Supplemental Materials

Comparison of 5′ and 3′ biases among the three S. cerevisiae *datasets*

The most pronounced deviations from expected nucleotide frequencies were observed at the 5′ ends of both fractions in all three datasets, though 3′ biases were also present in some cases (see below) (Fig. 2; Supplemental Figs S4-S6). The specific patterns of bias observed differed in a manner likely reflecting differences in library construction protocols used. Both the Artieri and Ingolia datasets performed circularization of the DNA products of *E. coli* poly-A polymerase mediated reverse transcription, adding spurious adenine nucleotides to the 3′ ends of most reads (Artieri and Fraser 2014; Ingolia et al. 2009). Therefore, mapping was accomplished by iteratively trimming the 3′ ends and retaining the longest mapping length within the acceptable range while discarding reads that mapped at either longer or shorter lengths. As a consequence, reads tended to map with greater length if they aligned to regions of the genome that happen to harbor adenines in the 3′ end of the mapping location, as revealed by an increasing relationship between 3′ adenine bias and read mapping length in some libraries (Supplemental Fig. S6). Unlike 5′ bias, this cannot affect the position of mapped reads, but it did increase the base-level coverage of adenine residues, which could create spurious signal in some analyses.

In contrast to the two other datasets, the McManus data employed a universal miRNA linker in order to prime reverse transcription, which appears to have mitigated the 3′ adenine bias (McManus et al. 2014) (Fig. 2, Supplemental Fig. S6). However, this may have introduced a more strongly pronounced bias against cytosine at the 4th nucleotide position as compared to the other datasets. We note that the McManus protocol also differed from the two other datasets in the manner of sucrose gradient mediated isolation of the 80s monosome, as well as in the specific details of the size-selection step subsequent to the addition of the miRNA linker. These may explain the more restricted mapping length distribution we observed in the 28 – 31 nt range in the this dataset (Supplemental Fig. S3).

Despite differences among protocols, a preference for adenine in the first nucleotide position remained a consistent bias among all datasets, indicating that some component of the library preparation protocol common to both fractions preferentially selected for certain fragments. However, the 'non-preferred' nucleotides varied among datasets – cytosine and guanine in the Artieri and McManus data vs. guanine and thymine in the Ingolia data – which could reflect changes in next-generation sequencing protocols, reagents, and/or equipment used (The Ingolia dataset was generated on the Illumina Genome Analyzer II instrument using reagents available at the time, whereas the two more recent datasets were generated using the Illumina HiSeq 2000 instrument).

It is also important to note that our approach of dividing the Ribo fraction enrichment by that of the mRNA fraction controls for shared biases between fractions. However, the method of mRNA isolation is not identical between the two fractions: the mRNA fraction is derived from alkaline hydrolysis of oligo-dT purified RNA, while the Ribo fraction is obtained by digesting non-ribosome bound mRNA by RNAse I (Ingolia et al. 2010). Such differences are likely to contribute to the fraction-specific differences observed in 5′ nucleotide enrichment (Figure 2) as well as the patterns of codon-specific bias observed at position 0 in Fig. 4 and Supplemental Figs. S9, S12, S16, S25, and S26. We believe that the enrichments at position 0 are unlikely to be biologically meaningful in explaining ribosomal stalling for the following reasons: First, enrichment observed at position 0 clustered based on the specific nucleotide at first, second, and third reading frame mappers rather than any clear codon type. Second, these patterns were highly specific to the 5′ ends of reads and were not observed at adjacent codons, as would be expected if, for instance, previously translated amino acids were hindering ribosomal progression due to interactions with the negatively charged exit tunnel (Lu et al. 2007) (see Supplemental Fig. S28).

Analysis of nucleotide and codon biases based on mapping read length

Previous studies have identified reads mapping at a length of 28 nt as being of particularly ‘high-quality’ as this corresponds to the fragment length that the ribosome should protect from nuclease digestion during library preparation (e.g., Ingolia et al. 2009; Qian et al. 2012). We therefore reanalyzed reads as in Fig. 2, but this time as a function of mapping length (Supplemental Fig. S6). While mapping read length had no substantial effect on biases in the mRNA fractions, 28 nt mapping reads from the Ribo fractions of all three datasets showed the strongest internal biases in codon representation (though internal enrichment of 27 nt reads were similar). Furthermore, 28 nt reads were enriched among first reading frame mappers in all three datasets, whereas third frame mappers were enriched for longer read lengths (Supplemental Fig. S7). Hence, if internal biases in codon representation represent a signal of ribosome stalling, reads ≤ 28 nt appeared to be optimal for detecting such a phenomenon.

Effect of allowing mismatches on interpretations of ribosomal occupancy

Our mapping approach allowed no mismatches between reads and the genomic sequence (see Methods), thus reducing the number of potential mapping reads available for analysis. Therefore we modified our mapping protocol to allow a single mismatch during mapping and compared reads mapping without mismatches to those containing a single mismatch (Supplemental Figs. S33 and S34; see Supplemental Methods). While we obtained between 15-62% more mapping reads depending on the library, we observed that for all Ribo fractions, the vast majority of mismatches occurred at the 5′ most base (Supplemental Fig. S33A), consistent with the known properties of untemplated addition of 5′ nucleotides by reverse transcriptases (Zajac et al. 2013). As a consequence mismatch mapping reads do not show the reading frame periodicity observed in non-mismatch reads, with second and third frame mapping reads showing increased frequency relative to first frame mapping reads (Supplemental Fig. S33B). Unlike the case of reads mapping without mismatches, the periodicity of reads mapping with a single mismatch is inconsistent among datasets. Therefore, given that reading frame and mapping length affect the results of ribosomal occupancy (see Main Text and above), we did not include reads mapping with mismatches as it is not possible to confidently assign them to a proper reading frame. However, we note that repeating our analysis of the Artieri and McManus datasets allowing at most one mismatch during mapping does not change the results of the analysis (Supplemental Fig. S34).

Positive amino acid codons are not enriched upstream of ribosomal active sites

Positively charged amino acids have been hypothesized to slow translation due to electrostatic interactions between their charged side-chains the negatively charged ribosomal exit tunnel (Lu et al. 2007; Lu and Deutsch 2008). Consequently, we would predict that positive amino acid encoding codons would be enriched upstream of codons putatively spanned by the ribosomal active sites, as they exert their influence only after having been translated. Furthermore, we should expect to observe a greater enrichment associated with amino acids with a stronger net-positive charge (lysine > arginine > histidine) (Charneski and Hurst 2013). However, we observed no evidence of a general enrichment of positive amino acid encoding codons upstream of the ribosomal active sites in any dataset (Figure 4, Supplemental Figures S9, S12, S16, S25, S26, and S28). We note that while there does appear to be a weak pattern of enrichment among two rare arginine codons (CGA and CGG) in the Artieri and McManus datasets (Supplemental Fig. S9), there is no compelling biochemical reason why the positive charges of only a subset of arginine codons, and no lysine codons, would exert a stalling effect. Therefore, these weak patterns are most likely due to sequence biases not associated with their positive charge.

No evidence of bias due to oligo-dT selection of the mRNA fraction

Enrichment of mRNA from total cellular mRNA via oligo-dT beads is known to bias read coverage towards the 3′ end of transcripts (Li et al. 2010; Zheng et al. 2011). In contrast, the Ribo fraction is not subject to the same position coverage bias (Ingolia et al. 2009). Because 3′ codons in the CDS will have greater sequencing depth as compared to 5′ codons, it is conceivable that our method normalizing Ribo fraction coverage by the that of the mRNA fraction could be sensitive to uneven codon usage at the ends of genes, in addition to translational stalling. Therefore, in order to test for this possibility, we analyzed patterns of position-specific codon enrichment in the Artieri data among either the first or last 250 codons of genes at least 500 codons in length (Supplemental Methods). Furthermore, because 3′ biases in the mRNA fraction are expected to be stronger for longer transcripts, we binned genes into three separate, equally sized, bins of increasing CDS length, performing the analysis on each length bin separately (Supplemental Fig. S11). As can be seen in the figure, patterns of enrichment are consistent among 5′ and 3′ ends across all three length categories – particularly in the case of the enrichment of proline codons (CCN) at position 4 – indicating that oligo-dT selection of mRNA did not influence our interpretation of position-specific codon enrichment.

Applying the corrected Ribo coverage method to the data of Ingolia

When the corrected Ribo coverage method was applied to the two datasets generated by Ingolia et al., we did not observe a consistent enrichment of proline codons at position 4 (the P-site), in contrast to the higher-coverage datasets (Supplemental Fig. S12). Furthermore, we could not attribute this difference to the lower coverage of these data as subsamples of first frame mappers of the Artieri or McManus data to the level of coverage of the Ingolia rich dataset continued to produce strong enrichment of the four proline codons (Supplemental Fig. S13). However, we note that features unique to the Ingolia dataset render it qualitatively different from the other two sets. First, stronger differences in 5′ nucleotide and codon bias were observed between the two biological replicates of the Ingolia data as compared to the other two sets, suggesting the presence of substantial batch effects across library preparation and/or sequencing (Supplemental

Fig. S5). Second, in addition to these differences, the magnitude of 5′ biases in all Ingolia Ribo fractions was substantially higher than their corresponding mRNA fractions, indicating that the Ribo fractions harbored more significant 5′ technical biases that could not be controlled by those observed in the mRNA fraction (in comparison, the 5′ biases are of comparable magnitude between fractions of the Artieri and McManus datasets) (Fig. 2, Supplemental Figs. S4-S6).

We believe that these biases are technical and not biological in nature because they were highly specific to the 5′ ends of Ribo reads and were not observed at adjacent nucleotide/codons sites, either upstream or downstream, as would be expected if, for instance, previously translated amino acids were hindering ribosomal progression due to interactions with the negatively charged exit tunnel (Lu et al. 2007) (see Supplemental Fig. S28). Therefore, we suggest that strong patterns of bias observed at the 5′ ends of the Ribo reads in the Ingolia data – as compared to the Artieri or McManus data – may have led to restricted patterns of mapping that interfered with the ability to detect patterns of over-represented codons at other sites.

Analysis of RNA secondary structure

Previous studies reported a correlation between ribosomal occupancy of the Ingolia data and the presence of secondary structure in mRNAs (Tuller et al. 2011; Charneski and Hurst 2013; Yang et al. 2014). Therefore we determined whether such a relationship was observed in the Artieri and McManus datasets. We obtained experimentally determined structures for 2,839 *S. cerevisiae* mRNAs (Ouyang et al. 2013) and tested whether the local secondary structure (from 10 codons upstream to 20 codons downstream of the read) was correlated with coverage of the 5′ ends in the Ribo, mRNA, or corrected Ribo reads (Supplemental Fig. S18; See Supplemental Methods).

We observed that the overall correlation between read occupancy and secondary structure was weak at all positions (Supplemental Fig. S18). However, the strongest correlations were observed at codons corresponding to the 5′ and 3′ ends of reads, coinciding with the locations of the most pronounced nucleotide biases (Fig. 2). For example, in the correlation with the mRNA fraction of the Artieri data, secondary structure at codons corresponding to the ends of reads were negatively correlated with occupancy, as expected from the over-representation of adenine-rich codons at these positions, which do not form strong base-pairing interactions. Despite a preference for adenines at the 3′ ends of reads, the Ribo fraction showed a slight increase in the correlation between secondary structure at codon position 9 and read occupancy (Supplemental Fig. S18); however, it remained weaker than the correlations observed in the mRNA fraction. The McManus et al. data showed the strongest correlations at the 5′ ends, also consistent with terminal nucleotide biases dominating the signal. Therefore, our results suggest that if there is a correlation between ribosomal occupancy and the presence of secondary structure, it is quite weak relative to the general variation in ribosomal occupancy across mRNAs – so much so that the signal is overwhelmed by the inherent biases introduced during the library construction process (see Discussion of main text).

The rate of translation is not correlated with codon optimality

We found no significant negative correlation between corrected Ribo occupancy and any of three different measures of codon optimality at either position 4 (Supplemental Fig. S20) or position 5 (Supplemental Fig. S21). This agreed with other riboprofiling-based analyses performed in bacteria (Li et al. 2012), yeast (Qian et al. 2012; Zinshteyn and Gilbert 2013), and mouse (Ingolia et al. 2011). In contrast, those studies that reported such a relationship focused primarily on the association between non-optimal codons and increased ribosomal occupancy at the 5′ ends of genes (Tuller et al. 2010b; 2011), rather than a direct association between Ribo fraction occupancy and non-optimal codons. In addition, a recent systematic study of translation in *E. coli* using synthetically designed N-terminal codon compositions found that the preference for non-optimal codons near start codons could be explained by their reduced secondary structure, enabling efficient translational initiation (Goodman et al. 2013). Therefore, our analysis supports the notion that the pool of tRNAs and transcriptome-wide CUB are adapted for efficient peptide synthesis *in vivo*, and that previous *in vitro* studies that found a strong relationship between CUB and translation rate may reflect circumstances that deviate significantly from those found in the cell (Plotkin and Kudla 2011; Qian et al. 2012).

Reanalysis of the approach of Charneski and Hurst

To illustrate the read coverage dependence of the method of Charneski and Hurst (2013), consider a situation where data are extremely sparse, such that no more than a single read maps to any 61 codon window (which is the size of the analysis space employed in their manuscript, representing 30 codons upstream of a putative stalling codon, to 30 codons downstream; Supplemental Fig. S29). Three possible mapped reads are shown to illustrate that read mapping position strongly influences its contribution to the average r_{pos}/r_{prec30} value. Importantly, averaging over all possible single read mapping positions produces a pattern that is highly characteristic of what is interpreted as the typical ‘stalling’ pattern observed in several of Charneski and Hurst’s figures. This bias towards producing a signature of stalling can be further illustrated by generating randomly positioned reads within the analyzed annotation, with length randomly chosen between 27 and 30 nt, and with increasing levels of coverage. When averaging over all possible 61 codon windows, a stalling pattern very similar to those observed in the Charneski and Hurst analysis manifests at low read coverage, but largely disappears at high coverage (Supplemental Fig. S29). Tellingly, averaging over all possible codon sites in the analyzed data produces a pattern very similar to that observed from 100 random reads assigned to each gene (note that the average read depth of the data is ~260 reads per gene and the reads are not randomly distributed as in the simulated data). It is also worth noting that the biases observed in this approach also explain the consistent pattern of increased coverage at the left end (near the -30 position) of almost all figures in the Charneski and Hurst manuscript, which the authors attribute to “some residual slowing […] due to slowing elements (e.g., positive charges) that may be encoded just upstream…” (pg. 4). It is clear from simulated data that this pattern results from the edge effect of low coverage reads that partially overlap the most upstream codons of the window being considered.

We also note that the most pronounced codon level biases observed in the Ingolia Ribo fractions are an enrichment of positive amino acid encoding codons at the 5′ end of reads, likely the result of the most abundant positive charge encoding codons being A rich (particularly among first frame mappers): lysine, AA[A/G], and arginine, AG[A/G] (Supplemental Fig. S28). Irrespective of the coverage sensitivity of the r_{pos}/r_{prec30} method, this bias could explain why the Ingolia data produce an accumulation of reads when positive amino acid encoding codons are at and downstream of the focal codon.

Supplemental Methods

Iterative mapping of reads allowing mismatches

Reads from both fractions of all datasets were mapped as detailed in the Materials and Methods of the main text, with the following modifications: We first excluded reads that mapped to the complete rDNA sequence of *S. cerevisiae* when trimmed to a length of 23 nt from the 5′ end using Bowtie version 0.12 (Langmead et al. 2009) allowing 3 mismatches and a maximum of 40 mapping locations. Remaining reads were mapped to the *S. cerevisiae* strain S288c genome (R61-1-1, 5th June 2008), retaining only reads with unique mapping locations and up to a single mismatch. Mapping reads were filtered such that no more than 30 bp (31 bp if the 3′ most nucleotide mapped with a mismatch), and no less than 27 bp mapped (28 bp if the 3′ most nucleotide mapped with a mismatch). Reads were assigned to the CDS as in the main text. Note that in the case of the Artieri data, we restricted our analysis to the libraries that were derived solely from *S. cerevisiae* (i.e., mRNA replicate 2 and Ribo Replicate 1) (Supplemental Table 3). For the other two datasets, replicates were combined for subsequent analysis.

Using the reads that mapped with a single mismatch, we determined the relative proportion of mismatches (# of reads with mismatches at site X/total # of reads mapping with a single mismatch) at each site among the first 27 nts of each read.

Analysis of potential biases introduced by oligo-dT selection of mRNA fraction

Genes of at least 500 codons in length were sorted by length and binned into 3 bins each containing an equal number of genes. Normalized position-specific codon enrichment (see Main Methods) was then performed separately on the first and last 250 codons of genes in each bin (Supplemental Fig. S11). As was the case for the whole CDS, the first and last 15 codons of each CDS were ignored due to previously observed patterns of increased occupancy due to translational initiation/disassociation (Ingolia et al. 2009). Consequently, the actual number of codons analyzed on each end was 235.

Analysis of S. paradoxus

We mapped the *S. paradoxus* parent riboprofiling data generated by McManus et al. (2014) as well as the mixed *S. cerevisiae*/*S. paradoxus* data of Artieri and Fraser (2014) using the same approach as in the case of *S. cerevisiae*, using the *S. paradoxus* reference genome generated by Scannell et al. (2011). Patterns of position-specific codon enrichment (Supplemental Figs. S15, S16, and S17) were assessed using the 4,640 high-quality *S. paradoxus* orthologs identified by Artieri and Fraser (2014) in order to avoid spurious or incomplete gene annotations in the Scannell et al. (2011) data. Biological replicates were combined for the purpose of analysis.

Analysis of mRNA secondary structure

We obtained the data of Ouyang et al. (2013), which includes binary designations for each base in 2,839 of the mRNA transcripts analyzed in this study, identifying it as single or double-stranded according to the transcriptome-wide measurements of Kertesz et al. (2010). Using the codon-specific occupancy for the 5′ end of first position mapping reads, we determined the correlation between occupancy at codon position 0, corresponding to the 5′ end of the read, and secondary structure from codon position -10 to +20. Codon-level secondary structure was scored as the average structural value of the three nucleotides in each codon, where a single-stranded nucleotide was given a value of 0 and a double-stranded nucleotide, 1. This correlation was determined independently using the mRNA and Ribo fractions in addition to the corrected Ribo coverage in order to determine whether biases in either fraction drove any patterns observed in the corrected data.

Analysis of riboprofiling data from additional species

Our approach to accounting for shared technical biases across fractions requires that size selection for the expected size of ribosome protected fragments (e.g., ~28 nt in yeasts) be applied to both the mRNA and Ribo fractions and that libraries from both fractions be generated using an identical protocol. In order to test the generality of the results observed in yeasts, we searched the literature for suitable datasets from other species identifying the *C. elegans* data of Stadler and Fire (2013) as well as the zebrafish data of Bazzini et al. (2014) as meeting the necessary conditions (Supplemental table S4 explains why additional riboprofiling datasets were unsuitable).

For the nematode analysis, we obtained the *C. elegans* data from Stadler and Fire (2013), GEO accession # GSE48140 consisting of three replicates of two separate conditions/timepoints each: starved L1 embryos and developing L1 embryos. Replicates were combined and mapped to the Ensembl *C. elegans* genome release WBcel235.75 using the same approach as was applied to the yeast data (see Methods). Only genes with complete CDSs were used for the codon level analysis, and for each gene with alternative isoforms, the longest isoform was used for analysis.

For the zebrafish analysis, we obtained all datasets prepared using the ARTseq Mammalian Ribosome Profiling Kit (Epicenter) from the dataset of Bazzini et al. (2014), GEO accession # GSE53693. Replicates and runs for each of stages 5, 12, and 24 hours post fertilization (hpf), corresponding to the three timepoints with the largest number of mapping reads in the Ribo fractions, were combined and mapped to Ensembl *D. rerio* genome release Zv9.63 again using the same approach as the yeast data in order to generate Supplemental Fig. S23. To generate Supplemental Fig. S26, all mRNA and Ribo fraction samples from all 5 timepoints (2, 5, 12, 24, and 48 hpf) were combined. As above, only genes with complete CDSs were used for the codon level analysis, and for each gene with alternative isoforms, the longest isoform was used for analysis.

Analysis of Ribo fraction codon-level biases

In order to generate Supplemental Fig. S18, we followed the method to generate corrected Ribo coverage, but without correction by the mRNA fraction: The 5′ ends of reads from the Ribo fraction were mapped as detailed in the methods section and the codon-level coverage was determined, retaining only codons with 5′ mapping data for analysis. Within each gene, codon-level coverage values were scaled by the mean codon-level coverage of analyzed codons in order to account for coverage differences among genes. These scaled values were then $\log_2$ transformed (e.g., $\log_2$[scaled Ribo coverage]) and then applied from -8 to +8 codons relative to the codon overlapped by the 5′ end (representing 17 codons in total). Performing this analysis over all positions with data within the coding transcriptome produced a distribution of $\log_2$(scaled Ribo coverage) values for each codon at each of the 17 positions, which were then combined into biochemical categories. The relative enrichment of each category at each position was determined by scaling its mean $\log_2$(scaled Ribo coverage) value by the mean value of the five categories at that position such that categories with positive $\log_2$ values were enriched relative to expectations and those with negative values were depleted. Error bars were calculated as the standard error of the mean among all measurements of codons within a biochemical category at each position.

Supplemental Tables

Supplemental Table S1. Number of reads in each dataset mapping to the *S. cerevisiae* genome used in the current analysis.

Sample	Fraction	Replicate	Mapped Reads
Artieri	Ribo	1	26,648,753
		2	17,102,202
	mRNA	1	10,128,586
		2	21,137,349
McManus	Ribo	1	6,246,282
		2	7,580,892
	mRNA	1	2,153,414
		2	2,050,551
Ingolia rich	Ribo	1	711,601
		2	958,424
	mRNA	1	288,061
		2	1,340,197
Ingolia starved	Ribo	1	350,787
		2	724,985
	mRNA	1	127,623
		2	796,049

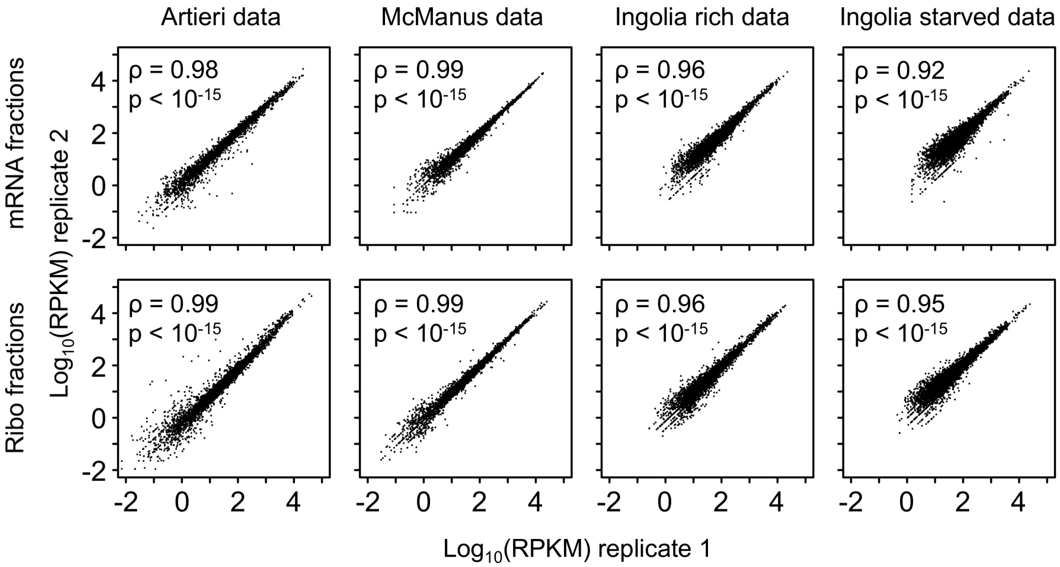
Supplemental Table S2. Proportion of adenine in the CDS as a function reading frame and expression level. The proportion of adenine within the CDS presented in the main text is calculated as the proportion among all nucleotides. As more highly expressed genes show greater CUB, this value will vary based on the subset of genes analyzed as well as the reading frame within each codon. However, the proportion of adenine varies within a narrow range, especially among first frame mappers. Expression quartiles were determined based on the mean RPKM among both replicates of the Artieri data.

Expression Quartile	Reading Frame	Percent Adenine
1	1	33.5
	2	33.1
	3	29.3
2	1	34.3
	2	36.1
	3	30.1
3	1	33.3
	2	35.9
	3	29.4
4	1	31.5
	2	35
	3	26.9

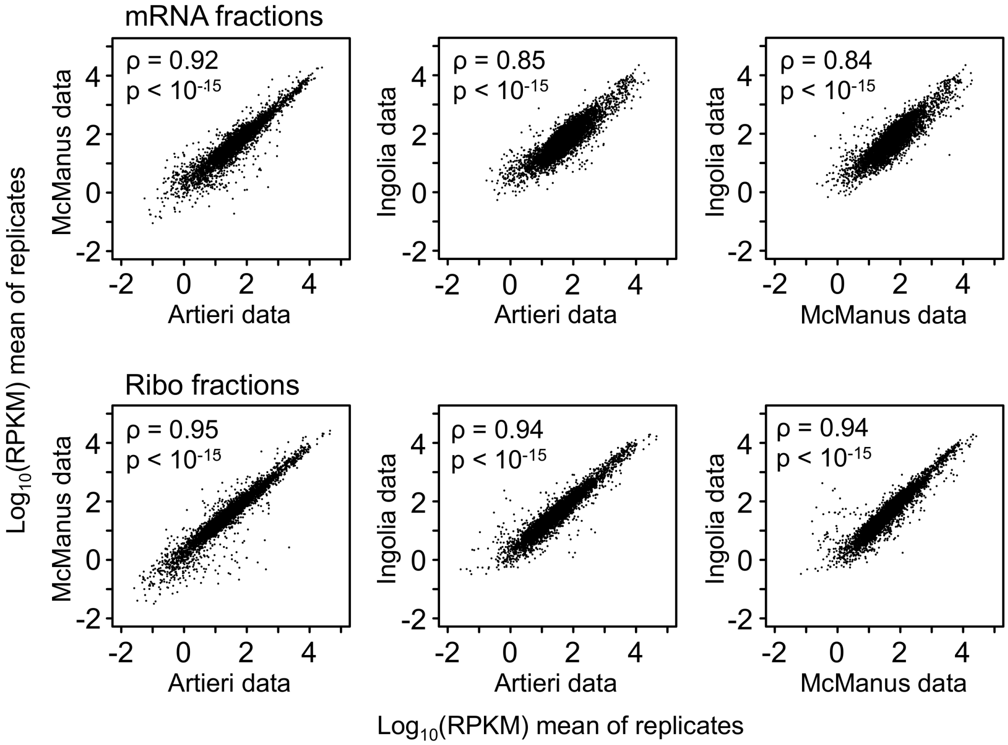
Supplemental Table S5. SRA sample ID numbers for Artieri and Fraser (2014) data used in the analysis.

Sample	Source	SRA Sample ID
mRNA Rep. 1	Artieri and Fraser mixed *S. cerevisiae*/*S. paradoxus* sample	SRS509272
mRNA Rep. 2	Artieri and Fraser *S. cerevisiae* sample	SRS469853
Ribo Rep. 1	*S. cerevisiae* sample generated for the present study	SRS514738
Ribo Rep. 2	Artieri and Fraser mixed *S. cerevisiae*/*S. paradoxus* sample	SRS509342 SRS509345

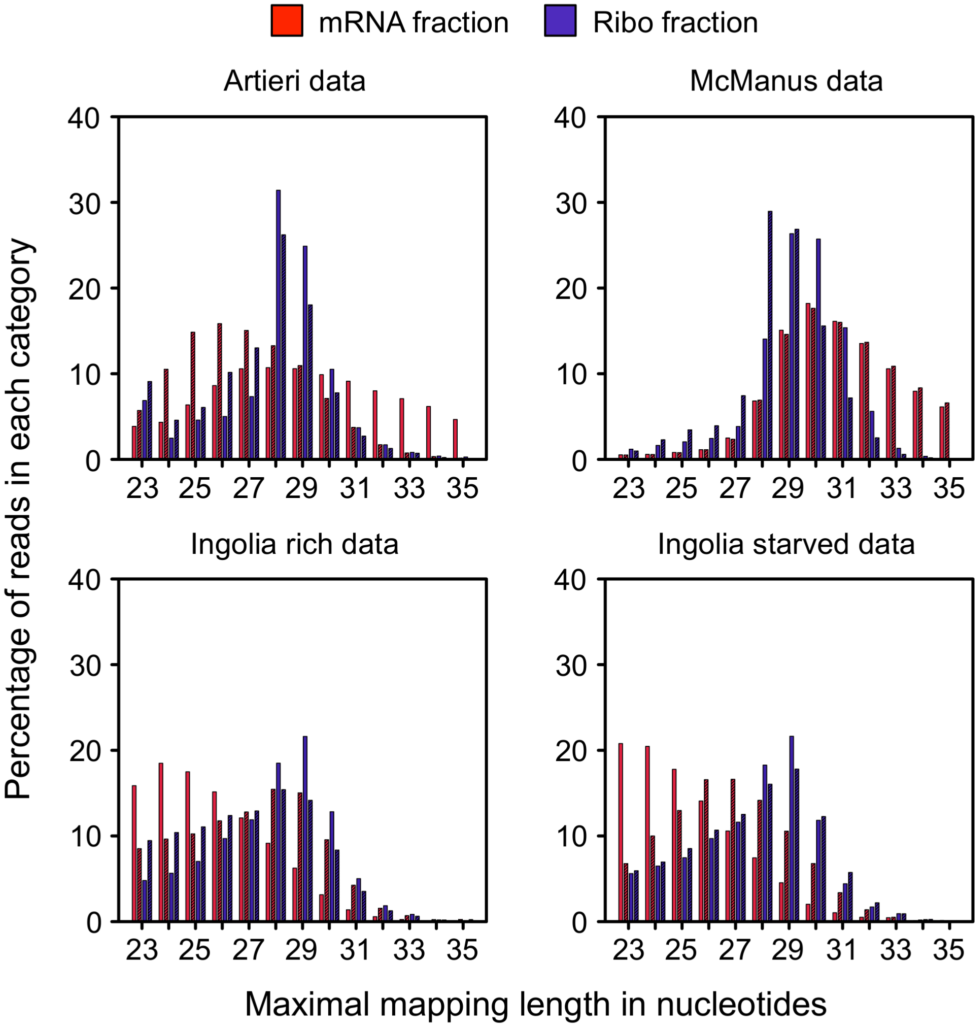
Supplemental Figures



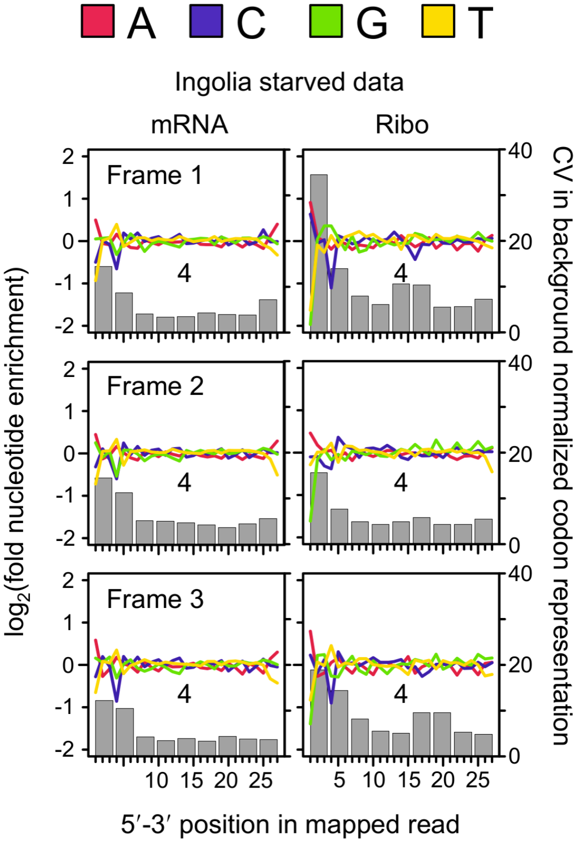
Supplemental Figure S1. Inter-replicate correlation of expression level estimates for the analyzed datasets. Only genes with Reads Per Kilobase per Million mapped reads (RPKM) > 0 are plotted. mRNA fractions are plotted in the row above, while Ribo fractions are indicated below. Spearman's ρ and associated p values are indicated in each panel. The lower ρ values in the Ingolia data likely reflect the lower number of mapping reads in that dataset (Table S1). Note that in the original Ingolia (2009) analysis, correlations were calculated using only genes with ≥ 128 mapping reads, which improves correlation coefficients.



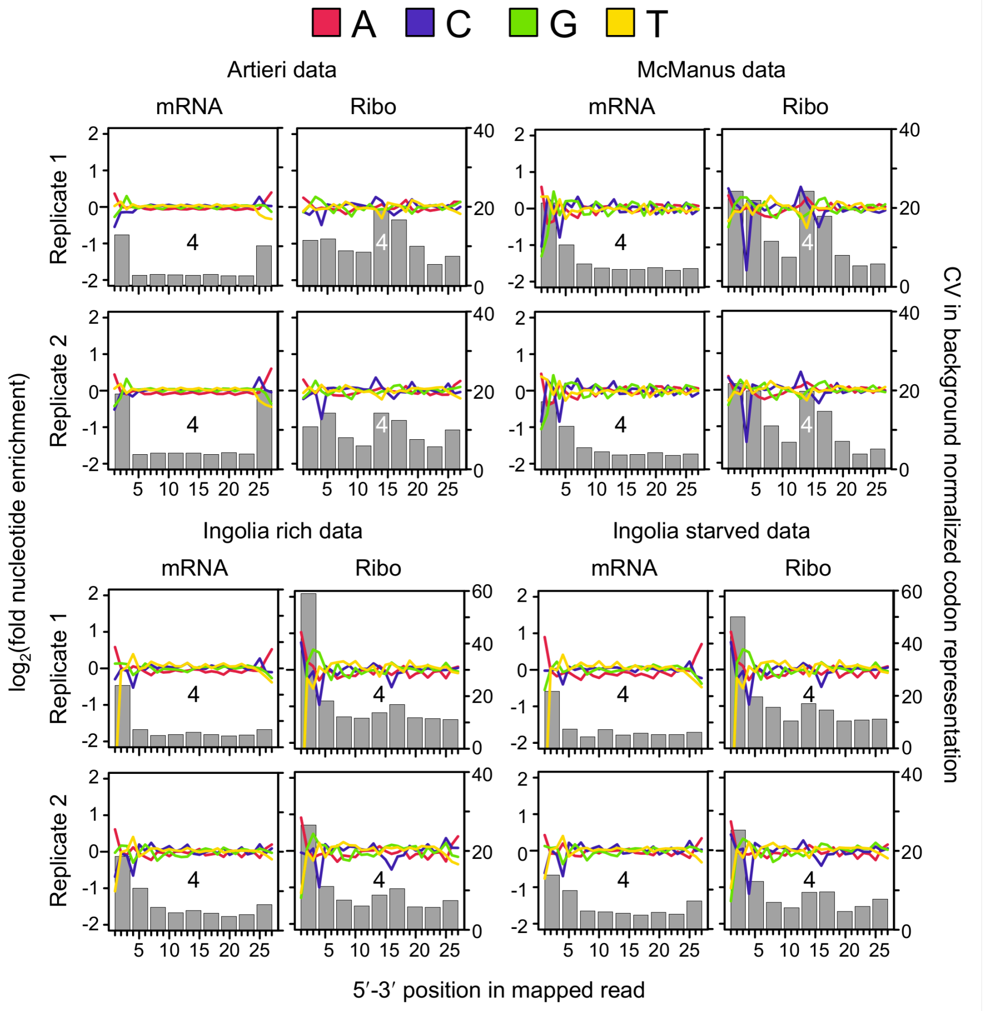
Supplemental Figure S2. Correlation of expression levels between the datasets used in the study. The mean RPKM of the two replicates generated in each study is plotted along with Spearman correlation coefficients, ρ, and associated p values. The slightly lower correlation for the mRNA data may reflect use of different methods to extract mRNA from the raw lysate (Ingolia et al. 2009; Artieri and Fraser 2014; McManus et al. 2014).



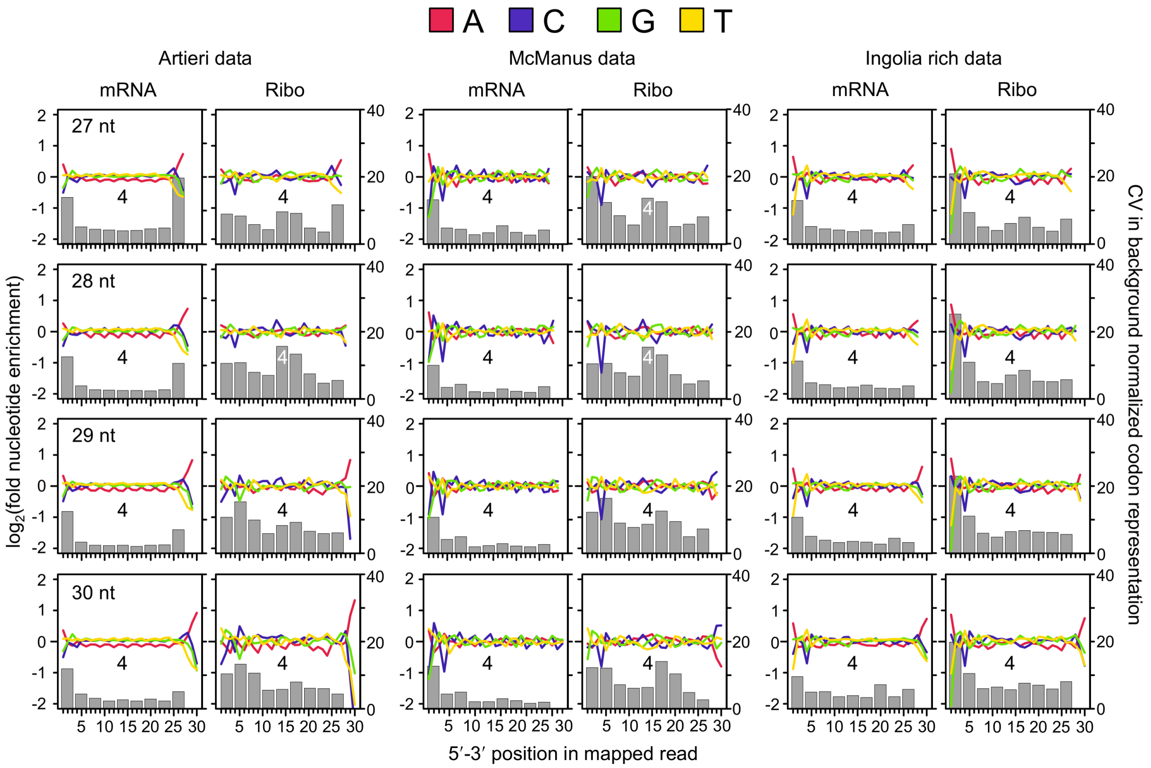
Supplemental Figure S3. Read mapping length distribution for each of the analyzed datasets. Variation in the length distribution of the mRNA fraction (red) is likely driven by the size and precision of the fragment excised from the denaturing SDS-PAGE gel during library construction (Ingolia 2010). In contrast, enrichment of ~28 nt fragments is expected in the Ribo fraction (blue) as this is the length of mRNA occupied by a translating ribosome (Ingolia et al. 2009). Biological replicates are plotted next to one another where replicate 1 is plotted without hashes and replicate 2 is hashed, resulting in a darker shade. The distribution of the McManus data is more compact and excludes short-length reads to a greater degree than other datasets. This is possibly due to their unique use of universal miRNA linkers followed by a modified size-selection protocol (see above) (McManus et al. 2014).



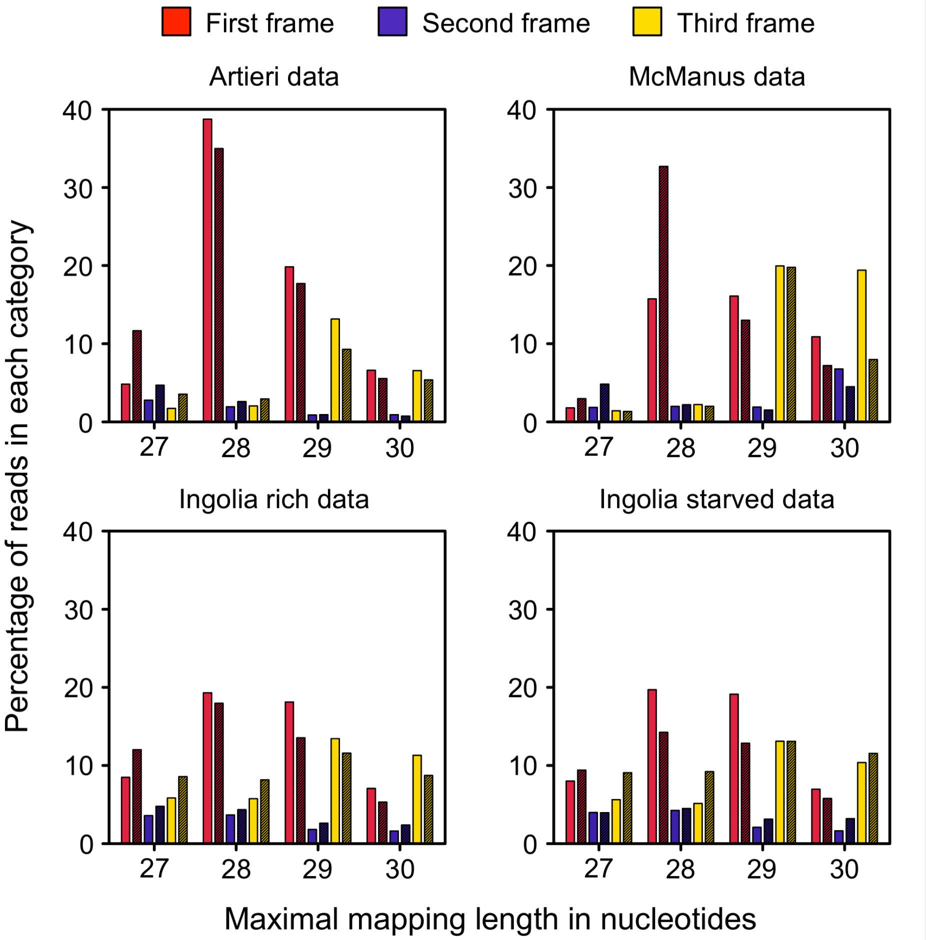
Supplemental Figure S4. Reproduction of Fig. 2 showing that the Ingolia starved data are qualitatively similar to the rich data.



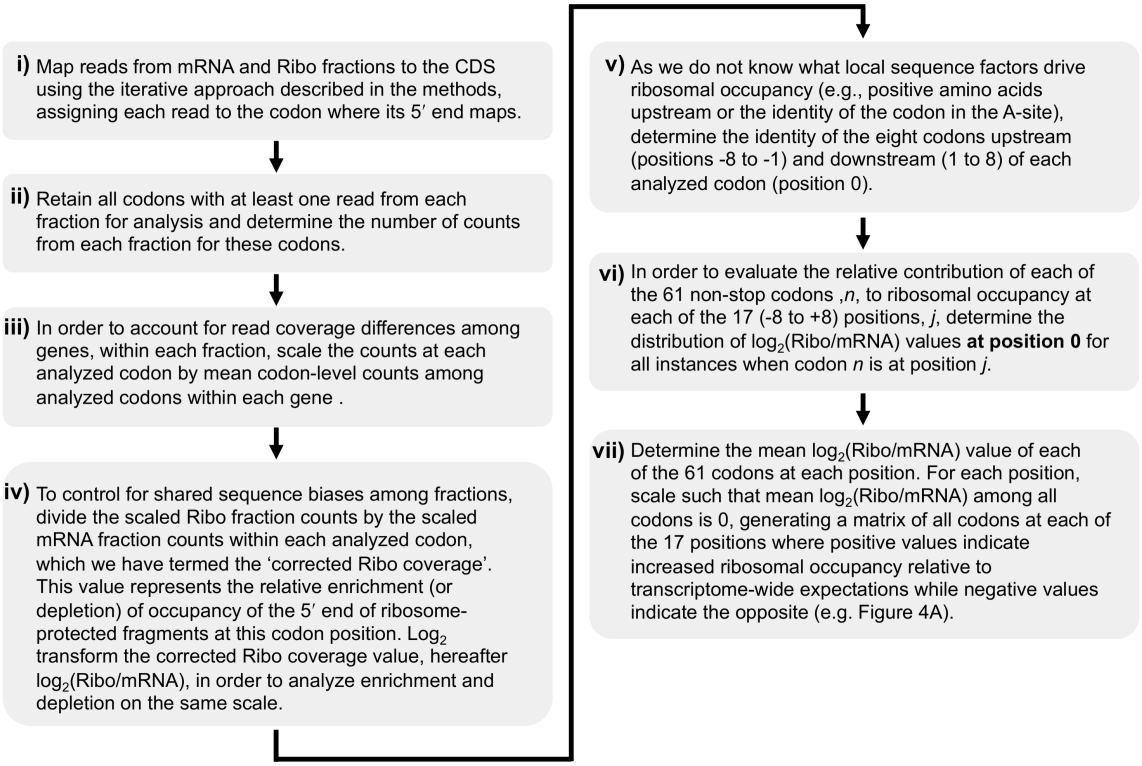
Supplemental Figure S5. Comparison of patterns of bias across replicates for first frame mappers. Patterns of bias were consistent across replicates of both fractions of the Artieri and McManus data (top). In the Ingolia data (bottom), 5′ biases were much more pronounced in the first as compared to the second replicate (note the difference in scale for the CV in background normalized codon representation used in replicate 1). In addition to differences between replicates in the Ingolia data, patterns of 5′ bias among the Ribo fractions were of greater magnitude relative to the mRNA fractions as compared to the higher-coverage datasets.



Supplemental Figure S6. Reanalysis of patterns of nucleotide and codon bias among reads binned by mapping length. The grey bars are plotted for the 9 codon positions (0-8) corresponding to 27 nt as in Fig. 2, with the fourth codon position indicated for reference. Patterns of 5′ bias are similar within fractions across mapping lengths. However, 3′ codon bias decreases with read length in the Artieri and Ingolia data due to biases in adenine content overlapping codon position 8 as a result of the use of poly-A polymerase to prime reverse transcription (see above). Significantly, in all three datasets, the strongest degree of internal codon bias in the Ribo fraction is observed among 28 nt mapping reads (with 27 nt reads often showing a similar pattern). This may be related to the general enrichment of first frame mapping reads among 28 nt mappers (Supplemental Fig. S7).



Supplemental Figure S7. Ribo fraction reads mapping at different lengths show differential enrichment among first, second and third frame mappers. All three fractions show an enrichment of 28-29 nt mapping reads among first frame mappers, which make up the majority of reads. Third frame mappers tend to be enriched for 29-30 nt mapping reads. Few reads map to the second reading frame in any dataset. The first, second, and third reading frames are shown in red, blue and yellow, respectively. Biological replicates are plotted next to one another where replicate 1 is plotted in a lighter shade and replicate 2 in a darker shade.



i) Map reads from mRNA and Ribo fractions to the CDS using the iterative approach described in the methods, assigning each read to the codon where its 5′ end maps.

ii) Retain all codons with at least one read from each fraction for analysis and determine the number of counts from each fraction for these codons.

iii) In order to account for read coverage differences among genes, within each fraction, scale the counts at each analyzed codon by mean codon-level counts among analyzed codons within each gene .

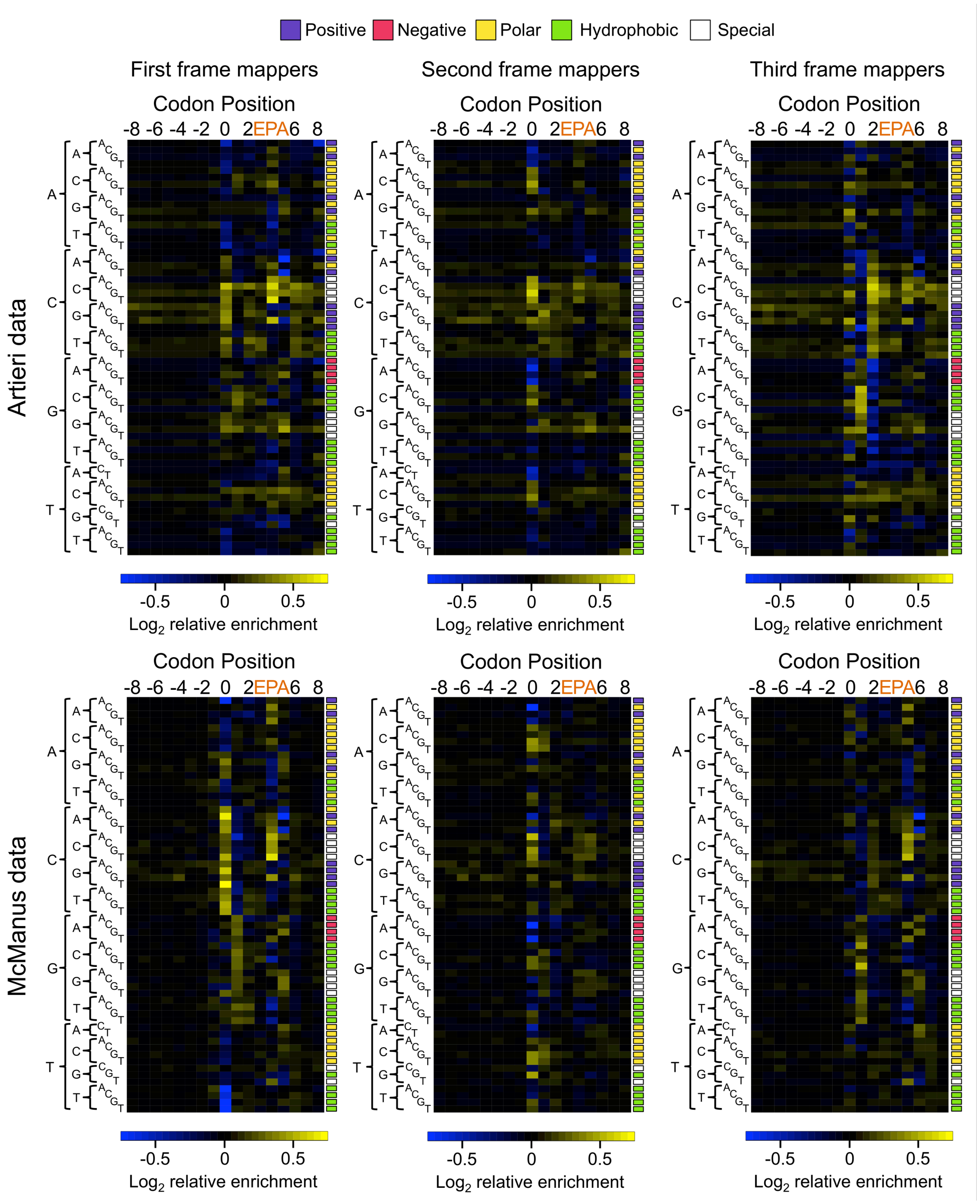
iv) To control for shared sequence biases among fractions, divide the scaled Ribo fraction counts by the scaled mRNA fraction counts within each analyzed codon, which we have termed the ‘corrected Ribo coverage’. This value represents the relative enrichment (or depletion) of occupancy of the 5′ end of ribosome-protected fragments at this codon position. Log_2 transform the corrected Ribo coverage value, hereafter $\log_2$(Ribo/mRNA), in order to analyze enrichment and depletion on the same scale.

v) As we do not know what local sequence factors drive ribosomal occupancy (e.g., positive amino acids upstream or the identity of the codon in the A-site), determine the identity of the eight codons upstream (positions -8 to -1) and downstream (1 to 8) of each analyzed codon (position 0).

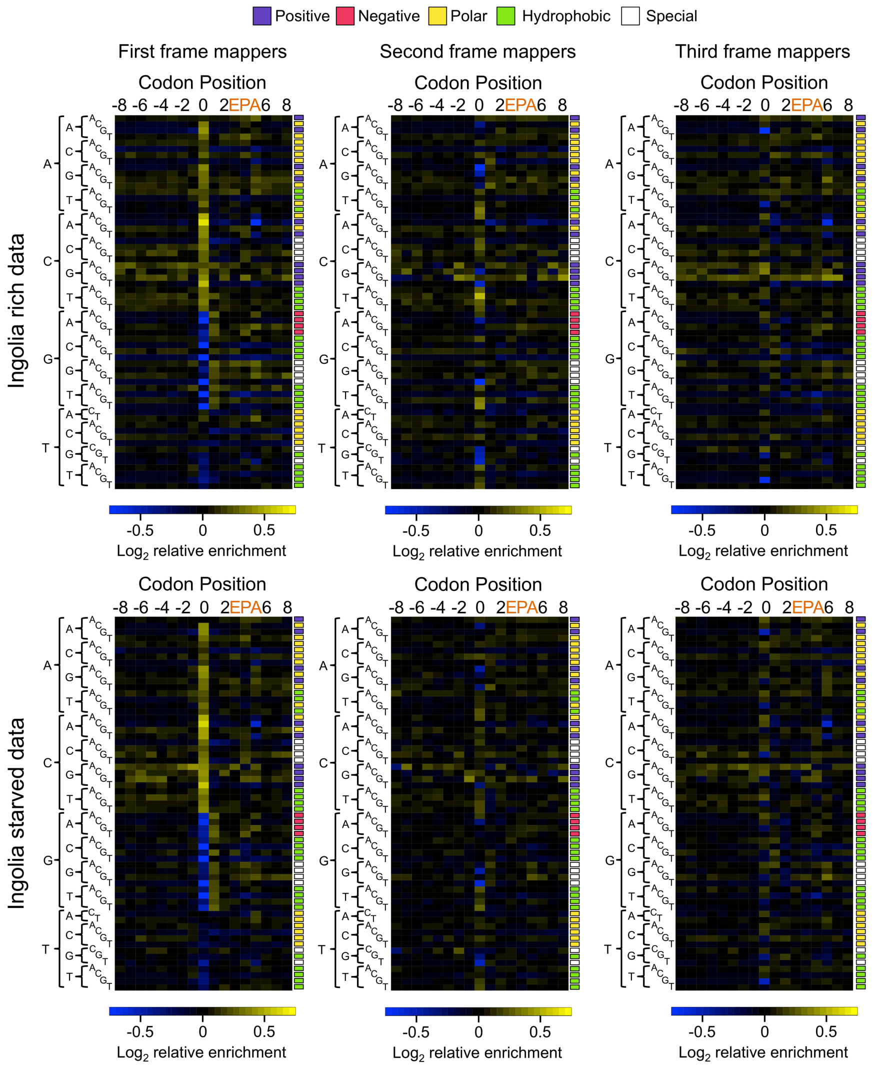
vi) In order to evaluate the relative contribution of each of the 61 non-stop codons ,*n*, to ribosomal occupancy at each of the 17 (-8 to +8) positions, *j*, determine the distribution of $\log_2$(Ribo/mRNA) values **at position 0** for all instances when codon *n* is at position *j*.

vii) Determine the mean $\log_2$(Ribo/mRNA) value of each of the 61 codons at each position. For each position, scale such that mean $\log_2$(Ribo/mRNA) among all codons is 0, generating a matrix of all codons at each of the 17 positions where positive values indicate increased ribosomal occupancy relative to transcriptome-wide expectations while negative values indicate the opposite (e.g. Figure 4A).

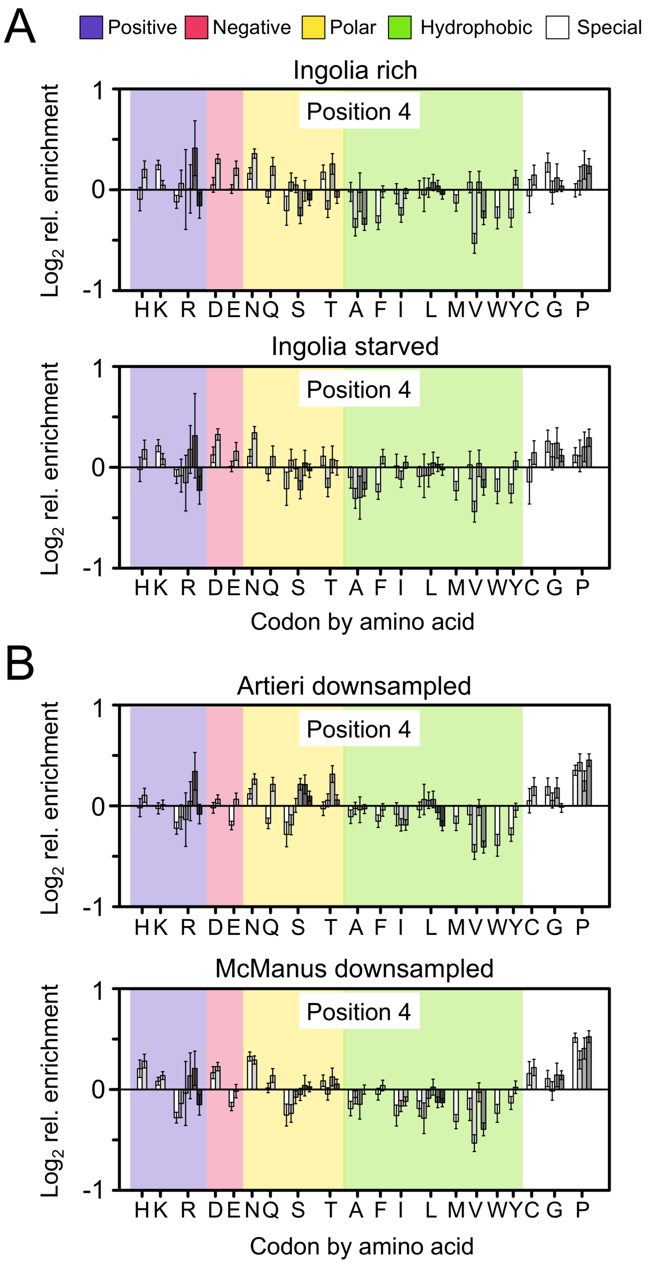
Supplemental Figure S8. Flow chart explaining how position-specific codon enrichment of corrected Ribo coverage was calculated.



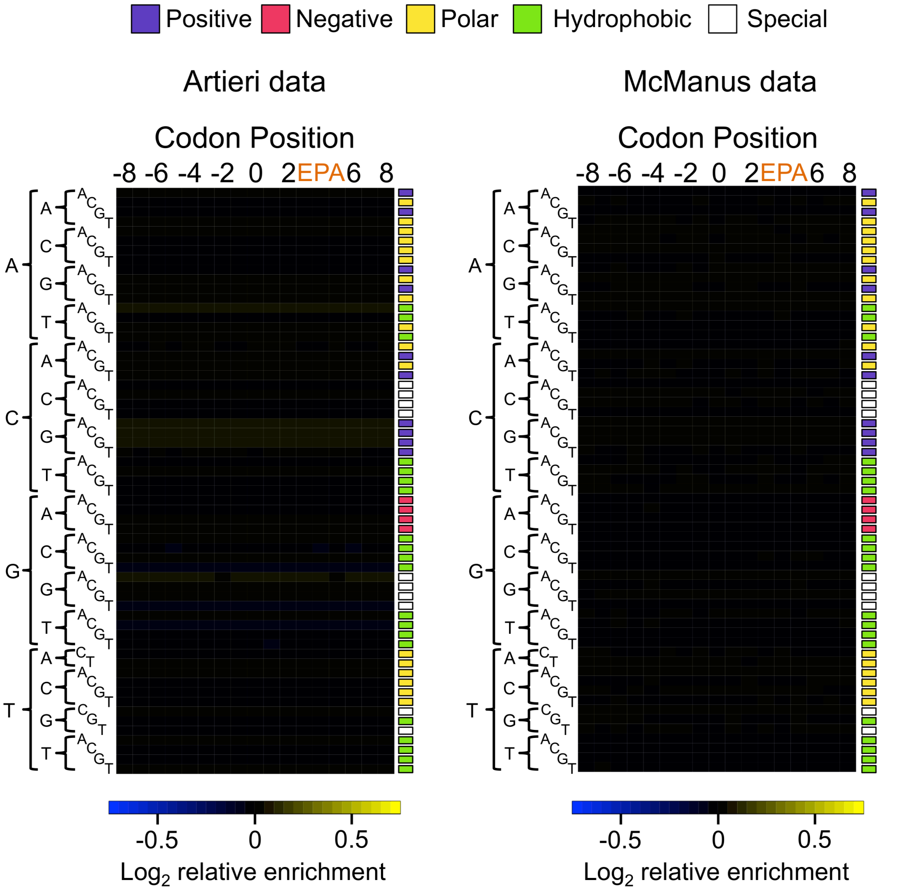
Supplemental Figure S9. Patterns of corrected codon enrichment at positions -8 to 8 for first, second, and third reading frame mapping reads for the Artieri and McManus data. All three frames continue to show substantial bias at codon position 0, corresponding to the 5′ end of the read, indicating that despite controlling for shared biases between fractions, substantial fraction-specific 5′ end biases remain. However, first position mappers of both datasets show a clear pattern of internal enrichment at position 4, previously defined as corresponding to the ribosomal P-site. Second frame mappers show a weaker pattern of enrichment corresponding to the same codons, but it is distributed among positions 4 and 5, suggesting that precise location of a potentially active ribosomal site is less clearly defined in relation to the 5′ end among reads mapping in this frame. Similarly, reads mapping to the third frame also show a qualitatively similar pattern of codon enrichment to first frame reads, but the most highly enriched codon position is 5. This supports our observation that codon positions of third frame mappers tend to be offset by +1 codon (see main text; Fig. 2). Third frame mapping reads also show stronger patterns of bias among codon positions 1 and 2, which are qualitatively consistent among the two datasets. Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



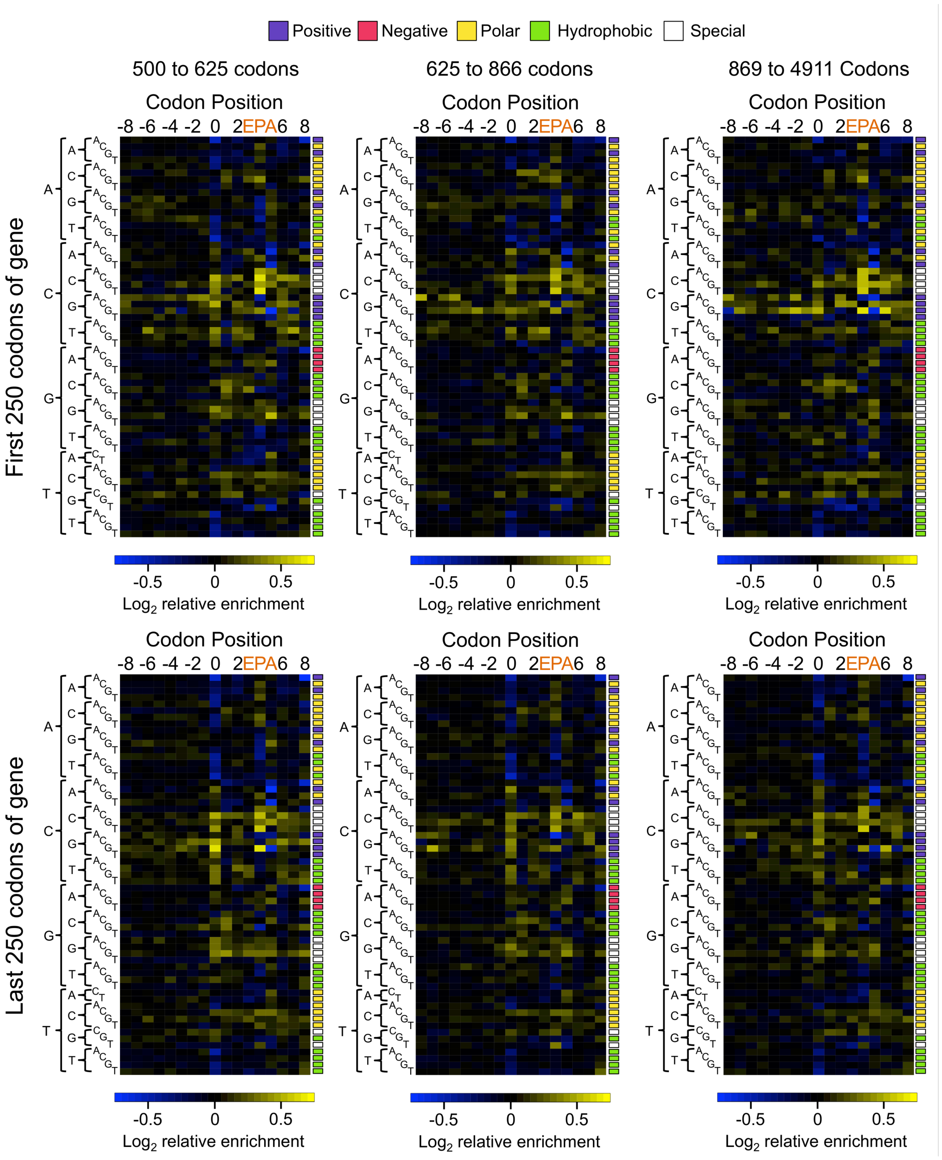
Supplemental Figure S10. Patterns of corrected codon enrichment at positions -8 to 8 for first, second, and third reading frame mapping reads for the Ingolia data. Patterns of enrichment in the Ingolia data (rich, above; starved, below) differ substantially from those observed in the higher coverage datasets. For instance, first frame mappers show overwhelming codon bias at position 0 (codons beginning with A or C are universally enriched, while those beginning with G and T are depleted) whereas no upstream, nor downstream positions show strong biases as was the case for position 4 in the Artieri and McManus data. Second and third frame mapping reads also show relatively high bias at position 0 relative to other codon positions; however, their magnitude appears to be much smaller, likely due to noise introduced by the small overall number of reads mapping in these frames (Supplemental Fig. S7). Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



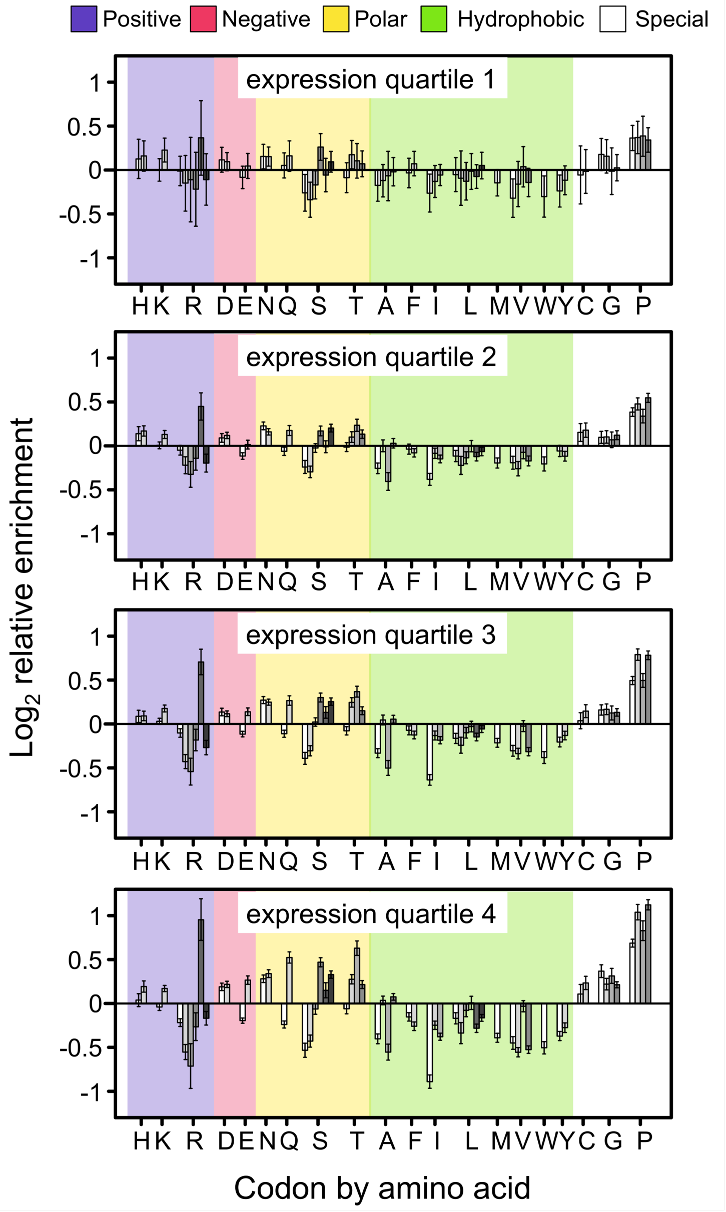
Supplemental Figure S11. Bar plots indicating the log$_2$ enrichment values at position 4 of first frame mappers. A) Combined replicates of the Ingolia datasets and B) Artieri and McManus datasets downsampled to the same number of reads as used to generate the Ingolia rich panel. Codons are organized by amino acid using single-letter designations below and grouped by biochemical type as indicated at the top of the panel. Individual codons for each amino acid are in alphabetical order. 95% confidence intervals around the scaled enrichment values are indicated for each bar. Subsets of the Artieri and McManus data still showed a clear enrichment of all proline-encoding codons, suggesting that the lack of proline codon enrichment in the Ingolia data was not explained by reduced coverage.



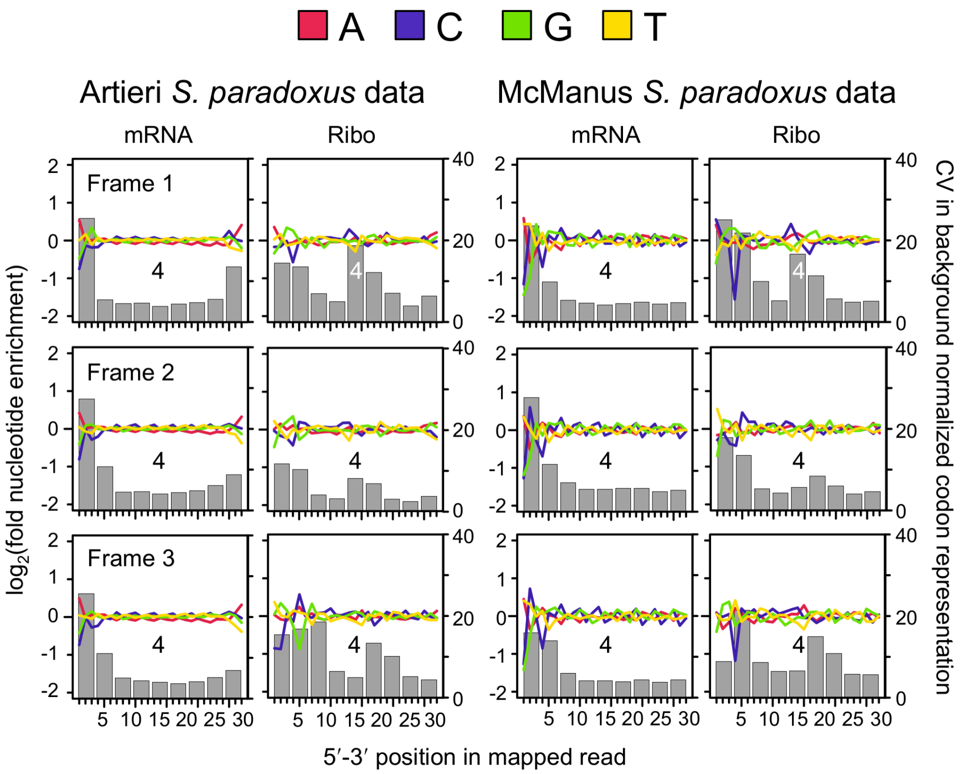
Supplemental Figure S12. Figure 4A redrawn from the mean of 100 permutations of the codon positions within transcripts while maintaining read mapping positions. Permutations are shown using either the Artieri or McManus data. All patterns observed in the actual data disappeared. The remaining very weak enrichment/depletion represents the underlying bias in the dataset that would be observed regardless of the biological factors influencing ribosomal occupancy. Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



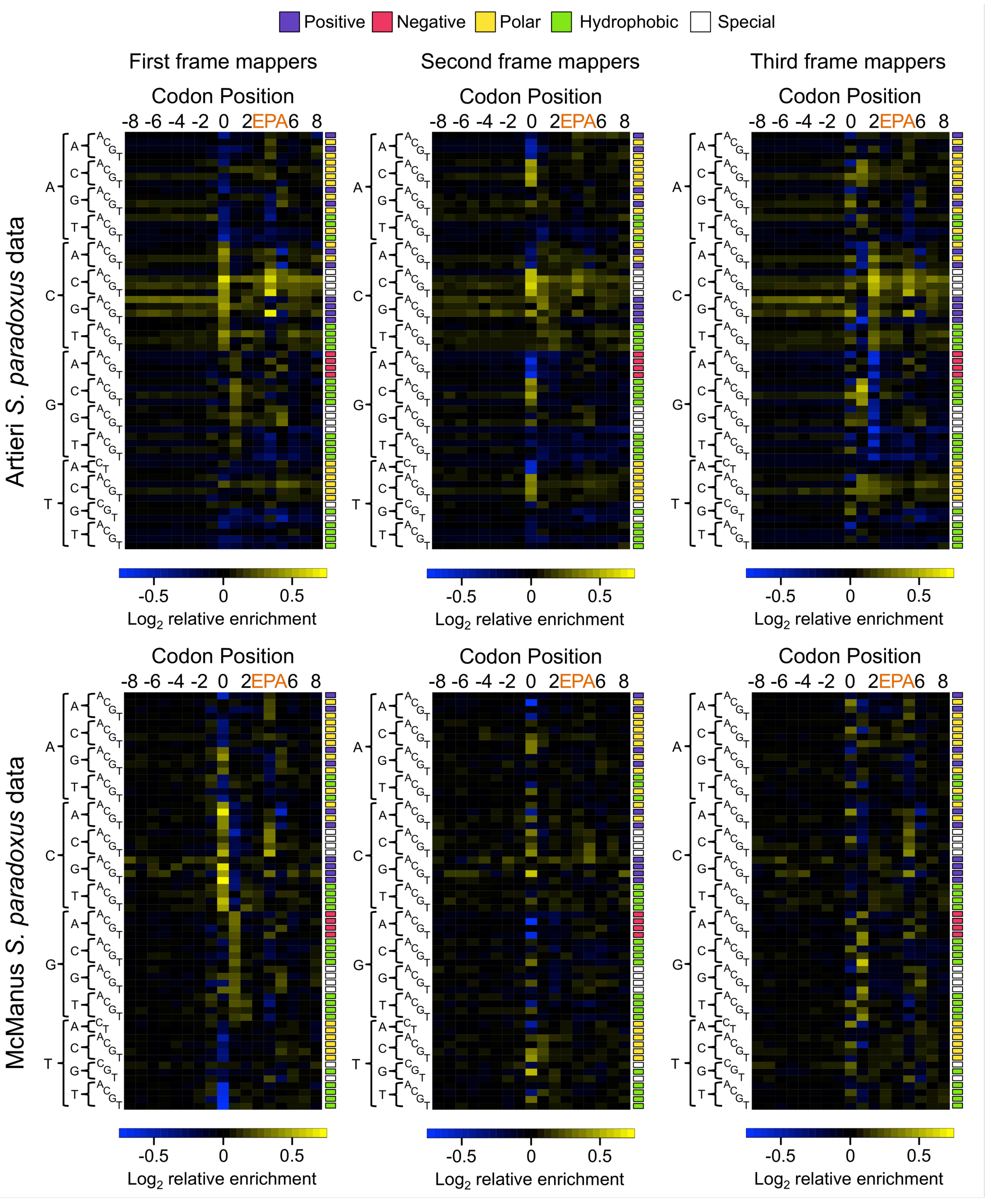
Supplemental Figure S13. Patterns of position-specific codon enrichment in the Artieri data are consistent when analyzing only the first or last 250 codons across different gene lengths. Because the 3′ coverage bias introduced by oligo-dT selection of the mRNA fraction is not shared by the Ribo fraction, this could produce artifactual patterns of enrichment if codon usage differs systematically between the 5′ and 3′ ends of the CDS. Therefore, genes of at least 500 codons in length were binned into three equally populated bins by CDS length (500 – 625, 625 – 866, and 869 – 4911 codons) and patterns of codon coverage were determined for the first or last 250 codons of each gene separately. This revealed that patterns of enrichment are consistent among 5′ and 3′ ends across all three length categories (particularly in the case of proline codon, CCN, enrichment at position 4) indicating that oligo-dT selection of mRNA did not influence our interpretation of position-specific codon enrichment. Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



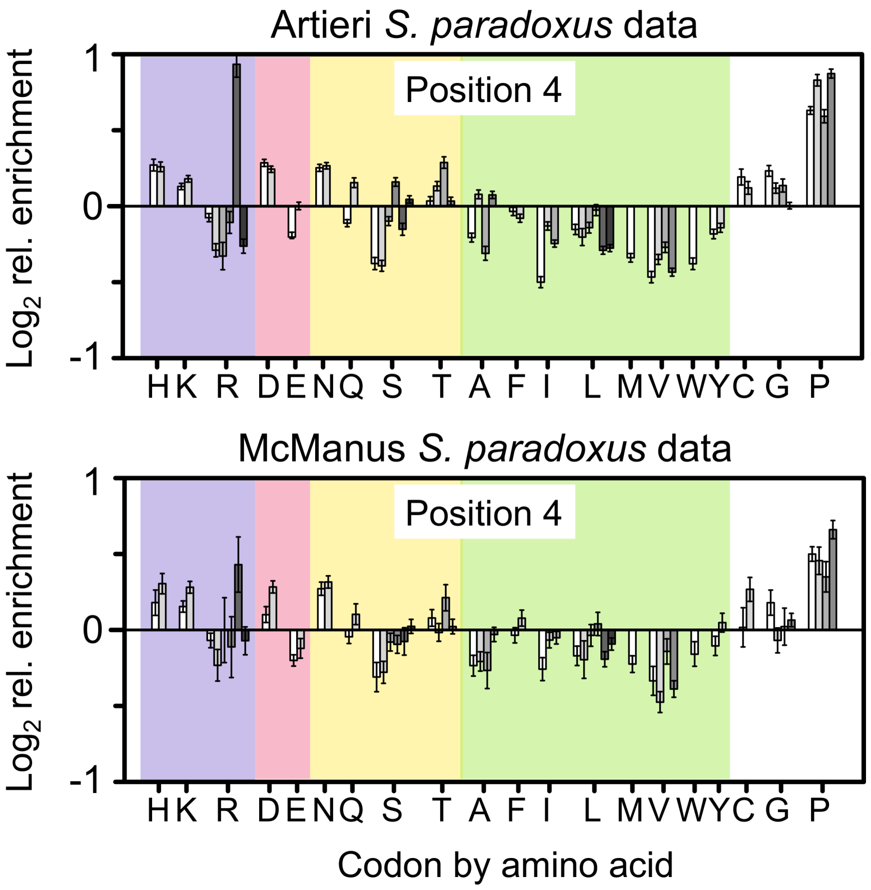
Supplemental Figure S14. Patterns of codon enrichment at position 4 are consistent across expression levels. Genes were divided into expression quartiles based on the inter-replicate mean RPKM estimated from the mRNA fraction of the Artieri data. Fewer codons are covered by reads in low expression genes leading to larger 95% confidence intervals, yet all proline codons remain enriched at all four quartiles. Results are qualitatively similar in the McManus dataset (not shown).



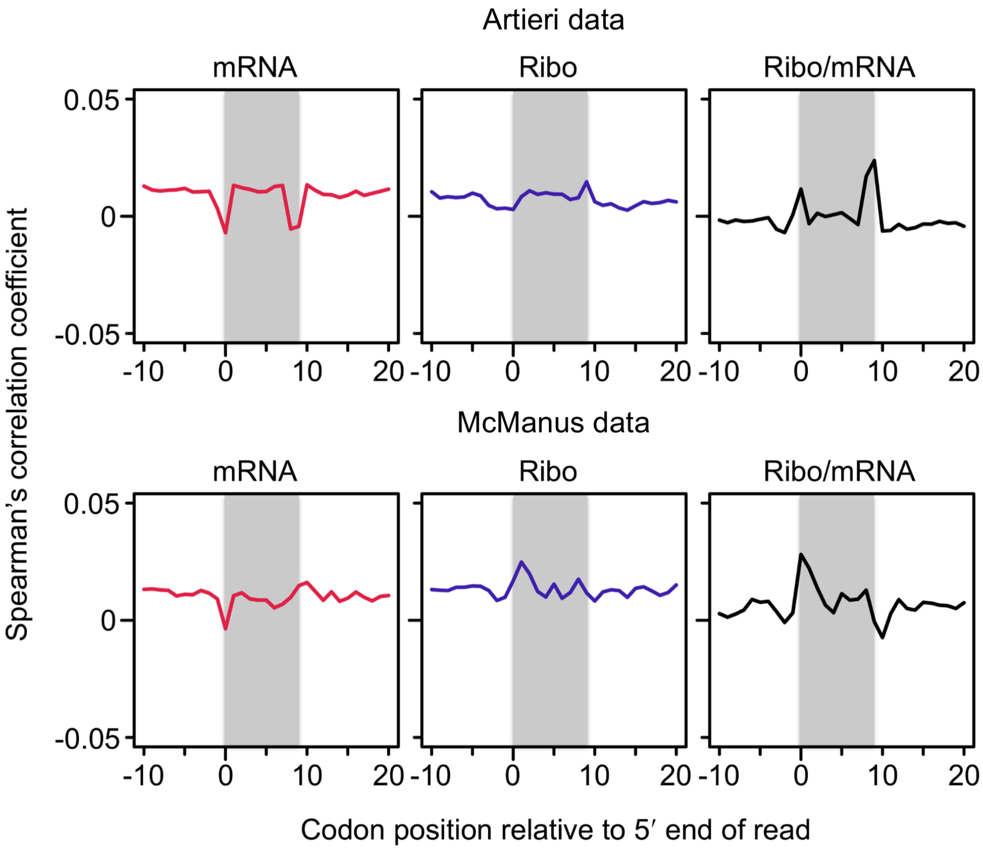
Supplemental Figure S15. Patterns of nucleotide and codon bias in the *S. paradoxus* data of Artieri and Fraser (2014) and McManus et al. (2014) are consistent with those observed in *S. cerevisiae*. Nearly identical biases are observed in the two species (compare to the *S. cerevisiae* data in Fig. 2).



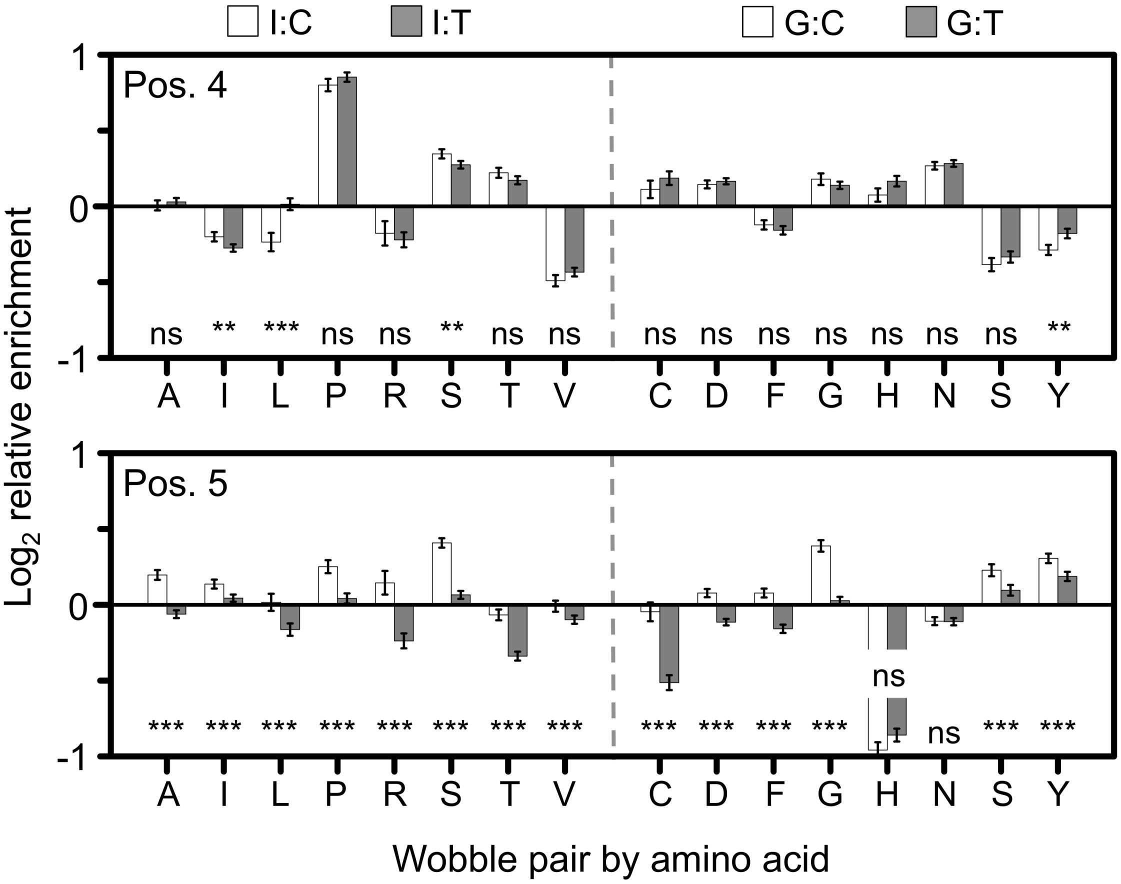
Supplemental Figure S16. Patterns of codon enrichment in the *S. paradoxus* data of Artieri and Fraser (2014) and McManus et al. (2014) are consistent with those observed in *S. cerevisiae*. Nearly identical enrichments are observed in the two species (compare to the *S. cerevisiae* data in Supplemental Fig. S9). Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



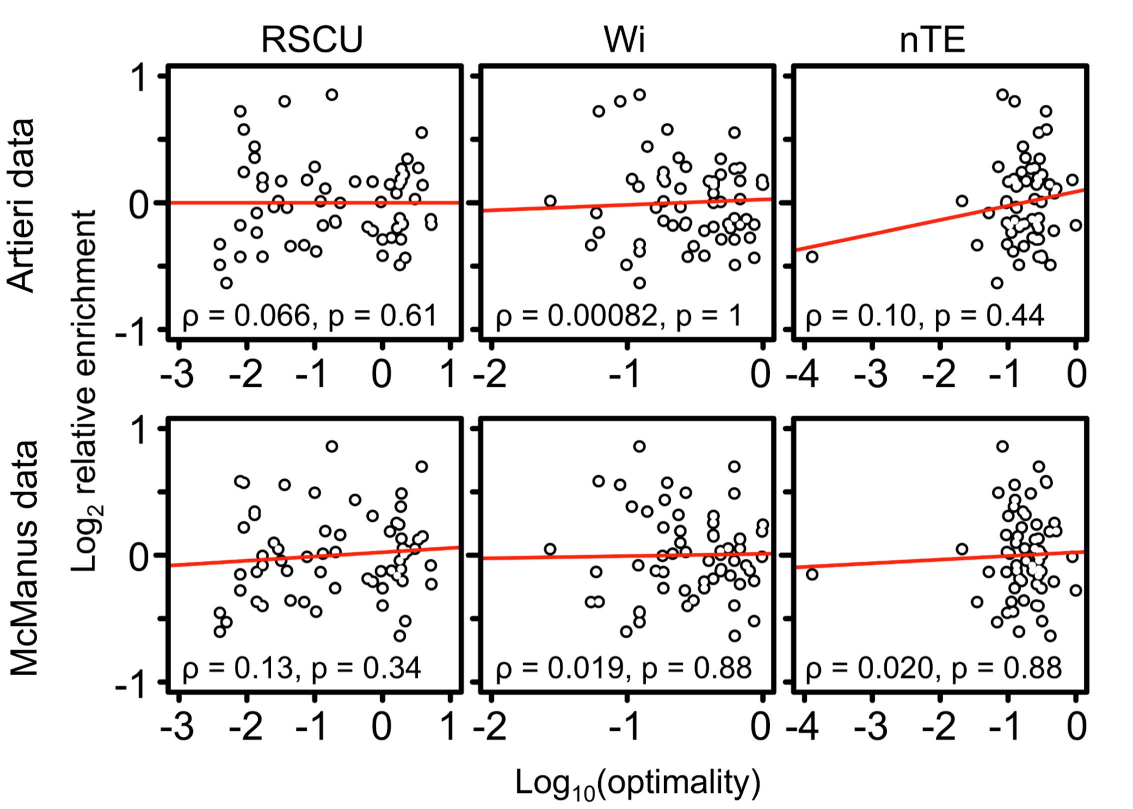
Supplemental Figure S17. Reproduction of Fig. 4B using the *S. paradoxus* data of Artieri and Fraser (2014) and McManus et al. (2014). Patterns of enrichment of enrichment at codon position 4 are consistent between the two species with regard to all four proline codons as well as the arginine codon, CGG.



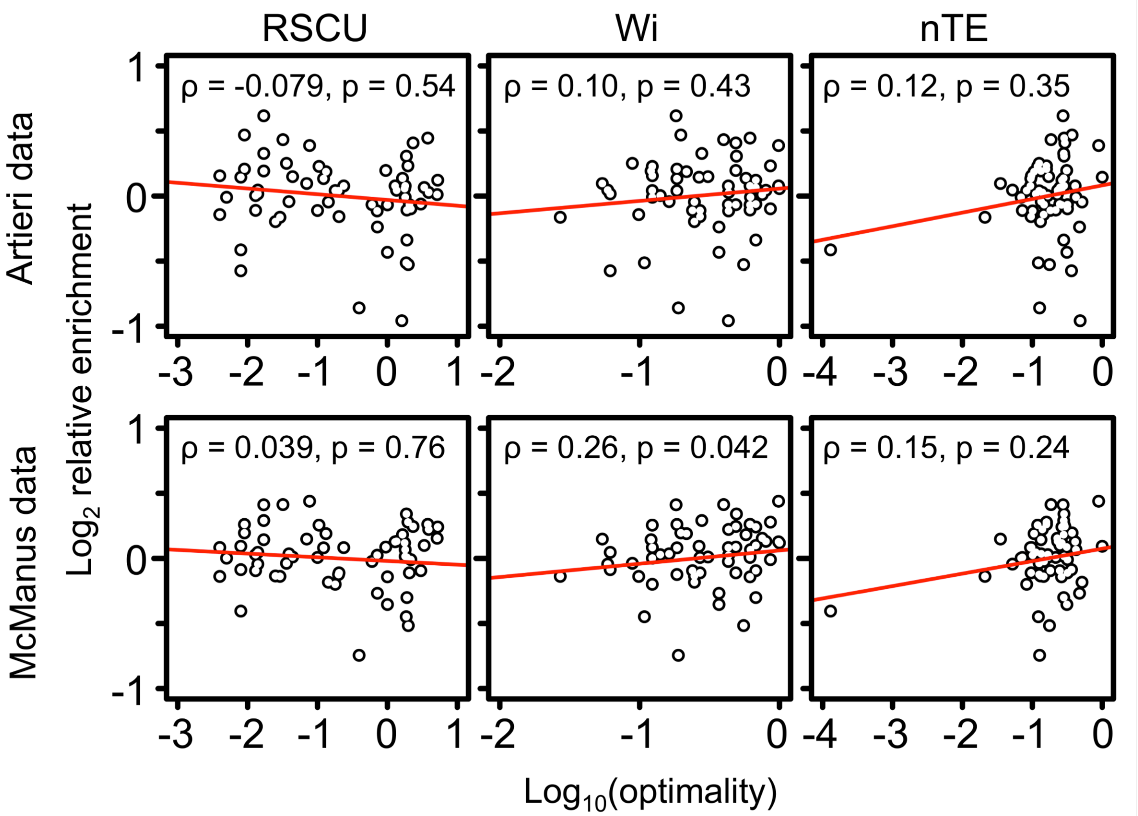
Supplemental Figure S18. Correlations between mRNA secondary structure and 5′ mapping codon coverage. Spearman correlations are shown for the mRNA and Ribo fractions, as well as the corrected Ribo coverage (Ribo/mRNA). The presence of secondary structure at each codon from 10 codons upstream to 20 codons downstream was independently correlated with occupancy at the 5′ mapping codon (position 0). The grey shading indicates the 9 codons completely overlapped by each read. The overall correlation between read occupancy and secondary structure was weak at all positions in both datasets. However, it was clear that terminal nucleotide biases, particularly 5′ over-representation of adenine, overwhelmed other signals. In the Artieri data, terminal adenine biases at both ends contributed to the positive correlations at positions 0 and 9 in the corrected Ribo coverage panel. The McManus data lacked the 3′ adenine bias contributing to an increased positive correlation only at the 5′ end of reads. This was unlikely to be caused by secondary structure hindering ribosome progression since the 5′ ends of reads represent the trailing edge of the ribosome footprint; secondary structure would be more likely to exert an effect at the leading edge.



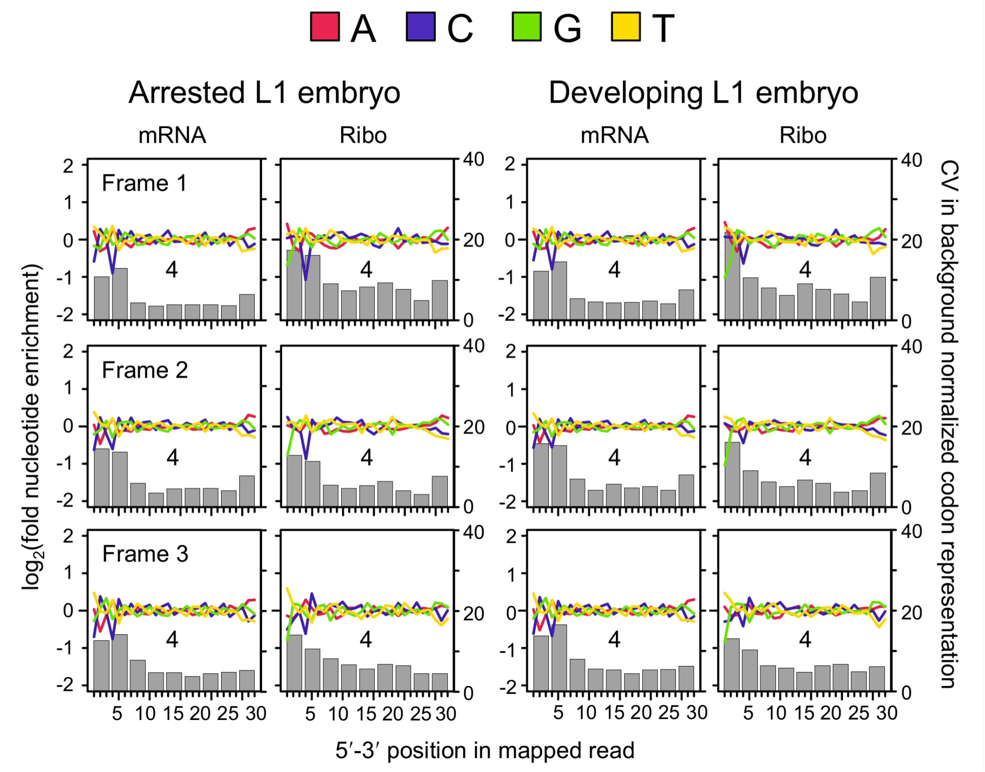
Supplemental Figure S19. Comparison of enrichment between Watson-Crick (I:C/G:C, white) and wobble pairing (I:U/G:U, grey) codon-pairs for the Artieri data. Amino acids using inosine (I, left) or guanine (G, right) wobble pairing are indicated. The significance of a Kruskal-Wallis rank sum test of the difference between enrichment of Watson-Crick and wobble codons is shown above each amino acid if p < 0.05 after Bonferroni correction for multiple tests (*, p < 0.05; **, p < 0.01; ***, p < 0.001). At position 4 (Pos. 4, top) there appeared to be an inconsistent general increase in enrichment of wobble pairing geometry, particularly in the case of the McManus data (not shown). However, at position 5 (Pos. 5, bottom) in all cases where there was a significant difference in the enrichment, Watson-Crick pairing was favored. Therefore, there did not appear to be a consistent favoring of either geometry. Furthermore, the relative enrichment differences between pairing geometries appeared to be overwhelmed by variation at individual codons themselves, irrespective of whether they use Watson-Crick or wobble pairing (e.g., Fig. 4A).



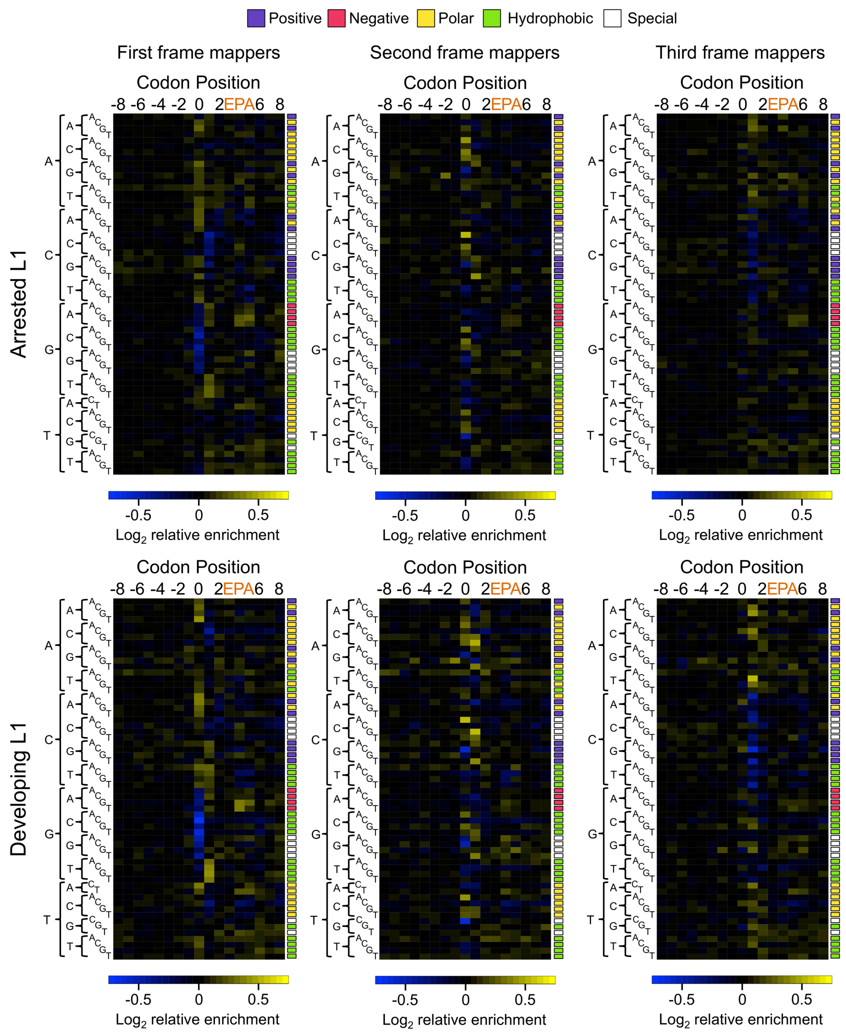
Supplemental Figure S20. No significant correlation among three different measures of codon optimality and log$_2$ corrected Ribo enrichment at codon position 4 (P-site). Spearman correlation coefficents and associated p-values are shown in each panel. A negative correlation would be expected if non-optimal codons slow ribosomes. RSCU, relative synonymous codon usage (Sharp and Li 1987); Wi, absolute adaptiveness (dos Reis et al. 2004); nTE, normalized translational efficiency scale (Pechman and Frydman 2013).



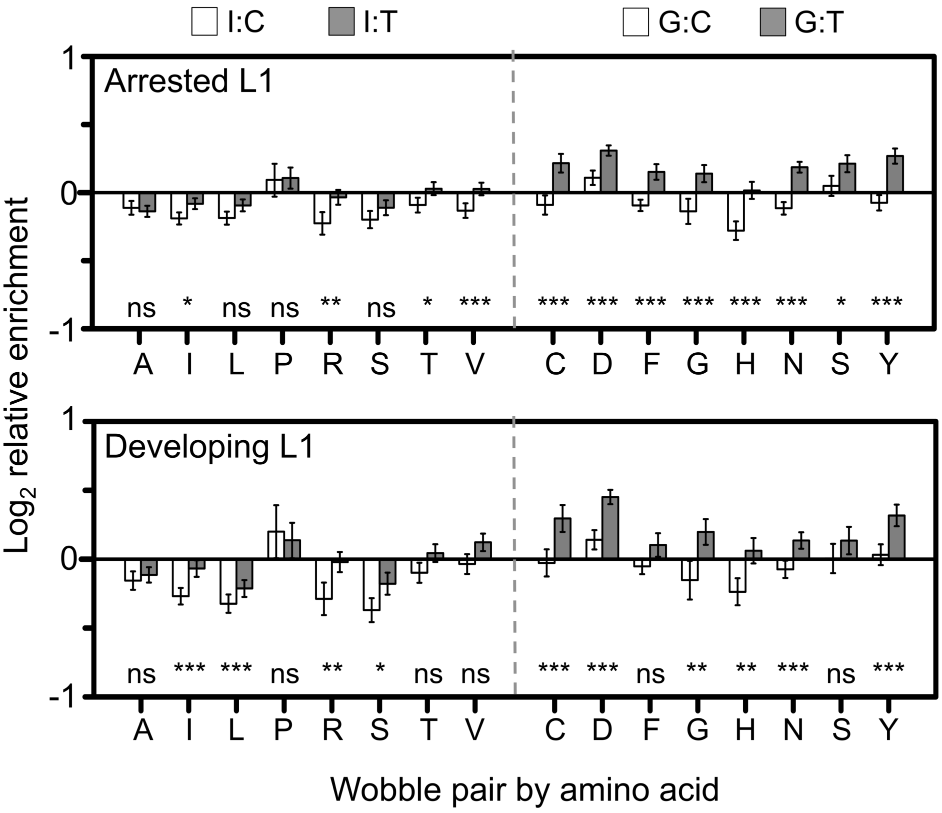
Supplemental Figure S21. No significant correlation among three different measures of codon optimality and $\log_2$ corrected Ribo enrichment at codon position 5 (A-site). Spearman correlation coefficents and associated p-values are shown in each panel. A negative correlation would be expected if non-optimal codons slow ribosomes. RSCU, relative synonymous codon usage (Sharp and Li 1987); Wi, absolute adaptiveness (dos Reis et al. 2004); nTE, normalized translational efficiency scale (Pechman and Frydman 2013). Note that the correlation between relative enrichment and Wi for the McManus data (p = 0.042) is not significant after correction for multiple tests.



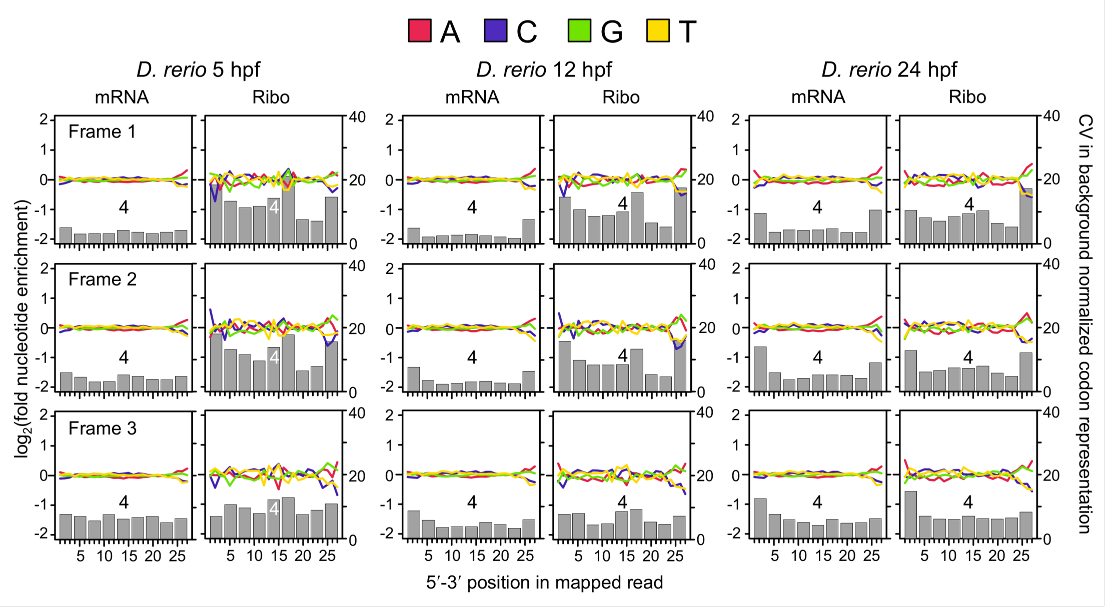
Supplemental Figure S22. Patterns of nucleotide and codon bias among the first 27 nt of reads in the *C. elegans* data of Stadler and Fire (2013). Replicates from each of the two profiled developmental stages (Arrested or Developing L1 embryos) were combined for analysis. As the nematode data was generated using similar methods to the yeast datasets, patterns of 5′ bias are consistent, including a bias towards adenines in the first position of first frame mappers as well as a paucity of cytosines at position 4. Though a slight signal of elevated codon bias at positions four and five is observed in first frame mappers of the Ribo fraction, it is overwhelmed by the biases observed a the 5′ end. However, the weak pattern of enrichment at internal positions may reflect a reduction of power as per-base coverage is ~20-fold lower than the Artieri yeast data (see also Supplemental Fig. S23).



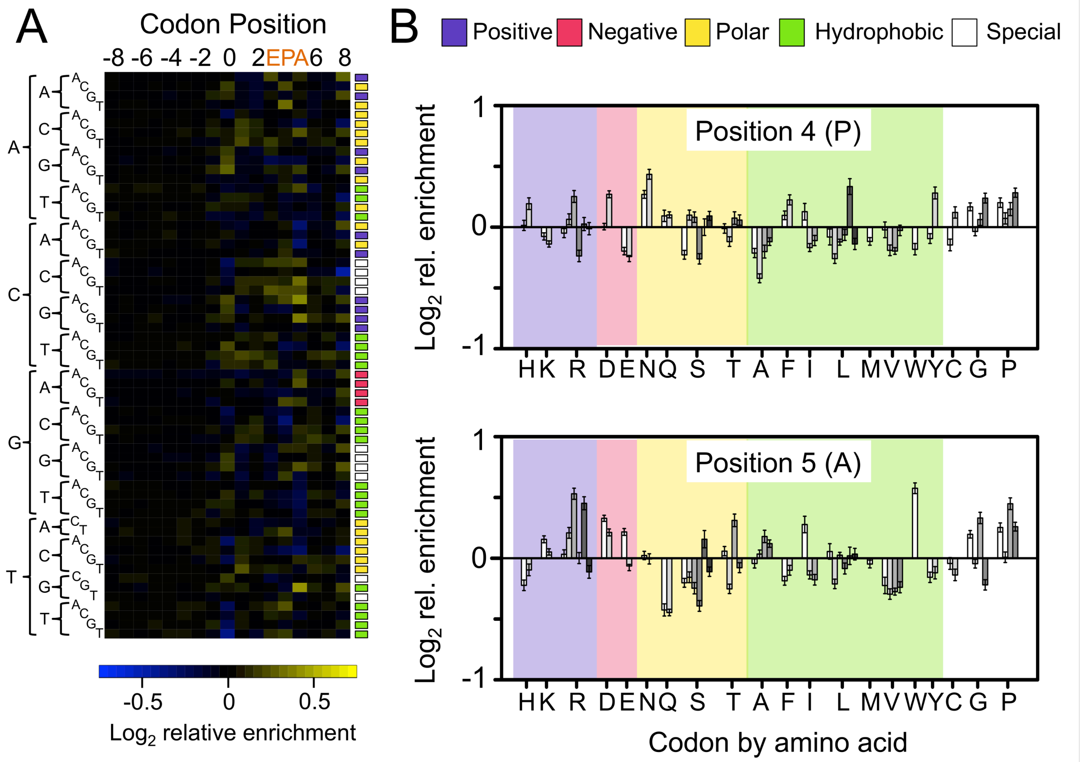
Supplemental Figure S23. Patterns of codon enrichment in the *C. elegans* data of Stadler and Fire (2013). Biases at position 0, corresponding to the 5′ end of reads, overwhelm all other positions in all three reading frames in both profiled developmental stages. Enrichment/depletion at all other positions is relatively weak, possibly due to the 20-fold lower per-base coverage than the Artieri data.



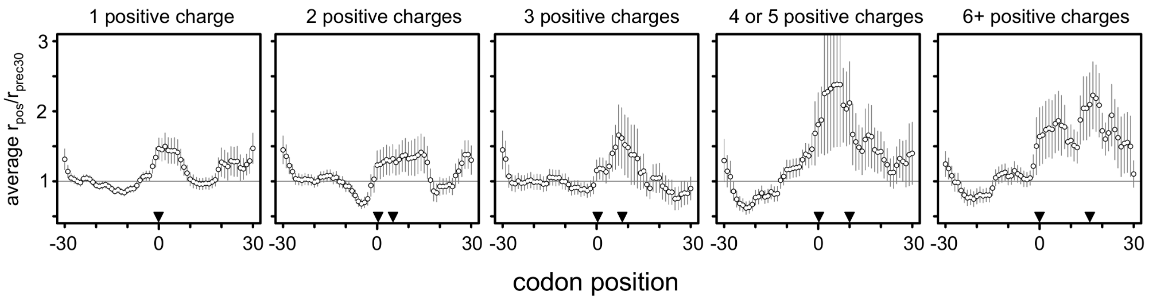
Supplemental Figure S24. Comparison of enrichment between Watson-Crick (I:C/G:C, white) and wobble pairing (I:U/G:U, grey) codon-pairs at position 4 (P-site) for the *C. elegans* data of Stadler and Fire (2013). Amino acids using inosine (I, left) or guanine (G, right) wobble pairing are indicated. The significance of a Kruskal-Wallis rank sum test of the difference between enrichment of Watson-Crick and wobble codons is shown above each amino acid if p < 0.05 after Bonferroni correction for multiple tests (*, p < 0.05; **, p < 0.01; ***, p < 0.001). First frame mapping reads from the combined replicates are shown for the Arrested L1 data (top) and Developing L1 data (bottom). In all cases where a significant difference was observed between the cognate- and wobble-pairing codons coding for the same amino acid, the corrected Ribo coverage was higher for the wobble-pairing codon, supporting the observations of Stadler and Fire (2011).



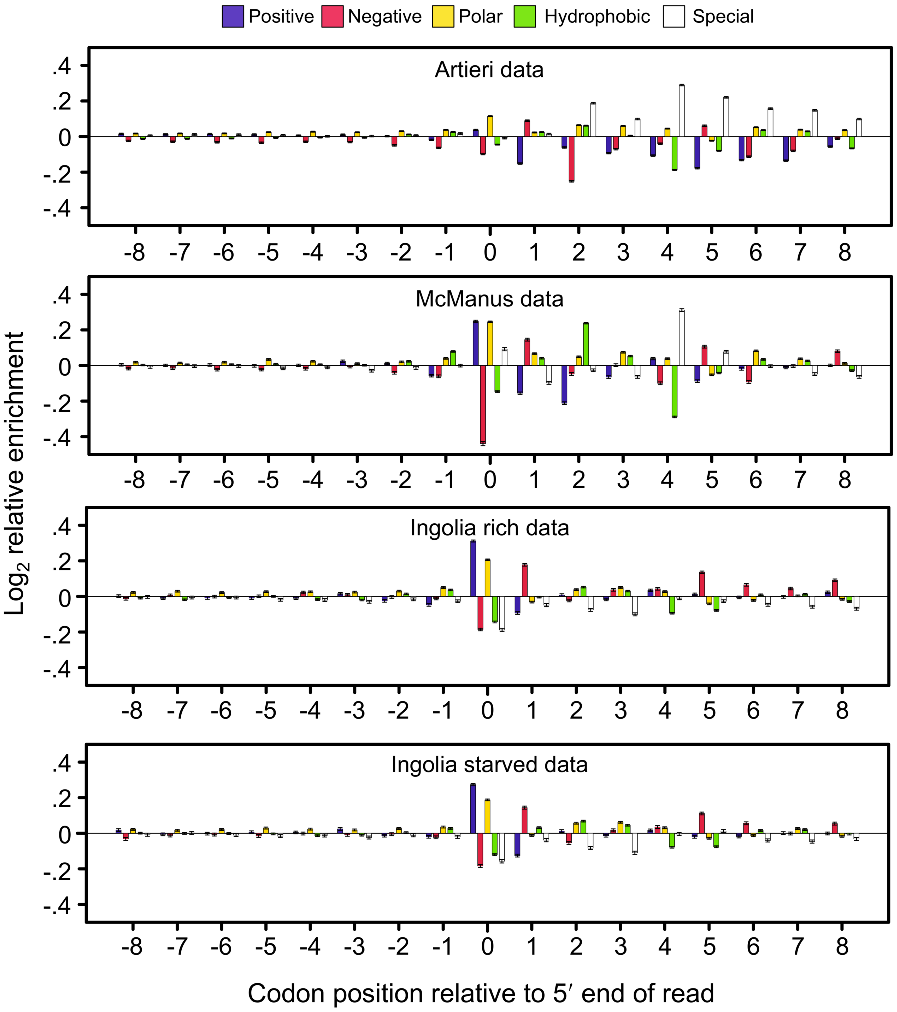
Supplemental Figure S25. Patterns of nucleotide and codon bias among the first 27 nt of reads in the *D. rerio* data of Bazzini et al. (2014). Replicates from the three (of five) profiled stages with the highest number of mapping reads were pooled for analysis. While the 5′ and 3′ most nucleotides and codons of most samples show substantial biases, the specifics of enrichment and depletion in the *D. rerio* data differ from those observed in the other datasets, likely reflecting Bazzini et al.'s (2014) use of the ARTseq riboprofiling library preparation kit (Epicenter). Unlike in the other samples, which were all generated with variants of the protocol developed by Ingolia et al. (2010), patterns of enrichment/depletion at the ends of reads generated using ARTseq appear to be mostly fraction-specific (though the 24 hpf [hours post-fertilization] samples appear to show some share 5′ enrichment). Nevertheless, patterns of internal codon bias appear to be restricted to the Ribo fraction as in the other species' data. Unlike in the yeast datasets, the position of highest internal bias appears to be codon 5, which may reflect differences between species in the size of the ribosome-protected fragment and/or the specific positioning of the ribosomal active sites (Stadler and Fire 2011).



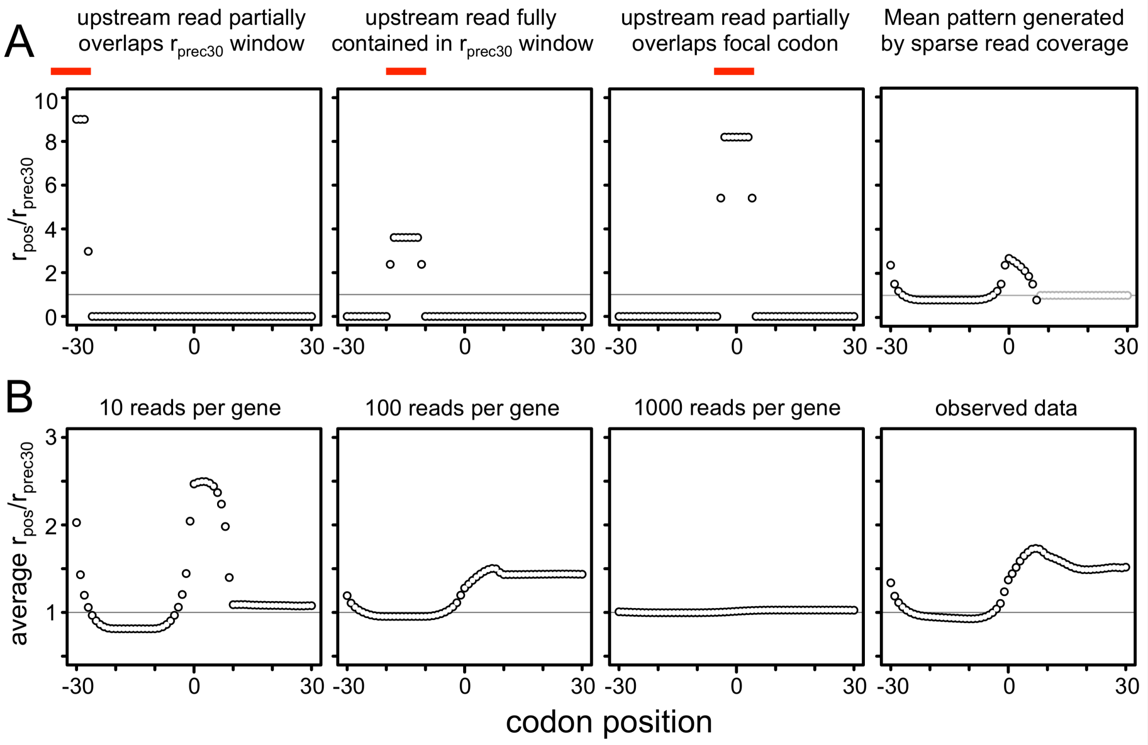
Supplemental Figure S26. The pooled *D. rerio* data of Bazzini et al. (2014) shows a signal of enrichment of proline codons at position 4 (P-site). We pooled all of the mRNA and Ribo fraction replicates separately in order to maximize our power to detect patterns position-specific codon enrichment in the genome of zebrafish (see Supplemental Methods). Specific patterns differ among most individual codons between yeast and *D. rerio*, possibly reflecting biological differences in the factors controlling the rate of translation. Nevertheless, similar to yeast, all four proline (P) codons show enrichment above mean levels at position 4 (P-site). Furthermore, three of the four codons also show enrichment above mean levels at position 5, corresponding to the codon position showing the strongest deviation from expected codon frequencies (Supplemental Figure S25). As in yeast, we detect very little codon enrichment upstream, and no evidence of enrichment of positive amino acid encoding codons.



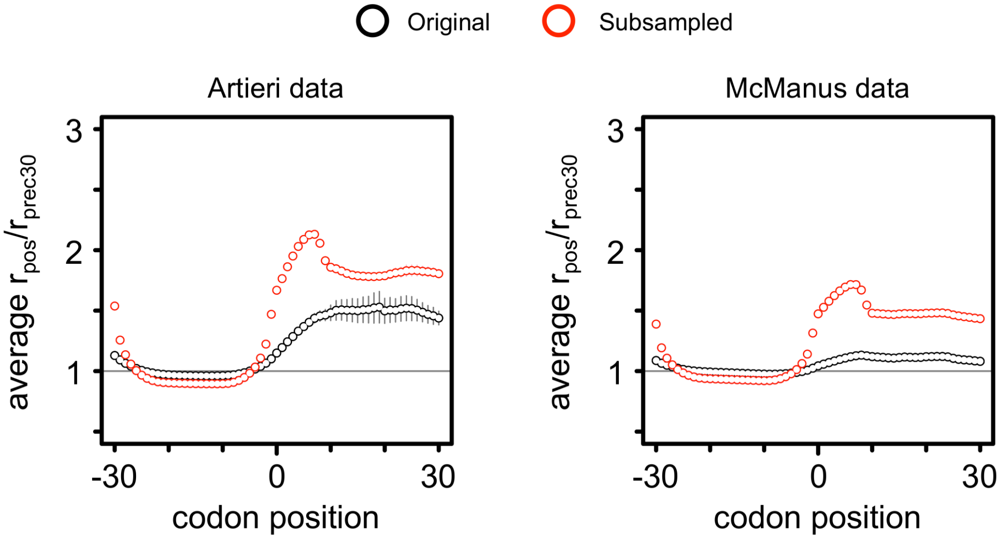
Supplemental Figure S27. Reproduction of the additive stalling effect observed in Figure 5 of Charneski and Hurst (2013) confirming that the same analysis method was used in the present study.



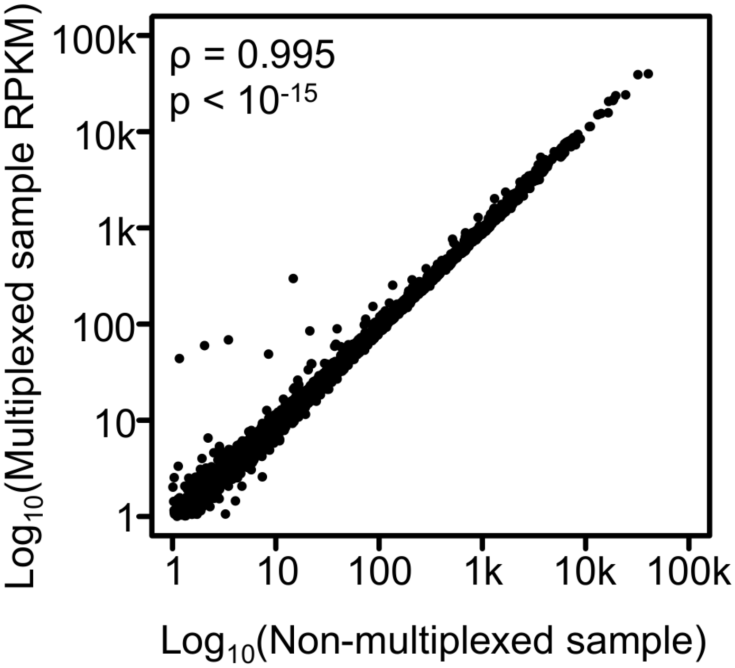
Supplemental Figure S28. Positive amino acids are not enriched among upstream codons in the uncorrected Ribo fractions in any of the datasets, as would be expected if ribosomes were slowed as these codons passed through the exit tunnel. Enrichment was determined at the level of biochemical class using the Ribo fraction alone without correcting by the mRNA fraction (see Supplemental Methods). Reads from all read lengths and reading frames were used as in the analysis of Charneski and Hurst (2013). Error bars indicate the standard error of the mean. Biochemical classes are indicated above. Positive amino acids were not enriched in positions -8 to -1 in any dataset. However, the strongest levels of enrichment in both Ingolia datasets are among positive amino acid encoding codons at position 0, which could lead to enrichment at and downstream of the focal codon used by Charneski and Hurst (Fig. 5; see above). The McManus dataset also showed enrichment among positive amino acids at position 0; however, this is not as strong as the enrichment of special codons at position 4, of which proline is a member.



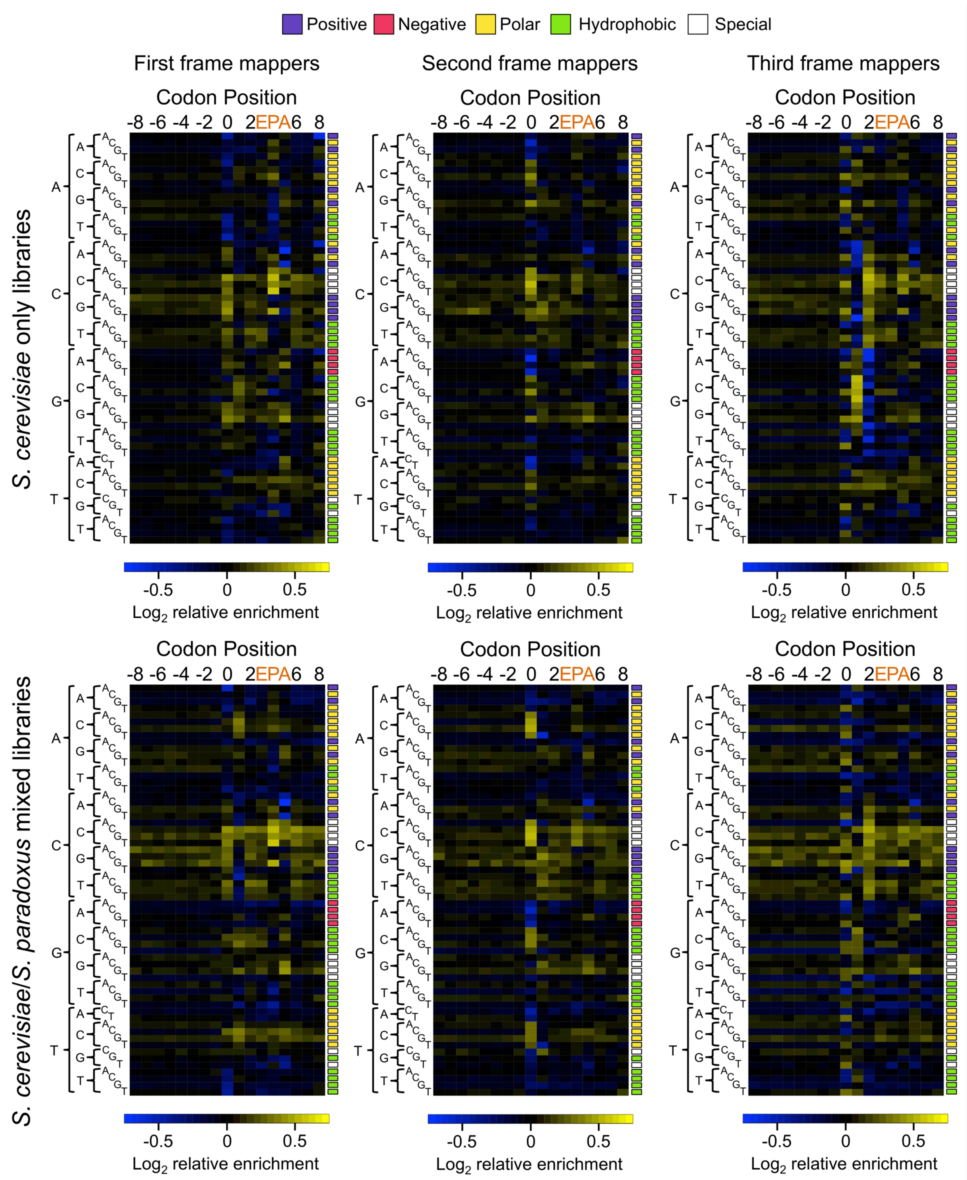
Supplemental Figure S29. The r_{pos}/r_{prec30} method of Charneski and Hurst (2013) is biased towards producing false signals of stalling when read coverage is sparse. (A) When only a single read overlaps the 61 codon window of investigation, its position (indicated by the red bar) influences its contribution to the overall mean r_{pos}/r_{prec30} value. Note that codon level coverage was calculated as the average coverage of its three nucleotides, allowing codons overlapping the ends of reads to have fractional coverage. Averaging over all possible 28 nt single read positions produced a characteristic 'saddle' pattern of stalling in the absence of any such an effect. Note that codon positions 8 to 30 are greyed out as windows containing single reads spanning them will produce r_{prec30} values of 0 leaving their r_{pos}/r_{prec30} values undefined. (B) Generating randomly positioned reads of length randomly chosen between 27 and 30 nt and averaging over all possible 61 codon windows produced signals of strong stalling that only disappeared at high coverage. Each gene was assigned the indicated number of reads. The observed data also showed an overall mean pattern of stalling when averaging over all available positions. Note that the observed data have a mean coverage of ~260 reads per gene; however, this coverage is not evenly distributed as in the simulated data, and therefore does not show a pattern intermediate between that of 100 and 1000 reads per gene.



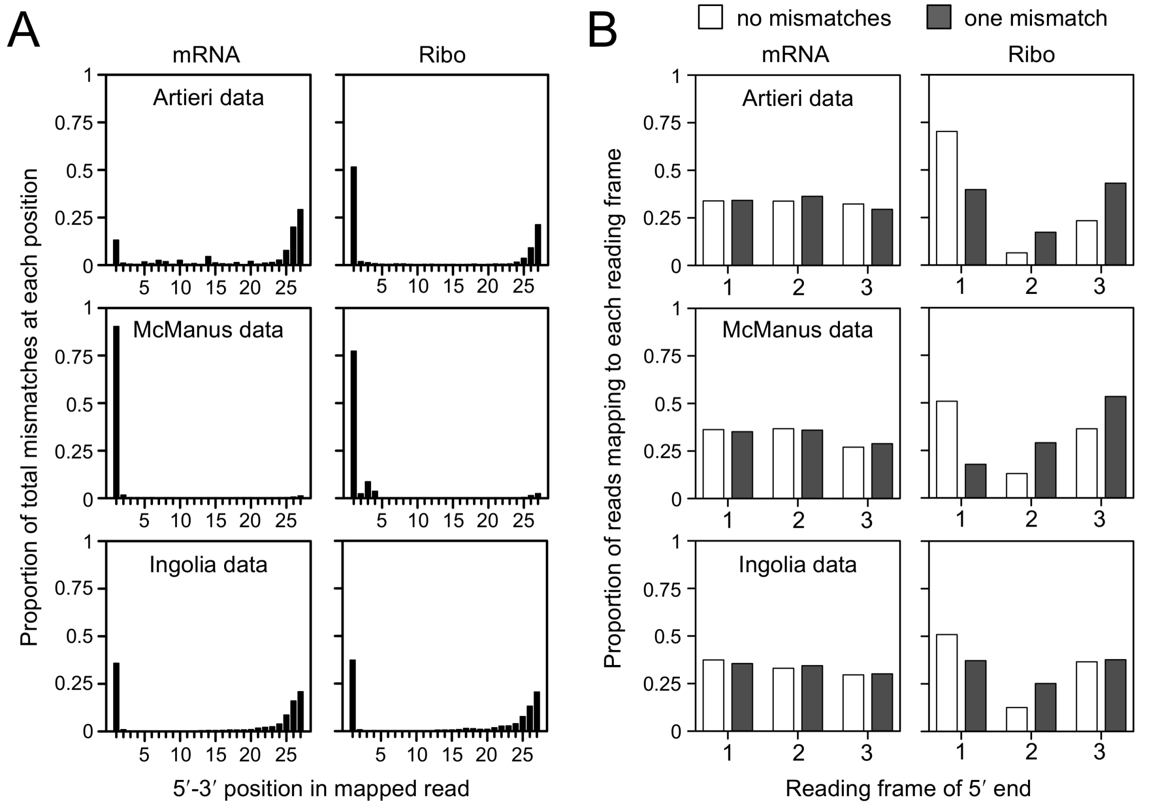
Supplemental Figure S30. Downsampling the Artieri or the McManus data increases the 'stalling' effect detected. Mapped reads from both replicates of the two datasets were randomly sampled down to the mean number of reads mapping among replicates of the Ingolia rich data. The average r_{pos}/r_{prec30} was determined across all 61 codon windows. In order to compare the same data directly, only sites that had mapping reads in both the original data and the subsampled data were used for plotting (black, original data; red, subsampled data). The average stalling pattern increases substantially in both datasets, highlighting the sensitivity of the method to sparse coverage.



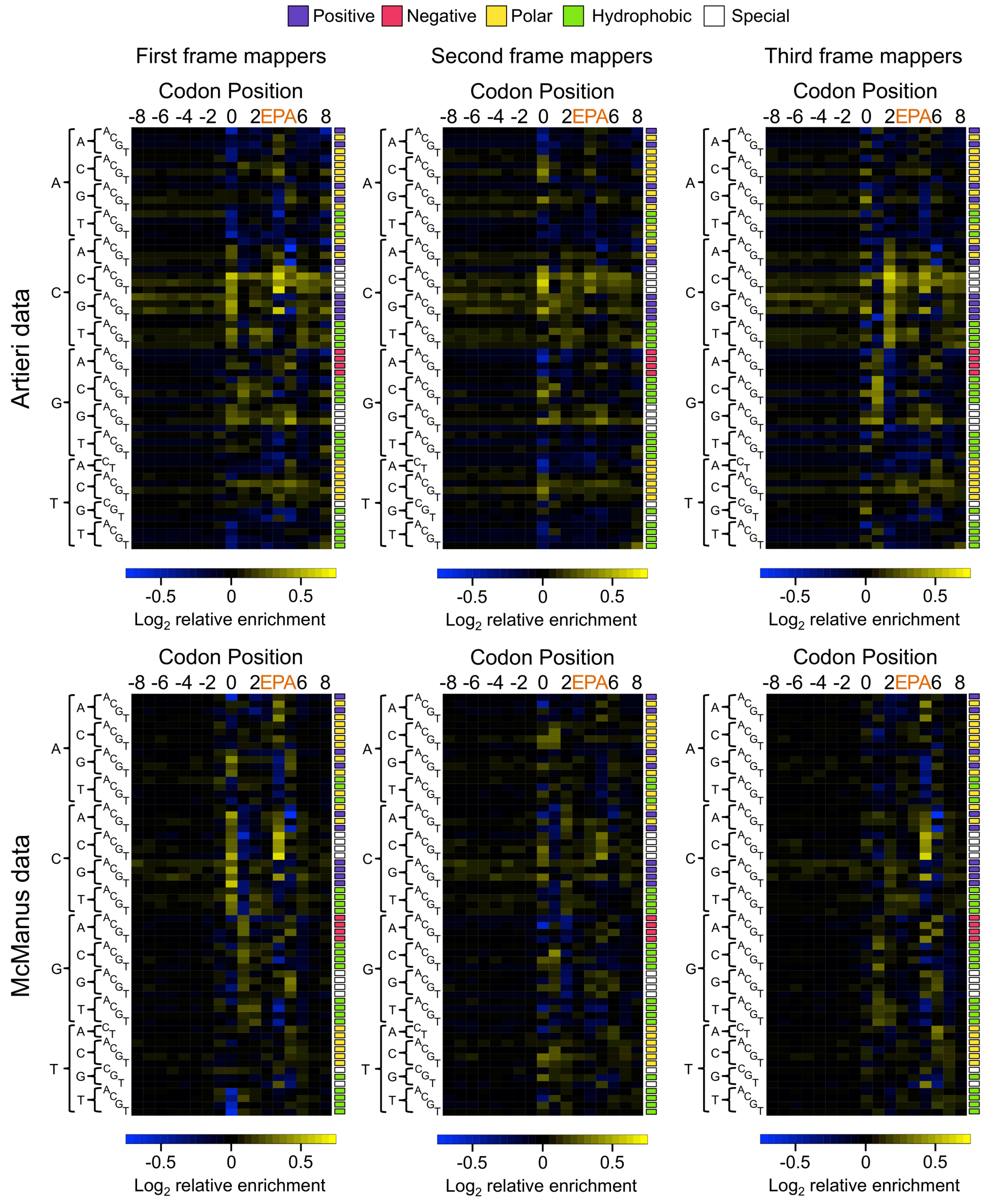
Supplemental Figure S31. Our mapping approach identified *S. cerevisiae* reads in the mixed high-coverage data of Artieri and Fraser (2014). The high correlation between expression levels (in RPKM) of *S. cerevisiae* genes in the *S. cerevisiae* + *S. paradoxus* mixed Ribo fraction sample and the *S. cerevisiae* only sample indicated that our mapping method robustly identified *S. cerevisiae*-specific reads. Only a small number of genes (< 10) show signs of higher expression in the mixed sample, indicating misallocation of reads. This could not have been a source of bias in our analysis given the high concordance between replicates, where the first replicate was not mixed, but the second was (Supplemental Fig. S5; Supplemental Table S2).



Supplemental Figure S32. Comparison of *S. cerevisiae* only and mixed *S. cerevisiae*/*S. paradoxus* Artieri data libraries shows no evidence of bias in mixed libraries. Analysis was performed by pairing either unmixed (Ribo fraction replicate 1/mRNA fraction replicate 2) or mixed (Ribo fraction replicate 2/mRNA fraction replicate 1) libraries (Supplemental Table S3). Patterns of enrichment appeared qualitatively similar in both cases, especially with respect to codon position 4. However, we do note that the magnitude of the patterns appears to be higher in the mixed libraries, which may reflect the lower coverage of the mixed libraries (Supplemental Table S1) and/or the differences in the degree of 5′ bias seen between the two replicate Ribo and mRNA fractions (Supplemental Fig. S5). Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.



Supplemental Figure S33. Allowing mismatches during mapping weakens the signal of periodicity in Ribo fraction reads. (A) The majority of mismatches occur at the 5′ most nucleotide in most libraries, including all Ribo fractions, consistent with the known property of reverse transcriptase to add untemplated bases the 5′ end during cDNA synthesis (Zajac et al. 2013). (B) No systematic differences between non-mismatch (white) and mismatch (grey) mapping reads are observed in the proportion of reads whose 5′ ends map to each reading frame in mRNA fractions. However, in Ribo fractions mismatch mapping reads no longer show the periodicity observed in non-mismatch mappers (first frame > third frame > second frame). Whereas the periodicity of reads mapping without mismatches is consistent among datasets, that observed among reads mapping with a single mismatch is not.



Supplemental Figure S34. Reproduction of Supplemental Fig. S9 using reads mapped allowing one mismatch. Patterns of enrichment remained qualitatively unchanged, indicating that our stringent mapping approach did not bias the analysis towards particular patterns of read coverage. Note that in the case of the Artieri data, we analyzed only Ribo fraction replicate 1 and mRNA replicate 2 as these were not mixed with *S. paradoxus* in the previous analysis of Artieri and Fraser (2014), and therefore could confidently be aligned to the *S. cerevisiae* genome despite mismatches (see Supplemental Methods). Codons associated with the E, P, and A active sites of the ribosome (3,4, and 5, respectively) are indicated.