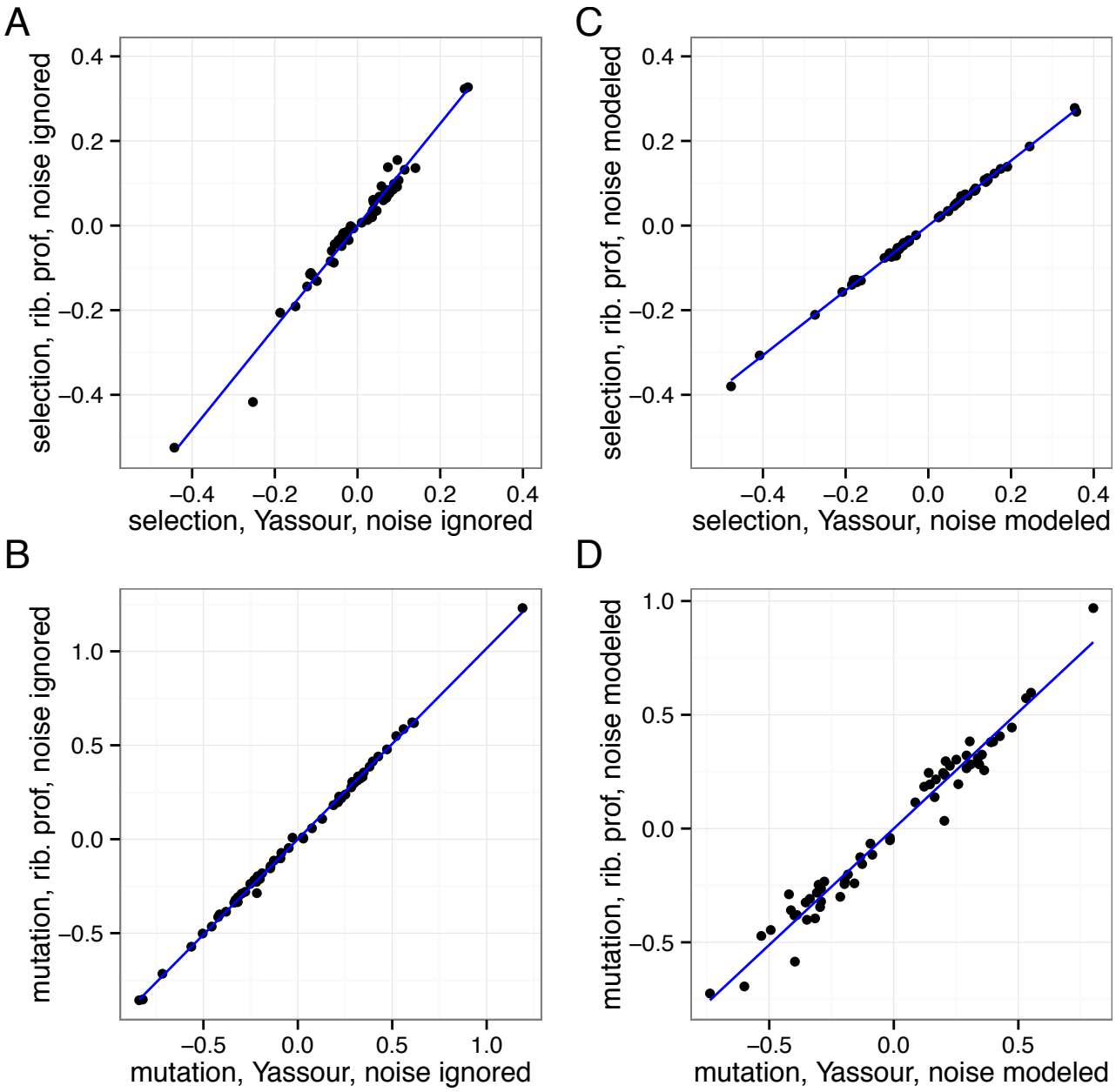

Figure S6: **Comparison of selection and mutation estimates from mRNA abundance and ribosome profiling datasets.** Each dot represents a codon, its x-coordinate is the selection coefficient (mutation bias) inferred from estimates of mRNA counts per cell in Yassour et al. (2009), and its y-coordinate is the selection coefficient (mutation bias) inferred from estimates of protein production rates in protein synthesis rates in proteins per cell per second from Ingolia et al. (2009). Ingolia et al's ribosome profiling dataset reported counts in terms of reads per kilobase per gene, and we converted to proteins produced per cell per second by normalizing the total to 13,000 (von der Haar, 2008); the numerical coincidence of the estimates being so similar despite different units is due to normalizing the mRNA counts per cell to an estimated total of 15,000 (Hereford and Rosbash, 1977; von der Haar, 2008).

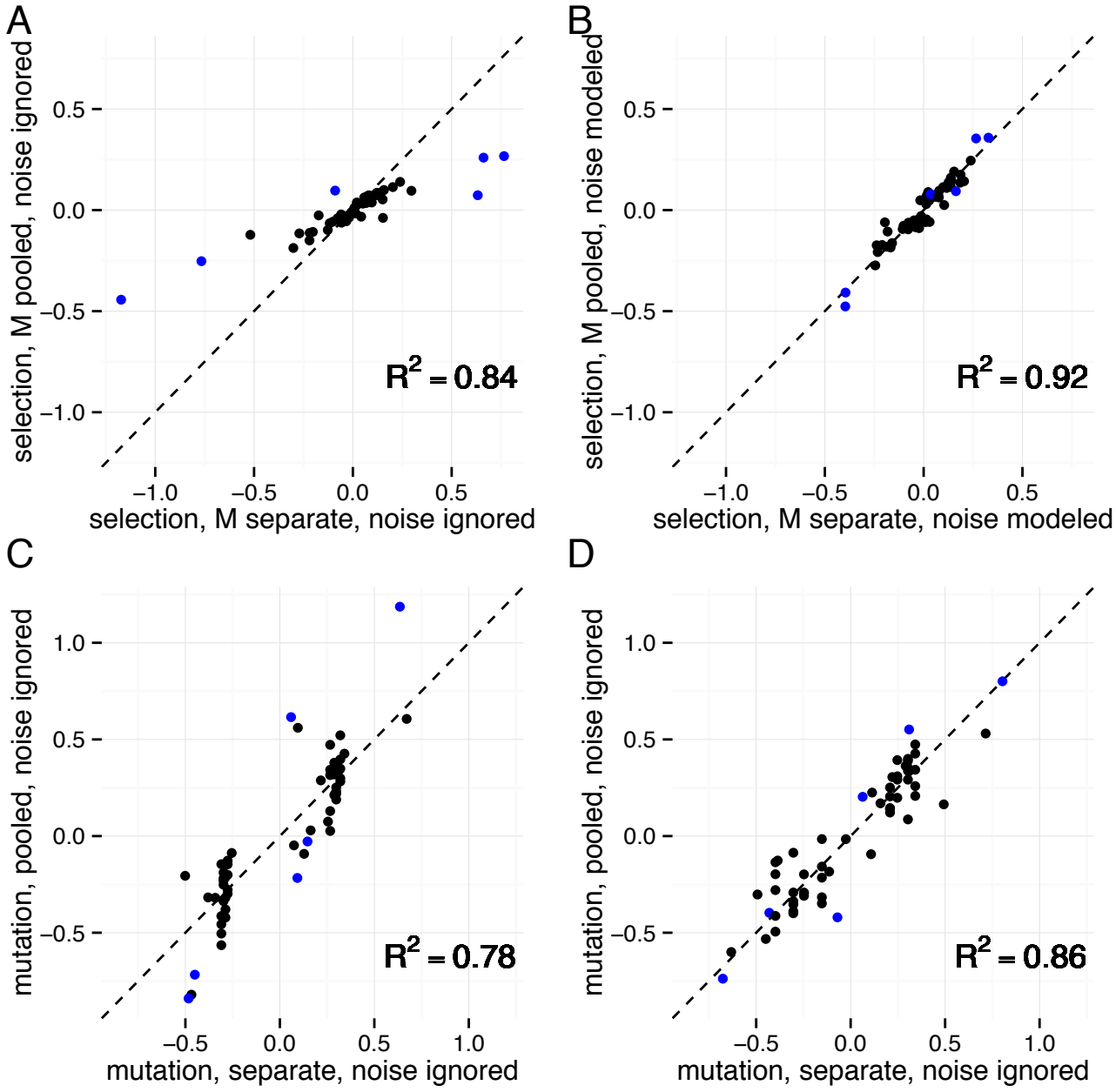

Figure S7: **Selection and mutation coefficients with mutation coefficients fitted separately, or pooled across amino acids.** A,B, selection coefficients compared with noise ignored (A) or modeled (B). C,D, mutation coefficients compared with noise ignored (A) or modeled (B). Again, each dot represents a codon, and fits are to the Yassour et al dataset; arginine codons are marked in blue.

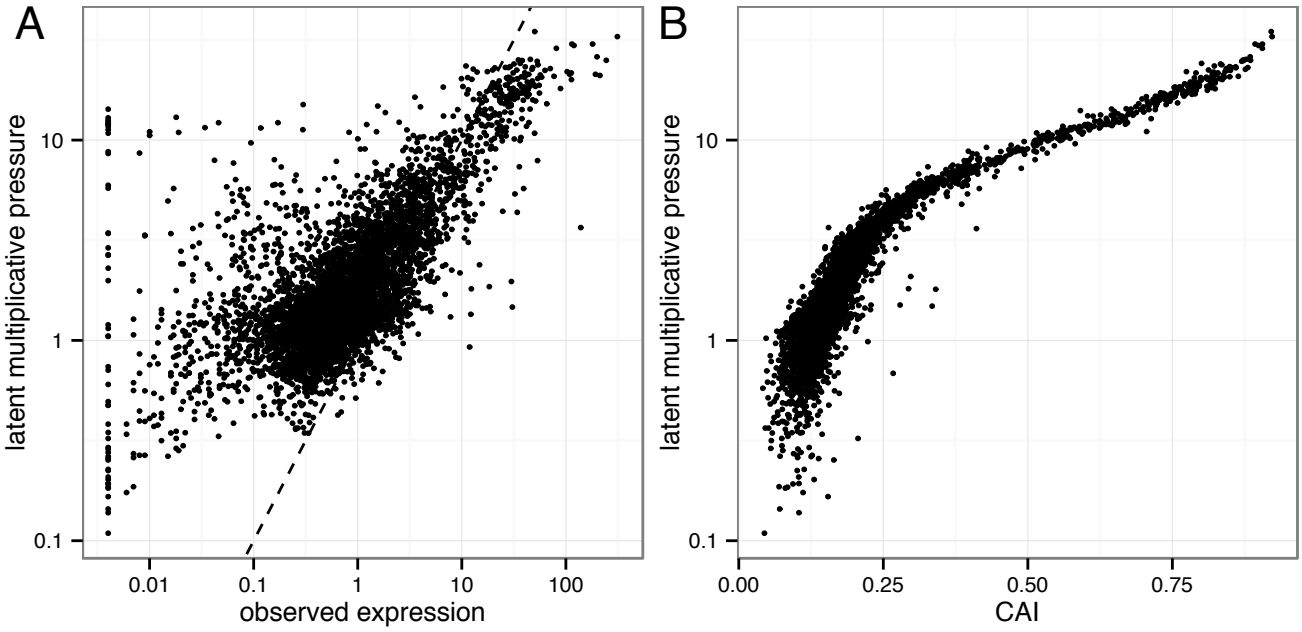

Figure S8: **Latent multiplicative pressure on translational selection compared with observed expression and CAI .** A, Latent versus observed expression for noise-modeled fit to Yassour dataset. B, Latent expression for noise-modeled Yassour dataset compared with CAI.

Supplementary Tables

Label	reference	name in reference
Yassour 1	Yassour et al. (2009)	ypd0.1
Yassour 2	Yassour et al. (2009)	ypd0.2
Yassour 3	Yassour et al. (2009)	ypd15.1
Yassour 4	Yassour et al. (2009)	ypd15.2
Lipson 1	Lipson et al. (2009)	chan7
Lipson 2	Lipson et al. (2009)	chan8
Lipson 3	Lipson et al. (2009)	chan9
Lipson 4	Lipson et al. (2009)	chan10
Lipson 5	Lipson et al. (2009)	chan11
Lipson 6	Lipson et al. (2009)	chan12
Lipson ma	Lipson et al. (2009)	MA
Ingolia 1	Ingolia et al. (2009)	mRNA-rich-1
Ingolia 2	Ingolia et al. (2009)	mRNA-rich-2
Ingolia 3	*	requantified
Ingolia 4	*	requantified, late
Ingolia 5	*	circular amplification
Ingolia 6	*	circular amplification, late

Table S1: **mRNA abundance measurements used.** (*) N. Ingolia and J. Weissman, unpublished data, 2010, used with permission of the authors. Datasets Ingolia 3 and Ingolia 4 use the same biological sample as Ingolia 2; datasets Ingolia 5 and Ingolia 6 also share a biological sample.

dataset	coefficient	corr	ρ	fold diff	abs. diff
Yassour	selection	0.98	0.98	1.68	0.04
Lipson 1	selection	0.96	0.96	2.10	0.04
Ingolia 1	selection	0.98	0.96	3.32	0.07
Yassour	mutation	0.88	0.82	0.88	0.13
Lipson 1	mutation	0.91	0.87	0.82	0.10
Ingolia 1	mutation	0.60	0.64	0.58	0.22

Table S2: **Comparing noise-modeled and noise-blind coefficient estimates.** For each of the datasets shown in figure 4, summary statistics comparing noise-modeled with noise-blind estimates of selection (mutation) coefficients. corr is Pearson correlation, ρ is Spearman's rank correlation, fold diff is the median of the fold difference (noise modeled divided by noise-blind estimate) and abs. diff is the median of absolute differences between the estimates.

Supplementary Files

MutSelCoefficientsAll.xls

Selection and Mutation coefficients for all datasets reported in paper. Reported selection coefficient is the average fitness effect per mRNA per cell of substituting the indicated codon for an equally-weighted mean of its synonyms, scaled by effective population size. The difference between reported selection coefficients for two synonymous codons is the selection coefficient of substituting one codon for the other. Likewise, the difference between reported mutation coefficients for two synonymous codons is the log ratio of mutation rates from one codon for the other. For comparison, mutation coefficients are reported normalized so that the mean per amino acid is zero; alternative mutation coefficients pooled by nucleotide are similarly reported so that the mean across nucleotides is zero.

orf-xObs-SCU-CAI-bygene.txt

Expression measurements and SCU (Selection on Codon Usage). Fields are orf, systematic ORF name for Saccharomyces cerevisiae; xObs, observed expression averaged from Yassour et al. (2009): $\exp(\text{mean}(\log(rpkb_g + 1)) - 1)$, normalized to total of 15,000 mRNAs per cell, then unverified orfs removed; CAI, codon adaptation index, calculated according to Sharp and Li (1987) from a reference list of proteins, see below; length is number of degenerate amino acids (non-Met and Trp amino acids in gene); SCU and mSCU are as described in text, displayed in figure 7.
Reference list of ribosomal and glycolytic proteins used to calculate CAI: YDR382W, YLR340W, YGR085C, YBL087C, YNL301C, YDL082W, YOR063W, YMR194W, YLR325C, YJL177W, YHR141C, YBR191W, YLR075W, YIL133C, YNL067W, YBR031W, YBL092W, YOL127W, YGR034W, YDR471W, YGL076C, YHL001W, YDR418W, YDL075W, YFR031C-A, YFL034C-A, YGL103W, YPL131W, YGL030W,

YDR500C, YLR448W, YMR230W, YJL190C, YPL090C, YOL040C, YNL096C, YHL015W, YGR118W, YLR441C, YNL302C, YGL123W, YJL191W, YDL083C, YJR123W, YGR027C, YOR369C, YER131W, YGR214W, YNL178W, YCR031C, YPL081W, YBL072C, YOR167C, YLR367W, YJL136C, YHR174W, YGL253W, YKL060C, YMR205C, YCR012W, YGR240C, YCL040W, YGR254W, YKL152C, YAL038W, YGR192C, YFR053C, YBR145W.

codonFits_0.4.tar.gz

An R package, codonFits, which implements all of the functions necessary to run the model. To install this, if you have a functional R installation, download the file, open a terminal window in the same directory and type

```
R CMD INSTALL codonFits_0.4.tar.gz
```

in the terminal window. This installation depends on several other R packages, listed in the `DESCRIPTION` file. Examples of how to use the package are in the pdf file `codonFits-vignette.pdf`, also accessible through `vignette('codonFits-vignette')` after installing the package in R.