Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level

Siyu Chen,[1,2,3] Ke Li,[1,2] Wenqing Cao,[1,2,3] Jia Wang,[1,2,3,4] Tong Zhao,[5] Qing Huan,[1,2] Yu-Fei Yang,[1,2,3] Shaohuan Wu,[1,2,3] and Wenfeng Qian*[,1,2,3,4]

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Sino-Danish Center for Education and Research, Beijing, China

[5]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

***Corresponding author:** E-mail: wfqian@genetics.ac.cn.

Associate editor: Deepa Agashe

Abstract

Codon usage bias (CUB) refers to the observation that synonymous codons are not used equally frequently in a genome. CUB is stronger in more highly expressed genes, a phenomenon commonly explained by stronger natural selection on translational accuracy and/or efficiency among these genes. Nevertheless, this phenomenon could also occur if CUB regulates gene expression at the mRNA level, a hypothesis that has not been tested until recently. Here, we attempt to quantify the impact of synonymous mutations on mRNA level in yeast using 3,556 synonymous variants of a heterologous gene encoding green fluorescent protein (*GFP*) and 523 synonymous variants of an endogenous gene *TDH3*. We found that mRNA level was positively correlated with CUB among these synonymous variants, demonstrating a direct role of CUB in regulating transcript concentration, likely via regulating mRNA degradation rate, as our additional experiments suggested. More importantly, we quantified the effects of individual synonymous mutations on mRNA level and found them dependent on 1) CUB and 2) mRNA secondary structure, both in proximal sequence contexts. Our study reveals the pleiotropic effects of synonymous codon usage and provides an additional explanation for the well-known correlation between CUB and gene expression level.

Key words: **codon usage bias, synonymous mutations, context-dependent effect, mRNA secondary structure, yeast.**

Article

Introduction

Eighteen of the 20 amino acids are each encoded by two to six synonymous codons, but these synonymous codons are not used with equal frequencies in a genome (Ikemura 1985; Sharp et al. 1988; Hershberg and Petrov 2009). This phenomenon is referred to as codon usage bias (CUB) (Ikemura 1985; Duret 2002; Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). Synonymous codons that are preferentially used in highly expressed genes in a genome usually have high concentrations of corresponding isoaccepting tRNAs (Ikemura 1981, 1982, 1985; Gouy and Gautier 1982; Moriyama and Powell 1997; Duret and Mouchiroud 1999; Duret 2000; Kanaya et al. 2001), and they are known as preferred codons. Other synonymous codons are called unpreferred codons. Synonymous mutations often lead to reduced protein level and fitness (Carlini and Stephan 2003; Carlini 2004; Agashe et al. 2013; Lampson et al. 2013), suggesting the vital role of synonymous codon usage in gene expression and adaptation.

The correlation between CUB and gene expression level cannot be explained by mutation and drift, and is usually attributed to selection, although the nature of such selection has not been fully understood (Bulmer 1991; Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). Because preferred codons are usually recognized by high-concentration isoaccepting tRNAs in translation (Ikemura 1981, 1982, 1985; Moriyama and Powell 1997; Duret 2000; Kanaya et al. 2001), it has been assumed that the phenomenon is attributable to natural selection on translation rather than on mRNA level (Bulmer 1991; Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). Two mutually nonexclusive hypotheses have been proposed. The first, termed the translational accuracy hypothesis, asserts that the probability of incorporating near/noncognate tRNAs is reduced for preferred codons, so translational accuracy of preferred codons is higher. Because translational errors can lead to the synthesis of nonfunctional or mis-folded proteins which are wasteful or toxic, increased translational

Open Access

accuracy is beneficial (Precup and Parker 1987; Akashi 1994; Drummond and Wilke 2008; Stoletzki and Eyre-Walker 2007; Zhou et al. 2009). The second hypothesis, termed the translational efficiency hypothesis, asserts that preferred codons have higher translational elongation rates (Varenne et al. 1984), which reduces the time for synthesizing a polypeptide and spares ribosomes for other genes. Although the relationship between codon optimality and translational elongation rate was under debate (Curran and Yarus 1989; Sorensen et al. 1989; Ingolia et al. 2011; Li et al. 2012; Qian et al. 2012; Charneski and Hurst 2013; Gardin et al. 2014; Pop et al. 2014; Yang et al. 2014), the results of recent studies (Hussmann et al. 2015; Weinberg et al. 2016) suggest that the earlier inconsistent observations are likely due to the use of the eukaryotic translation inhibitor cycloheximide, and optimal codons are indeed decoded faster in flash-frozen experiments. Consistently, based on approaches independent of ribosome profiling, it was also observed that preferred codons are decoded faster (Yu et al. 2015). Since ribosome is a limited resource in a cell (Warner 1999), the elevated elongation rate benefits the organism by promoting the translational initiation rates of all genes (Andersson and Kurland 1990; Bulmer 1991; Akashi 2001; Kudla et al. 2009). Regardless of the relative importance of translational accuracy and translational efficiency, both impacts scale with expression level, so that natural selection on synonymous codon usage is stronger in more highly expressed genes, leading to a positive correlation between CUB and mRNA level.

Current theories on the origin of CUB focus on the role of mRNA level in determining the strength of natural selection on synonymous codon usage (Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011), while the opposite possibility that synonymous codon usage influences mRNA level has not been examined until recently. In 2015, a correlation between mRNA degradation rate and CUB among yeast genes was reported (Presnyak et al. 2015). Further, the impact of synonymous codon usage on mRNA degradation rate was validated with a small number of synonymous variants of the same gene (Presnyak et al. 2015; Bazzini et al. 2016; Mishima and Tomari 2016). Recently, it was also reported that synonymous codon usage affected mRNA concentration through transcription rate (Newman et al. 2016; Zhou et al. 2016). In addition, the impact of codon usage on protein level was examined in *Escherichia coli* with regression models (Boel et al. 2016). However, because the numbers of synonymous variants of the same gene were small in these studies, these authors could not quantify the effects of individual synonymous mutations through a direct comparison between a pair of synonymous variants that are different by only one synonymous codon. More importantly, it remains unclear whether the effect of a synonymous mutation is dependent on its sequence contexts. In this study, we generated libraries of hundreds or thousands of synonymous variants of the same gene in the budding yeast *Saccharomyces cerevisiae* and attempted to quantify the effects of individual synonymous mutations.

Results and Discussion

High-Throughput Quantification of the mRNA Levels of *GFP* Synonymous Variants

To systematically examine the effect of synonymous codon usage on mRNA level, we generated synonymous variants in a 12-amino-acid region (amino acid 41–52, LTLKFICTTGKL) of GFP (fig. 1*A* and supplementary fig. S1*A*, Supplementary Material online), by synthesizing DNA oligos with doped nucleotides added to the third nucleotide site of each codon (TTR-ACN-TTR-AAR-TTY-ATY-TGY-ACN-ACN-GGN-AAR-TTR). Twelve is a reasonable length given the constraint on the total length of nucleotides (80) in synthesizing high-quality degenerate oligonucleotides (~20 fixed nucleotides on each end for PCR amplification + 36 nucleotides in the variable region; supplementary table S1, Supplementary Material online). We obtained full length *GFP* variants with fusion PCR and transformed these variants into a haploid yeast strain to replace the coding sequence of *GAL1*. In this haploid yeast strain, the coding sequence of *GAL7* has been replaced by *dTomato*, a gene encoding red fluorescent protein. Thus, the *GFP* variants are expressed from the chromosomal DNA under the *GAL1* promoter and *dTomato* are expressed under the *GAL7* promoter (fig. 1*A*). Because both promoters are regulated by the transcription factor GAL4 upon the induction with 2% galactose, *dTomato* can be used to normalize the expression level of *GFP* to control for cell-to-cell variation in cell size and the level of galactose induction. A total of 1,124 yeast strains of *GFP* synonymous variants were generated in this region.

We pooled the yeast strains of *GFP* variants, induced the expression of *GFP* with 2% galactose, and harvested the cells in mid-log phase (fig. 1*B*). To obtain the mRNA level of these variants, we extracted total RNA from the harvested cells, performed reverse transcription to obtain cDNA, PCR-amplified the variable region of *GFP* from the cDNA, and used Illumina sequencing to quantify the relative frequencies of *GFP* variants in the harvested cells (*R*). To control the impact of cell number variation among yeast strains with different *GFP* sequences and the potential bias in Illumina sequencing on the quantification of mRNA frequencies, we further PCR amplified the variable region of *GFP* from the genomic DNA, performed Illumina sequencing, and calculated the relative genomic content of each *GFP* variant (*D*). The *R*/*D* ratio of each *GFP* variant reflects the average mRNA level of a *GFP* variant among cells (fig. 1*B* and supplementary fig. S1*B*, Supplementary Material online). We performed two biological replicates by independently inducing the expression of *GFP* variants and observed that *R*, *D*, and mRNA level were all highly correlated between two replicates ($r = 0.99$, 0.97, and 0.83, respectively, $P < 10^{-100}$ for all three Pearson's correlations; fig. 1*C*–*E*).

The Impact of Synonymous Codon Usage on mRNA Level

To investigate the genetic basis of the variation in mRNA level among *GFP* variants (fig. 1*E*), we first counted the number of preferred codons (see Materials and Methods) in the

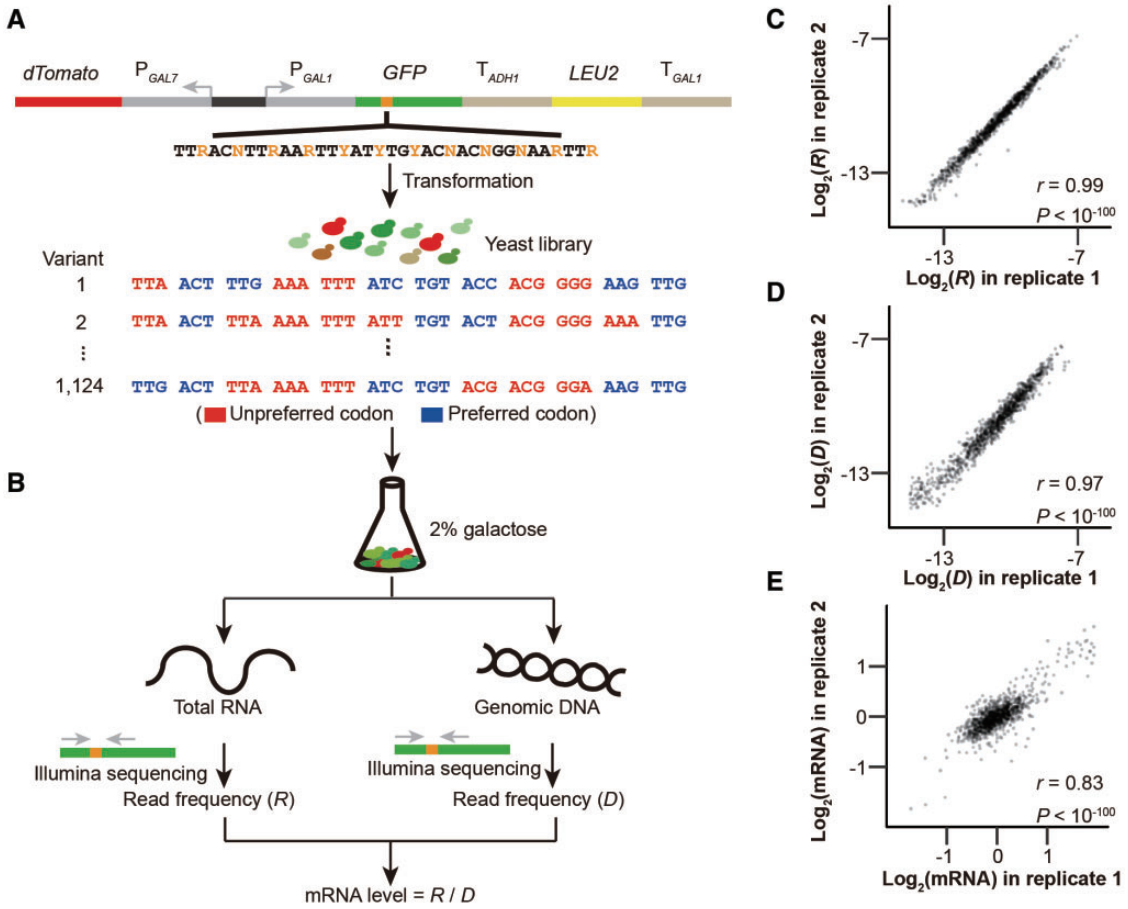

Fig. 1. High-throughput generation and mRNA level measurement of *GFP* synonymous variants. (*A*) The flowchart of the experimental design. DNA oligos were synthesized with doped nucleotides and yeast transformation was performed to generate the variants where *GFP* is expressed from the *GAL1* promoter and *dTomato* from the *GAL7* promoter on chromosome II. Examples of *GFP* variants are shown. (*B*) The variants were pooled and the expression of *GFP* was induced with 2% galactose. DNA and RNA were extracted from the pooled cells and the frequency of each variant was calculated from Illumina sequencing read counts. The ratio between the read frequency of a variant in the mRNA library and that in the DNA library reflects the mRNA level. (*C*–*E*) The frequency of mRNA (*R*), DNA (*D*), and mRNA level (*R*/*D*) were highly correlated between two biological replicates.

12-codon region for each *GFP* variant and observed that this number was positively correlated with mRNA level ($\rho = 0.39$, $P = 1.5 \times 10^{-42}$, Spearman's correlation; fig. 2*A*). Then, we calculated the codon adaptation index (*CAI*), which measures the overall tendency of a gene to use preferred codons (Sharp and Li 1987). Consistently, we observed a positive correlation between *CAI* and mRNA level in both replicates ($\rho = 0.33$ and 0.31, $P = 1.6 \times 10^{-29}$ and 1.0×10^{-25}, respectively; fig. 2*B* and *C*). We further repeated the analysis with the tRNA adaptation index (*tAI*), another measure of codon usage preference without defining a reference set of highly expressed genes (dos Reis et al. 2004), and again observed a positive correlation between CUB and mRNA level in both replicates ($\rho = 0.36$ and 0.32, $P = 8.5 \times 10^{-35}$ and 4.6×10^{-29}, respectively; supplementary fig. S2, Supplementary Material online). In addition, Presnyak *et al.* estimated the correlation coefficient between the frequency of a codon in a gene and the mRNA stability of the gene (the codon stabilization coefficient, *CSC*) (Presnyak et al. 2015). We also observed a positive correlation between *CSC* and mRNA level among our synonymous variants ($\rho = 0.42$ and 0.32, $P = 1.2 \times 10^{-48}$ and 3.2×10^{-28}, in replicates 1 and 2, respectively; supplementary fig. S3, Supplementary Material online). It is important to note that the correlation observed here is different from the genome-wide correlation mentioned earlier (Ikemura 1981, 1982, 1985; Gouy and Gautier 1982; Duret and Mouchiroud 1999), because it cannot be explained by the natural selection on translational accuracy or efficiency. Rather, it suggests a direct effect of synonymous codon usage on mRNA level.

It is worth noting that synonymous codon usage not only affects *CAI* but can also change mRNA secondary structures, which may have influence on transcription or mRNA decay (Wan et al. 2012; Zamft et al. 2012). To control for this potential confounding effect, we calculated the minimum free energy (MFE) to estimate the stability of mRNA secondary structure in the variable region (Lorenz et al. 2011). We found that MFE was only weakly correlated with mRNA level ($\rho = 0.077$, $P = 0.01$; fig. 2*D*), suggesting the negligible influence of mRNA secondary structure on mRNA level in this region. Nevertheless, we calculated the partial correlation between *CAI* and mRNA level controlling for mRNA secondary

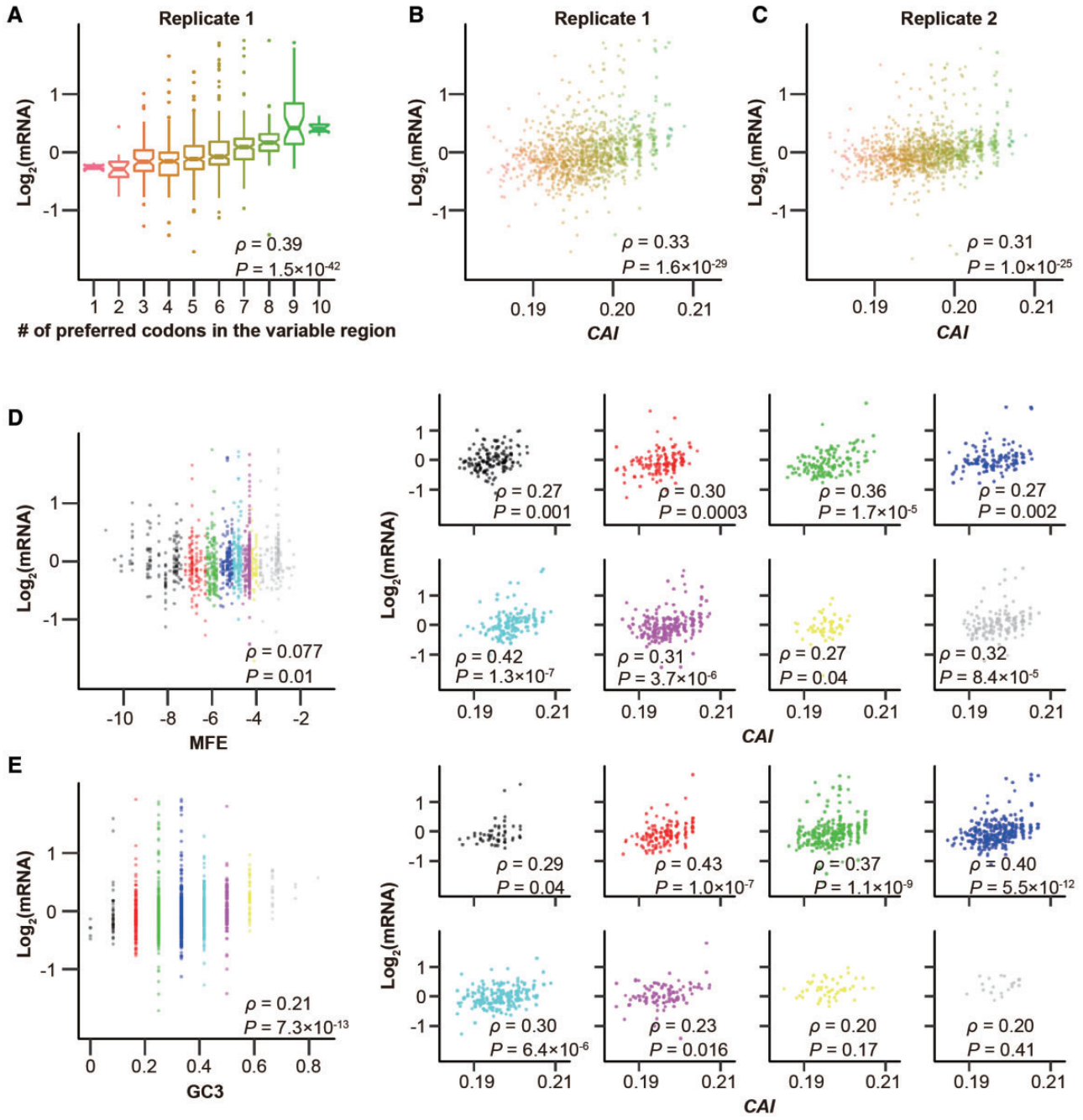

Fig. 2. Synonymous codon usage affected mRNA level among *GFP* synonymous variants. (*A*) mRNA level increased with the number of preferred codons in a *GFP* variant. Spearman's correlation coefficient was calculated from the raw data. Normalized $\log_2$(mRNA) was used on *y* axis. (*B*) mRNA level was positively correlated with *CAI*. Each dot represents a *GFP* variant, and the color of the dot indicates the number of preferred codons as shown in (*A*). *CAI* was calculated based on the full length *GFP*. (*C*) The result in the biological replicate. Similar to (*B*). (*D*) mRNA level was marginally correlated with mRNA secondary structure quantified by MFE. After dividing *GFP* variants into bins according to MFE, the correlation between *CAI* and mRNA level persisted in each bin. (*E*) mRNA level was positively correlated with GC content in the third nucleotide site of codons (GC3). After dividing *GFP* variants into bins according to GC3, the correlation between *CAI* and mRNA level persisted in each bin.

structure, and found the correlation virtually unchanged (partial correlation $\rho = 0.32$, $P = 4.6 \times 10^{-28}$). Furthermore, we divided the *GFP* synonymous variants into eight equal-sized groups based on MFE, and still observed a robust correlation between *CAI* and mRNA level in each group (fig. 2*D*).

Similarly, synonymous codon usage can also affect the GC content in a region, which may regulate mRNA level through nucleosome positioning or other unknown molecular mechanisms (Kudla et al. 2006; Kaplan et al. 2009; Tillo and Hughes 2009). Nevertheless, after controlling for GC content in the third nucleotide position of codons (GC3), the correlation between *CAI* and mRNA level remained unchanged (partial correlation $\rho = 0.34$, $P = 1.8 \times 10^{-31}$). Furthermore, we divided the *GFP* synonymous variants into eight groups based

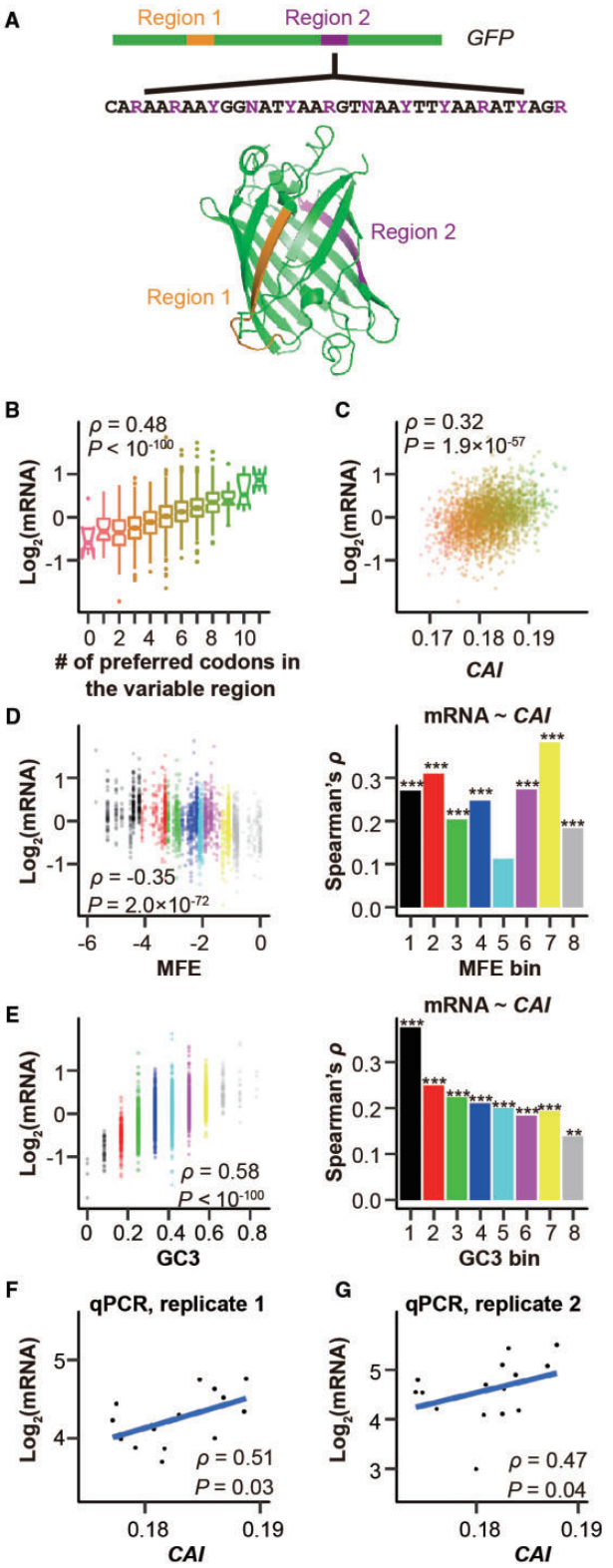

Fig. 3. Synonymous codon usage affected mRNA level in the second region of GFP. (*A*) The second variable region of GFP was selected (shown in purple), which covers the C terminus of a loop and the N terminus of a beta sheet. (*B*) mRNA level increased with the number of preferred codons in the *GFP* variants of region 2. Spearman's correlation coefficient was calculated from the raw data. (*C*) mRNA level was positively correlated with *CAI* in region 2 variants. (*D*) After dividing *GFP* variants into bins according to MFE, the positive correlation between *CAI* and mRNA level persisted in seven out eight bins. **, $P < 0.01$; ***, $P < 0.001$. (*E*) Similar to (*D*), *GFP* variants were divided into bins according to GC3. (*F*) The correlation between *CAI* and mRNA level was validated by qPCR-based mRNA quantification for individual strains. These strains were randomly chosen from region 2 library. The mRNA level of *GFP* was normalized by that of *dTomato* for each variant. (*G*) Similar to (*F*), 15 additional variants were chosen from region 2 library.

on GC3, and still observed positive correlations between *CAI* and mRNA level in most groups (fig. 2*E*).

The Impact of Synonymous Codon Usage on mRNA Level in a Second GFP Region

The synonymous variants we described so far are localized in codon 41–52, which covers the C terminus of a beta sheet and the N terminus of a loop (fig. 3*A*, region 1, in orange). To investigate if the influence in mRNA level by synonymous codon usage is specific in this region, we examined another 12-amino-acid region (amino acid 156–167, QKNGIKVNFKIR), which covers the C terminus of a loop and the N terminus of a beta sheet (fig. 3*A*, region 2, in purple). To this end, we synthesized oligonucleotide CAR-AAR-AAY-GGN-ATY-AAR-GTN-AAY-TTY-AAR-ATY-AGR with doped nucleotides and measured the mRNA levels of 2,432 *GFP* synonymous variants. Again, we observed positive correlations between *CAI* and mRNA level in both biological replicates ($\rho = 0.32$ and 0.24, $P = 1.9 \times 10^{-57}$ and 8.5×10^{-33}, respectively; fig. 3*B* and *C*; supplementary fig. S4, Supplementary Material online), suggesting that the impact of synonymous codon usage on mRNA level is not region-specific. Furthermore, although mRNA secondary structure and GC content were both correlated with mRNA level in this region, the correlation between *CAI* and mRNA level persisted after controlling for mRNA secondary structure (partial correlation $\rho = 0.28$, $P = 2.0 \times 10^{-46}$; fig. 3*D*) and GC content (partial correlation $\rho = 0.39$, $P = 1.0 \times 10^{-91}$; fig. 3*E*).

Validation of High-Throughput Experiment with Quantitative PCR (qPCR)

To validate the impact of synonymous codon usage on mRNA level identified from our high-throughput experiments, we randomly chose 15 *GFP* variants from the library of region 2 (supplementary table S2, Supplementary Material online), induced the expression of *GFP* individually with 2% galactose, and measured the mRNA level of each variant with qPCR. Again, we observed that mRNA level was positively correlated with *CAI* ($\rho = 0.51$, $P = 0.03$; fig. 3*F*). A replicate with 15 different *GFP* variants exhibited a similar pattern ($\rho = 0.47$, $P = 0.04$; fig. 3*G*; supplementary table S2, Supplementary Material online).

Synonymous Codon Usage Influences mRNA Level at Least Partly by Affecting mRNA Stability

The underlying mechanism by which synonymous codon usage may influence mRNA level is still under debate. A few recent papers reported that synonymous codon usage could regulate mRNA degradation rate (Presnyak et al. 2015;

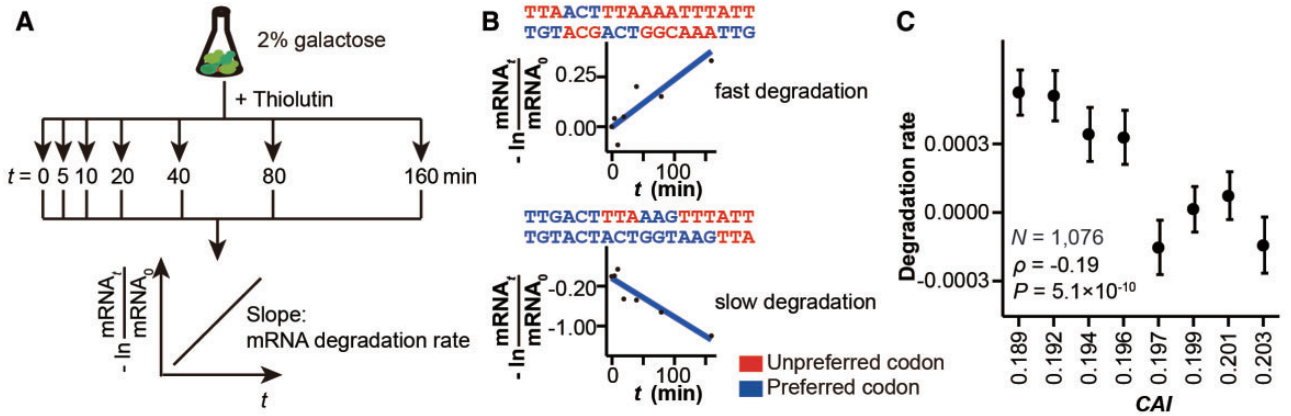

FIG. 4. Codon usage affected mRNA degradation rate. (*A*) The measurement of mRNA degradation rate. The slope of the regression line in the linear model $-\ln(\mathrm{mRNA}_t/\mathrm{mRNA}_0) \sim t$ reflected mRNA degradation rate. Note that mRNA level was calculated as the read count of a variant normalized by the total count of mapped reads in the sequencing library, so that both positive and negative slopes could be observed. (*B*) As an illustration, two synonymous variants in the region 1 library of *GFP* exhibited different mRNA degradation rates. Unpreferred codons are marked in red and preferred codons are marked in blue. (*C*) *CAI* and mRNA degradation rate were negatively correlated among synonymous variants in region 1 library of *GFP* ($\rho = -0.19$, $P = 5.1 \times 10^{-10}$, $N = 1{,}076$, Spearman's correlation). Variants were divided into eight bins with a similar size. The median *CAI* in each bin is shown on *x* axis. The means and standard errors of degradation rates are plotted.

Bazzini et al. 2016; Boel et al. 2016; Mishima and Tomari 2016), whereas it was also reported that synonymous codon usage did not consistently influence mRNA degradation rate (Newman et al. 2016; Zhou et al. 2016). To resolve this dispute, we quantified mRNA degradation rates of our synonymous variants. Specifically, we measured mRNA levels at seven time points ($t = 0$, 5, 10, 20, 40, 80, and 160 min; fig. 4*A*) after the addition of a transcriptional inhibitor, thiolutin (Jimenez et al. 1973; Herrick et al. 1990). We estimated mRNA degradation rate from the change of mRNA level over time in the library of region 1 (fig. 4*A* and *B*). We observed a negative correlation between *CAI* and mRNA degradation rate ($\rho = -0.19$, $P = 5.1 \times 10^{-10}$, $N = 1{,}076$; fig. 4*C*), which persisted after controlling for potential confounding factors such as mRNA secondary structure (partial correlation $\rho = -0.16$, $P = 1.8 \times 10^{-7}$; supplementary fig. S5*A*, Supplementary Material online) and GC content (partial correlation $\rho = -0.20$, $P = 6.2 \times 10^{-11}$; supplementary fig. S5*B*, Supplementary Material online). These observations suggest that codon usage regulates mRNA level at least partly through modulating mRNA degradation rate of *GFP*. Additional mechanisms such as the effect of codon usage on transcription rate as reported in previous studies (Newman et al. 2016; Zhou et al. 2016), may also play a role.

The Effects of Individual Synonymous Mutations on mRNA Level

Although we observed an overall positive correlation between *CAI* and mRNA level among our synonymous variants, it remained unclear whether this was true for each individual synonymous mutation. The large number of synonymous variants in our library provides a unique opportunity to investigate the effects of individual synonymous mutations, which could add important details on top of the global relationship between *CAI* and mRNA level. To this end, for each of the 12 codon sites in region 1, we divided the 1,124 *GFP* variants into two groups according to the synonymous codon category (preferred or unpreferred), and calculated the average normalized $\log_2$(mRNA) within each group (fig. 5*A*, see Materials and Methods). We termed this value individual codon effect (*ICE*) because it reflected the effect of a single synonymous codon at a particular site on mRNA level. We found that for 8 out of the 12 codon sites, preferred codons exhibited significantly higher *ICE* than unpreferred codons ("Method 1", *P* values were calculated with *t* tests). By contrast, only one codon site exhibited the opposite pattern (codon site #2 in fig. 5*A*), which was significantly smaller than 8 ($P = 0.04$, binomial test; $G = 6.2$, $df = 1$, $P = 0.01$, *G*-test). We also estimated *ICE* values for preferred or unpreferred codons defined by *tAI* or *CSC*, and the result kept unchanged (supplementary fig. S6*A* and *B*, Supplementary Material online). This is not unexpected, because although the correlations among these measures (*CAI*, *tAI*, and *CSC*) are only moderate (Presnyak et al. 2015), the identities of preferred codons are almost the same.

To further explore the effects of individual synonymous mutations, we performed a more rigorous analysis, by comparing between pairs of *GFP* variants that are different by only one synonymous codon ("Method 2"; fig. 5*B*). Among the 95 pairs of *GFP* variants that are different only in the first codon (TTA/G) of region 1 (fig. 5*B*), the variants that contain the preferred codon (TTG) exhibited higher mRNA levels in 62 pairs and lower mRNA levels in 33 pairs, demonstrating that the preferred codon (TTG) increased mRNA level on significantly more occasions ($P = 0.004$, binomial test). Among 12 sites in this region, 6 exhibited significantly more variant pairs with $\mathrm{mRNA}_{\mathrm{preferred}} > \mathrm{mRNA}_{\mathrm{unpreferred}}$ than variant pairs with $\mathrm{mRNA}_{\mathrm{preferred}} < \mathrm{mRNA}_{\mathrm{unpreferred}}$, whereas no site showed the opposite pattern ($P = 0.03$, binomial test; $G = 8.3$, $df = 1$,

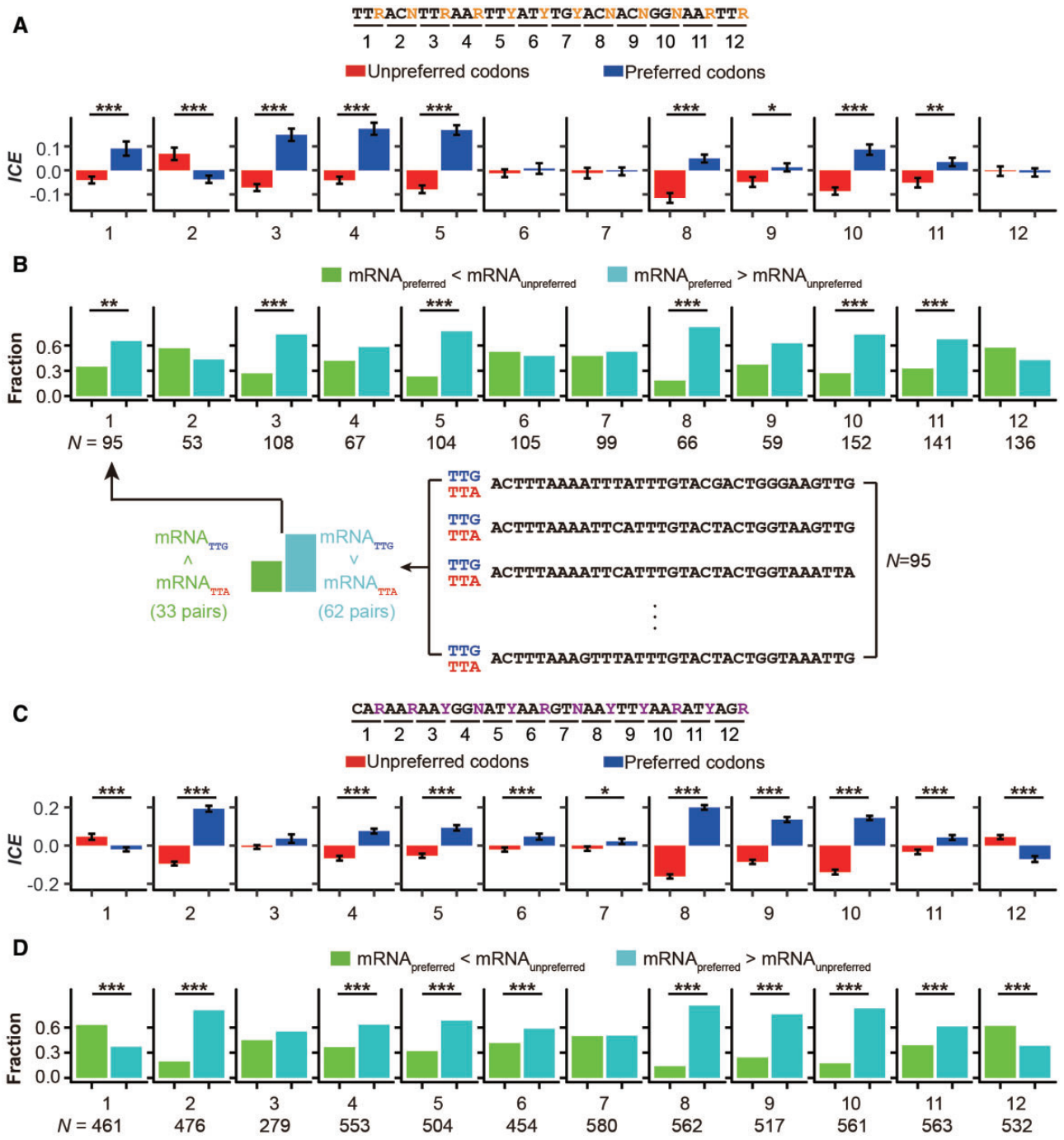

Fig. 5. Impact of individual codons on mRNA level. (*A*) Variants were divided into two groups based on the category (preferred or unpreferred) of a specific codon (from site #1 to site #12). mRNA levels were normalized within a library as described in Materials and Methods. The *ICE* value was estimated as the average normalized $\log_2$(mRNA) within a variant group. Error bars represent standard errors. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. (*B*) Fractions of $mRNA_{preferred} > mRNA_{unpreferred}$ and $mRNA_{preferred} < mRNA_{unpreferred}$ pairs. The mRNA levels were compared within the pairs of variants different by only one synonymous codon. *P* values were calculated with binomial test. (*C*) The *ICE* values in the library of region 2. Similar to (*A*). (*D*) Fractions of variant pairs in the library of region 2. Similar to (*B*).

$P = 0.004$, *G*-test; fig. 5*B*). A similar pattern was observed in region 2 (fig. 5*C* and *D*; supplementary fig. S6*E* and *F*, Supplementary Material online). When two regions were combined, *P* values of repeated *G*-tests were 0.003 (total $G = 11.0$, $df = 2$) with Method 1 (fig. 5*A* and *C*) and 0.002 (total $G = 12.2$, $df = 2$) with Method 2 (fig. 5*B* and *D*), respectively.

The Impact of an Individual Synonymous Mutation Is Sequence-Context-Dependent

We examined whether *ICE* quantified in our experiments agreed with the codon indices estimated in previous studies. We observed positive correlations between *ICE* and the relative synonymous codon usage (*RSCU*, $\rho = 0.38$ and 0.49, $P = 0.03$ and 0.008, for regions 1 and 2, respectively;

supplementary fig. S7*A* and *B*, Supplementary Material online), as well as between *ICE* and *CSC* ($\rho = 0.39$ and 0.67, $P = 0.03$ and 0.0001, for regions 1 and 2, respectively; supplementary fig. S7*C* and *D*, Supplementary Material online). Albeit significant, these correlations were only moderate, probably because the variable effects of the same synonymous mutation at different sites attenuated these correlations.

A codon-resolution examination revealed that the effect of a synonymous mutation could vary among different sequence contexts. For example, among the 95 variant pairs in which the only difference is at the first codon of region 1, although the preferred-codon-containing variants exhibited higher mRNA levels in 62 variant pairs, they exhibited lower mRNA levels in 33 pairs where the sequence contexts are different (fig. 5*B*). Furthermore, the same synonymous codon exhibited variable effects at different codon sites. For example, amino acid threonine is encoded by four synonymous codons (ACT, ACC, ACA, and ACG), among which the first two are preferred and the rest are unpreferred. Threonine appeared in three codon sites in region 1 (sites #2, #8, and #9). Whereas the preferred codons of threonine exhibited higher *ICEs* at sites #8 and #9, they exhibited a lower *ICE* at site #2 (fig. 5*A*). Importantly, the *ICE* values of synonymous codons were largely consistent between two biological replicates (supplementary fig. S8, Supplementary Material online), suggesting that the impact of a codon on mRNA level is indeed influenced by its sequence context. And it was more likely proximal rather than distal sequence contexts that mattered, because the length of the variable region was 12 codon sites in our experiments.

The Effect of an Individual Synonymous Mutation Is Modulated by Codon Usage in Proximal Sequence Contexts

We next investigated potential features in proximal sequence contexts that may modulate the effect of a synonymous mutation. It was observed that the correlation between CUB and mRNA degradation rate disappeared after adding cycloheximide, suggesting that this correlation depends on active translation (Bazzini et al. 2016). Furthermore, the difference of the effects on mRNA degradation rate among synonymous codons was reduced when a translation-related DEAD-box gene *Dhh1* was knocked out (Radhakrishnan et al. 2016). Therefore, we speculated that the dependence on sequence contexts observed above (fig. 5 and supplementary fig. S8, Supplementary Material online) might also be related to the features influencing translation, such as CUB and mRNA secondary structure, which are separately discussed below.

In spite of the overall positive correlation between *CAI* and mRNA level among synonymous variants (figs. 2 and 3), two sites of region 2 (sites #1 and #12) exhibited a reproducible opposite pattern ($ICE_{unpreferred} > ICE_{preferred}$; fig. 5*C*; supplementary fig. S6*E*, Supplementary Material online). Furthermore, among variant pairs where the only difference is at site #1, we identified 222 variant pairs that exhibited $mRNA_{preferred} < mRNA_{unpreferred}$ in both replicates and only 81 variant pairs exhibited the opposite pattern ($P = 2.5 \times 10^{-16}$, binomial test). Intriguingly, the former

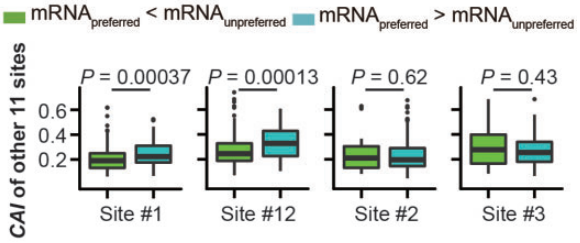

Fig. 6. Codon usage in proximal sequence contexts modulated the relationship between codon optimality and mRNA level. For variant pairs where the only difference is at site #1 of *GFP* region 2 library, we compared the *CAI* of other 11 sites between variant pairs showing $mRNA_{preferred} < mRNA_{unpreferred}$ and those showing $mRNA_{preferred} > mRNA_{unpreferred}$. The results of three more sites in the library are shown. *P* values were calculated with Mann–Whitney *U* test.

exhibited lower *CAI* in the rest 11 codon sites ($P = 4 \times 10^{-4}$, Mann–Whitney *U* test; fig. 6). A similar pattern was observed at site #12 ($P = 1 \times 10^{-4}$, Mann–Whitney *U* test; fig. 6). By contrast, this pattern was not observed at other sites of the region (e.g., sites #2 and #3, fig. 6). When the difference in mRNA level between a pair of variants ($mRNA_{unpreferred} - mRNA_{preferred}$) was considered, the patterns in figure 6 became more apparent (supplementary fig. S9, Supplementary Material online). These observations suggest that codon usage in proximal sequence contexts may influence the impact of a synonymous mutation on mRNA level.

The Variable Impact of an Individual Synonymous Mutation Is Unlikely Caused by the Interaction between Neighboring Codons or tRNA Recycling

It has been proposed that two mechanisms, the interaction between neighboring codons and the reuse of synonymous codons recognized by the same tRNA (tRNA recycling), can regulate translational elongation (Gutman and Hatfield 1989; Irwin et al. 1995; Buchan et al. 2006; Coleman et al. 2008; Cannarozzi et al. 2010; Gamble et al. 2016). Since the impact of synonymous codon usage on mRNA level is associated to translation (Bazzini et al. 2016; Radhakrishnan et al. 2016), we sought to examine whether these mechanisms are related to the variable impact of an individual synonymous mutation on mRNA level. To this end, we first defined an *i*nteraction *i*ndex (*II*; fig. 7*A*) to examine whether codons at two sites influence mRNA level independently. Specifically, we divided *GFP* variants in our libraries into four classes (p-p, p-u, u-p, and u-u, where u and p stand for unpreferred and preferred codons, respectively) based on the codon category for each codon site pair, and calculated the average mRNA level for each class. We defined *II* as the ratio between ($mRNA_{p\text{-}p} \times mRNA_{u\text{-}u}$) and ($mRNA_{p\text{-}u} \times mRNA_{u\text{-}p}$). If synonymous codons at two sites influence mRNA level independently, the logarithm of the *II* should be equal to 0; otherwise, it should deviate from 0 (fig. 7*A*). In addition, we applied a linear regression model to further quantify pair-wise interactions between codons at different sites, and the results were largely identical to the $\log_2(II)$ ($r > 0.99$, $P < 10^{-100}$ for both regions,

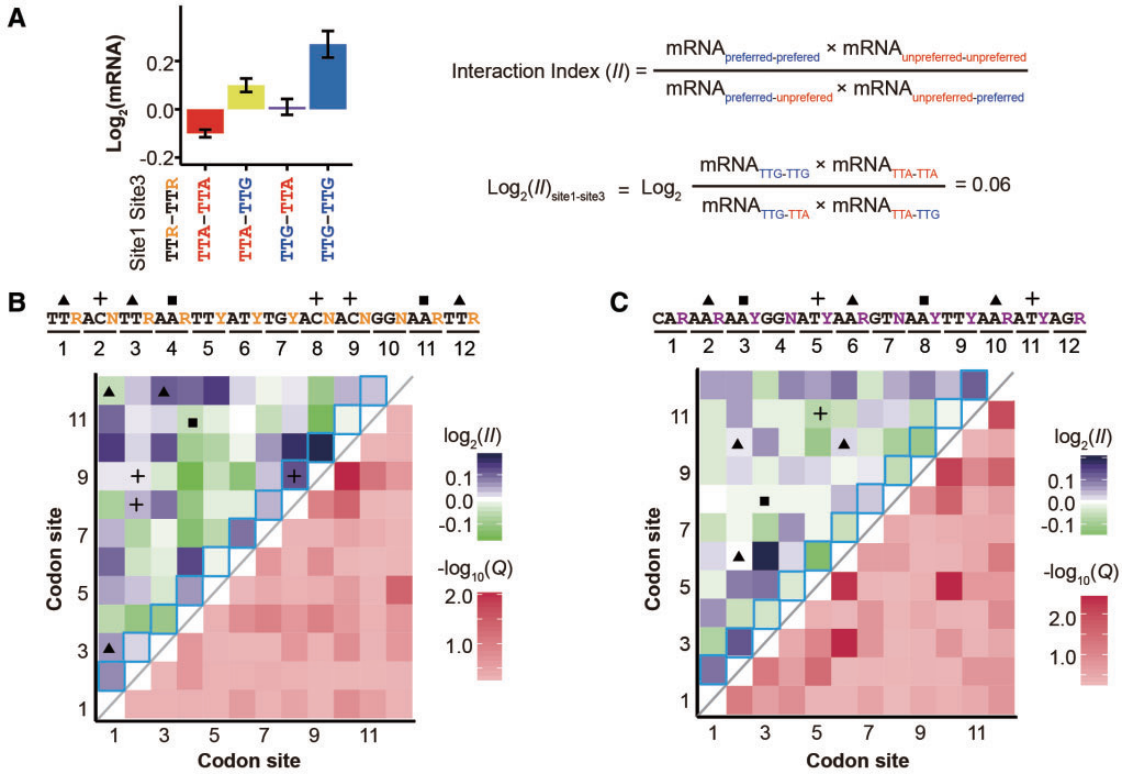

Fig. 7. Interaction between codons at different sites. (*A*) Interaction index (*II*) was defined. Preferred codons are marked in blue, and unpreferred codons are marked in red. Error bars represent the standard errors of the means. (*B*) The pairwise interactions between codons at different sites. $Log_2(II)$ are shown on the upper left corner, and the corresponding *Q* values are shown on the lower right corner. *P* values were estimated from bootstrapping (resampled *GFP* variants with replacement for 10,000 times). Codon sites encoding the same amino acids are marked with the same sign (triangle, square, or plus) above the sequence, and their interactions are marked in the heat map. Interactions of neighboring codons are surrounded by cyan lines. (*C*) Similar to (*B*), the result in the library of region 2.

Pearson's correlation; supplementary table S3, Supplementary Material online). Therefore, we used *II* in the rest of this study.

We observed that some codons at two sites significantly interacted with each other. For example, $log_2(II)$ between codon sites #9 (ACN) and #10 (GGN) in region 1 was 0.18 (bootstrapping $P = 2 \times 10^{-4}$, false discovery rate $Q = 0.008$; fig. 7*B*). We further calculated the *II*s of all pairs of neighboring codons, and observed that they were not significantly different from the *II*s of other codon pairs ($P = 0.06$ and 0.47 for regions 1 and 2, respectively, Mann–Whitney *U* test; fig. 7*B* and *C*; supplementary fig. S10*A* and *B*, Supplementary Material online), suggesting that the effect of a synonymous codon on mRNA level is unlikely modulated by the identities of its neighboring codons. Furthermore, *II* and the distance between two codon sites were not correlated ($\rho = 0.01$ and 0.04, $P = 0.92$ and 0.74, for regions 1 and 2, respectively). Some amino acids appeared multiple times in the variable regions of our *GFP* libraries (figs. 1*A* and 3*A*), which provides us an opportunity to examine whether the reuse of tRNA affects mRNA level. Again, we did not observe a significant difference in *II* between site pairs with the same amino acid and others ($P = 0.31$ and 0.70 for regions 1 and 2, respectively, Mann–Whitney *U* test; fig. 7*B* and *C*; supplementary fig. S10*C* and *D*, Supplementary Material online), suggesting that tRNA recycling is unlikely an important mechanism modulating the effect of a synonymous mutation on mRNA level.

Consistently, our additional analyses suggest that the impact of an individual synonymous mutation is likely modulated by multiple codon–codon interactions, each with a small effect (supplementary Results and Discussion and supplementary fig. S11, Supplementary Material online). Importantly, a preferred codon elevates mRNA level when it is surrounded by preferred codons and sometimes an unpreferred codon also elevates mRNA level when it is surrounded by unpreferred codons (supplementary Results and Discussion and supplementary fig. S11 and table S4, Supplementary Material online). In agreement with this observation, preferred or unpreferred codons form clusters within genes of *S. cerevisiae* and other species (Cannarozzi et al. 2010; Clarke and Clark 2008) (see supplementary Results and Discussion and supplementary fig. S12, Supplementary Material online). Together, these observations are in agreement with the previous studies that a synonymous mutation which changes an unpreferred codon to a preferred codon did not always lead to elevated gene expression or fitness (Agashe et al. 2013; Zhou et al. 2015).

The Effect of an Individual Synonymous Mutation Is Also Modulated by Proximal mRNA Secondary Structure

The same synonymous codons, TTA and TTG which encode leucine, exhibited different impacts on mRNA level when

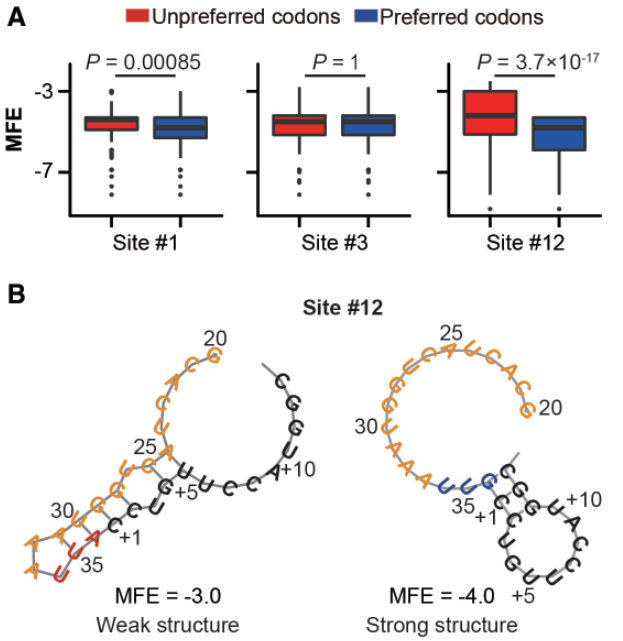

FIG. 8. mRNA secondary structure in the proximal sequence contexts modulated the relationship between codon optimality and mRNA level. (*A*) The difference of MFE between pairs of variants where the only difference is at a single site. The results of sites #1, #3, and #12 of region 1 library are shown. Codons at all three sites encode leucine. *P* values were calculated with paired Mann-Whitney *U* test. (*B*) mRNA secondary structures of a pair of variants. The only difference between these two variants is at site #12. The structures were predicted with RNAfold.

they were localized on sites #1, #3, and #12 of region 1 (fig. 5*A* and *B*). *ICE* values of the preferred codon TTG were higher at sites #1 and #3, but the difference in *ICE* value was insignificant at site #12 (fig. 5*A* and *B*), which needs an explanation. Intriguingly, among the 136 variant pairs where the only sequence difference was at site #12 (last bar plot in fig. 5*B*), preferred-codon-containing variants exhibited lower MFE (stronger secondary structure) than unpreferred-codon-containing variants ($P = 3.7 \times 10^{-17}$, paired Mann–Whitney *U* test; fig. 8*A*). It suggests that codons TTA and TTG at site #12 may pair with other bases and thus alter mRNA secondary structures. Indeed, the base G in the preferred codon TTG at site #12 could pair with the base C in the 13th nucleotide position of the downstream nonvariable region, forming a relatively stable mRNA secondary structure (MFE = −4.0; fig. 8*B*). By contrast, the variants containing the unpreferred codon TTA could not form this secondary structure because the GC paring was destroyed. Instead, a less stable mRNA structure might form (MFE = −3.0; fig. 8*B*). Since stable mRNA secondary structures have been suggested to influence translational elongation (Tuller et al. 2011; Yang et al. 2014), which is related to mRNA decay (Bazzini et al. 2016; Radhakrishnan et al. 2016), proximal sequence contexts can modulate the impact of a synonymous mutation by forming mRNA secondary structures with the focal codon. Note that this effect is different from the observation in figure 2*D* where we showed that the relationship between mRNA level and *CAI* persisted after controlling for MFE.

The mRNA Level Difference Caused by Synonymous Codon Usage Was Reflected at the Protein Level

To further examine if the impact of synonymous codon usage on mRNA level is reflected at the protein level, we measured the GFP level of each synonymous variant in our libraries (fig. 9*A*). To this end, we induced the expression of *GFP* in the pooled library with 2% galactose and harvested yeast cells in mid-log phase. The harvested cells were further sorted into seven bins according to the GFP level (normalized by dTomato level) by fluorescence-activated cell sorting (FACS; fig. 9*A*). We PCR-amplified the variable regions of *GFP* variants and then the PCR products were subject to high-throughput sequencing (Dvir et al. 2013). Based on the respective fractions of cells in seven bins that were estimated from the read frequencies in the high-throughput sequencing, we obtained the distribution of cells in seven bins for each variant (fig. 9*B* and *C*). The protein level of each *GFP* variant was calculated as the average GFP level weighted by the distribution of cells in seven bins (fig. 9*A*–*C*). For example, cells of variant #1 were mainly distributed in bins two and three (fig. 9*B*), while cells of variant #2 were mainly distributed in bins 4–7 (fig. 9*C*). Therefore, the estimated expression level of variant #2 was higher than that of variant #1. The correlations between *CAI* and protein level in both regions ($\rho = 0.31$ and 0.39, $P = 1.2 \times 10^{-25}$ and 5.5×10^{-90}, for regions 1 and 2, respectively; fig. 9*D* and *E*) suggest that the influence of synonymous codon usage on mRNA level was reflected in the protein level, which partly explains the relationship between synonymous codon usage and the protein level of a heterologously expressed gene (Welch et al. 2009; Supek and Smuc 2010; Boel et al. 2016). Previous studies failed to detect the correlation between *CAI* and protein level in *Escherichia coli* (Kudla et al. 2009; Goodman et al. 2013), potentially because mRNA secondary structures around the start codon dominate translational initiation rate, and thus, masked the correlation between *CAI* and protein level (Supek and Smuc 2010). It is also possible that *CAI* is not the optimal index of codon usage for heterologously overexpressed genes (Welch et al. 2009; Supek and Smuc 2010; Guimaraes et al. 2014; Boel et al. 2016).

The Impact of Codon Usage on mRNA Level Was Also Observed in an Endogenous Gene

So far, we presented observations in a heterologous gene *GFP*. To examine the impact of codon usage on mRNA level in endogenous genes, we further constructed a synonymous variant library in a 12-amino-acid region of an *S. cerevisiae* gene (*TDH3*) by synthesizing its degeneracy sequence (GAR-GTN-TCN-CAY-GAY-GAY-AAR-CAY-ATH-ATH-GTN-GAY; fig. 10*A*). *TDH3* encodes a glyceraldehyde-3-phosphate dehydrogenase, is highly expressed, and mainly uses preferred codons. Again, we observed a positive correlation between *CAI* and mRNA level ($\rho = 0.39$, $P = 2.6 \times 10^{-20}$, $N = 523$; fig. 10*B*), which persisted after controlling for mRNA secondary structure and GC content (supplementary fig. S13*A* and *B*, Supplementary Material online). We further examined the impact of individual codons, and observed significant higher *ICE* values for six preferred codons and two unpreferred

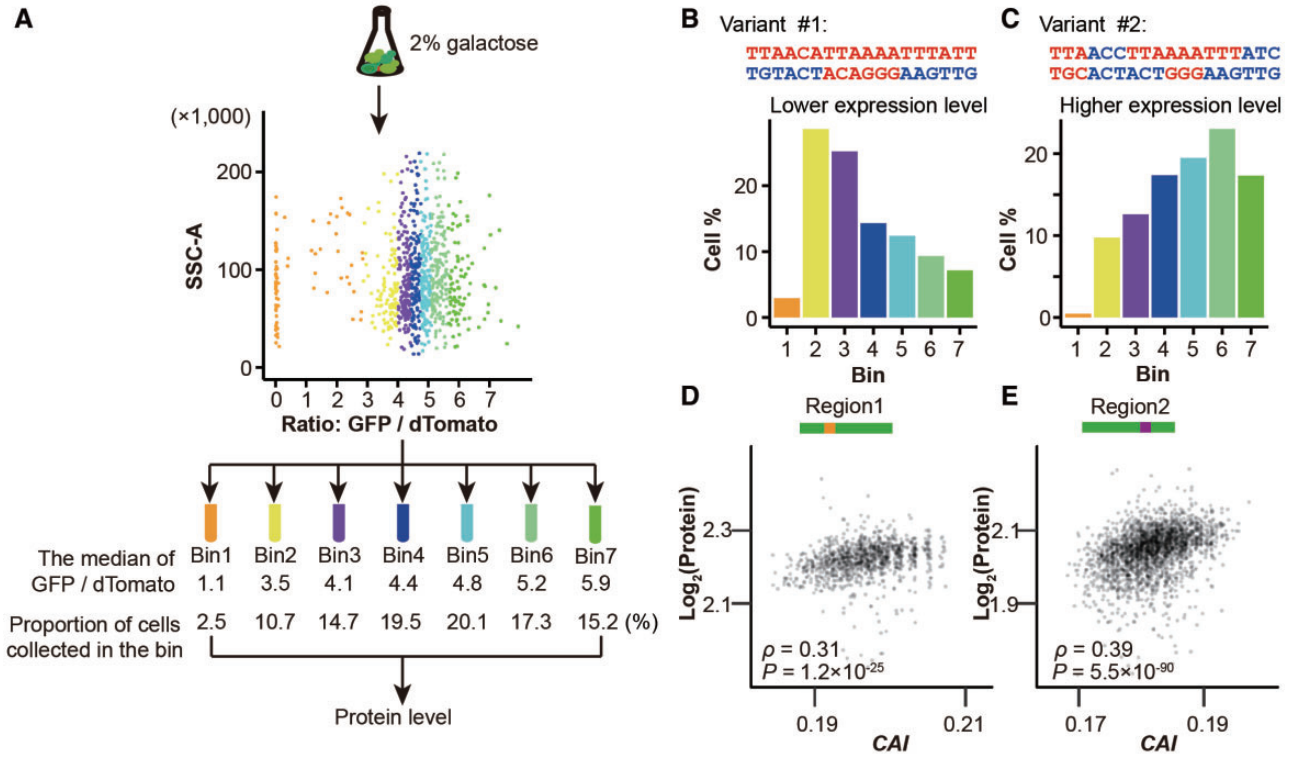

Fig. 9. The impact of codon usage on mRNA level was reflected in the protein level. (*A*) The flowchart of the experimental design of protein level measurement. Induced cells were sorted into seven bins based on the GFP/dTomato fluorescent ratio quantified by the flow cytometer. High-throughput sequencing was performed within each bin. The distribution of the numbers of cells in seven bins was estimated based on the fraction of cells collected in each bin in FACS and the read frequency of a variant in each bin. (*B*, *C*) The distribution of the numbers of cells in seven bins for variant #1 and variant #2 in region 1 library, which have lower and higher protein levels, respectively. Codons marked in red are unpreferred codons and those marked in blue are preferred ones. (*D*) The correlation between *CAI* and protein level in region 1 library. A few synonymous variants with protein levels out of the *y* axis range are not shown. (*E*) Similar to (*D*), the result in the library of region 2.

codons (fig. 10C). Together with the results of two *GFP* variant libraries (figs. 5A and C), we confirmed that individual preferred codons more often increase than decrease mRNA level (total $G = 13.1$, $df = 3$, $P = 0.004$, repeated *G*-tests).

The Evolution of Codon Usage Bias

It was reported previously that the activities of core promoters were a major determinant of gene expression level in yeast (Lubliner et al. 2015). Our findings based on reporter genes (*GFP* and *TDH3*), however, indicate that in addition to promoter activity, synonymous codon usage may also contribute to the evolution of gene expression level through regulating mRNA stability. In a previous study, 859 yeast promoters were cloned and their activities were accurately measured with the same fluorescent reporters (Keren et al. 2013). We plotted *CAI* against promoter activity (supplementary fig. S14A, Supplementary Material online) and found them highly correlated ($\rho = 0.73$, $P < 10^{-100}$). More importantly, mRNA level increases with both promoter activity and *CAI* (from green dots to blue dots in supplementary fig. S14A, Supplementary Material online), suggesting that both contribute to the evolution of yeast transcriptome. Indeed, adding *CAI* to the linear model $\log_2(\text{mRNA}) \sim \log_2(\text{promoter activity})$ (Akaike information criterion, AIC = 2,346) significantly improved the predictive power ($\log_2(\text{mRNA}) \sim \log_2(\text{promoter activity}) + CAI$, AIC = 2,115).

These observations provide an additional explanation to the correlation between CUB and expression level among genes, which has been reported for decades and in multiple species (Ikemura 1981, 1982, 1985; Gouy and Gautier 1982; Moriyama and Powell 1997; Duret and Mouchiroud 1999; Duret 2000; Kanaya et al. 2001; Krisko et al. 2014). This correlation has long been explained by 1) stronger natural selection on translational accuracy and/or efficiency in more highly expressed genes (Bulmer 1991; Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). In this and other recent studies (Presnyak et al. 2015; Boel et al. 2016; Zhou et al. 2016), 2) a direct impact of synonymous codon usage on mRNA level was reported. Therefore, the well-known correlation between *CAI* and mRNA level among genes is likely contributed by both 1) and 2) (the new model in supplementary fig. S15, Supplementary Material online).

A positive feedback exists in this new model. That is, natural selection optimizes mRNA levels partly through synonymous codon usage and the altered mRNA level in turn affect the evolution of synonymous codon usage of this gene (which in turn, will again have impact on mRNA level). This positive feedback can result in a group of genes with ultrahigh *CAI* and mRNA level, which was indeed observed (the top-right corner

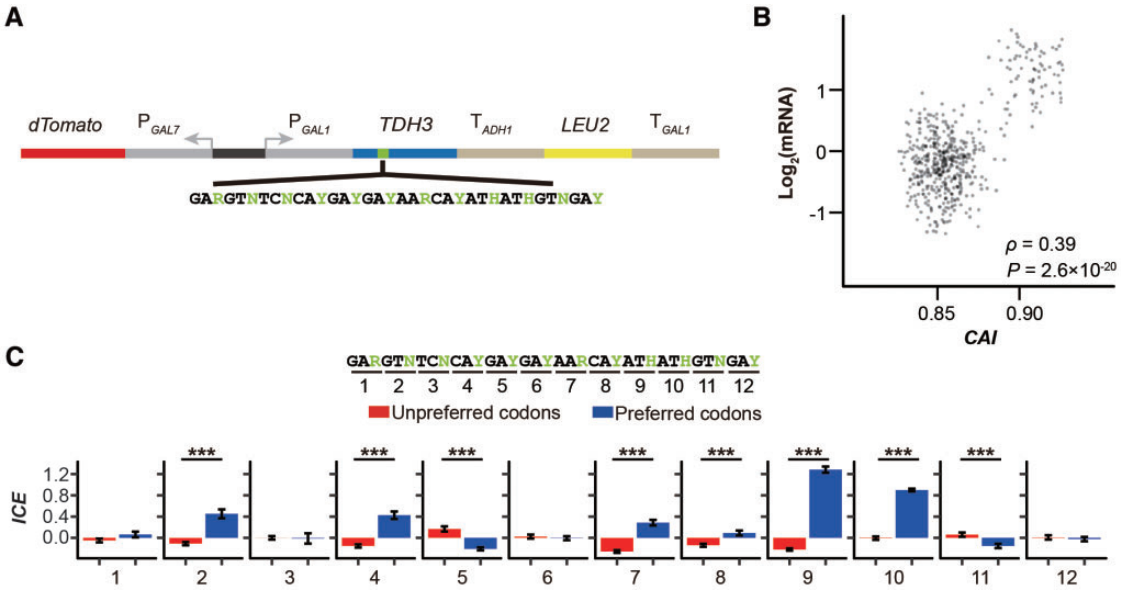

Fig. 10. Synonymous codon usage affected mRNA level in the library of *TDH3* variants. (*A*) Construction of synonymous variants of a region in *TDH3*. The variable region of *TDH3* is shown in green. (*B*) mRNA level was positively correlated with *CAI* among *TDH3* variants. (*C*) The *ICE* values in the *TDH3* library, similar to figure 5*A*.

in supplementary fig. S14*A*, Supplementary Material online). Because *CAI* exhibits larger variation among these genes, we speculate that *CAI* plays a more important role in regulating the mRNA levels of these genes (supplementary fig. S14*B*, Supplementary Material online).

Our codon-resolution analysis reveals a direct and context-dependent impact of individual synonymous mutations on mRNA level. Needless to say, synonymous codon usage also plays an important role in optimizing translational efficiency and/or accuracy (Hershberg and Petrov 2008; Gingold and Pilpel 2011; Plotkin and Kudla 2011). In addition, synonymous substitution may occur as a by-product of natural selection on mRNA secondary structures (Yang et al. 2014) or other mRNA features. Therefore, this work, together with previous studies, revealed the pleiotropic effects of synonymous codon usage and the multifaceted selective forces during the evolution of synonymous codons. It would be of importance to develop a quantitative model in the future to understand how these forces together drive the evolution of synonymous codon usage.

Materials and Methods

Construction of *GFP* and *TDH3* Variant Libraries

In order to generate the synonymous variants of *GFP*, we first constructed a vector with P_{GAL1}-*GFP*-T_{ADH1}-*LEU2*-T_{GAL1} inserted into the backbone of pUC19, with recombination-based cloning (supplementary fig. S1, Supplementary Material online). Here, the auxotrophic marker *LEU2* was used to select successful transformants on the yeast synthetic drop-out media. The sequence of the insert was confirmed by Sanger sequencing. Variable regions were synthesized with doped nucleotides, and were integrated into the full length *GFP* with fusion PCR, with 25 overlapping oligonucleotides (supplementary fig. S1, Supplementary Material online). All the primers used to construct *GFP* variant libraries are listed in supplementary table S1, Supplementary Material online.

We modified the laboratory strain BY4742 by replacing the coding sequence of *GAL7* with *dTomato* (*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 gal7Δ0::dTomato*). Then, the PCR amplicons above were transformed into this strain, and successful transformants were selected on the agar plates with leucine dropped out (8 g/L synthetic medium minus leucine, 2% glucose, and 2% agar) at 30 °C for 2 days. 1,344 and 4,394 colonies were collected and pooled for region 1 and region 2 of *GFP*, respectively. The pooled libraries were stored at −80 °C.

A library containing synonymous variants in codon 57–68 of *TDH3* was constructed similarly. We confirmed that synonymous codon occurrences are independent among codon sites in our libraries (supplementary figs. S16 and S17, Supplementary Material online).

Library Preparation for Illumina Sequencing

Each of the pooled libraries was inoculated into 200 mL YPGEG media (1% yeast extract, 2% peptone, 2% galactose, 2% ethanol, and 2% glycerol) at OD_{660} ~0.05 and was harvested at OD_{660} ~0.5. Galactose was used to induce the expression of *GFP* (or *TDH3*) and *dTomato*, while ethanol and glycerol were used as the carbon source. For the *GFP* libraries, harvested cells were split into three aliquots, which were used for FACS-seq, RNA-seq, and DNA-seq, respectively. For the *TDH3* library, cells were split into two aliquots, which were used for RNA-seq and DNA-seq.

All the primers used to prepare the Illumina sequencing libraries above are listed in supplementary table S5, Supplementary Material online. RNA library, DNA library and FACS library were sequenced with Illumina HiSeq 2500 (PE125, paired-end 2 × 125 bp). A detailed protocol is described in the supplementary Methods, Supplementary Material online.

Quantification of mRNA and Protein Abundances

Sequencing read pairs were sorted into samples according to the barcodes introduced during the library preparation, and the number of read pairs in each sample is listed in supplementary table S6, Supplementary Material online. Because the lengths of the inserts are 117 and 116 base pairs in the *GFP* libraries of "region 1" and "region 2", respectively, the full sequences of the inserts can be obtained from the sequencing reads of both ends. For a pair of reads, if the barcode sequences were different from each other or different from that in the primer, both reads were discarded. Then, sequences in the variable region were extracted from both reads. Again, if the sequences were different in both reads or the sequence was different from that of the designed variant, both reads were discarded. To further remove variant sequences containing PCR errors, we kept only the variants that appeared in both replicates of all three libraries (FACS, RNA, and DNA), with the cutoff of at least 64 reads in both replicates of the DNA library. For the *TDH3* library, we kept only the variants that have at least 64 reads in both RNA and DNA libraries.

To calculate the mRNA level of each synonymous variant, read frequencies of the variant in both RNA and DNA libraries were calculated. Read frequency of variant i in the RNA library $R_i = \frac{r_i}{\sum r_i}$, and read frequency of variant i in the DNA library $D_i = \frac{d_i}{\sum d_i}$, where r_i and d_i are the read counts of variant i in the RNA library and DNA library, respectively. Therefore, the normalized $\log_2$(mRNA) of variant i was calculated as $\log_2 mRNA_i = \log_2 mRNA_i' - \overline{\log_2 mRNA'}$, where $mRNA_i' = \frac{R_i}{D_i}$ and $\overline{\log_2 mRNA'}$ is the average of all $\log_2$-transformed $mRNA'$ values in a library. mRNA levels in two replicates were highly correlated ($r = 0.83$, 0.73, and 0.72, $P < 10^{-100}$ for both regions of *GFP* and the region of *TDH3*, Pearson's correlation).

The protein level of each variant was calculated as the weighted mean of the median GFP/dTomato ratios in seven bins. The weight of variant i in bin j is

$$c_{ij} = p_{ij} \times n_j$$

where p_{ij} is the read frequency of variant i in bin j, and n_j is the number of cells collected in bin j by FACS. Thus, c_{ij} reflects the fraction of cells containing variant i falling into bin j in FACS. The protein level of variant i

$$P_i = \frac{\sum_{j=1}^{7} G_j \times c_{ij}}{\sum_{j=1}^{7} c_{ij}}$$

where G_j is the median GFP/dTomato ratio in bin j. The information of synonymous variants is provided in supplementary tables S7–S9, Supplementary Material online.

Quantification of mRNA Levels by qPCR

The pooled yeast library was spread onto an agar plate, and 30 colonies in the synonymous library of *GFP* region 2 were randomly chosen. The sequences of the variable region were determined with Sanger sequencing. The mRNA levels of *GFP* were quantified with Mx3000P qPCR System (Agilent Technologies), which was further normalized by that of *dTomato*, to control for the potential variation in the level of galactose induction. The sequences of the primers used for qPCR are listed in supplementary table S10, Supplementary Material online.

Calculation of Codon Adaption Index (*CAI*) and tRNA Adaptation Index (*tAI*)

RSCU values were retrieved from Sharp and Li (Sharp and Li 1987). *CAI* was calculated following a previous study (Sharp and Li 1987). The relative *RSCU* of a codon was defined as the ratio between *RSCU* of the codon and the maximum *RSCU* of all synonymous codons encoding the same amino acid. Preferred codons were defined as those with a relative *RSCU* larger than 0.9. The rest codons were defined as unpreferred codons. All the codons in *GFP* were used in calculating the *CAI* of *GFP* variants. *tAI* was calculated with codonR (dos Reis et al. 2004), in which the copy numbers of tRNA genes in yeast were obtained from a previous study (Percudani et al. 1997).

Authors' Contributions

S. C., Y.-F. Y., and W.Q. conceived the research; S. C., K. L., W. C., J. W., Q. H., and T. Z. performed the experiments, S. C., S. W., Y.-F. Y., and W.Q. analyzed the data; S. C., K. L., Y.-F. Y., S. W., and W. Q. drafted the manuscript, and all authors contributed to the manuscript writing.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Lucas Carey, Xiao Chu, Ya-Nan Han, Bin Z. He, Jianrong Yang, Jianzhi "George" Zhang, and Taolan Zhao for discussion. This work was supported by National Natural Science Foundation of China (91331112 and 31571308). The sequence data reported in this study have been deposited in the genome sequence archive (GSA) of Beijing Institute of Genomics, Chinese Academy of Sciences, (http://gsa.big.ac.cn/, accession no. PRJCA000227).

References

Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. 2013. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol*. 30(3):549–560.

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*. 11(6):660–666.

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

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev*. 54(2):198–210.

Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, Yao J, Khokha MK, Giraldez AJ. 2016. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J*. 35(19):2087–2103.

Boel G, Letso R, Neely H, Price WN, Wong KH, Su M, Luff JD, Valecha M, Everett JK, Acton TB, et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* 529:358–363.

Buchan JR, Aucott LS, Stansfield I. 2006. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res.* 34(3):1015–1027.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129(3):897–907.
Cannarozzi G, Cannarrozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* 141(2):355–367.
Carlini DB. 2004. Experimental reduction of codon bias in the Drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. *J Evol Biol.* 17(4):779–785.
Carlini DB, Stephan W. 2003. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics* 163(1):239–243.
Charneski CA, Hurst LD. 2013. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.* 11(3):e1001508.
Clarke TFt, Clark PL. 2008. Rare codons cluster. *PLoS One* 3(10):e3412.
Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol.* 209(1):65–77.
dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32(17):5036–5044.
Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev.* 12(6):640–649.
Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16(7):287–289.
Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A.* 96(8):4482–4487.
Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A.* 110(30):E2792–E2801.
Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. 2016. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell* 166(3):679–690.
Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. 2014. Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* 3. doi:10.7554/eLife.03735.
Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol.* 7:481.
Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342(6157):475–479.
Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10(22):7055–7074.
Guimaraes JC, Rocha M, Arkin AP. 2014. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic Acids Res.* 42(8): 4791–4799.
Gutman GA, Hatfield GW. 1989. Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A.* 86(10):3699–3703.
Herrick D, Parker R, Jacobson A. 1990. Identification and comparison of stable and unstable mRNAs in Saccharomyces cerevisiae. *Mol Cell Biol.* 10(5):2269–2284.
Hershberg R, Petrov DA. 2009. General rules for optimal codon choice. *PLoS Genet.* 5(7):e1000556.
Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet.* 42:287–299.
Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. 2015. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet.* 11(12):e1005732.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2(1):13–34.
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(4):573–597.
Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4):789–802.
Irwin B, Heck JD, Hatfield GW. 1995. Codon pair utilization biases influence translational elongation step times. *J Biol Chem.* 270(39):22801–22806.
Jimenez A, Tipper DJ, Davies J. 1973. Mode of action of thiolutin, an inhibitor of macromolecular synthesis in *Saccharomyces cerevisiae*. *Antimicrob Agents Chemother.* 3(6):729–738.
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(4–5):290–298.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458(7236):362–366.
Keren L, Zackay O, Lotan-Pompan M, Barenholz U, Dekel E, Sasson V, Aidelberg G, Bren A, Zeevi D, Weinberger A, et al. 2013. Promoters maintain their relative activity levels under different growth conditions. *Mol Syst Biol.* 9:701.
Krisko A, Copic T, Gabaldon T, Lehner B, Supek F. 2014. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol.* 15(3):R44.
Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4(6):e180.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924):255–258.
Lampson BL, Pershing NL, Prinz JA, Lacsina JR, Marzluff WF, Nicchitta CV, MacAlpine DM, Counter CM. 2013. Rare codons regulate KRas oncogenesis. *Curr Biol.* 23(1):70–75.
Li GW, Oh E, Weissman JS. 2012. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484(7395):538–541.
Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol.* 6:26.
Lubliner S, Regev I, Lotan-Pompan M, Edelheit S, Weinberger A, Segal E. 2015. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* 25(7):1008–1017.
Mishima Y, Tomari Y. 2016. Codon usage and 3' UTR length determine maternal mRNA stability in zebrafish. *Mol Cell* 61(6):874–885.
Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol.* 45(5):514–523.
Newman ZR, Young JM, Ingolia NT, Barton GM. 2016. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proc Natl Acad Sci U S A.* 113(10):E1362–E1371.
Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol.* 268(2):322–330.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet.* 12(1):32–42.
Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol.* 10:770.
Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem.* 262(23):11351–11355.

Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. 2015. Codon optimality is a major determinant of mRNA stability. *Cell* 160(6):1111–1124.

Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* 8(3):e1002603.

Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* 167(1):122–132.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res.* 16(17):8207–8211.

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(3):1281–1295.

Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in Escherichia coli. *J Mol Biol.* 207(2):365–377.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24(2):374–381.

Supek F, Smuc T. 2010. On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. *Genetics* 185(3):1129–1134.

Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* 10:442.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* 12(11):R110.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol.* 180(3):549–576.

Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, Makino DL, Nutter RC, Segal E, Chang HY. 2012. Genome-wide measurement of RNA folding energies. *Mol Cell* 48(2):169–181.

Warner JR. 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*. 24(11):437–440.

Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. 2016. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.* 14(7):1787–1799.

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009. Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* 4(9):e7002.

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

Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol Cell* 59(5):744–754.

Zamft B, Bintu L, Ishibashi T, Bustamante C. 2012. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci U S A.* 109(23):8948–8953.

Zhou M, Wang T, Fu J, Xiao G, Liu Y. 2015. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol Microbiol.* 97(5):974–987.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 26(7):1571–1580.

Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A.* 113:E6117–E6125.