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
1State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China
2Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China
3University of Chinese Academy of Sciences, Beijing, China
4Sino-Danish Center for Education and Research, Beijing, China
5Institute 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.

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; 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...
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, LTLKFICTTGGKL) of GFP (fig. 1A and supplementary fig. S1A, Supplementary Material online), by synthesizing DNA oligos with doped nucleotides added to the third nucleotide site of each codon (TT–ACN–TTR–AAR–TTY–ATY–ACN–ACN–AGG–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. 1A). 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. 1B). 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. 1B and supplementary fig. S1B, 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. 1C–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. 1E), we first counted the number of preferred codons (see Materials and Methods) in the
12-codon region for each GFP variant and observed that this number was positively correlated with mRNA level ($\rho = 0.39$, $P = 1.5 \times 10^{-82}$, Spearman’s correlation; fig. 2A). 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 $1.0 \times 10^{-25}$, respectively; fig. 2B 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 tAI and mRNA level in both replicates ($\rho = 0.36$ and $0.32$, $P = 8.5 \times 10^{-35}$ and $4.6 \times 10^{-29}$, 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 $3.2 \times 10^{-28}$, 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; Zamfte 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. 2D), 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 structure.
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. 2D).

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

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. 3A, 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. 3A, 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 $8.5 \times 10^{-33}$, respectively; fig. 3B 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. 3D) and GC content (partial correlation $\rho = 0.39$, $P = 1.0 \times 10^{-91}$; fig. 3E).

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. 3F). A replicate with 15 different GFP variants exhibited a similar pattern ($\rho = 0.47$, $P = 0.04$; fig. 3G; 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;...
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. 4A 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. 4A 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. 4C, 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. S5A, Supplementary Material online) and GC content (partial correlation \( \rho = -0.20, P = 6.2 \times 10^{-11}; \) supplementary fig. S5B, 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 log2(mRNA) within each group (fig. 4A, 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. 4A), which was significantly smaller than \( 6 (P = 0.04, \) binomial test; \( G = 62, df = 1, P = 0.01, \) G-test). We also estimated ICE values for preferred or unpreferred codons defined by tAI or CSC, and the result kept unchanged (supplementary fig. S6A 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. 5B). Among the 95 pairs of GFP variants that are different only in the first codon (TTA/G) of region 1 (fig. 5B), 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 mRNA\(_{\text{preferred}} > \)mRNA\(_{\text{unpreferred}}\) than variant pairs with mRNA\(_{\text{preferred}} < \)mRNA\(_{\text{unpreferred}}\), whereas no site showed the opposite pattern \( (P = 0.03, \) binomial test; \( G = 8.3, df = 1, \)

---

**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 \(-\ln(\text{mRNA}_{\text{t}}/\text{mRNA}_{\text{0}}) \sim t\) 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.
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, \( p = 0.38 \) and 0.49, \( P = 0.03 \) and 0.008, for regions 1 and 2, respectively).

**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 log2(mRNA) within a variant group. Error bars represent standard errors. *\( P < 0.05 \); **\( P < 0.01 \); ***\( P < 0.001 \).

(B) Fractions of mRNA\(_{\text{preferred}}\) > mRNA\(_{\text{unpreferred}}\) and mRNA\(_{\text{preferred}}\) < mRNA\(_{\text{unpreferred}}\) 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. 5B). A similar pattern was observed in region 2 (fig. 5C and D; supplementary fig. S6E 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. 5A and C) and 0.002 (total \( G = 12.2, df = 2 \) with Method 2 (fig. 5B and D), respectively.

\( P = 0.004, \) G-test; fig. 5B). A similar pattern was observed in region 2 (fig. 5C and D; supplementary fig. S6E 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. 5A and C) and 0.002 (total \( G = 12.2, df = 2 \) with Method 2 (fig. 5B and D), respectively.
Codon-resolution analysis 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. S8). 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. 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. S5A). 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 CUB and mRNA level among synonymous variants (figs. 2 and 3), two sites of region 2 (sites #1 and #12) exhibited a reproducible opposite pattern (ICES\textsubscript{unpreferred} > ICES\textsubscript{preferred}; fig. S5C, supplementary fig. S6E, Supplementary Material online). Furthermore, among variant pairs where the only difference is at site #1, we identified 222 variant pairs that exhibited mRNA\textsubscript{preferred} < mRNA\textsubscript{unpreferred} in both replicates and only 81 variant pairs exhibited the opposite pattern (P = 2.5 × 10^{-16}, binomial test). Intriguingly, the former exhibited lower CAI in the rest 11 codon sites (P = 4 × 10^{-4}, Mann–Whitney U test; fig. 6). A similar pattern was observed at site #12 (P = 1 × 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\textsubscript{unpreferred} − mRNA\textsubscript{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 interaction index (I; fig. 7A) 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 I as the ratio between (mRNA\textsubscript{p-p} × mRNA\textsubscript{u-u}) and (mRNA\textsubscript{p-u} × mRNA\textsubscript{u-p}). If synonymous codons at two sites influence mRNA level independently, the logarithm of the I should be equal to 0; otherwise, it should deviate from 0 (fig. 7A). 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_{10}(I) (r > 0.99, P < 10^{-100} for both regions, Supplementary fig. S7, Supplementary Material online) might also be related to the variable impact of an individual synonymous mutation on mRNA level.

**The Variables of Other Distal Sequence Contexts**

Although these distal sequence contexts were not significant, the variable region was 12 codon sites in our experiments. The variable region was 12 codon sites in our experiments. Supplementary fig. S7 and B, Supplementary Material online), as well as between ICE and CSC (p = 0.39 and 0.67, P = 0.03 and 0.0001, for regions 1 and 2, respectively; supplementary fig. S7C and D, Supplementary Material online). Although significant, these correlations were only moderate, probably because the variable effects of the same synonymous mutation at different sites attenuated these correlations.

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 interaction index (I; fig. 7A) 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 I as the ratio between (mRNA\textsubscript{p-p} × mRNA\textsubscript{u-u}) and (mRNA\textsubscript{p-u} × mRNA\textsubscript{u-p}). If synonymous codons at two sites influence mRNA level independently, the logarithm of the I should be equal to 0; otherwise, it should deviate from 0 (fig. 7A). 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_{10}(I) (r > 0.99, P < 10^{-100} for both regions, Supplementary fig. S7, Supplementary Material online) might also be related to the variable impact of an individual synonymous mutation on mRNA level.

**The Variables of Other Distal Sequence Contexts**

Although these distal sequence contexts were not significant, the variable region was 12 codon sites in our experiments. The variable region was 12 codon sites in our experiments. Supplementary fig. S7 and B, Supplementary Material online), as well as between ICE and CSC (p = 0.39 and 0.67, P = 0.03 and 0.0001, for regions 1 and 2, respectively; supplementary fig. S7C and D, Supplementary Material online). Although significant, these correlations were only moderate, probably because the variable effects of the same synonymous mutation at different sites attenuated these correlations.

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 interaction index (I; fig. 7A) 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 I as the ratio between (mRNA\textsubscript{p-p} × mRNA\textsubscript{u-u}) and (mRNA\textsubscript{p-u} × mRNA\textsubscript{u-p}). If synonymous codons at two sites influence mRNA level independently, the logarithm of the I should be equal to 0; otherwise, it should deviate from 0 (fig. 7A). 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_{10}(I) (r > 0.99, P < 10^{-100} for both regions, Supplementary fig. S7, Supplementary Material online) might also be related to the variable impact of an individual synonymous mutation on mRNA level.

**The Variables of Other Distal Sequence Contexts**

Although these distal sequence contexts were not significant, the variable region was 12 codon sites in our experiments. The variable region was 12 codon sites in our experiments. Supplementary fig. S7 and B, Supplementary Material online), as well as between ICE and CSC (p = 0.39 and 0.67, P = 0.03 and 0.0001, for regions 1 and 2, respectively; supplementary fig. S7C and D, Supplementary Material online). Although significant, these correlations were only moderate, probably because the variable effects of the same synonymous mutation at different sites attenuated these correlations.

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 interaction index (I; fig. 7A) 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 I as the ratio between (mRNA\textsubscript{p-p} × mRNA\textsubscript{u-u}) and (mRNA\textsubscript{p-u} × mRNA\textsubscript{u-p}). If synonymous codons at two sites influence mRNA level independently, the logarithm of the I should be equal to 0; otherwise, it should deviate from 0 (fig. 7A). 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_{10}(I) (r > 0.99, P < 10^{-100} for both regions, Supplementary fig. S7, Supplementary Material online) might also be related to the variable impact of an individual synonymous mutation on mRNA level.
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. 7B). 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. 7B and C; supplementary fig. S10A 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. 1A and 3A), 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. 7B and C; supplementary fig. S10C 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
they were localized on sites #1, #3, and #12 of region 1 (fig. 5A 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. 5A 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. 5B), preferred-codon-containing variants exhibited lower MFE (stronger secondary structure) than unpreferred-codon-containing variants (P = 3.7 × 10^{-17}, paired Mann-Whitney U test; fig. 8A). 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. 8B). By contrast, the variants containing the unpreferred codon TTA could not form this secondary structure because the GC pairing was destroyed. Instead, a less stable mRNA structure might form (MFE = -3.0; fig. 8B). 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 2D 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. 9A). 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. 9A). 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. 9B 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. 9A–C). For example, cells of variant #1 were mainly distributed in bins two and three (fig. 9B), while cells of variant #2 were mainly distributed in bins 4–7 (fig. 9C). 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 (ρ = 0.31 and 0.39, P = 1.2 × 10^{-25} and 5.5 × 10^{-90}, for regions 1 and 2, respectively; fig. 9D 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 (GARGTN-TCN-CAY-GAY-GAY-AAR-CAY-ATH-ATH-GTN-GAY; fig. 10A). 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 (ρ = 0.39, P = 2.6 × 10^{-29}, N = 523; fig. 10B), which persisted after controlling for mRNA secondary structure and GC content (supplementary fig. S13A 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
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 ($\rho = 0.73$, $P < 10^{-10}$). 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 $\log_2($mRNA$) \sim \log_2($promoter activity$) + $CAI, $AIC = 2,346$) significantly improved the predictive power ($\log_2($mRNA$) \sim \log_2($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 affects 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. 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.

---

**Synonymous Mutations Affect mRNA Level.**

*doi:10.1093/molbev/msx229*
CAI exhibits larger variation among these genes, we speculate that CAI plays a more important role in regulating the RNA level of these genes (supplementary fig. S14B, 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 pGAL-TDHC-LEU2-TGAL1 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 (MATa 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 YPEG media (1% yeast extract, 2% peptone, 2% galactose, 2% ethanol, and 2% glycerol) at OD660 0.05 and was harvested at OD660 ~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 x 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 cutoff of at least 64 reads in both replicates of the DNA 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

where r_i and d_i are the read counts of variant i in the RNA library and DNA library, respectively. Therefore, the normalized log2(mRNA) of variant i was calculated as

where mRNA_i' and log2(mRNA)' is the average of all log2-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

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_j reflects the fraction of cells containing variant i falling into bin j in FACS. The protein level of variant i

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 (Bazzini et al. 1997).

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


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.


