Modulation of Translation Efficiency in *S. cerevisiae* by Codon Pairs and mRNA Structure

Caitlin E. Gamble

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Stanley Fields, Chair

Maitreya J. Dunham

Adam P. Geballe

Program authorized to offer degree:

Molecular and Cellular Biology

University of Washington

Abstract

Modulation of Translation Efficiency in *S. cerevisiae* by Codon Pairs and mRNA Structure

Caitlin E. Gamble

Chair of Supervisory Committee:
Professor Stanley Fields
Department of Genome Sciences

Synonymous codon choice modulates translation, but the properties of codons or codon combinations that result in impaired translation are not understood. We scored expression of 35,811 three-codon insertions in GFP in *Saccharomyces cerevisiae* and evaluated these variants for codon usage and RNA structure effects on GFP fluorescence levels. We have established that codon pairs affect translation elongation and efficiency in yeast in a manner distinct from the effects of individual codon tRNA abundance. Also, similar to previous studies in bacteria, we have found that the base-pairing status of nucleotides near the translation start site is likely to impair translation initiation. Both inhibitory codon pairs and 5' mRNA structure can impose substantial limitations on translation efficiency through synonymous variation. For 17 inhibitory codon pairs, we show that it is the pair, rather than the dipeptide, the 6-base sequence, or the two individual codons, that is responsible for inhibition. Variants from the GFP insertion library that

had an inhibitory pair had significantly lower expression than variants in which: the 6 base sequences were out of frame; the two codons were present but separated; or one of the codons of the pair was instead an optimal codon. We find that the inhibitory pairs act in translation, based on both suppression of inhibition by over expressed tRNA (11/12 tested) and the reduction in translation speeds relative to synonymous dipeptide sequences, as observed from ribosome occupancies along yeast transcripts. Furthermore, for 12 of the 17 pairs, preserving the order of codons in the pair was required for strong inhibitory effects. Thus the position of inhibitory pairs within the ribosome is likely a key factor in translation efficiency. Moreover, the identity of codons in inhibitory pairs is inconsistent with an inhibition mechanism governed primarily by limited tRNA supply. Rather, our data implicates wobble decoding and interactions between adjacent sites in the ribosome. The high-throughput experimental analysis described here has resulted in the direct and extensive identification of multiple inhibitory codon pairs, a quantitative analysis of their relative effects on translation in vivo, and tests of their activity as modulators of translation.

Table of Contents

List of Figures and Tables

Acknowledgements

Thank you to my advisor, Stan Fields. I appreciate the tremendous opportunities and support Stan has provided me over the years. Through Stan's steadfast, observant, and thoughtful mentorship I've learned a great deal as both a scientist and person. I'm especially grateful for his patience in seeing me through challenging times.

Thank you to everyone in the Fields Lab, both past and present. I feel lucky to have worked with such bright, wonderful people and to have made dear friends among my co-workers. I'm glad for the encouraging and fun lab environment. Thank you, especially, to Lea Starita for sharing wisdom in science and life.

Throughout my time as graduate student, I have appreciated the interest, involvement, and advice of many great advisors. Thank you to my committee members: Maitreya Dunham, Adam Geballe, David Morris, and Christine Queitsch. I've also had the good fortune to be part of a collaborative and inspiring community in the Department of Genome Sciences. Thank you to the faculty, post-docs, students, and staff of the Department.

Finally, thank you to my family, friends, and my significant other, Gabe, for providing much needed balance, care, and love. I love you all dearly.

CHAPTER 1: Introduction

Protein Production is Central to Cellular Life

There are many steps in the process of a cell producing a protein including: generating mRNA transcripts, processing the mRNA, exporting the mRNA to the cytosol, initiating translation by ribosomes, elongating and terminating the amino acid chain, adding post-translational modifications, degrading proteins with potentially hazardous mistakes, achieving a functional protein fold, and localizing the protein to its proper place the cell environment. At each step, the efficiency and accuracy of the process carries consequences for the final protein product. As a key cellular process, protein production touches on many aspects of biology, from disease predisposition in individuals to the cost and efficacy of protein-based diagnostics, therapies, and vaccines.

In the following chapters, I will focus on one major component of protein production: the process of translating mRNA into a sequence of amino acids. Translation elongation shapes the cellular proteome, influencing the amount of protein produced per mRNA and folding of nascent proteins (Dinman, 2012; Gingold and Pilpel, 2011; Ingolia et al., 2009a; Thanaraj and Argos, 1996a). Translation in eukaryotic cells typically begins once the small ribosomal subunit has entered the 5' end of an mRNA transcript and scanned through the 5' untranslated region (UTR) to identify the first occurrence of an AUG start codon. Upon pausing of the small subunit at the AUG site, GTP hydrolysis releases the initiation factors and the 60S subunit joins to form a complete 80S initiation complex (Kozak, 2005). With the arrival of a tRNA matching for the subsequent codon, the ribosome begins elongating the amino acid chain. During each translation elongation cycle, the ribosome coordinates the interaction of each codon in the mRNA with the

anticodon of a cognate tRNA, resulting in the insertion of an amino acid, followed by a precise three-base translocation of the mRNA with the cognate tRNA to maintain the reading frame.

Structure is Reduced Near the Translation Start Site of Many mRNA Transcripts

Within an mRNA molecule, numerous intramolecular hydrogen bonds form between the nucleotides: three bonds between each guanine (G) and cytosine (C) pair, two bonds between adenine (A) and uracil (U) pairs, as well as two bonds between U and G pairs. Through the stacking of base pairs to form stems of paired bases and loops of unpaired bases, an mRNA moves toward energetically favorable secondary structure formations. Translation of mRNA requires separation of these base pairs for decoding by the ribosome and for the single-stranded RNA molecule to pass through the ribosome translation complex.

Relatively high levels of structure in downstream portions of coding sequences (Mortimer et al., 2014) and the helicase activity of elongating ribosomes (Satchidanandam and Shivashankar, 1997; Takyar et al., 2005) support the idea that ribosomes move through moderately stable structures with little disruption during elongation cycles. The 3-codon periodicity of base-pairing within coding sequences, which has been observed in yeast open reading frames (ORFs), further suggests that the presence of some structural features may even contribute to translation efficiency (Kertesz et al., 2010; Mortimer et al., 2014). However, in contrast to moderate structure within coding sequences, the 5' ends of open reading frames (ORFs) in many organisms have a reduced tendency to form stable mRNA structure (Mortimer et al., 2014; Shabalina, 2006; Wan et al., 2012). This observation suggests selective pressures act against formation of strong structure near the translation start site, perhaps because of its interference with translation initiation.

The Genetic Code Has Redundancies in Amino Acid Specifications

Both mRNA structure and translation elongation are influenced by the choice of synonymous codons, which specify insertion of the same amino acid. Synonymous codons differ from each other in their relative use in the genome, in the abundance of the tRNAs that decode them, and in the requirement of some codons for wobble interactions (non-Watson Crick base pairing) between the third base of the codon and the first base of the tRNA anticodon (Gingold and Pilpel, 2011). Thus, although a set of synonymous codons encodes the identical amino acid, variation in a gene's synonymous codons can lead to subtle alterations in protein production and can exert significant phenotypic effects.

In the universal genetic code, 64 codons encode 20 amino acids and a translation termination signal. One to six synonymous codons code for the same amino acid. Specifically, in the genetic code two amino acids are specified by a single codon. Nine amino acids are specified by two different codons. One amino acid is specified by three codons. Five amino acids have four synonymous codon possibilities, and three amino acids have 6 synonymous codon possibilities.

Some tRNAs decode more than one codon. Francis Crick first proposed the Wobble Hypothesis for how tRNAs could read more than one codon, based on insights into the stereochemistry of anticodon-codon pairing (Crick, 1966). He predicted five non-canonical base pairs (G:U, U:G, I:U, I:C and I:A), called "wobble" base pairs. Wobble decoding is universal and obligatory at some codons since all organisms use 49 or fewer tRNAs to decode the 61 codons (Grosjean et al., 2010). However, the particular codons that exclusively rely on wobble base pairing for their decoding varies between species (Chan and Lowe, 2015).

The yeast *S. cerevisiae* has 272 tRNA genes in its genome. Gene copy number for a given tRNA anticodon species ranges from 1 to 16 copies. For sets of tRNA isoacceptors, which bind to alternate codons for the same amino acid, tRNA gene copy number ranges from 4 to 21 (Chan and Lowe, 2015) (GtRNAdb: http://gtrnadb.ucsc.edu/). A total of 20 codons, coding for 15 amino acids, rely on a wobble mechanism for decoding (**Figure 1.1**). Specifically, there are 9 codons that rely on a G:U (codon base : anticodon base) wobble. For 6 of these 9 codons, the codon has only one other synonymous codon, and thus, a single tRNA anticodon species is used to decode the amino acid. A total of seven codons rely on an I:C wobble. Three codons rely on a U:G

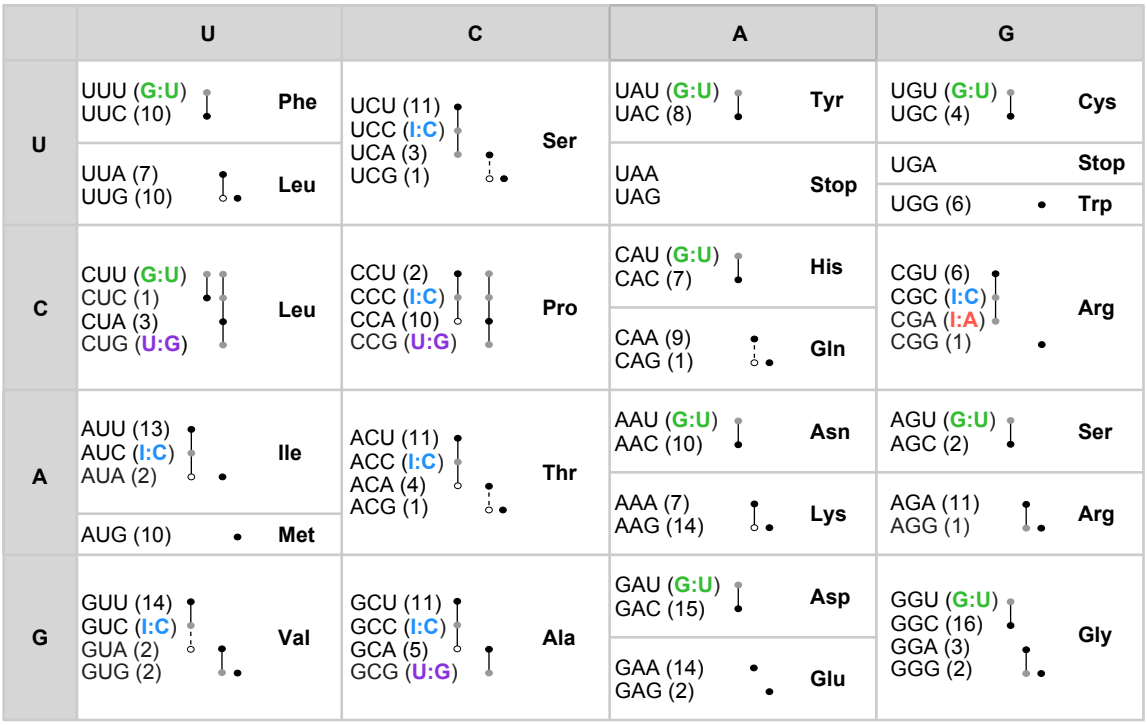

Figure 1.1 The Universal Genetic Code and *S. cerevisiae* tRNA
Parentheses next to each codon indicate either tRNA gene copy number or wobble type (anticodon:codon). Each black dot is a tRNA with exact base-pairing to the codon; lines connecting to gray dots indicate wobble decoding possibilities; white dots indicate rare decoding possibilities. Those connected by dashes generally only occur if the tRNA is over expressed (Johansson et al., 2008).

wobble, and one codon, CGA, is decoded by a rare I:A wobble.

Selective Forces Shape Synonymous Codon Usage in the Genomes of Many Organisms

Although synonymous codons are expected to produce the same amino acid sequence, synonymous codons for an amino acid are not used equally within genomes. Early studies calculating the frequency of codon use revealed a species-specific bias towards certain codons (Bennetzen and Hall, 1982; Grantham et al., 1981). Biases toward certain codons are magnified in highly expressed genes (Sharp and Li, 1987) and correlate with gene expression levels on a genome-wide scale (Hiraoka et al., 2009). In addition to the biased use of individual codons, organisms in all three kingdoms of life display various forms of bias with respect to codon context, the nucleotide sequence surrounding the use of an individual codon (Fedorov et al., 2002; Moura et al., 2005; Plotkin and Kudla, 2010).

In the genomes of bacteria, yeast, worms, flies, and plants, highly expressed genes tend to use a subset of preferred codons (Plotkin and Kudla, 2010). The approximate tRNA abundance correlates with the frequency of codon use in highly expressed genes (Ikemura, 1981; 1982) and with tRNA gene copy number (Percudani et al., 1997; Tuller et al., 2010a). Yet, a central question remains about the direction of causality in these relationships (Novoa and de Pouplana, 2012; Plotkin and Kudla, 2010). The mechanisms by which synonymous codon usage relates to gene expression are not fully understood.

Given the importance of protein production to the cell, researchers have debated since the 1960's how the codon usage in an organism's coding sequences reflects selective pressures (Ames and Hartmann, 1963) on protein production efficiency and accuracy at both local (gene-specific) and genome-wide scales. The observation that the frequency of codon use in highly expressed genes correlates with tRNA abundance in bacteria and yeast (Ikemura, 1981; 1982;

Tuller et al., 2010a) has shaped much of the thought on this relationship. Differences in tRNA abundance imply that codons corresponding to abundant cognate tRNAs are recognized and translated more quickly. From the postulate that decoding rates are variable and governed by tRNA abundance, it follows that the additive effects of faster decoding should lead to more protein product (Zhou et al., 2004), facilitate efficient allocation of limited translation resources (Stoebel et al., 2008), and/or reduce synthesis errors as a result of reduced opportunity for erroneous, near-cognate binding to occur (Drummond and Wilke, 2008; Zhou et al., 2009). Yet the relationship between evolutionary forces and codon usage bias, as well as the mechanisms by which codon usage affects translation efficiency and accuracy are not fully understood and remain topics of much debate (Novoa and de Pouplana, 2012; Plotkin and Kudla, 2010).

Synonymous Variation Gives Rise to Diverse Outcomes

The compelling case that codon choice modulates translation efficiency is substantiated by the significance of genome-wide correlations between codon use and translation efficiency (Brockmann et al., 2007; Ghaemmaghami et al., 2003; Ishihama et al., 2008; Tuller et al., 2010b), by correlations between codon use and differential production of operon-encoded components of protein complexes (Quax et al., 2013), and by numerous examples in which recoding genes with optimal or suboptimal codons changes expression as predicted (Chu et al., 2013; Gustafsson et al., 2004; Presnyak et al., 2015; Welch et al., 2009). On the other hand, *how* codon choice modulates translation efficiency is not fully understood.

While recoding a gene with optimal codons can sometimes lead to gene expression improvements, as seen in numerous *E. coli* (Gustafsson et al., 2004) and yeast examples (Keppler-Ross et al., 2008; Quartley et al., 2008), the extent to which expression is improved

varies, and overall improvement is not guaranteed. In comparing the results of several efforts to improve the expression of heterologous genes by recoding the genes with host optimal codons, Gustafsson et al. (Gustafsson et al., 2004) found that the results were highly variable, with gene expression improving from not at all to 10^5-fold. Thus, simply recoding genes with optimal codons is not sufficient to reliably improve expression.

Determining the relationship between synonymous codon use and translation efficiency is complicated by the fact that the effects of codons on translation are almost certainly dependent upon additional sequence parameters including interactions with adjacent codons (Boycheva et al., 2003; Gutman and Hatfield, 1989; Irwin et al., 1995; Letzring et al., 2010; Moura et al., 2005), nucleotide context (Fedorov et al., 2002; Moura et al., 2005; Plotkin, 2011; Yarus and Folley, 1985), location in the gene (Letzring et al., 2010; Pechmann and Frydman, 2012; Tuller et al., 2010a; 2010b; Wolf and Grayhack, 2015), and biased co-occurrence of codons decoded by the same tRNA isoacceptor (Cannarozzi et al., 2010).

In addition to the complexity introduced by sequence context factors, determining the relationship between synonymous codon usage and translation efficiency is further complicated by the fact that changes in synonymous codon usage impact sequence features not directly related to the efficiency with which codons are translated, including mRNA structure, binding sites for RNA-binding proteins and microRNAs, splicing signals, and even transcription factor binding sites in the DNA (Weatheritt and Babu, 2013).

Still further complication arises from the fact that variation in translation efficiency itself can give rise to diverse outcomes beyond the quantity of protein synthesized. Sometimes these outcomes may be beneficial and at other times detrimental to protein production. For instance, suboptimal codon use can affect feedback on translation initiation rates (Chu et al., 2013; Hersch

et al., 2014) or cause recruitment of quality control systems (Letzring et al., 2013). There are numerous examples in which codon use modulates protein folding (Thanaraj and Argos, 1996b; Xu et al., 2013; Zhang and Ignatova, 2009; Zhou et al., 2013). These examples include the outcome of folding competition within a fluorescent reporter (Sander et al., 2014) and the substrate specificity of the P-glycoprotein molecular pump encoded by the *MDR1*/*ABCB1* gene (Kimchi-Sarfaty et al., 2007). In some cases, reduced translation rates can increase functional protein yield by facilitating proper folding of a protein subdomain before further extension of the amino acid chain (Zhang et al., 2009). Additionally, studies on the association of codon choice with features of the amino acid suggest that codon choice at key positions may influence accuracy of translation (Akashi, 1994; Stoletzki and Eyre-Walker, 2006). Accurate translation of residues in the protein core (Zhang and Ignatova, 2009) could prevent the most harmful instances of protein misfolding. Finally, although the mechanism is not understood, suboptimal codon use may drive mRNA decay. This conclusion is based on a genome-wide correlation between mRNA half-lives and codon use and substantiated by evidence that changing codon use alters mRNA decay rates as expected (Presnyak et al., 2015).

In summary, synonymous codon choice is implicated in critical roles in translation efficiency, accuracy, protein folding, mRNA folding, motif recognition, and even mRNA decay. Given the gene-specific nature of these effects, they are challenging to identify. Without a solid mechanistic understanding of the primary effects of codon choice on translation elongation efficiency, we are limited in our ability to both detect and interpret the functional implications of coding sequence variation.

Challenges in Reaching a Comprehensive Understanding of the Genetic Code

A major impediment to understanding how codons affect translation efficiency has been the lack of a direct means to identify the specific codons and contexts that reduce translation efficiency. Neither the identity of the codons or codon combinations that inhibit effective elongation nor the parameters that modulate these effects are fully understood. This information is crucial to determine the location of codon-mediated regulatory sites in the genome, as well as the mechanisms by which synonymous codon choice affects translation elongation.

Three main challenges to understanding how codons affect translation efficiency are described above: 1) the contribution of sequence parameters outside of the individual codon (e.g. adjacent codons and location in the transcript); 2) the impact of synonymous variation on other functional aspects of the coding sequence not directly related to the translation of specific codons (e.g. mRNA structure and protein-binding motifs); and 3) the diverse possible outcomes beyond absolute protein quantity that may arise from variation in translation rates (e.g. feedback on initiation and modulation of protein folding). Many of our current theories on synonymous codon usage and fitness derive from empirical observation of genomes. The identification of modulatory sequences based on evolutionary conservation is difficult, because selection acts on many features in addition to the efficiency with which codons are translated. Given that genome-wide analysis of systematic biases cannot fully resolve how synonymous codons affect translation efficiencies, there is a need for experimental comparisons of codon usage factors in order to validate codon usage theories. However, since there are many synonymous coding possibilities (for instance, 3,721 possible codon pairs encode 400 dipeptides), even the experimental identification of inhibitory codon pairs is problematic. Mid-throughput approaches that examine 20 to 154 variants of a single amino acid sequence (Kudla et al., 2009; Welch et al.,

2009) are limited in the conclusions that can be drawn. High-throughput, experimental approaches (Goodman et al., 2013) are needed both to explore a fuller range of coding sequence possibilities as well as to compare and weigh conflicting hypothesis about the nature of observed effects.

Dissertation Objectives

I sought to further our understanding of the selective forces behind synonymous codon usage and the mechanistic impacts on cellular translation dynamics. In the following chapters, I have taken advantage of the sensitivity and massively parallel capabilities of current DNA sequencing technology to investigate the relationship between synonymous codon usage and translation efficiency and to pursue the following aims:

1) **Identify specific codons and pairs of codons likely to reduce translation efficiency**

   We reasoned that extensive analysis of a small region could be used to identify codons or codon combinations that reduce gene expression in the yeast *S. cerevisiae*. By using fluorescence activated cell sorting (FACS) and deep sequencing to measure expression of >35,000 variants of GFP, in which three codons near the 5' end of the coding sequence were randomized, I have identified contiguous 6-base sequences enriched in low expression variants.

2) **Evaluate the ribosome footprints from yeast transcripts for evidence of reduced translation speed at inhibitory codon pair sites**

Ribosome profiling, which uses sequences derived from ribosome-protected mRNA fragments (footprints) to infer ribosome location, provides a relative measure of *in vivo* decoding rates. I have developed a computational analysis to examine inhibitory codon pair sites for evidence of ribosome pausing.

3) **Determine the impact of wobble decoding on inhibition by codon pairs and assess whether pair inhibition is likely to occur through sequential individual codon effects or a concerted pair effect.**

   Using the GFP library results and ribosome profiling datasets, I have further examined the hypotheses that inhibitory effects of codon pairs arise largely from wobble-decoding interactions, rather than limited tRNA, and that inhibition is a result of a pair effect, rather than the sum of individual codon effects.

4) **Characterize the relationship between 5'-end mRNA secondary structure and expression level within the library of GFP variants**

   I have applied mRNA structure prediction tools to investigate the impact of structure on GFP expression levels to and to characterize when structure is most likely to impact expression.

Through the experiments undertaken to accomplish these aims, I and my collaborators at the University of Rochester have identified 17 inhibitory codon pairs associated with low expression and examined their effects on translation. We find 12 codon pairs that substantially reduce *in vivo* translation rates along yeast transcripts, 10 of which work in only a single orientation, as

expected if inhibition is due to the codon pair and is not the sum of effects by individual codons. In addition, we describe predicted mRNA structural features near the start codon that increase the likelihood of low expression. With regard to translation initiation, we conclude that expression levels are impacted when base pairs immediately downstream of the start codon are incorporated into strong base pair stems. With respect to the impact of synonymous codon use on translation elongation, we conclude that elongation rates are modulated by the combined effects of the tRNA:codon interactions in two adjacent sites in the ribosome.

CHAPTER 2: Adjacent Codons Act in Concert to Modulate GFP Translation in Yeast

2.1 BACKGROUND

Interactions between neighboring codons were first implicated in translation efficiency based on observations of non-randomness in the mRNA sequence surrounding individual codons within sets of high and low expression *E. coli* genes (Gutman and Hatfield, 1989; Yarus and Folley, 1985). The composition of nucleotides and adjacent codons immediately surrounding a given codon is refered to as codon context. A variety of mechanisms are likely to contribute to codon context bias, including translation efficiency and accuracy as well as mutation bias, GC content bias, DNA polymerase slippage, and constraints imposed by both mRNA structure and protein composition. Although codon context bias has received comparatively little attention in the literature compared to individual codon bias, organisms in all three kingdoms of life display various forms of biased codon context (Fedorov et al., 2002; Moura et al., 2005; Plotkin, 2011). In particular, non-random use of codon pairs is independent of individual codon-usage and dipeptide biases. It correlates with gene expression in *E. coli* (Boycheva et al., 2003; Irwin et al., 1995), and it is found in other organisms (Buchan, 2006).

Initial experimental evidence implicating codon context in translation accuracy came from findings that neighboring nucleotides influence the suppression of missense mutations and termination of translation in *E. coli*, yeast, and human cells (Bonetti et al., 1995; Bossi, 1983; Murgola et al., 1984; Phillips-Jones et al., 1993). For example, asparagine AAU codons are frequently misread as lysine in starved *E. coli* cells, the context of an AAU codon had as much as a two fold impact on the frequency of misreading (Precup and Parker, 1987).

Similarly, evidence for codon context influences on translation efficiency have come from several studies. Chevance et al. demonstrated that nucleotides and codon pairs neighboring a UCA codon in the HisT leader peptide impact the rate of translation elongation in *Salmonella enterica* (Chevance et al., 2014). They further identified arginine-arginine pairs with varying degrees of translation efficiency (Chevance et al., 2014). Letzring et al. found that adjacent arginine CGA codons inhibit translation in *S. cerevisiae* more effectively than individual CGA codons (Letzring et al., 2010). In human cells, Coleman et al. found that recoding viral genes with hundreds of underrepresented codon pairs reduced expression and lead to attenuated viruses (Coleman et al., 2008).

These observations provide evidence that interactions between sites in the ribosome play important roles in regulating translation. However, such examples are limited to only a few experimentally tested cases. The extent to which codon context influences translation has remained virtually unknown. A major impediment to understanding translational control mediated by codon choice has been the lack of an unambiguous method to specifically identify codons or codon combinations that reduce translation efficiency. We reasoned that experimental analysis of an extensive set of variants in a small region could identify codon combinations that reduce gene expression in the yeast *S. cerevisiae*. Large synthetic libraries of a reporter gene provide a robust tool for evaluating functional impacts of sequence variation (Goodman et al., 2013; Kudla et al., 2009; Welch et al., 2009) and allow for rigorous alternative hypothesis testing. Thus, we sought to identify and analyze an extensive set of inhibitory codon pairs by using fluorescence activated cell sorting (FACS) and deep sequencing to measure the expression of 35,811 GFP variants in which three adjacent codons near the 5' end of the coding sequence

were randomized. We identified 17 codon pairs associated with low expression and examined their effects on translation.

2.2 RESULTS

We created two libraries of yeast strains, with either $(VNN)_3$ [V= A, C, or G] or $(NNN)_3$ sequences inserted at amino acids 6-8 of a fusion protein encoding superfolder GFP (**Figure 2.1**). The $(VNN)_3$ library was created to limit the number of nonsense codons. We used the chromosomally integrated yeast RNA-ID reporter (Dean and Grayhack, 2012), in which a bidirectional *GAL1,10* promoter separately drives expression of both the GFP variant and RFP, to normalize GFP expression to that of RFP and thereby control for transcriptional effects. This approach seemed likely to comprehensively define interactions between adjacent codons, since each codon pair, the reverse of each codon pair, and the two individual codons would be represented many times in different contexts.

Analysis of 35,811 GFP Variants Reveals 21 Codon Pairs Linked to Reduced Expression

To detect differences in GFP expression and identify sequences that substantially inhibit yeast translation, we used fluorescence-activated cell sorting (FACS) to separate yeast cells into three fluorescence bins. For the $(NNN)_3$ library, we made and separately sorted two independent yeast libraries. We estimated that the assay detected expression levels in the high bin from ~75-100% of a no insert reference GFP, in the intermediate bin from ~25-75% (median 43% of the high bin median), and in the low bin from ~2.5-25% (median 6% of the high bin median). Following FACS, we sequenced the variable three-codon insertions from cells in each bin, carried out quality filtering, and determined the relative distribution of sequences. We estimated mean

expression (GFP^{SEQ}) for each sequence based on the sequence's distribution across bins and applying the median fluorescence of a bin to all reads counts in that bin. GFP^{SEQ} scores correlated across the 3 yeast sorting libraries (r = 0.91 to 0.93) (**Figure 2.2A**) and with mean

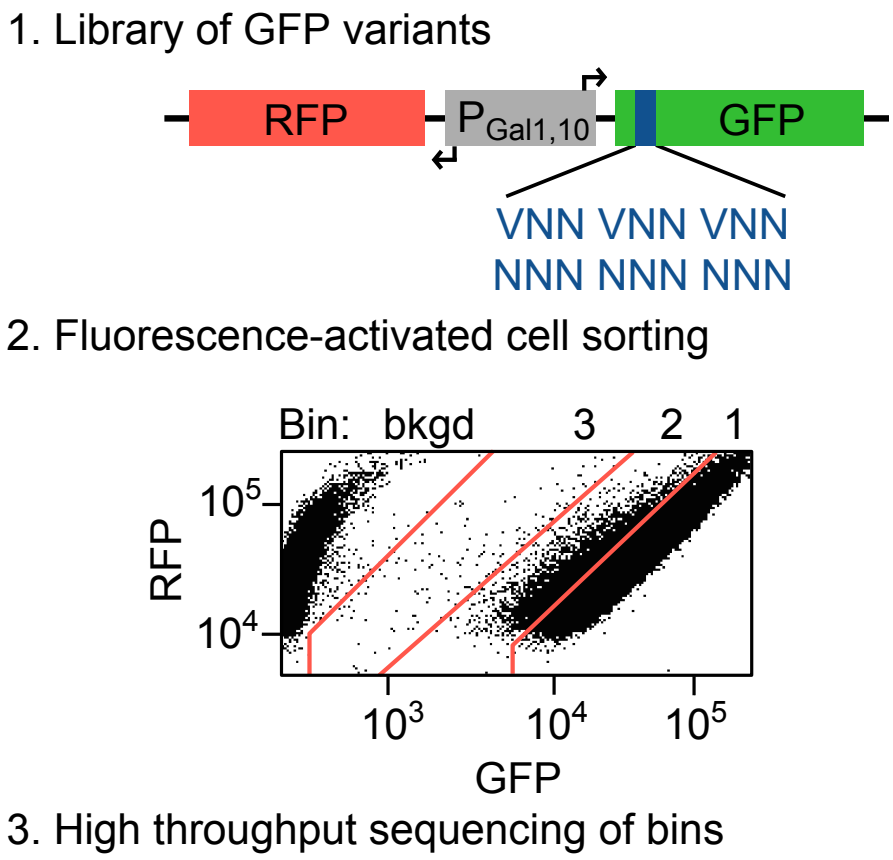

Figure 2.1 Method to Assay GFP Library Expression
Schematic of method to examine effects of three randomized codons on superfolder GFP expression, using the RNA-ID reporter. The randomized codons were inserted at codon position 6. FACS of yeast was used to sort GFP variants into GFP/RFP fluorescence bins, followed by deep sequencing of the variants in each bin. Shown is the FACS sort of (NNN)3 Library 1.

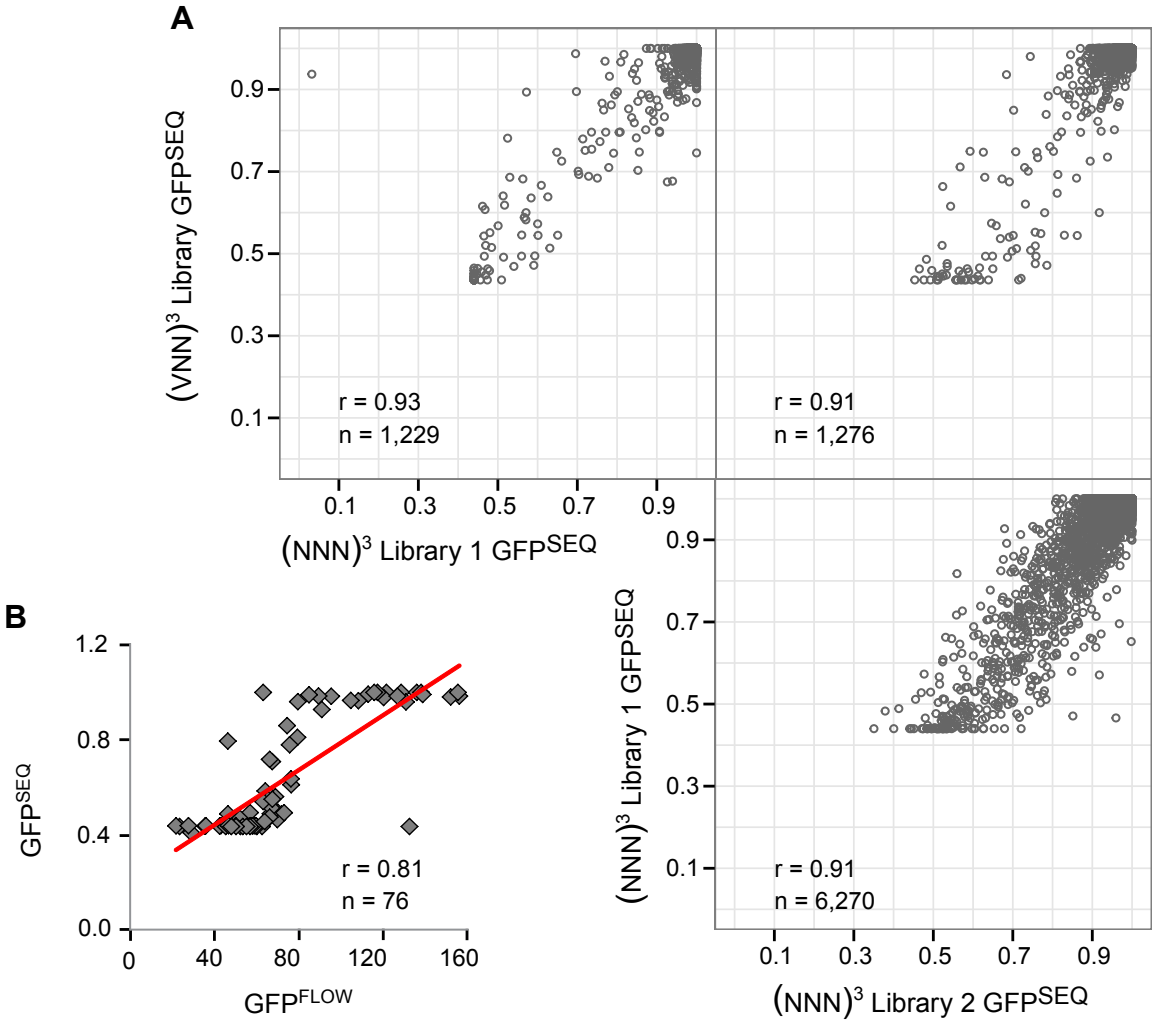

Figure 2.2 GFP^{SEQ} Correlation Between Libraries and with GFP^{FLOW}
(A) GFP^{SEQ} Pearson correlation between the 3 yeast sorting libraries.
(B) Correlation between GFP/RFP flow cytometry (GFP^{FLOW}) of 76 individually constructed variants and GFP^{SEQ} of library variants.

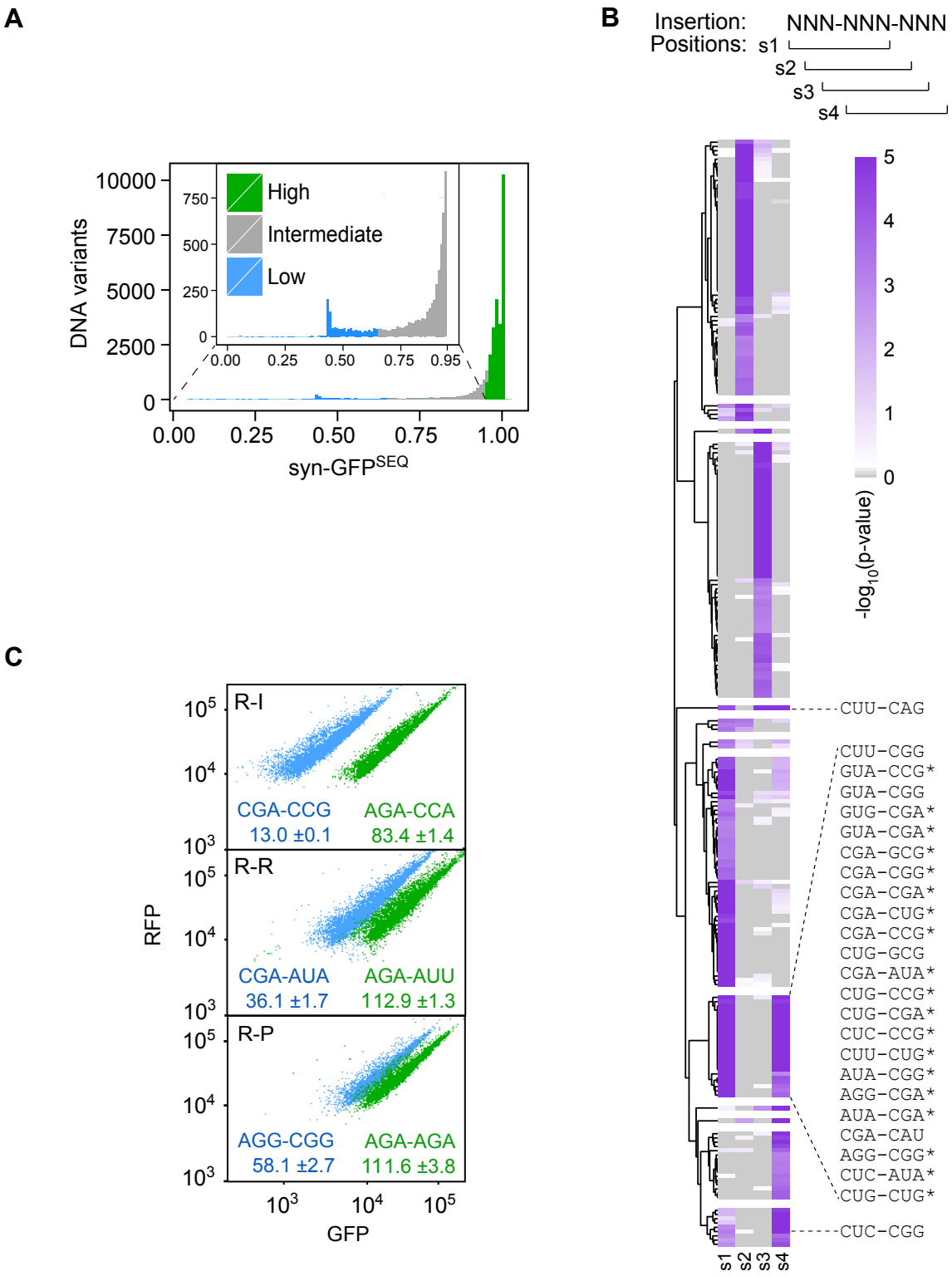
A
DNA variants
10000
7500
5000
2500
0
High
Intermediate
Low
750
500
250
0
0.00
0.25
0.50
0.75
0.95
0.00
0.25
0.50
0.75
1.00
syn-GFP^SEQ
B
Insertion: NNN-NNN-NNN
Positions: s1
s2
s3
s4
5
4
3
2
1
0
-log10(p-value)
CUU-CAG
CUU-CGG
GUA-CCG*
GUA-CGG
GUG-CGA*
GUA-CGA*
CGA-GCG*
CGA-CGG*
CGA-CGA*
CGA-CUG*
CGA-CCG*
CUG-GCG
CGA-AUA*
CUG-CCG*
CUG-CGA*
CUC-CCG*
CUU-CUG*
AUA-CGG*
AGG-CGA*
AUA-CGA*
CGA-CAU
AGG-CGG*
CUC-AUA*
CUG-CUG*
CUC-CGG
s1
s2
s3
s4
C
R-I
CGA-CCG
13.0 ±0.1
AGA-CCA
83.4 ±1.4
R-R
CGA-AUA
36.1 ±1.7
AGA-AUU
112.9 ±1.3
R-P
AGG-CGG
58.1 ±2.7
AGA-AGA
111.6 ±3.8
RFP
GFP
10^3
10^4
10^5

Figure 2.3 Some Synonymous Variants Have Substantially Lower Expression
(A) Distribution of syn-GFPSEQ scores. Variants were assigned to low (blue; n = 1119), intermediate (gray; n = 5127), and high (green; n = 24417, excluding high expression synonymous references) expression categories. (B) Significance of 6-mer enrichment in low expression variants at each 6-mer starting position (s1-s4) in the 9-base variable region. Values are plotted based on hierarchical clustering of permutation p-values for 6-mers with a p-value ≤ 0.001. Fifty-seven 6-mers with a significant value are not plotted due to missing values. 6-mers with a p-value ≤ 0.001 at both in-frame start positions are labeled (CUG-AGG, CUG-AUA*, and CUU-AGG not plotted because they form stop codons in another frame). Stars indicate candidate inhibitory pairs that remain enriched in a reduced structure set (C) Flow cytometry scatter plots showing GFP^{FLOW} of three individual variants with a putative inhibitory codon pair (blue) is reduced in each case compared to a synonymous variant with an optimized pair (green).

GFP expression of 76 individual constructs measured by flow cytometry (GFP^{FLOW}) (r = 0.81), although binning limits the resolution (**Figure 2.2B**).

We considered that the amino acid sequence encoded by the insert could affect GFP stability, although most of these effects should be mitigated by using superfolder GFP, which has robust fluorescence when fused to several insoluble proteins (Pédelacq et al., 2005). Thus for downstream analysis, we included only the 35,811 unique DNA sequences specifying one of 5,148 tripeptides that had at least one synonymous sequence above the mean of all GFP^{SEQ} scores (leaving out 4.1% of tripeptides and 6.2% of DNA variants). For each DNA sequence, the highest scoring sequence encoding a synonymous peptide served as its synonymous reference. We scored expression due to codon usage (syn-GFP^{SEQ}) as the GFP^{SEQ} ratio of a given sequence and its synonymous reference.

As expected, most synonymous variants had similar expression (**Figure 2.3A**), with a mean syn-GFP^{SEQ} of 0.954. However, 1,119 DNA sequences (low variants) showed extreme expression differences, with syn-GFP^{SEQ} ranging from 0.059 to 0.648 (3 standard deviations or more from the mean). Intermediate variants comprised 5,127 sequences between 0.648 and

0.954; and high variants comprised 24,417 non-reference (as well as 5,148 reference sequences) with syn-GFP^{SEQ} greater than 0.954.

There were no examples in which use of an individual codon consistently reduced expression to a degree detectable in our assay. The syn-GFP^{SEQ} medians for each set of variants containing 1 or more copies of a given codon ranged from 0.97 to 1.00, but the use of broad expression bins limited our ability to detect relative differences in GFP^{FLOW} values between 75% and 100% of the reference GFP. A subset of codons occurred frequently in low variants (**Table 2.1**), suggesting that combined use of particular codons may dramatically reduce expression.

Codon	Amino Acid	Frequency	CAI	Wobble
CGA	R	0.092	0.002	I•A
CUG	L	0.077	0.003	U•G
CGG	R	0.071	0.002	-
AGG	R	0.060	0.003	-
CUU	L	0.047	0.006	U•U
GUA	V	0.043	0.002	-
GUG	V	0.041	0.018	-
CCG	P	0.035	0.002	U•G
AUA	I	0.034	0.003	-
CUC	L	0.032	0.003	-

Table 2.1 Frequency of Codon Use in Low Variant Insertions
The top 10 frequencies are shown.

To identify inhibitory codon pair candidates, we looked for combinations of adjacent codons enriched in the low variant category. We found 293 six-base sequences (non-gapped 6-mers) enriched at one or more of the four possible starting positions in the low variants of the 9-base insertion libraries (permutation p-value ≤ 0.001); 28 of these 6-mers (0.75% of all possible 6-mers without a stop codon) were enriched at both in-frame start positions (permutation p-value ≤ 0.001 for each position) and comprised our initial list of inhibitory codon pair candidates (**Figure 2.3B**).

Since strong RNA secondary structure in the 5' end of an open reading frame can reduce translation efficiency in *E. coli* and *S. cerevisiae* (Goodman et al., 2013; Kudla et al., 2009; Shah et al., 2013; Tuller et al., 2010b), we investigated whether enrichment of each 6-mer in low expression variants was explicable primarily by formation of strong secondary structure. We identified a reduced-structure subset of variants as those with a similar degree of structure to the majority of high expression variants, based on both local and global structure predictions (see **2.5 Methods**). We then evaluated whether each candidate pair remained enriched among low expression variants present in the reduced-structure subset. We found 20 of the 28 candidates remained enriched at in-frame start positions (permutation p-value ≤ 0.055), and revised our candidate list to include only these (**Figure 2.3B** and **Table 2.2**). We conclude that structure is unlikely to account for most of the reduced expression by these 20 candidates.

Expression of GFP variants with a candidate inhibitory pair was substantially reduced, with syn-GFP^{SEQ} medians ranging from 0.44 to 0.82. At least one of the candidate pairs was present in 29% (n = 319) of all low expression variants. We validated the inhibitory effects of the 20 inhibitory codon pair candidates by flow cytometry of individual constructs (**Figure 2.3C**; **Table 2.2**; and **Appendix A**). For each pair, we assessed expression due to codon usage by comparing GFP^{FLOW} of two synonymous variants, one with the inhibitory codon pair and the other with an optimized pair based on codon adaption index (CAI), which assesses the representation of codons in highly expressed genes (Sharp and Li, 1987) . All variants with an inhibitory pair candidate had lower GFP^{FLOW}, ranging from 14-76% that of their synonymous optimized variants.

Candidate Pair	AA	s1 pvalue	s2 pvalue	s3 pvalue	s-4 pvalue	Median syn-GFP^{SEQ}	IQR	n	Inhib. / Opt. GFP^{FLOW}
AGG-CGA	RR	1E-05	1E+00	1E+00	1E-05	0.48	0.31	30	0.42 ±0.02
AGG-CGG	RR	1E-05	1E+00	5E-01	2E-04	0.82	0.46	36	0.52 ±0.03
AUA-CGA	IR	1E-05	1E+00	1E+00	7E-05	0.58	0.30	11	0.39 ±0.01
AUA-CGG	IR	1E-05	1E+00	1E+00	1E-05	0.65	0.43	27	0.64 ±0.02
CGA-AUA	RI	1E-05	1E+00	1E+00	1E-05	0.51	0.29	27	0.34 ±0.02
CGA-CCG	RP	1E-05	1E+00	1E+00	1E-05	0.44	0.05	22	0.15 ±0.01
CGA-CGA	RR	1E-05	1E+00	1E+00	1E-05	0.44	0.06	25	0.19 ±0.01
CGA-CGG	RR	1E-05	1E+00	1E+00	1E-05	0.48	0.13	38	0.35 ±0.01
CGA-CUG	RL	1E-05	1E+00	1E+00	1E-05	0.47	0.31	21	0.46 ±0.01
CGA-GCG	RA	1E-05	1E+00	1E+00	1E-05	0.44	0.03	30	0.26 ±0.01
CUC-AUA	LI	1E-05	1E+00	1E+00	3E-04	0.70	0.45	12	0.45 ±0.06
CUC-CCG	LP	1E-05	1E+00	1E+00	1E-05	0.44	0.04	15	0.14 ±0.01
CUG-AUA	LI	1E-05	1E+00	NA	1E-05	0.71	0.30	22	0.61 ±0.06
CUG-CCG	LP	1E-05	1E+00	1E+00	1E-05	0.49	0.39	30	0.39 ±0.01
CUG-CGA	LR	1E-05	1E+00	1E+00	1E-05	0.50	0.48	25	0.37 ±0.01
CUG-CUG	LL	1E-05	1E+00	1E+00	2E-04	0.66	0.31	25	0.76 ±0.08
CUU-CUG	LL	1E-05	1E+00	1E+00	1E-05	0.74	0.32	27	0.66 ±0.06
GUA-CCG	VP	2E-05	1E+00	1E+00	2E-05	0.80	0.42	25	0.42 ±0.03
GUA-CGA	VR	1E-05	1E+00	1E+00	1E-05	0.53	0.21	36	0.39 ±0.04
GUG-CGA	VR	1E-05	1E+00	1E+00	1E-05	0.60	0.33	30	0.43 ±0.01

Table 2.2 Candidate Inhibitory Codon Pairs
For each candidate, the in-frame dipeptide sequence (AA); permutation p-value for enrichment in low variants at each starting position in the variable region (s1 pvalue, s2 pvalue, s3 pvalue, and s4 pvalue); median syn-GFP^{SEQ} ; syn-GFP^{SEQ} IQR; and total number of in-frame variants (n); is shown with inhibitory:optimal ratio from GFP^{FLOW} (GFP/RFP) measurement of two individual constructs. See APPENDIX A for individual construct sequences and GFP^{FLOW}

Codon Pairs Mediate Frame-Dependent Inhibition

To assess the likelihood that inhibition is mediated by translation, we examined the properties of the candidate pairs. If inhibition is coupled to translation, then the enriched 6-base sequences would likely inhibit expression only when the two codons were in-frame and not when at frame-shifted positions where mechanisms de-coupled from translation might explain enrichment. For the 20 candidates, we compared the syn-GFP^{SEQ} distribution of variants with these candidates at in-frame positions (**Figure 2.4A, blue**) to variants with the candidates at +1 and +2 positions

(**Figure 2.4A, gray)**; 19 pairs had lower syn-GFP^{SEQ} scores when the codon pairs were in-frame (corrected Wilcoxon p-values ≤ 0.008; CUC-AUA not significant). Additionally, we compared syn-GFP^{SEQ} distributions between variants with an inhibitory pair to variants with these codons at non-adjacent positions (**Figure 2.4A, purple)**. If the codon pair mediates translation inhibition, rather than additive single codon effects, then we would expect greater inhibition by adjacent codons. Seventeen of the 20 candidates had lower syn-GFP^{SEQ} scores, when the codons were adjacent (corrected Wilcoxon p-values ≤ 0.008; CUC-AUA**,** CUG-CUG and CUU-CUG not significant). Thus, inhibition by these 17 pairs is dependent upon both frame and adjacent positioning of the two codons.

If inhibition by codon pairs is a general function of their translation by the ribosome, then they should reduce expression when positioned at diverse locations within the coding sequence. However, the magnitude by which codons affect expression can depend upon their location relative to the start of translation; for example, CGA codon repeats are more inhibitory near the start of the coding sequence (Letzring et al., 2010; Wolf and Grayhack, 2015). Therefore, we tested whether inhibition occurs at internal locations by inserting 3 copies of an inhibitory codon pair at amino acid 100 (between an N-terminal *GLN4*$_{(1\text{-}99)}$ domain and GFP) and at amino acid 318 (between *Renilla* luciferase and GFP) (Letzring et al., 2010; Wolf and Grayhack, 2015); we carried out this test for the 12 pairs with the lowest syn-GFP^{SEQ} medians. In each case, GFP^{FLOW} with the inhibitory pairs was lower than with optimized pairs (from 20-67%; **Figure 2.5A**). We also showed that increasing the copy number of the codon pairs from 1 to 3 copies resulted in greater inhibition (three pairs tested at amino acid 6 in **Figure 2.5B** and two pairs tested at amino acid 100 in **Figure 2.5C**). Thus, each of these inhibitory codon pairs mediates reduced expression at internal coding sequence locations.

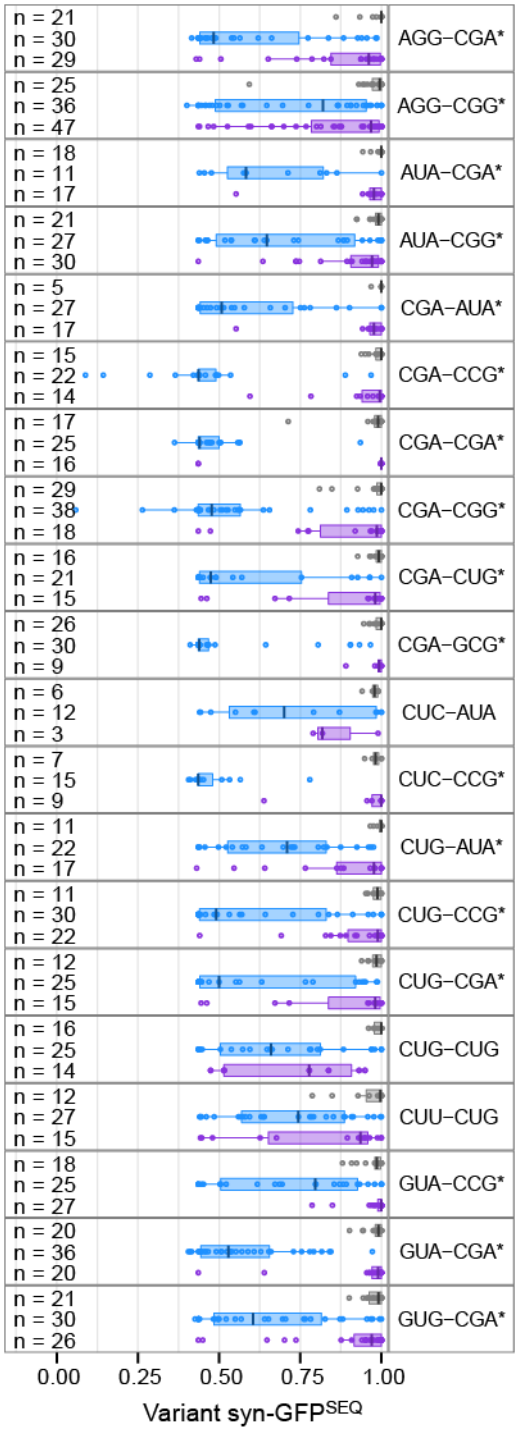

Figure 2.4 Adjacent Codons Mediate Frame-Dependent Inhibition
syn-GFP^{SEQ} distributions of variants with one of the 20 inhibitory codon pair candidates. Variants with an adjacent and in-frame codon pair (blue) are compared to those with the out-of-frame 6-mer (gray) and non-adjacent, in-frame codons (purple). Boxplot shows median centerline and edges mark the first and third quartiles. Inhibitory pairs that depend upon both frame and adjacent positioning (corrected Wilcoxon p-values ≤ 0.008) are indicated with a star.

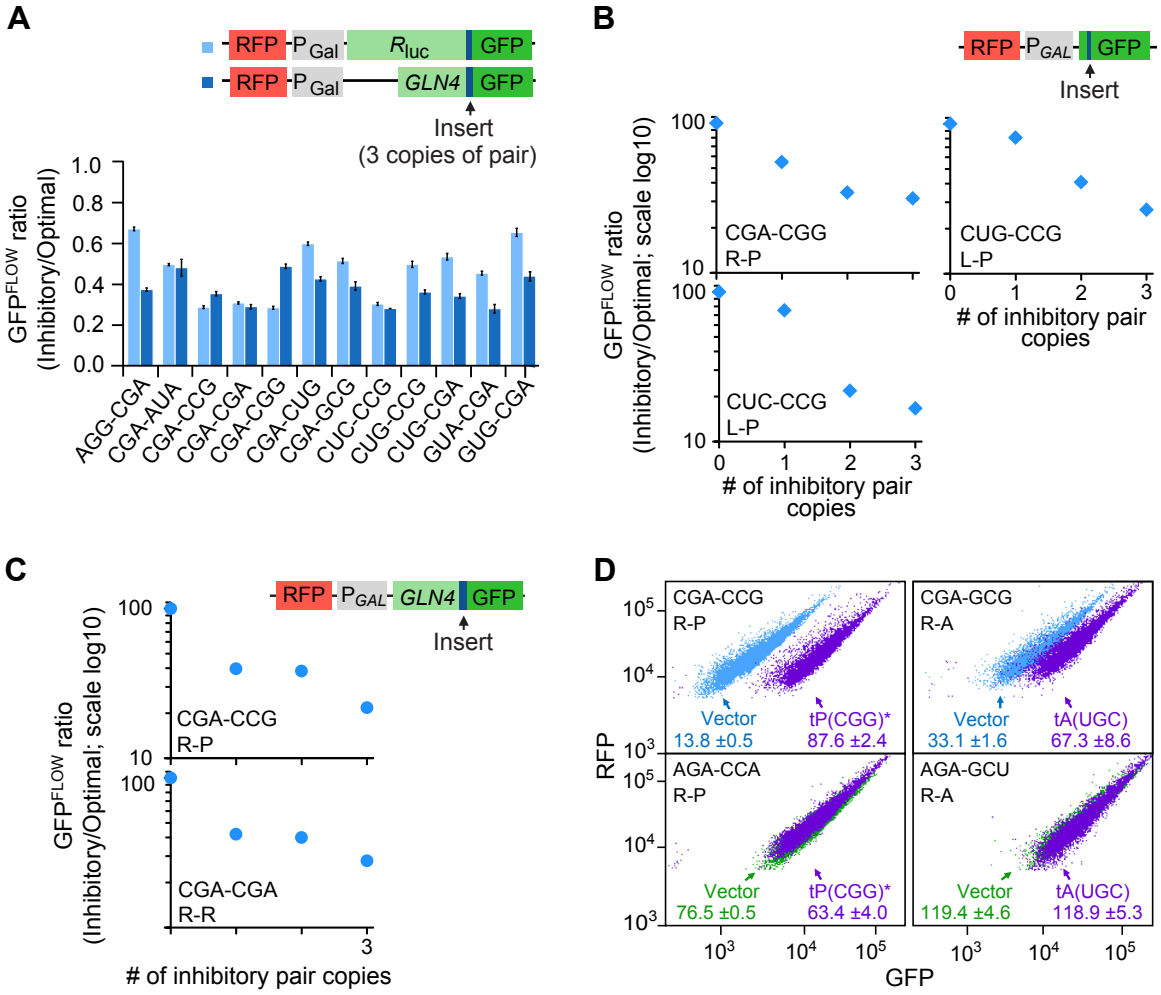

Figure 2.5 Codon Pairs Mediate Translation Inhibition
(A) GFP^{FLOW} from variants with three copies of an inhibitory pair is reduced compared to synonymous variants with three copies of the optimized pair. Codons are inserted at amino acid 318 in Renilla luciferase-GFP (light blue) or at amino acid 100 in GLN4(1-99)-GFP (dark blue).
(B) Inhibition by codon pair copy number for three inhibitory codon pairs at amino acid 6. All points are normalized to GFP^{FLOW} of a synonymous construct with 3 optimized pairs (plotted at 0 codon pairs)
(C) Inhibition by codon pair copy number for two inhibitory pairs. Inserts are positioned at amino acid 100. All points are normalized to GFP^{FLOW} of a synonymous construct with 3 optimized pairs (plotted at 0 codon pairs)
(D) Scatter plots illustrate fluorescence for two sets of synonymous variants with an inhibitory codon pair (top panels) and a synonymous, optimized pair (bottom panels); cells have either with an empty tRNA vector (blue and green) or with a vector expressing the indicated tRNA (dark purple); non-native tRNA is labeled with a star.

Codon Pairs Inhibit Translation Efficiency

To determine if defects in decoding inhibitory codon pairs are responsible for low expression, we evaluated the ability of over-expressed tRNAs to suppress the low GFP^{FLOW} of variants with inhibitory codon pairs. We initially examined suppression by tRNAs that decode the 3' codon of inhibitory pairs, since the 3' codon is likely to occupy the ribosomal A site during the inhibitory reaction. The expression defect for variants with 10 of 12 pairs tested was suppressed either by increasing the abundance of a native tRNA or by expressing a non-native tRNA that enables decoding by Watson-Crick base pairing at all three bases (exact matching) (**Figure 2.5D**; **Table 2.3**). Maximal suppression ranged from a 1.8- to 7.7-fold improvement in GFP^{FLOW} (relative to a synonymous optimized variant) (**Table 2.3**), implying that translation of the inhibitory pair limits expression. For one of the pairs (CGA-CGG) in which tRNA for the 3' codon did not suppress, the expression defect was strongly suppressed by a non-native exact matching tRNA that decodes the 5' codon (**Table 2.3**). Thus, for these 11 tRNA-suppressible pairs, inhibition is due to a translation defect.

2.3 DISCUSSION

We have shown that codon context is an important component of translation efficiency, and we have identified 17 codon pairs that inhibit translation efficiency in yeast. For these pairs of adjacent codons, we show that inhibition is specific to the codon pair and not the dipeptide, 6-base sequences, or presence of the two individual codons. Most variants with an inhibitory codon pair had dramatically lower syn-GFP^{SEQ} scores compared to variants with the same 6 base

sequences in out-of-frame positions, the same two codons with an intervening codon in the middle; or other synonymous pairs with a single codon from the inhibitory pair. Furthermore, we provide three additional lines of evidence that many of these pairs act in translation. First, we

Inhibitory Codon Pair	tRNA Vector	GFP^{FLOW} ratio (Inhibitory/ Optimal)	Fold change
AGG-CGA	Empty	0.55 ±0.05	1.18 ±0.14
	tR(UCG)*	0.65 ±0.05	
CGA-AUA	Empty	0.31 ±0.01	1.92 ±0.12
	tI(UAU)	0.60 ±0.03	
CGA-CCG	Empty	0.18 ±0.01	7.69 ±0.61
	tP(CGG)*	1.38 ±0.10	
CGA-CGA	Empty	0.19 ±0.03	4.60 ±0.61
	tR(UCG)*	0.89 ±0.01	
CGA-CGG	Empty	0.25 ±0.00	2.39 ±0.20
	tR(UCG)*	0.61 ±0.05	
CGA-CUG	Empty	0.36 ±0.02	2.43 ±0.13
	tL(CAG)*	0.88 ±0.03	
CGA-GCG	Empty	0.28 ±0.02	2.04 ±0.30
	tA(UGC)	0.57 ±0.08	
CUC-CCG	Empty	0.10 ±0.00	6.44 ±1.09
	tP(CGG)*	0.67 ±0.11	
CUG-CCG	Empty	0.36 ±0.05	2.81 ±0.48
	tP(CGG)*	1.03 ±0.09	
CUG-CGA	Empty	0.34 ±0.04	2.25 ±0.32
	tR(UCG)*	0.75 ±0.06	
GUA-CGA	Empty	0.40 ±0.02	2.11 ±0.11
	tR(UCG)*	0.84 ±0.03	
GUG-CGA	Empty	0.42 ±0.04	1.77 ±0.27
	tR(UCG)*	0.75 ±0.10	

Table 2.3 tRNA Suppresses Codon Pair Inhibition
Each tRNA vector is a LEU2 2μ plasmid. With the exception of CGA-CGG (in which case the 3' tRNA did not suppress) the expressed tRNA matches the 3' codon of the pair. With the exception of CGA-GCG, the expressed tRNAs are exact Watson:Crick base-pair matches to the codon. Stars indicate non-native tRNA. Fold change refers to the ratio of tRNA vector GFP^{FLOW} to empty vector GFP^{FLOW}. See APPENDIX A for sequences and GFP^{FLOW} of individual constructs.

show for 12 tested pairs that when multiple copies of the pair are placed in an internal sequence location, GFP expression is reduced relative to constructs with synonymous pairs. Second, for 5

tested pairs the degree of inhibition increases with pair copy number. Lastly, for 11/12 tested pairs we show suppression of inhibition by overexpression of tRNA.

The fact that pairs of codons are major modulators of translation efficiency may explain why effects of synonymous codon variation on translation rates and efficiency have remained unpredictable (Plotkin and Kudla, 2010). The effects of individual codons can differ considerably in different pair contexts. For example, eight CGA-NNN codon pairs had syn-GFP^{SEQ} medians between 0.44 and 0.73, while the remaining 53 such pairs had medians > 0.91. Moreover, the existence of strong inhibitory pairs calls into question the idea that suboptimal codons mediate a continual series of individually small events that sum to a substantial effect. Instead, a few inhibitory codon pairs may act as discrete regulatory signals and could be as strongly selected as miRNA recognition sequences.

The 17 pairs we have identified are those with the strongest experimental evidence of substantial inhibitory effect. However, syn-GFP^{SEQ} medians for all possible codon pairs fall across a wide range, with 240 pairs (6%) having a median from 0.95 down to 0.44. This broad range could reflect a spectrum of codon pair-mediated inhibitory effect sizes. Since our FACS bins were set up to measure relatively large effect sizes and some pair combinations were underrepresented, we think there are likely additional inhibitory codon pairs that were not identified in the current assay.

Overall, pair-mediated inhibition reflects a complex phenomenon. In no case did overexpression of a single tRNA result in full suppression of inhibition, nor was the degree of suppression consistent (across either different pairs or constructs with the same pair). Rather, tRNA suppression for individual constructs fell into a wide range from no improvement to as high as a 7.7-fold improvement in expression. Given the unique biochemical and biophysical

properties of each codon pair combination and their associated tRNAs, we believe codon pairs may inhibit translation kinetics in unique, pair-specific manners. These interactions may even encompass greater complexity when considering other surrounding nucleotides or tRNA modifications.

There are at least two possible mechanisms by which inhibitory codon pairs could influence protein expression. One mechanism is by reducing the abundance of mRNA transcripts. Measures of translation efficiency that normalize protein levels to transcript abundance ignore potential feedback mechanisms between translation and transcript maintenance. In the case of the CUC-CCG inhibitory pair, we have found that tRNA overexpression increases both mRNA and protein levels (data not shown). The possibility of codon pair mediated mRNA decay is supported by the observed correlation of mRNA decay rates with codon usage (Presnyak et al., 2015) and by the existence of surveillance pathways that rapidly degrade mRNA transcripts upon detection of stalled translation complexes (Graille and Séraphin, 2012). Another mechanistic possibility is that slow elongation through some codon pairs facilitates nascent protein folding or interactions between nascent proteins, with effects on translation efficiency a secondary consequence of the need for translation to slow down (Thanaraj and Argos, 1996b; Zhang and Ignatova, 2009).

2.4 METHODS

Library Construction, FACS, and Flow Cytometry

The RNA-ID GFP construct is a fusion protein encoding a site for 3C protease, an HA epitope, and His6, followed by superfolder GFP. The $(NNN)_3$ and $(VNN)_3$ libraries of GFP variants in *E. coli* were derived from those made by Dean and Grayhack (Dean and Grayhack, 2012). See

APPENDIX B for details on yeast library preparation. Fluorescence-activated cell sorting (FACS) of ~3 million to ~9.5 million cells was performed as previously described (Dean and Grayhack, 2012).

Sequencing of GFP 3-Codon Insertions

From genomic DNA samples, we amplified *GFP* library fragments through 25 PCR cycles, using primers specific to the flanking regions and containing a FACS bin-specific index. We then pooled the amplified fragments and sequenced on an Illumina GAII sequencer with single-end reads. For quality control, we required each read to have accurately called 6 bases ('AACGCA') immediately downstream of our variable region and for each of the 9 variable base calls to have a score of Q30 or better. To compare read counts across bins, we corrected for the number of cell sorting events in a given bin.

Quality Filtering of GFP Library Sequences

We used PRINSEQ (Schmieder and Edwards, 2011) to trim reads and require that each of the 9 variable base calls had a quality score of at least Q30. We also applied a read depth cutoff based on the maximum number of possible variants in each bin (the total number of cells sorted). To ensure a dataset of only the highest-quality variant expression scores, we applied a minimum threshold for total number of variant read counts across FACS bin samples. For the $(NNN)_3$ libraries, we determined these thresholds empirically, based on the drop in stop codon-containing variants with a high proportion of spurious, high-expression bin counts. This threshold was 30 read counts for $(NNN)_3$ Library 1 variants, and 60 read counts for $(NNN)_3$ Library 2 variants. We then used stop codon-containing variants with reads above these thresholds to estimate the

average degree of spread into distant fluorescent bins. We removed variants with bimodal-like distributions, where the variant had more than an average spread in both the background (no expression) and the high expression bin. These bimodal-like variants constituted 4% to 5% of each library. For the $(VNN)_3$ Library, we imposed the stricter of each $(NNN)_3$ Library's threshold values. Furthermore, we removed any $(VNN)_3$ Library variants with a "T" base call in the first nucleotide position of a codon. Variants with ≥ 75% of reads in the background bin were removed from all libraries.

GFP^{SEQ} and syn-GFP^{SEQ} Expression Scores

We used the median GFP fluorescence value of each FACS bin and the proportion of variant read counts in each bin to calculate a mean expression score (GFP^{SEQ}) for each 9-base sequence variant, relative to 100% high bin expression. After checking for a high degree of correlation between libraries, we combined the variant data from each library. For identical sequences measured in separate FACS libraries, we treated each library's measurement as a biological replicate and took the average. To obtain syn-GFP^{SEQ} scores we normalized to the highest GFP^{SEQ} score among a set of synonymous variants. If there were no synonymous variants, or if the highest GFP^{SEQ} fell below the mean of all scores (0.9547), then these sequences were not included in the downstream analysis.

Statistical Analysis

To assess the significance of each 6-mer sequence's frequency in low GFP variants, we tracked occurrences of the 6-mer in low variants across 100,000 permutations. Variants were assigned to one of 10 pools based on GC count, and we shuffled the expression categories within each pool.

From this analysis we derived p-values to ≤ 0.00001 for each 6-mer, based on the probability of obtaining as many, or more, low variant counts by chance.

We calculated Benjamini-Hochberg false discovery rates (FDR) to control for the number of false positives. The 28 candidate pairs reached significance in the full dataset at a FDR of 3%, while the revised list of 20 candidate pairs reached significance in the reduced structure set of variants at a FDR of 7%. In evaluating the reduced structure set, we determined permutation p-values based on occurrences at the combination of s1 and s4 positions; as opposed to at each position independently.

We ran one-sided Wilcoxon rank sum tests to compare syn-GFP^{SEQ} distributions of candidate inhibitory pairs and related sequences. Wilcoxon rank sum tests were carried out in R. For each of the 20 candidate pairs, we compared the distribution of variants with an inhibitory pair to variants with the 6-mer sequence in an out-of-frame position as well as to variants with the two codons present but separated. For 12 inhibitory pairs we compared the distribution of variants with the inhibitory pair to the distribution for variants with the codons in reverse order. We corrected Wilcoxon p-values for 67 tests (including tests described in **Chapter 4**) using the Holms-Bonferroni procedure.

Heatmap Generation

The heatmap shows permutation p-values for enrichment of 6-mers in low expression variants at each of the four possible 6-mer positions in the 9-base variable region. It includes all 6-mers that have a permutation p-value ≤ 0.001 at one or more of the four positions and numeric p-values at all four positions. Fifty-seven significant sequences with missing data at one or more of the

positions are not plotted. Most of these 6-mers (79%) had a 3-nucleotide stop codon sequence in one of the reading frames. We clustered and plotted the data using pheatmap package in R.

RNA Structure Prediction and Reduced Structure Subset

To assess RNA structure across all 35,811 sequence variants of our library, we found the global free energy (ΔG) for each variant and compared local, paired nucleotide probabilities between synonymous sequences in a manner similar to Goodman et al. (Goodman et al., 2013). See **Chapter 5** methods for a detailed description of the prediction methods and calculation of relative paired nucleotide probabilities.

To identify a subset of variants with similar degrees of structure, we first identified all window positions, after the translational start, where relative paired nucleotide probabilities had a significant negative correlation with expression (p-value < 0.001; positions: 1-10, 17-35, 51-54, 57-59, 78-87, and 94-96). Then we removed variants with probabilities more than 1 standard deviation away from the high category mean. We also removed variants for which the global free energy (as measured by NUPACK), for the region running from the transcription start site to +102, fell more than 1 standard deviation away from the high variant category mean. This reduced structure subset had 13,061 variants in total, with 183 low variants, 1,133 intermediate variants, and 6,597 non-reference high variants.

Additional Methods

See APPENDIX B for information on strains, plasmids, oligonucleotides, and the analysis of individual variants by flow cytometry.

2.5 NOTES AND CONTRIBUTIONS

This chapter is part of the publication, "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast" by Caitlin E. Gamble, Christina E. Brule (University of Rochester), Kimberly M. Dean (University of Rochester), Stanley Fields, and Elizabeth J. Grayhack (University of Rochester); in press at *Cell* (2016).

I carried out amplification and sequencing of library GFP from genomic DNA samples. I filtered the sequencing data; wrote custom software for analysis and interpretation; and I made Figures 2.2, 2.3A, 2.3B, and 2.4. I wrote the associated manuscript together with EJG, SF, and CEB. KMD created the *GFP* libraries and transformed them into *E. coli*. CEB transformed these libraries into yeast; carried out FACS runs; constructed individual variants; took flow cytometry measurements; and made Figures 2.1, 2.3C, and 2.5.

CHAPTER 3: Ribosomes Tend to Slow-Down at Inhibitory Codon Pair Sites in Yeast Transcripts

3.1 BACKGROUND

At the start of each translation elongation cycle, the aminoacyl-site (A-site) of the ribosome presents an mRNA codon for decoding by aminoacyl-tRNA. Once an appropriate tRNA match is found, conformational changes accommodate the tRNA in the A-site. A new peptide bond is formed, attaching the amino acid to a growing polypeptide chain, and the ribosome translocates 3 nucleotides along the mRNA transcript, moving the codon and peptidyl-tRNA (with the polypeptide chain) into the peptidyl-site (P-site) and beginning a new cycle. Upon a second translocation, the codon and now uncharged tRNA move into the exit site (E-site) before disassociating from each other and the ribosomal complex.

Despite a rich body of theoretical work on the selective forces shaping synonymous codon usage and the influence of codon usage on translation speed, our ability to test synonymous codon usage theories has been limited by lack of a direct means to measure *in vivo* decoding rates. In 2009, Ingolia et al. introduced ribosome profiling (Ingolia et al., 2009a), a technology development that has brought the field significantly closer to quantitative measurement of translation speeds at single-codon resolution. Ribosome profiling applies high throughput sequencing to the detection of nuclease-treated mRNA fragments. Based on fragment size, ribosome profiling identifies the mRNA fragments (ribosome footprints) protected from nuclease digestion by the presence of a translating ribosome. The overall protein synthesis rate for each open reading frame (ORF) in the transcriptome can be measured by the ratio of footprints along a transcript to the transcript abundance. In addition, the specific location of

ribosomes is inferred from the footprint sequences. Since the precise location of footprints on each ORF is stochastic, one can infer the rate of translation from footprint frequency. In theory, higher footprint frequencies from particular regions and codons along the transcript correspond to longer dwell times during translation. Taken across a population of cells, fragment frequencies within each ORF represent the average distribution of ribosomes during steady-state conditions.

In actual practice, different profiling experiments have led to seemingly contradictory results. Variation in the decoding rates of individual codons has been detected in some ribosome profiling studies (Dana and Tuller, 2014a; Gardin et al., 2014; Lareau et al., 2014; Stadler and Fire, 2011), but not in others (Charneski and Hurst, 2013; Ingolia et al., 2009a; Pop et al., 2014; Qian et al., 2012). Much of this discrepancy can be traced to differences in the experimental protocols (Hussmann et al., 2015; Weinberg et al., 2015). Most yeast ribosome profiling experiments have followed the original Ingolia et al. (2009) protocol (Artieri and Fraser, 2014; Brar et al., 2011; Gerashchenko et al., 2012; Hussmann et al., 2015; Ingolia et al., 2009a; McManus et al., 2014; Zinshteyn and Gilbert, 2013), which pretreats cells with a small-molecule translation inhibitor, cycloheximide, to arrest translating ribosomes. However, cycloheximide pretreatment leads to the accumulation of ribosomes near the start site (Gerashchenko and Gladyshev, 2014) since it does not inhibit translation initiation. Furthermore, it impacts the length distribution of fragments, which are thought to correspond to alternative ribosome confirmation states (Lareau et al., 2014). An alternative protocol uses flash freezing to arrest translation, rather than cycloheximide (Gardin et al., 2014; Lareau et al., 2014; Pop et al., 2014; Weinberg et al., 2015). Comparison of the use of these two protocols in yeast datasets has found consistent differences between the two (Hussmann et al., 2015). Furthermore, these comparisons suggest that cycloheximide pretreatment fails to immediately halt translation and perturbs

measures of the translation rate dynamics for each codon, particularly so for the CGA, CGG, and CCG codons (Hussmann et al., 2015). Recently, data from an alternative technological approach, 5'P sequencing (5PSeq), further substantiated the finding that cycloheximide perturbs translation dynamics and causes loss of codon-specific signals (Pelechano et al., 2015). Thus, conclusions from studies using the pretreatment will need to be reexamined with respect to the specific elongation dynamics within transcripts.

In addition to divergent experimental protocols, researchers have also applied a variety of computation approaches to the ribosome profiling data (Artieri and Fraser, 2014; Charneski and Hurst, 2013; Dana and Tuller, 2014b; Pop et al., 2014; Qian et al., 2012; Stadler and Fire, 2011). Diverse computational approaches have arisen in part as a result of the controversial and seemingly contradictory conclusions drawn by separate studies and the many challenges that ribosome profiling data presents for accurate quantitative measurement. As stated by Gardin et al., "Defining the right normalizations to compensate for differences in gene expression, gene length, sequence composition, etc, is complicated and problematic" (Gardin et al., 2014). Similar to the original Ingolia et al. (2009) paper, most studies have calculated ribosome occupancy for individual codons by first normalizing footprint counts to the mean count for each ORF (Dana and Tuller, 2014b; Ingolia et al., 2009a; Lareau et al., 2014; Weinberg et al., 2015), thereby normalizing to each ORF's overall level of translation and providing a relative measure of decoding speed that is comparable across ORFs. In a few exceptions to this approach, others have normalized to neighboring windows of codons within each ORF (Charneski and Hurst, 2013; Gardin et al., 2014), to the average frequency of a given codon in positions flanking the inferred ribosomal A, P, and E sites and within the ribosome footprints (Nedialkova and Leidel,

2015; Stadler and Fire, 2011), as well as to expected frequencies estimated from the corresponding total mRNA abundance (Qian et al., 2012).

There are at least three major challenges in applying ribosome-profiling data to analyze codon pairs and codon context. One, the data for many ORF transcripts are sparse. The number of footprints per codon in a transcript (coverage) varies widely across different ORFs. In the Lareau et al. (2014) yeast dataset, coverage falls across 5 orders of magnitude, with 73% of ORFs (n = 3,202) having less than 100% coverage (median = 23% coverage for the subset). In part because the data are sparse, frequency distributions are not normally distributed. A substantial number of codons and sequence windows have no footprints, and the mode of a distribution is often 0 (Gardin et al., 2014). Density distributions are not representative when the total number of footprints is very low. To deal with this issue, studies have applied various coverage filters in pretreatment of the data, thereby excluding most low expression genes from the analysis (Dana and Tuller, 2014a; Gardin et al., 2014). However a second major challenge arises from the unequal representation and distribution of sequence features within the transcriptome. While coverage and representation may not appear to present huge issues in measuring the decoding speeds of individual codons, these issues become much more prevalent when considering codon pairs and other forms of codon context, where there are fewer representative sites. Additionally, if the context is detrimental to expression, many of these sites may occur predominately in low expression genes. Finally, the third major challenge in applying profiling data to analyze codon context is that there is no obvious *a priori* expectation for the distribution of footprints across yeast transcripts, which would provide a null hypothesis in accessing the strength of ribosome profiling evidence for variable translation rates. Combined with the issues of coverage and representation, differential analysis within profiling datasets is

complicated and difficult. While many studies have looked at overall correlations, for example between tRNA abundance and ribosome occupancy at particular codons, or identified codons at the extreme of an effect size range, to our knowledge Gardin et al. is currently the only study to have assigned p-values to the ribosome occupancy measures for specific codons (Gardin et al., 2014).

Here, we evaluate yeast ribosome profiling data for evidence that specific inhibitory codon pairs, rather than their individual codons or dipeptides, significantly slow translation elongation. Explicitly, we have evaluated 17 inhibitory codon pairs identified by their prevalence in low expression variants of a GFP library (**Chapter 2**). To avoid the confounding effects of cycloheximide, we chose the Lareau et al. (Lareau et al., 2014) untreated yeast data for our analysis. In addition to this dataset being one of the few available experiments in which pre-treatment of the cells with cycloheximde was omitted, it is currently the only dataset to include both long and short ribosome footprints. Lareau et al. found evidence of slow translation at three wobble codons (CGA, CUG, and CCG) (Lareau et al., 2014). These three codons occur at high frequency among the 17 inhibitory codon pairs, suggesting that pair effects may contribute to the overall degree of slower translation detected for these codons.

Given the challenges of ribosome profiling data and the unique issues that arise in evaluating inhibitory codon context, we have developed a computational approach that combines footprints from local position windows surrounding codon pair sites, controls for distribution noise, and accounts for low footprint counts through direct hypothesis testing on categorical counts data. With this combined strategy, we find 12 of the 17 inhibitory codon pairs show significant evidence of slower translation times relative to when the pair is outside of a ribosomal site position.

3.2 METHODS

For each codon in the Lareau et al. (2014) untreated datasets, we took the tally of footprints with the codon in the A-site position of the ribosome (**Figure 3.1**). Since we had no *a priori* expectation of how codon pairs might impact specific ribosome confirmations and in order to obtain higher coverage, we used the combined total of both short (20-22 nucleotides) and long (28-30 nucleotides) mRNA fragment sizes. The Lareau et al. untreated datasets excluded ORFs with fewer than ten footprints. The datasets also excluded footprints for the first 50 codons of each open reading frame (ORF) due to the scarcity of footprints from these regions. ORF coverage (footprints/codon) varied widely across the three biological replicates. Replicate 1 (GSM1406453) had 65% coverage of all codons (footprints per codon from the pooled dataset), whereas Replicate 2 (GSM1406454) had 54% and Replicate 3 (GSM1406455) only 16%. Since the inhibitory pairs identified in our GFP assay occurred relatively infrequently in the sampled transcriptome and often in transcripts with relatively low expression (where footprints were rarer) (**Figure 3.2**), we considered coverage critical to our analysis, and thus, we pooled A-site codon footprint counts from each replicate.

Footprint Count Window at Codon Pair Sites

Most profiling studies have calculated ribosome occupancy by normalizing footprints within each ORF to the mean footprint count per codon in the ORF (Artieri and Fraser, 2014; Dana and Tuller, 2014b; Ingolia et al., 2009a; Lareau et al., 2014). By normalizing to the mean of an entire transcript, larger effect sizes within the transcript will dominate the ribosome occupancy signal. If globally large effects occur at regions of the transcript far away from a given codon pair site,

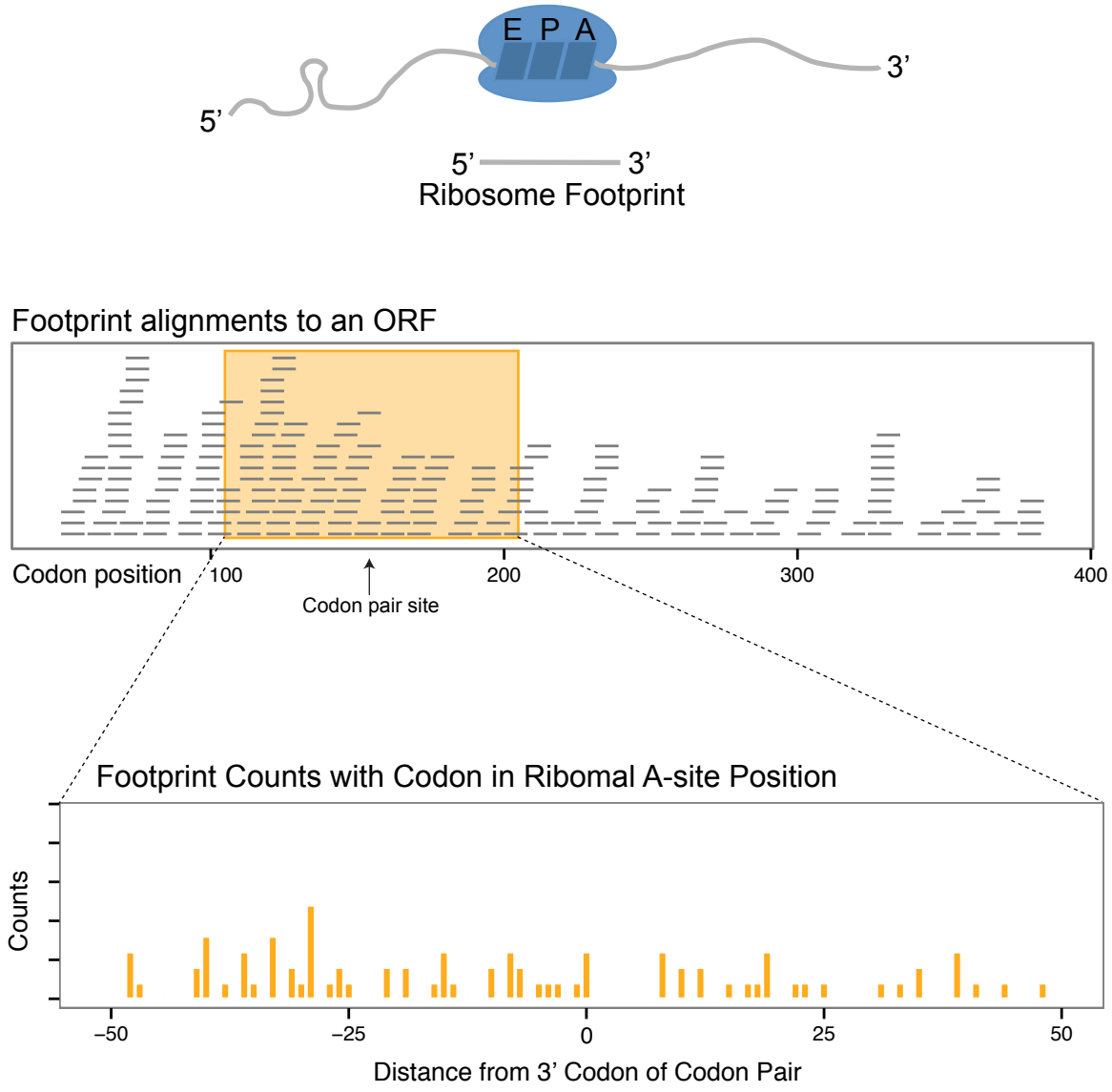

Figure 3.1 Example of Ribosome Profiling Data for an ORF

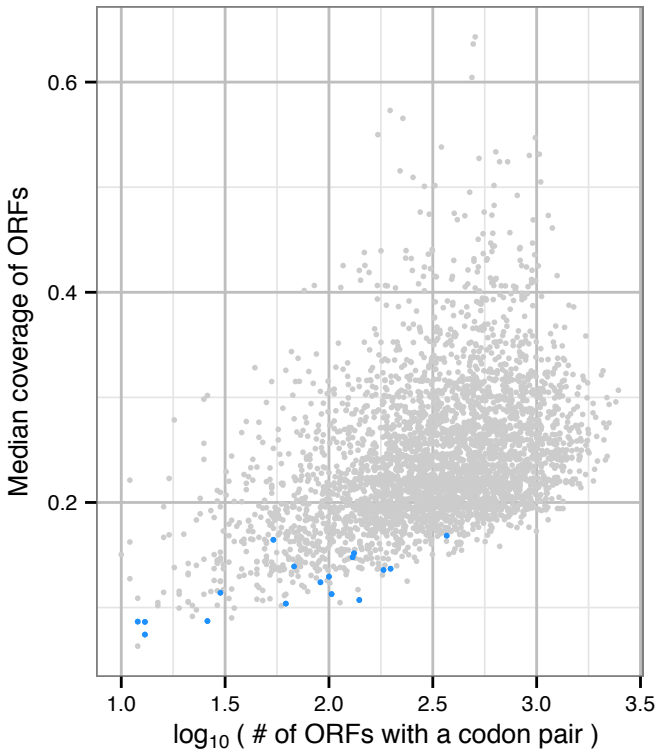

Figure 3.2 Ribosome Footprint Coverage of ORFs with Inhibitory Pairs
Median coverage for ORFs with a given codon pair and total number of ORFs with the pair; for inhibitory codon pairs (blue) and all other codon pairs (gray) in the Lareau et al. pooled replicates.

occupancy measurements potentially lose sensitivity to local translation differences on a codon-by-codon level. To increase sensitivity to local, rather than global, translation differences within ORFs, we considered footprint counts in only a window of codons surrounding the pair (**Figure 3.3**). Balancing the ideal of a local window with the need to sample the surrounding baseline ribosome occupancies at many positions, we decided upon a window length of 100 codons. After trying both shorter and longer window sizes, we observed that a length of 100-codon positions provided a long enough window to minimize sampling noise in measurement of baseline occupancy, while still limiting the total window length.

Ribosome Occupancy at Codon Pairs

Whether using footprint counts across an entire open reading frame or within a more limited window of positions, most studies have normalized footprint counts on a gene-by-gene level. The occupancies measured at individual codon sites are then averaged across all sites to assign "typical" ribosome occupancies to the individual codons (Dana and Tuller, 2014a; Gardin et al., 2014). By this approach, each qualifying gene or gene window serves as an independent speed trial and the approach achieves statistical robustness by averaging over the hundreds to thousands of codon sites (Gardin et al., 2014). However, when a substantial proportion of ORFs have very low coverage (and therefore inaccurate occupancy measures), the approach is less reliable. Furthermore, the approach losses statistical robustness when instead of having hundreds or thousands of examples for sequence feature, such as with individual codons, the total number of sites is more limited, as it is for certain codon pairs. Thus, to limit the inaccuracies from taking individual window density measurements, for each pair we combined footprint counts from all pair site windows into a meta window of joint counts at each position in the window, and then we have converted the joint footprint counts into densities (ribosome occupancies) (**Figure 3.3**). With this approach, genes with higher footprint coverage were weighted more heavily in the final occupancy calculation.

In applying this approach, for each pair we aligned 100-codon windows from ORFs containing a particular pair with the pair site at the center of the window. Then we took the sum of ribosome footprints at aligned codon positions to obtain the joint footprint count at each position, up to 50 codons away from the center (48 positions 5' and 50 positions 3'). Within the meta-window of joint counts, we converted counts at each position into a density (ribosome occupancy) (**Figure 3.3**). Thus ribosome occupancy is a number between 0 and 1, and

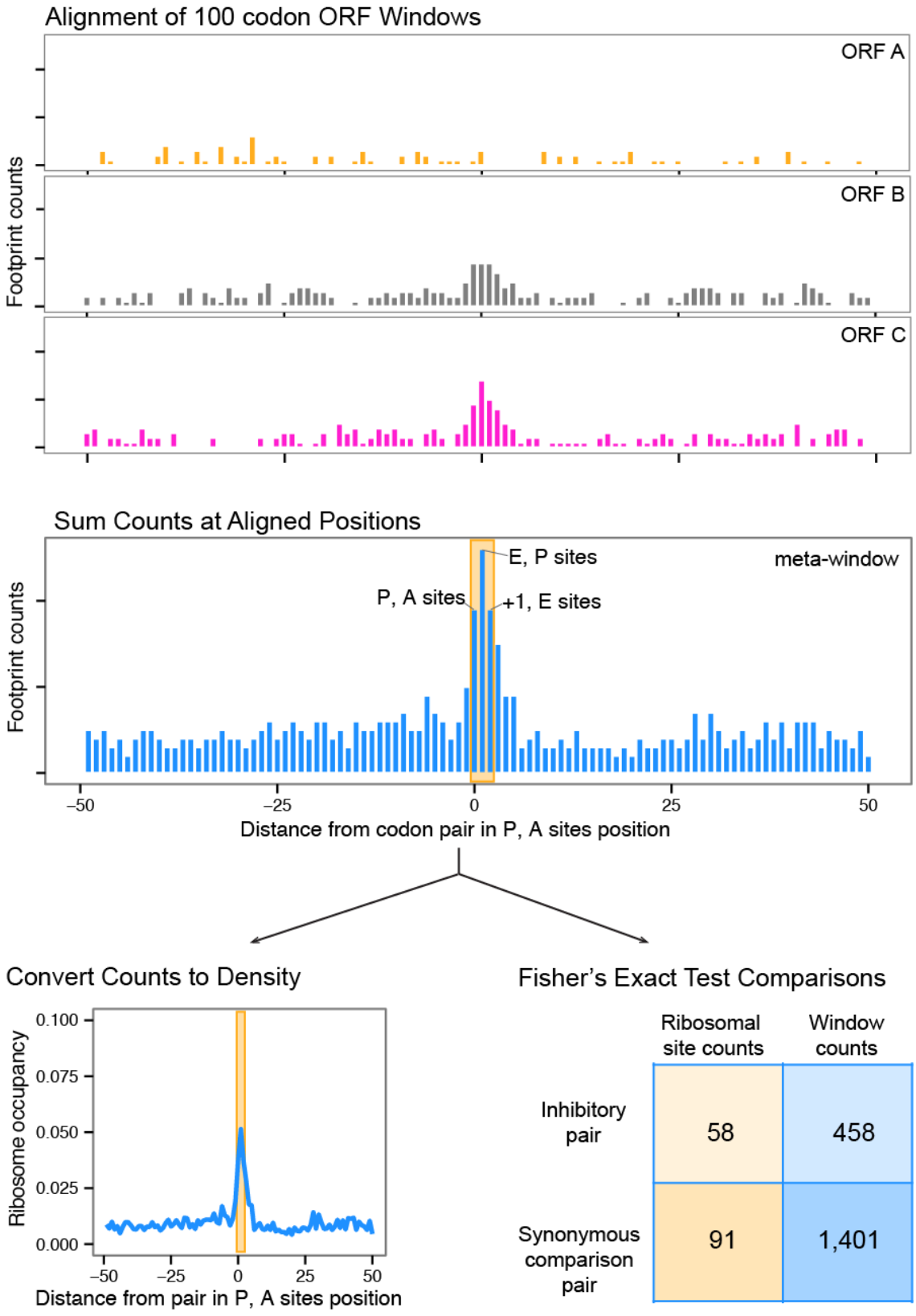

Figure 3.3 Schematic of Computational Analysis

occupancy reflects the typical ribosome occupancy when the pair is positioned at a given distance away from a P and A sites position.

To assess each codon pair for evidence of reduced translation rates, we found the cumulative ribosome occupancy of the codon pair by summing the occupancies corresponding to 3 positions with the codon pair in the ribosomal P, A sites (window distance of 0); E, P sites (window distance of +1); and +1, E site (window distance of +2) (**Figure 3.3**). We included the +1, E-site position after having observed that many pairs with elevated ribosome occupancies at the first two positions also had higher densities at the +1, E-site position. This residual effect could be due to imprecision in assignment of the ribosomal site position (due to heterogeneity in the 5' fragment cut); others have reported such an effect in analyzing individual codon sites (Gardin et al., 2014; Weinberg et al., 2015).

Significance of Ribosome Occupancies Given Background Noise

In the absence of decoding speed differences within a codon pair window, there should, in theory, be a uniform distribution of footprints across positions. Thus as a baseline expectation for a 100-position window with a total of 100 footprints, one might expect ~1 footprint per codon position (an occupancy of 0.01). Across 3 ribosomal site positions the expected baseline occupancy becomes 0.03. On the other hand, if the ribosome were to slow down while translating the codon pair, one would expect to see a higher concentration of footprints at positions where the pair is located in ribosomal sites (center of the window). Ribosome profiling data are not normally distributed, however. Thus, we sought to control for noise in the distribution of footprints by evaluating the significance of occupancy at codon pairs. We did this given that footprint counts from just a few windows could hold substantial sway in the overall

occupancy, especially in cases with few codon pair sites present in the genome. To this end, we estimated the probability of obtaining ribosome occupancies as high as the actual occupancies by chance within the set of ORFs associated with each pair. For each pair, we carried out 10,000 permutations, in which we shuffled the A-site codon footprint counts within each ORF and re-calculated ribosome occupancy. Permutation analysis p-values were then corrected for 17 inhibitory codon pair tests using the Holms-Bonferroni procedure.

Dipeptide and Individual Codon Comparisons

Converting footprint counts to occupancies (window densities) enables comparison between pairs despite differences in representation and coverage. However, while such comparisons are valid when the occupancies are based on a large number of footprints and pair sites, one must exercise caution in making these direct comparisons with inhibitory pair codons, since there are fewer pair sites and very low coverage within some pair site windows. Thus, we sought to gain more assurance that higher occupancies reflected meaningful differences between pairs and not an inflated value due to low coverage and limited sampling. To this end, we applied Fisher's exact test, taking into account the observed footprint count contributing to each pair's ribosome occupancy and the likelihood that the distribution of counts between ribosomal site positions and other window positions differs from that of a comparison pair (**Figure 3.3**). Thus, to determine if each inhibitory codon pair had a significantly higher occupancy than related, synonymous comparison pairs, we made 2 x 2 contingency tables. Each table had the footprint counts for each pair at positions with the pair occupying a ribosome site and at positions in the remainder of the window. We ran one-sided Fisher's exact tests in R, and we corrected the Fisher's exact p-values for 46 tests using a Holms-Bonferroni procedure (including 12 tests for the reverse order pair;

Chapter 5). From each Fisher's exact test we also derived an odds ratio (calculated by conditional maximum likelihood estimation) and confidence interval.

3.3 RESULTS

Genes with Inhibitory Codon Pairs Tend to Have Low Ribosome Footprint Coverage

Due to the dynamic range of gene expression within a cell's transcriptome, some ORF transcripts are present in much lower abundance within cells. Correspondingly, these transcripts often have fewer overall ribosome footprint sequencing reads per codon than higher abundance transcripts translated at a similar rate. Aside from filtering for transcripts that meet a threshold level of footprint counts, most profiling approaches have not specifically dealt with coverage issues in measuring the decoding rates of codons, since even the most rarely used codons are present at a large number of sites within the transcriptome. However, in seeking to evaluate decoding rates based on context, this issue becomes much more prominent.

We observed that ORFs with inhibitory codon pairs tended to have low footprint coverage in the Lareau et al. profiling data (**Table 3.1**). For example, the median coverage for ORFs with the CGA-CGA pair was only 9% (compared to a 24% average across all pairs). Low coverage could result from rapid translation of the associated ORF mRNA. However, it may also reflect low transcript abundance in the cell. Most ORFs with an inhibitory pair (71%) had mRNA levels falling below the median of all quantified ORFs (based on RNAseq from Presnyak et al., 2015), suggesting that low coverage in the Lareau et al. dataset in part reflects the prevalence of inhibitory pairs in low expression genes. Consistent with inhibitory codon pairs having a negative effect on translation efficiency, most pairs are present at relatively few sites in the yeast genome. The CGA-CGA codon occurs in only 12 ORFs within the Lareau et al. profiling data (**Table 3.1**). Overall, the list of 17 inhibitory codon pairs from our GFP assay had a wide range

Pair	ORFs (Lareau et al., 2014)	Median ORF Coverage	Coverage IQR Limits (Q1, Q3)	ORFs (Presnyak et al., 2015)	Median mRNA	mRNA IQR Limits (Q1, Q3)
AGG-CGA	91	0.12	0.07, 0.25	140	20.55	14.86, 33.37
AGG-CGG	68	0.14	0.08, 0.22	102	19.64	14.62, 31.80
AUA-CGA	183	0.14	0.07, 0.20	284	20.77	13.93, 29.90
AUA-CGG	103	0.11	0.05, 0.19	176	18.73	13.08, 29.93
CGA-AUA	140	0.11	0.05, 0.17	237	20.64	13.98, 28.81
CGA-CCG	13	0.09	0.06, 0.12	18	19.27	15.27, 24.62
CGA-CGA	12	0.09	0.07, 0.14	24	15.97	5.81, 23.49
CGA-CGG	13	0.07	0.06, 0.27	19	15.64	10.33, 28.47
CGA-CUG	62	0.10	0.06, 0.16	92	21.00	15.76, 29.79
CGA-GCG	26	0.09	0.04, 0.20	36	15.24	11.26, 26.89
CUC-CCG	30	0.11	0.06, 0.18	61	17.91	10.68, 27.43
CUG-AUA	369	0.17	0.09, 0.31	547	22.99	14.41, 36.25
CUG-CCG	132	0.15	0.08, 0.29	187	22.77	15.50, 34.72
CUG-CGA	100	0.13	0.08, 0.25	161	19.05	12.60, 28.44
GUA-CCG	198	0.14	0.07, 0.27	284	22.75	15.11, 33.37
GUA-CGA	130	0.15	0.09, 0.23	208	20.39	13.72, 28.73
GUG-CGA	54	0.16	0.06, 0.32	78	20.49	13.81, 29.66
Mean of all possible pairs	447	0.24	0.11, 0.64	625	28.62	15.98, 55.40

Table 3.1 Coverage and mRNA Abundance of ORFs with Inhibitory Codon Pairs

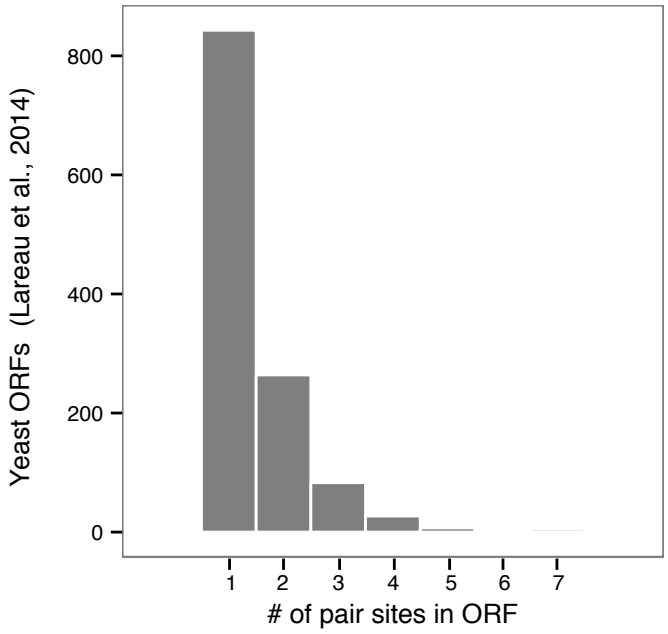

Figure 3.4 Total Number of Inhibitory Pair Sites by ORF

of representation within ORFs from the Lareau et al. dataset. One pair, CUG-AUA, was present in 369 ORFs, a similar number of ORFs as many 6-mers that were not identified as inhibitory in the GFP assay. The remaining inhibitory codon pairs were represented in between 13 to 198 ORFs (**Table 3.1**). The vast majority of ORFs had a single inhibitory pair site, although 388 ORFs had from 2 to 8 inhibitory pair sites in total (**Figure 3.4**).

15 Inhibitory Codon Pairs Have High Ribosome Occupancies on Yeast Transcripts

Given the low representation and coverage of inhibitory codon pairs, we developed a tailored computational approach. Specifically, for each pair, we summed footprint counts from windows at each pair site into a meta-window for all pair sites. Then we calculated the relative ribosome occupancy with the pair at ribosomal site positions and positions up to 50 codons away. From this analysis, we observed that all but two of 17 inhibitory codon pairs (AUA-CGG and AGG-CGG) had higher ribosome occupancy than expected by chance (corrected permutation p-value < 0.002), given the coverage and footprint distributions in each pair's set of ORFs. Ribosome occupancies for the 15 pairs ranged from 5.6% to 25%, compared to a baseline expectation of 3% and a mean of 3.7% in the population of all possible pairs. Nine inhibitory pairs were in the top 1% of all ribosome occupancies (more than 3 standard deviations away from the mean; **Figure 3.5**). In particular, four pairs with the highest occupancies were also in the top five occupancies of all 6-mers and were the four most inhibitory codon pairs from the GFP assay (**Figure 3.5**).

A number of other pairs had higher ribosome occupancies than the baseline expectation, but no detectable inhibition in the GFP assay. Some of these pairs may reflect dipeptide effects, which are internally controlled for in the syn-GFP^{SEQ} score (**Chapter 2**). For example, others

have reported that proline dipeptides have a slow decoding rate (Artieri and Fraser, 2014), and we see that 7 out of the 16 possible Pro-Pro codon pairs were more than 2 standard deviations above the mean pair occupancy. However, 91% of the pairs falling 2 standard deviations or more above the mean ribosome occupancy for all 6-mers had one of the 3 individual codons (CGA, CUG, and CCG) originally reported by Lareau et al. to have higher occupancy. Thus the baseline occupancy for 6-mers that contain one of these codons is generally higher, and the inhibitory codon pairs tend to have even more extreme occupancies (**Figure 3.6**). Individual codon effects may have been missed in the GFP assay, which was not sensitive to reductions in expression less than ~25%. Conversely, for those pairs that were strongly inhibitory in the GFP assay but had lower ribosome occupancies (i.e. AUA-CGG), we believe that inhibition may be due to other mechanisms.

12 Inhibitory Pairs Have Significantly Higher Occupancies than Synonymous Pairs

Given that many pairs encoding the same dipeptides or using similar codons to the 17 inhibitory codon pairs also had higher ribosome occupancy, we next sought to test the alternative hypothesis that higher occupancy at each of these pairs was due to amino acid or single codon effects. To directly compare the ribosome occupancy of each inhibitory pair with that of individual codons and dipeptides, we compared footprint counts with an inhibitory pair in ribosomal site positions and non-ribosome site positions to the footprint counts for two synonymous pairs. One of these comparison pairs was made up of the 5'-codon of the inhibitory pair and a CAI-optimized 3'-codon (pink). The other comparison pair had a CAI-optimized 5'-codon and the 3'-codon of the inhibitory pair (orange) (**Figure 3.7** and **Table 3.2**). We observed for 12 inhibitory pairs that the proportion of footprints at the inhibitory pair was higher than the

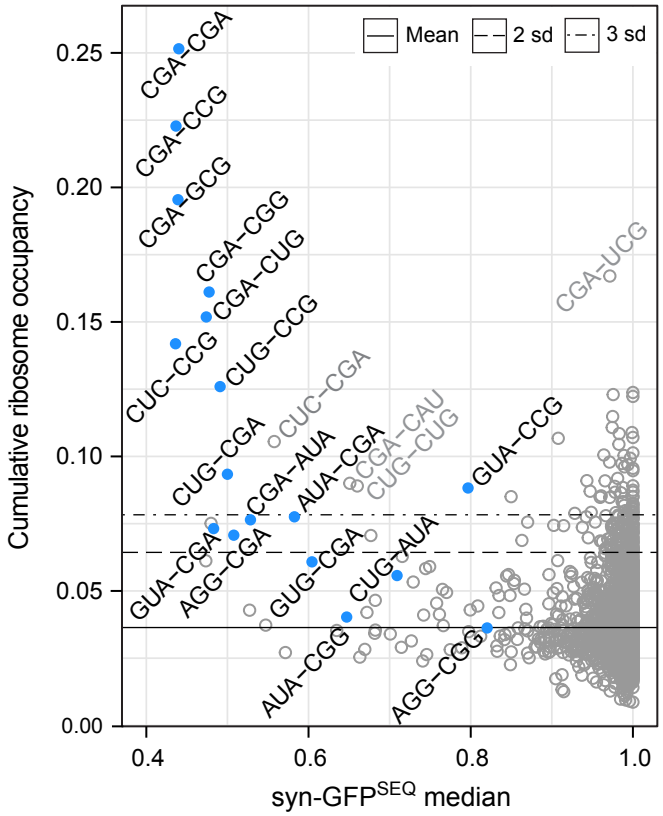

Figure 3.5 Inhibitory Pairs in Yeast ORFs Have Ribosome Occupancy Peaks
Median syn-GFP^{SEQ} of variants with each codon pair versus cumulative ribosome occupancy with the codon pair at 3 positions: the ribosomal P, A-sites, E, P-sites, and +1, E-site. Identified inhibitory codon pairs from the GFP assay are shown in blue, while other codon pairs are in gray.

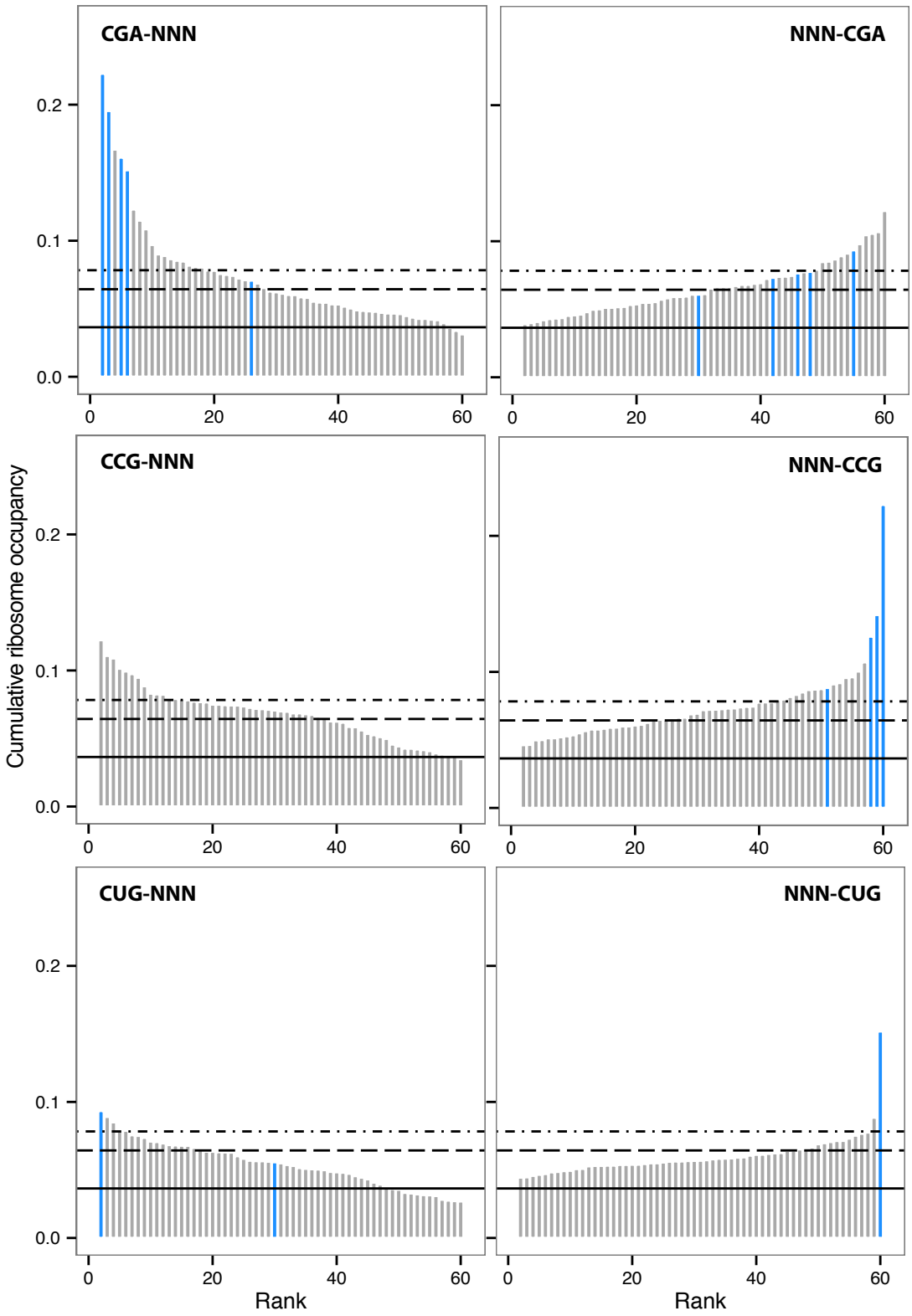

Figure 3.6 Pairs with a CGA, CCG, or CUG Codon Ranked by Ribosome Occupancy
Pairs identified as inhibitory codon pairs in the GFP assay are indicated in blue.

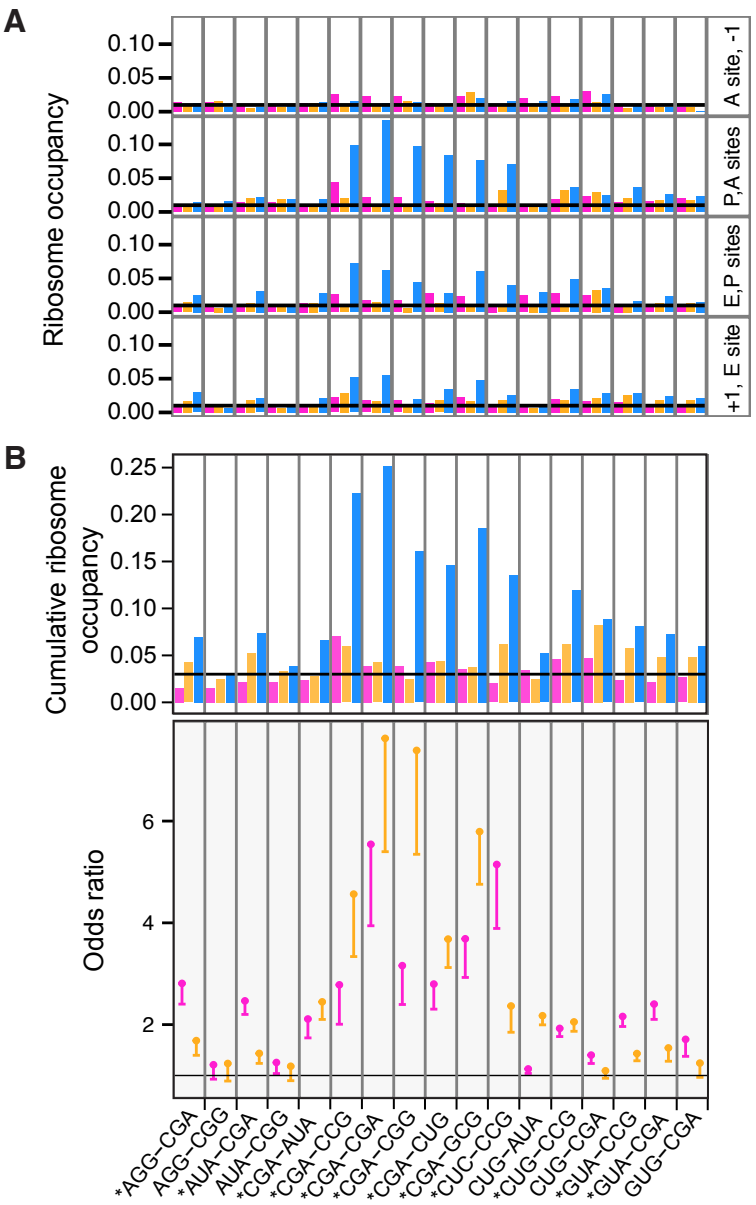

Figure 3.7 Ribosome Occupancy at Inhibitory Pairs is Greater than at Synonymous Pairs
Proportion of footprints (within 100 codon window) at positions with pairs occupying ribosomal sites. Black line indicates the expected proportion, based on an even distribution of footprints. Inhibitory pair proportions (blue) are shown with two synonymous pairs: a pair with an optimized 3' codon (pink) and a pair with an optimized 5' codon (orange).
(A) Proportion of footprints at specific positions
(B) Proportion of footprints at 3 positions combined (P, A-sites; E, P-sites; and +1, E-site; upper panel) and the conditional odds ratio (inhibitory pair footprint proportions relative to comparison pair proportions; bottom panel). Error bar indicates the lower 95% confidence interval limit. Pairs with significant differences (one-sided Fisher's exact corrected p-value ≤ 0.05) in both comparisons are indicated by a star.

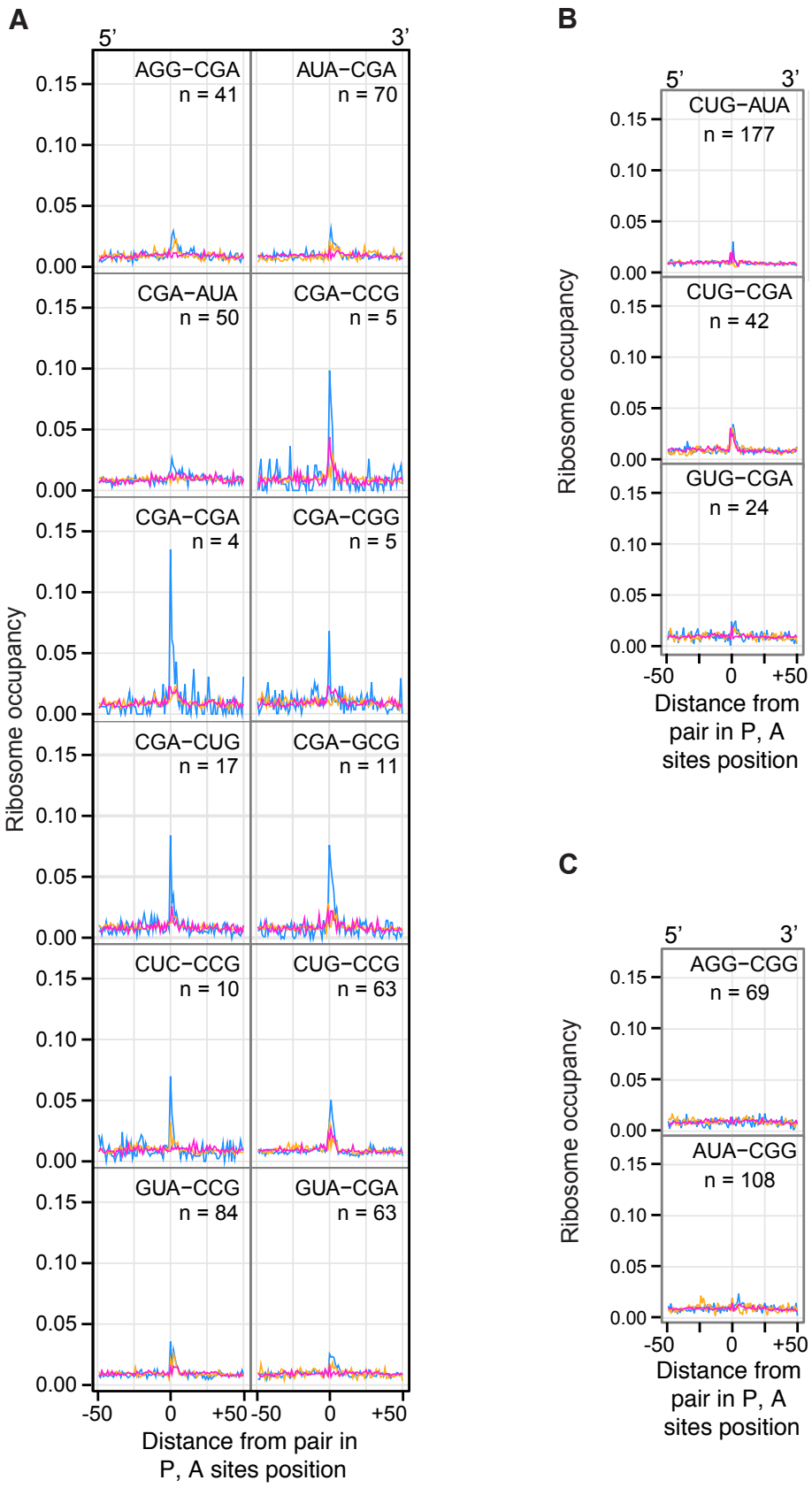
A
5'
3'
AGG−CGA
n = 41
AUA−CGA
n = 70
CGA−AUA
n = 50
CGA−CCG
n = 5
CGA−CGA
n = 4
CGA−CGG
n = 5
CGA−CUG
n = 17
CGA−GCG
n = 11
CUC−CCG
n = 10
CUG−CCG
n = 63
GUA−CCG
n = 84
GUA−CGA
n = 63
0.15
0.10
0.05
0.00
-50
0
+50
Ribosome occupancy
Distance from pair in P, A sites position
B
CUG−AUA
n = 177
CUG−CGA
n = 42
GUG−CGA
n = 24
Distance from pair in P, A sites position
C
AGG−CGG
n = 69
AUA−CGG
n = 108

Figure 3.8 Comparison of Synonymous Pair Ribosome Occupancies Across Window Positions
Ribosome occupancies across window positions for each inhibitory pair (blue) together with occupancies for two synonymous pairs: a pair with an optimized 3' codon (pink) and a pair with an optimized 5' codon (orange). A codon distance of 0 marks the position with each pair occupying the inferred ribosomal P and A-sites.
(A) Occupancies for 12 inhibitory pairs (blue) with significant differences from both synonymous comparisons.
(B) Occupancy comparisons of 3 inhibitory codon pairs with higher than expected occupancy, based on permutation analysis, but not higher occupancy relative to the synonymous comparison pairs.
(C) Occupancy comparisons for 2 inhibitory pairs with no evidence of elevated ribosome occupancy.

proportion for each comparison pair, even when both pairs had an occupancy peak at these positions (corrected Fisher's p-value $\leq 6.31 \times 10^{-4}$; **Figure 3.8** and **Table 3.3**). Therefore, we conclude that ribosomes tend to translate through 12 of the inhibitory codon pairs more slowly than either of the individual codons, across synonymous pair sites.

3.4 DISCUSSION

We present evidence of fluctuating and varied elongation within yeast transcripts, whereby the identity of codons in adjacent ribosomal site positions can negatively impact the translation speed of ribosomes. Based on ribosome occupancy along yeast transcripts, we show that 15 of 17 inhibitory pairs had higher occupancies than expected by chance, and we find for 12 of these pairs that the codon pair reduced translation rates to a greater extent than synonymous pairs. While previous reports have detected variation in the decoding rates of individual codons, ours is the first to identify codon context influences on translation speed within genome-wide experimental data.

To address challenges in evaluating inhibitory codon pair influences within sparse ribosome profiling data, we have developed a computational approach with the following goals and characteristics: a focus on local translation differences through utilization of footprint counts in only a 100-codon window surrounding each inhibitory pair site; robustness to noise from low coverage windows by allowing higher coverage windows to carry more weight; estimation of ribosome occupancy significance through permutation analysis; and application of counting statistics to test alternative hypotheses (e.g. synonymous pairs) for elevated counts. While some characteristics of this approach, such as focus on a limited codon distance window and the unequal weighting of ORFs by coverage, may not serve as ideal approaches for every profiling application, these approaches are well suited to the unique challenges of context analysis. Thus, despite the relatively few number of sites and low coverage of transcripts, for 12 pairs we were able to detect significantly higher occupancies.

Ribosome occupancies were not higher than expected when the 5' codon of the pair was present in the A-site of the ribosome (and the 3' codon at the +1 position), as would be expected if the higher ribosome occupancies at inhibitory pairs were due to the slow decoding of each individual codon. Some of this observed depletion may be explained by the fact that window density measures are not independent of each other (and thus a position with the 5' codon in the A-site may appear more depleted due to the presence of large relative effect sizes within the same window). Nevertheless, we have observed that the primary inhibitory effects across window positions occur once the 5' codon of an inhibitory pair has entered the P-site of the ribosome. For 6 out of 12 inhibitory pairs with elevated ribosome densities, CGA is present in the 5' position of the pair. The prevalence of CGA at this position suggests that an intrinsic property of the CGA codon, in particular, negatively impacts accommodation of adjacent codons

into the ribosomal A-site. A spectrum of P-site effects may occur in combination with a variety of 3'-codons. While making direct density comparisons is problematic, given coverage and representation issues, in a ranking of CGA-NNN pairs by ribosome occupancy, inhibitory codon pairs identified in the GFP assay represent the more extreme examples.

In this analysis both small and large ribosome footprint fragments were present within the dataset. While we have used the combined total of short and long footprints to achieve the highest overall coverage, these different fragment sizes are postulated to represent alternative conformations of the ribosome during an elongation cycle (Gardin et al., 2014; Lareau et al., 2014). In particular, long footprints are thought to report on decoding processes, and short footprints are thought to report on events after the decoding step. Lareau et al. reported substantial variation between codons in the overall distribution of associated fragment sizes. For example, the CGG codon, present in the two inhibitory codon pairs that did not reach significance for higher overall footprint densities, fell into one extreme of this distribution, with 87 +/- 9% of A-site codon reads from small footprints (Lareau et al., 2014). Thus, future studies will need to address the degree to which inhibitory codon pairs may differentially impact footprint size distributions to help differentiate between potential mechanisms of inhibition within the translation machinery. Footprint size distribution information also has the potential to facilitate identification of ORF positions where codon pair-mediated inhibition may play a regulatory role in protein expression.

Pair	A, -1	P, A	E, P	+1, E	Ribosomal site counts	Window counts	Ribosome occupancy	Pair sites	p-value
AGG-CGA	30	33	60	69	162	2214	0.0732	91	0.0017
AGG-CGG	10	24	14	11	49	1352	0.0362	70	0.4326
AUA-CGA	28	86	125	83	294	3791	0.0776	193	0.0017
AUA-CGG	21	45	22	26	93	2308	0.0403	109	0.2538
CGA-AUA	31	41	62	48	151	2135	0.0707	145	0.0017
CGA-CCG	3	19	14	10	43	193	0.2228	13	0.0017
CGA-CGA	2	22	10	9	41	163	0.2515	13	0.0017
CGA-CGG	5	35	16	7	58	360	0.1611	13	0.0017
CGA-CUG	7	80	27	33	140	922	0.1518	62	0.0017
CGA-GCG	12	46	36	29	111	568	0.1954	27	0.0017
CUC-CCG	7	32	18	12	62	437	0.1419	30	0.0017
CUG-AUA	187	113	350	147	610	10948	0.0557	405	0.0017
CUG-CCG	89	173	235	169	577	4581	0.1260	141	0.0017
CUG-CGA	72	67	95	79	241	2581	0.0934	105	0.0017
GUA-CCG	43	200	87	163	450	5098	0.0883	206	0.0017
GUA-CGA	21	72	67	68	207	2708	0.0764	135	0.0017
GUG-CGA	1	28	18	26	72	1184	0.0608	54	0.0017

Table 3.2 Ribosome Footprint Counts at Inhibitory Codon Pair Sites
Footprint counts are tallied for 4 separate positions with the given codon pair at either: the ribosomal A-site and the codon position immediately 3' (A, -1); the ribosomal P-site and A-site (P, A); the ribosomal E-site and P-site (E, P); or the E-site and the codon position immediate 5' (+1, E). The ribosomal sites column is the cumulative count of the pair at ribosomal P, A sites; E, P sites; and +1, E site positions, while window counts include the remaining positions within a 100-codon window centered on the pair. Ribosome occupancy is the proportion of ribosomal site counts out of the window total. The pair sites column refers to the sites present in ORFs within the untreated Lareau et al. datasets (Lareau et al., 2014), and the p-value is based on permutations of the A-site codon counts within each ORF with a given pair.

Pair	Comparison	Odds ratio	Lower CI	p-value
AGG-CGA	AGA-CGA	1.69	1.40	3.36E-05
	AGG-AGA	2.81	2.40	2.99E-23
AGG-CGG	AGA-CGG	1.24	0.89	7.47E-01
	AGG-AGA	1.21	0.93	7.16E-01
AUA-CGA	AUU-CGA	1.44	1.24	3.32E-04
	AUA-AGA	2.47	2.20	3.01E-33
AUA-CGG	AUU-CGG	1.18	0.90	5.02E-01
	AUA-AGA	1.26	1.04	2.22E-01
CGA-AUA	AGA-AUA	2.45	2.10	2.81E-18
	CGA-AUU	2.11	1.74	1.98E-09
CGA-CCG	AGA-CCG	4.57	3.34	2.47E-12
	CGA-CCA	2.78	2.01	5.03E-06
CGA-CGA	AGA-CGA	7.63	5.40	1.08E-17
	CGA-AGA	5.54	3.94	1.58E-13
CGA-CGG	AGA-CGG	7.39	5.35	1.04E-21
	CGA-AGA	3.16	2.40	7.20E-10
CGA-CUG	AGA-CUG	3.68	3.12	5.05E-31
	CGA-UUG	2.80	2.30	1.15E-16
CGA-GCG	AGA-GCG	5.79	4.76	1.41E-38
	CGA-GCU	3.69	2.93	4.63E-19
CUC-CCG	UUG-CCG	2.37	1.85	4.19E-07
	CUC-CCA	5.15	3.89	2.30E-18
CUG-AUA	UUG-AUA	2.17	2.00	4.47E-46
	CUG-AUU	1.13	1.04	7.42E-02
CUG-CCG	UUG-CCG	2.05	1.87	3.20E-34
	CUG-CCA	1.93	1.77	1.88E-32
CUG-CGA	UUG-CGA	1.09	0.95	6.42E-01
	CUG-AGA	1.40	1.24	9.50E-05
GUA-CCG	GUU-CCG	1.43	1.29	1.86E-07
	GUA-CCA	2.16	1.96	7.45E-36
GUA-CGA	GUU-CGA	1.54	1.28	6.31E-04
	GUA-AGA	2.40	2.10	4.82E-23
GUG-CGA	GUU-CGA	1.24	0.96	5.63E-01
	GUG-AGA	1.71	1.38	5.54E-04

Table 3.3 Synonymous Comparison Sequences and Significance of Footprint Comparison
The odds ratio is the conditional odds ratio of each inhibitory pair footprint proportion relative to the synonymous pair (Comparison) proportion. The lower 95% confidence interval limit (Lower CI) is provided along with the corrected, one-sided, Fisher's exact p-value.

3.5 NOTES AND CONTRIBUTIONS

This chapter is part of the publication, "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast" by Caitlin E. Gamble, Christina E. Brule (University of Rochester), Kimberly M. Dean (University of Rochester), Stanley Fields, and Elizabeth J. Grayhack (University of Rochester); in press at *Cell* (2016).

The work in this chapter is my own. I thank my co-authors for discussion and suggestions. I would also like to thank Ron Haus for his thoughtful commentary and discussion regarding both the analysis and this section of the manuscript.

CHAPTER 4: Codon Pair Characteristics Implicate Wobble Decoding and Ribosomal Site Interactions

4.1 BACKGROUND

The ribosome selects a matching aminoacyl-tRNA (aa-tRNA) from the pool of competing aa-tRNAs during each elongation cycle. Upon finding a correct match, peptide bond formation occurs, followed by translocation of the A-site tRNA to the P-site and leading the E-site tRNA to exit. Availability of aa-tRNA for codons rarely used in highly expressed genes is generally thought to govern differences in translation rates between synonymous codons. In biotechnology settings where a heterologous gene (such as a GFP reporter) is overexpressed from strong promoters, tRNA availability is thought to impose rate limitations (Elf, 2003; Ikemura, 1981; Plotkin and Kudla, 2010; Robinson et al., 1984; Varenne et al., 1984; Welch et al., 2009), either directly, through starvation for particular tRNAs, or indirectly through sequestration of ribosomes along the transcript. Clusters of codons that are rarely used in highly expressed genes may exacerbate limitations imposed by tRNA availability, since, for all copies of the transcript, the presence of an additional rarely used codon in an adjacent position sequesters already limited tRNA in the ribosomal P-site for another slow decoding cycle (Kane, 1995; Rosenberg et al., 1993; Varenne and Lazdunski, 1986).

Differences in relative tRNA abundance has dominated much discussion and thought as to how synonymous codon use impacts translation efficiency (Dana and Tuller, 2014b; Gingold and Pilpel, 2011; Plotkin and Kudla, 2010; Tuller et al., 2010a). However, another factor influencing decoding speeds is anticodon:codon base pairing (Stadler and Fire, 2011). Some synonymous codons are decoded using the same tRNA but with an alternative base pair

geometry. Rather than exact, Watson:Crick base pairing at all positions, these codons are decoded by a "wobble" pairing between the first nucleotide of the anticodon and the third nucleotide of the codon. Within the A-site, rRNA rapidly assesses the minihelix geometry resulting from the anticodon-codon base-pairing interaction. When mismatched base pairs are present, as in the case of when a tRNA for the wrong amino acid binds to the mRNA, fewer hydrogen bonds are produced, resulting in dissociation of the aa-tRNA from the ribosome (Demeshkina et al., 2012; Rozov et al., 2015). In cases of wobble decoding, tRNA modifications often must pre-structure the anticodon to reduce conformational dynamics and to facilitate formation of a minihelix with few geometric differences from the canonical Watson:Crick pairs (Agris et al., 2007), thus facilitating accommodation into the A-site and subsequent translocation. Negative impacts on translation efficiency as a result of wobble decoding have been demonstrated in *E. coli* for specific sets of codons and sequence contexts (Curran and Yarus, 1989; Kato et al., 1990; Thomas et al., 1988). Recently, ribosome profiling datasets have also provided systematic evidence of reduced translation speeds at G:U and I:U wobbles in *C. elegans* and humans (Stadler and Fire, 2011) and for U:G and I:A in yeast (CUG, CCG, and CGA codons) (Lareau et al., 2014).

From the perspective that tRNA availability is limited and that it governs decoding rates in conjunction with wobble base pairing factors, the influence of synonymous codons on translation elongation efficiency has been considered in terms of an additive model of small, individual effect sizes, whereby the sum of each codon's typical decoding speed across the length of the transcript determines the overall efficiency with which a given message is translated. Several codon indices (CAI, tAI, and nTE) have been developed and used to model the additive effects of individual codons (Pechmann and Frydman, 2012; Reis, 2004; Sharp and

Li, 1987; Tuller et al., 2010b). In particular, the tRNA adaptation index (tAI) assigns an adaptiveness value to each codon based on the gene copy number of tRNAs that decode the codon together with a weight for different codon-anticodon base pairs at the codon's third nucleotide (Reis, 2004; Tuller et al., 2010a). However, these indices remain poor predictors of translation efficiency. In yeast, they explain less than 7% of the variance in protein levels per mRNA transcript (Ingolia et al., 2009b; Man and Pilpel, 2007; Tuller et al., 2010b). Many factors potentially contribute to low predictive values. Nevertheless, the low predictive values suggest the need for an extended mechanistic understanding of codon usage and incorporation of factors beyond individual decoding speeds based primarily on tRNA abundance. The ribosome is a highly coordinated machine. Interactions between tRNAs/codons in the A, P, and E sites could be mediated by numerous protein and rRNA contacts (Demeshkina et al., 2010).

We have shown that codon sequence context in the form of adjacent codons can reduce both expression of a GFP reporter and translation speed along yeast transcripts (**Chapter 2**). Here, we further explore characteristics of the 17 inhibitory codon pairs identified in the GFP assay to assess the likely contributions of tRNA availability and wobble decoding in pair-mediated inhibition. We then evaluate whether inhibition is most likely to arise from the sequential occurrence of individual codons with slow decoding speeds or a concerted pair effect, which would imply interactions between tRNA/codon sites in the ribosome. We find that I:A and U:G wobble decoding contributes to inhibition by these codon pairs and that in contrast to an additive model of individual effects, for many of the pairs inhibition is most likely to arise from pair effect involving interactions between adjacent tRNA sites within the ribosome. Our findings are inconsistent with a model of codon pair inhibition based primarily on limited tRNA availability.

4.2 RESULTS

I:A and U:G Wobble Decoding Contribute to Inhibition by Codon Pairs

The codon composition of the 17 inhibitory pairs identified in the GFP expression assay is consistent with the idea that these pairs inhibit translation and that wobble decoding mechanisms contribute to inhibitory effects. Each of the codons that are found three or more times in the 17 pairs are implicated in poor translation by their infrequent use in highly expressed genes, as measured by CAI (Sharp and Li, 1987). Four of the codons found in the inhibitory pairs are exclusively decoded via wobble by tRNA that is present at moderate to high abundance in yeast cells (3, 5, 6 and 10 tRNA gene copies; gene copy number correlates with tRNA abundance (Tuller et al., 2010a) and ranges from 1 to 16). While these tRNAs decode more than one codon, only the codons decoded by U:G and I:A wobbles are found in the identified inhibitory pairs. This observation suggests that properties of the specific anticodon:codon pairing contribute to inhibition by these pairs, rather than limited availability of the tRNA. In particular, the Arg CGA codon, the only codon in yeast decoded via a purine-purine (I:A) wobble, is found in more than half (11) of the candidate pairs (**Figure 4.1A**). All 3 of the codons that are decoded by U:G wobble are also present (**Figure 4.1**). There are six codons present in the inhibitory pairs that do not rely on a wobble decoding mechanism, since an exact-matching tRNA is present in the yeast genome. However, for two of these six codons (AGG and GUG), a low abundance (1-2 gene copies), exact-matching tRNA species competes with another tRNA species that can decoded the codon through U:G wobble (**Figure 4.1B**) (Johansson et al., 2008). (In *S. cerevisiae* this type of tRNA competition is found for one other codon, GGG, which was not present among the

A

Amino Acids	Codon Pairs
R-R	AGG-CGA
R-R	AGG-CGG
I-R	AUA-CGA
I-R	AUA-CGG
R-I	CGA-AUA
R-P	CGA-CCG
R-R	CGA-CGA
R-R	CGA-CGG
R-L	CGA-CUG
R-A	CGA-GCG
L-P	CUC-CCG
L-I	CUG-AUA
L-P	CUG-CCG
L-R	CUG-CGA
V-P	GUA-CCG
V-R	GUA-CGA
V-R	GUG-CGA

6 amino acids
11 dipeptides

10 codons
17 pairs

6 with exact match tRNA

Amino Acids	Codon Pairs
R-R	**AGG**-CGA
R-R	**AGG**-CGG
I-R	**AUA**-CGA
I-R	**AUA**-CGG
R-I	CGA-**AUA**
R-P	CGA-CCG
R-R	CGA-CGA
R-R	CGA-**CGG**
R-L	CGA-CUG
R-A	CGA-GCG
L-P	**CUC**-CCG
L-I	CUG-**AUA**
L-P	CUG-CCG
L-R	CUG-CGA
V-P	**GUA**-CCG
V-R	GUA-CGA
V-R	**GUG**-CGA

4 decoded by wobble

Amino Acids	Codon Pairs
R-R	AGG-**CGA**
R-R	AGG-CGG
I-R	AUA-**CGA**
I-R	AUA-CGG
R-I	**CGA**-AUA
R-P	**CGA**-CCG
R-R	**CGA**-**CGA**
R-R	**CGA**-CGG
R-L	**CGA**-CUG
R-A	**CGA**-GCG
L-P	CUC-CCG
L-I	**CUG**-AUA
L-P	**CUG**-CCG
L-R	**CUG**-**CGA**
V-P	GUA-CCG
V-R	GUA-**CGA**
V-R	GUG-**CGA**

7 present only in 5' or 3'

Amino Acids	Codon Pairs
R-R	**AGG**-CGA
R-R	**AGG**-CGG
I-R	AUA-CGA
I-R	AUA-**CGG**
R-I	CGA-AUA
R-P	CGA-**CCG**
R-R	CGA-CGA
R-R	CGA-**CGG**
R-L	CGA-CUG
R-A	CGA-**GCG**
L-P	**CUC**-CCG
L-I	CUG-AUA
L-P	CUG-**CCG**
L-R	CUG-CGA
V-P	**GUA**-CCG
V-R	GUA-CGA
V-R	**GUG**-CGA

3 occur in both positions

Amino Acids	Codon Pairs
R-R	AGG-**CGA**
R-R	AGG-CGG
I-R	**AUA**-**CGA**
I-R	**AUA**-CGG
R-I	**CGA**-**AUA**
R-P	**CGA**-CCG
R-R	**CGA**-**CGA**
R-R	**CGA**-CGG
R-L	**CGA**-**CUG**
R-A	**CGA**-GCG
L-P	CUC-CCG
L-I	**CUG**-**AUA**
L-P	**CUG**-CCG
L-R	**CUG**-**CGA**
V-P	GUA-CCG
V-R	GUA-**CGA**
V-R	GUG-**CGA**

B

	U		C		A		G	
U	UUU (G:U) UUC (10)	**Phe**	UCU (11) UCC (I:C) UCA (3) UCG (1)	**Ser**	UAU (G:U) UAC (8)	**Tyr**	UGU (G:U) UGC (4)	**Cys**
	UUA (7) UUG (10)	**Leu**			UAA UAG	**Stop**	UGA	**Stop**
							UGG (6)	**Trp**
C	CUU (G:U) **CUC (1)** CUA (3) **CUG (U:G)**	**Leu**	CCU (2) CCC (I:C) CCA (10) **CCG (U:G)**	**Pro**	CAU (G:U) CAC (7)	**His**	CGU (6) CGC (I:C) **CGA (I:A)** **CGG (1)**	**Arg**
					CAA (9) CAG (1)	**Gln**		
A	AUU (13) AUC (I:C) **AUA (2)**	**Ile**	ACU (11) ACC (I:C) ACA (4) ACG (1)	**Thr**	AAU (G:U) AAC (10)	**Asn**	AGU (G:U) AGC (2)	**Ser**
	AUG (10)	**Met**			AAA (7) AAG (14)	**Lys**	AGA (11) **AGG (1)**	**Arg**
G	GUU (14) GUC (I:C) **GUA (2)** **GUG (2)**	**Val**	GCU (11) GCC (I:C) GCA (5) **GCG (U:G)**	**Ala**	GAU (G:U) GAC (15)	**Asp**	GGU (G:U) GGC (16) GGA (3) GGG (2)	**Gly**
					GAA (14) GAG (2)	**Glu**		

Figure 4.1 Composition of 17 Inhibitory Codon Pairs
(A) Tables of the 17 inhibitory codon pairs identified in the GFP assay with color coding to highlight the presence of specific codons and amino acids.
(B) Parentheses next to each codon indicate either tRNA gene copy number or wobble type. Each black dot is a tRNA with exact base-pairing to the codon; lines connecting to gray dots indicate wobble decoding possibilities; white dots indicate rare decoding possibilities. Those connected by dashes generally only occur if the tRNA is over expressed (Johansson et al., 2008).

identified inhibitory pairs). Overall, all but 2 of the 17 inhibitory pairs have a codon that exclusively relies on wobble decoding. Thus, the prevalent codons decoded by I:A and U:G wobble in the 17 inhibitory codon pairs suggest that wobble contributes to the inhibitory effect of these pairs.

To directly assess the contribution of limited tRNA availability versus wobble decoding on inhibition by codon pairs, we compared the expression levels (syn-GFP^{SEQ}; **Chapter 2**) of those variants with a given inhibitory codon pair to those with a synonymous codon pair, decoded by the same tRNAs as the inhibitory pair. Two inhibitory pairs (AGG-CGG and AUA-CGG) were excluded from this analysis, because each codon is decoded primarily by a tRNA species that decodes just a single codon (via exact Watson and Crick base-pairing). In all other cases, one or more wobble-decoded codons in the inhibitory pair were substituted with a codon that is decoded by an exact Watson and Crick base-pair match to the anticodon. We found that for the remaining 15 inhibitory codon pairs, the syn-GFP^{SEQ} distributions were significantly lower than syn-GFP^{SEQ} distributions for synonymous pairs decoded by exact base pairing (**Figure 4.2**, corrected Wilcoxon p-value ≤ 0.03). Thus anticodon:codon wobble base-pairing, rather than limited tRNA availability, contributes to inhibition by 15 pairs.

We reported in a previous chapter that overproduction of tRNA matching one of the inhibitory pair codons partially suppressed inhibition by nearly all of the inhibitory codon pairs (**Chapter 2**). This finding linked codon pair-mediated inhibition to translation and showed that supplementing the existing tRNA pool can, at least partially, overcome inhibition by these pairs. Since substantially increased tRNA abundance is likely to facilitate faster decoding, even if decoding is otherwise impaired, the finding does not necessarily imply that inhibition arises from limited tRNA availability. Many of the suppressing tRNA species were non-native, exact base-

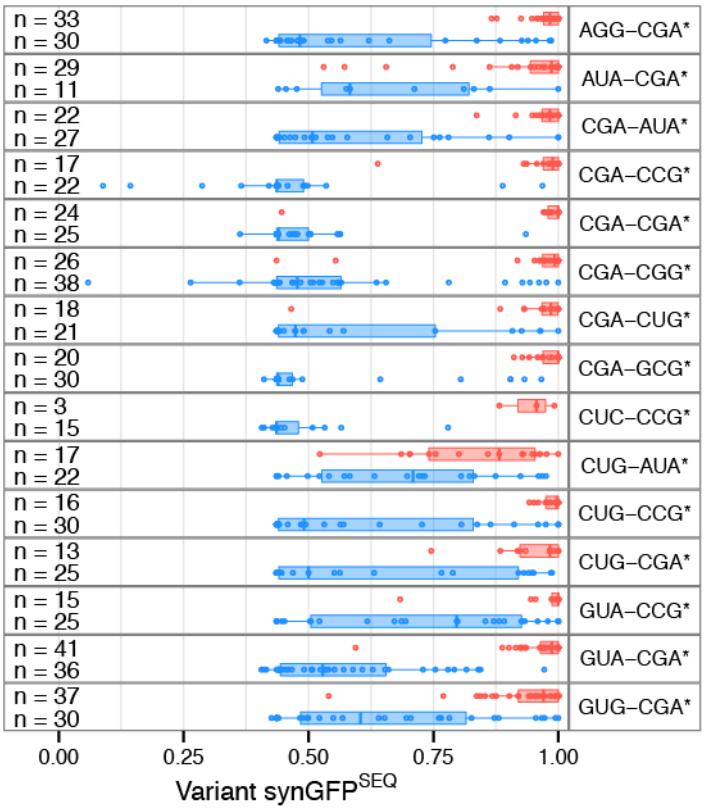

Figure 4.2 Pairs that Rely on I:A and U:G Wobble Decoding Have Reduced Expression Compared to Watson:Crick Matches
syn-GFP^{SEQ} distributions of variants with one of 15 inhibitory codon pairs (blue) compared to a pair using an identical set of tRNAs but decoded by Watson:Crick pairing (red). Boxplot shows median centerline and edges mark the first and third quartiles.

pairing tRNA species that may have suppressed inhibitory effects due to wobble. Thus, to further evaluate the impact of wobble decoding, we expressed different tRNA species for the 3' codon of an inhibitory pair, and we compared the degree of suppression achieved by expressing native, wobble-decoding tRNA to the degree of suppression by expressing exact-base pairing tRNA. Generally, native, wobble-decoding tRNA for the 3' codon of an inhibitory pair was a weak suppressor of inhibition (imparting 1.35 to 2.20 fold increases in expression; **Table 4.1**), supporting the notion that inhibition by pairs is not primarily due to limited tRNA. Exact base-pairing tRNA was a more effective suppressor (imparting 2.11 to 7.69 fold increases in

expression) for seven tested pairs, including three inhibitory pairs with a 3' Pro CCG codon and three pairs with a 3' Arg CGA codon (**Table 4.1**). For the CGA-CUG pair, suppression by the exact base-pairing tRNA was only marginally better than the wobble decoder. Overall, since correcting wobble decoding improved translation more effectively than increased amounts of the native tRNA and since synonymous pairs decoded by exact base pairing with identical tRNAs did not show inhibition in the GFP assay, we conclude that I:A and U:G wobble decoding contributes to codon pair-mediated inhibition.

Inhibitory Codon Pair	3' Codon tRNA Vector	Match	Fold Change	Fold Change Difference
CGA-CCG	Empty	None		5.59
	tP(CGG)*	Exact	7.69 ±0.61	
	tP(UGG)	U:G wobble	2.10 ±0.23	
CGA-CUG	Empty	None		0.23
	tL(CAG)*	Exact	2.43 ±0.13	
	tL(UAG)	U:G wobble	2.20 ±0.24	
CUC-CCG	Empty	None		4.7
	tP(CGG)*	Exact	6.44 ±1.09	
	tP(UGG)	U:G wobble	1.74 ±0.51	
CUG-CCG	Empty	None		1.46
	tP(CGG)*	Exact	2.81 ±0.48	
	tP(UGG)	U:G wobble	1.35 ±0.23	
CUG-CGA	Empty	None		0.93
	tR(UCG)*	Exact	2.25 ±0.32	
	tR(ICG)	I:A wobble	1.32 ±0.19	
GUA-CGA	Empty	None		0.65
	tR(UCG)*	Exact	2.11 ±0.11	
	tR(ICG)	I:A wobble	1.46 ±0.12	
GUG-CGA	Empty	None		0.39
	tR(UCG)*	Exact	1.77 ±0.27	
	tR(ICG)	I:A wobble	1.38 ±0.14	

Table 4.1 Wobble-Decoding tRNAs are Weaker Suppressors of Inhibition
Each tRNA vector is a LEU2 2u plasmid. Stars indicate non-native tRNA. Fold change refers to the ratio of tRNA vector GFP^{FLOW} to empty vector GFP^{FLOW}. See APPENDIX A for sequences and GFP^{FLOW} of individual constructs.

Codon Order is a Critical Factor for Inhibition

In a previous chapter, we found that the adjacency of each codon in the pair is important to inhibition, demonstrating that codon context influences inhibition; otherwise inhibition by two independent codons would not have changed with the presence of an intervening codon (**Chapter 2**). However, this finding did not differentiate between a model of context inhibition based on sequential occurrence of two independent (but inefficient) elongation cycles and a model in which inhibition arises from specific properties of the pairing within the ribosome, perhaps during a single elongation cycle. We reasoned that if pairs limit translation by a pair effect within the ribosome (rather than during sequential reactions), then the position of each codon in the ribosome would be critical for its effects, and thus the order of codons in an inhibitory codon pair should be important. Of the 17 pairs identified in this study, the CGA-CGA pair is composed of identical codons, and two sets of pairs are inhibitory with the codons present in either order (AUA-CGA, CGA-AUA; and CUG-CGA, CGA-CUG), leaving 12 pairs in which a single order of the codons is included in the list of inhibitory codon pairs. For each of these pairs, we compared the GFP expression of variants with the inhibitory pair to variants with the same two codons in reverse order. For all 12 pairs, we observed that variants with the inhibitory pair tended to have lower syn-GFP^{SEQ} scores than variants with the reverse pair (corrected Wilcoxon p-value ≤ 0.009; **Figure 4.3A**). We further validated these findings by showing that two distinct variants with the CUC-CCG inhibitory pair had lower GFP^{FLOW}, relative to a synonymous variant with the optimized pair, whereas two variants with the reverse pair (CCG-CUC) had high GFP^{FLOW} (**Figure 4.3B**). Pairs with significant differences from the reverse pair included the two pairs (AGG-CGG and AUA-CGG) for which the contribution of tRNA availability to inhibition could not be evaluated by comparison to a synonymous pair using the

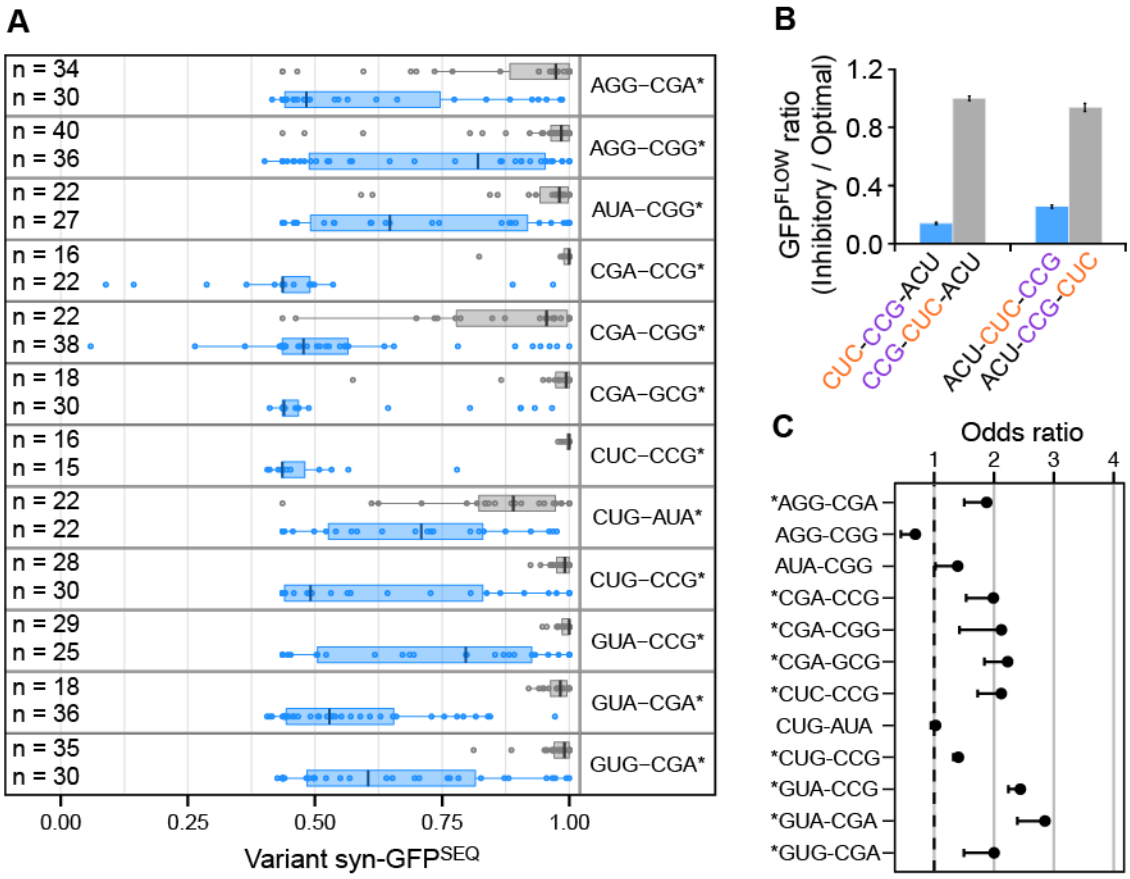

Figure 4.3 Inhibition Depends on Codon Order and Pair Effect
(A) syn-GFP^{SEQ} distribution for variants with each of the 17 inhibitory codon pairs (blue) compared to variants with the same pair of codons in reverse order (gray). Boxplot edges mark the first and third quartiles. Stars indicate a corrected Wilcoxon p-value ≤ 0.008.
(B) Inhibitory:optimal GFP^{FLOW} for variants with the CUC-CCG inhibitory pair (blue) and reverse pair (gray) in two different positions (amino acid 6,7 and amino acid 7,8). Error bars represent ±SD.
(C) Conditional odds ratio for each inhibitory pair's proportion of ribosome footprints (in a 100-codon window) relative to the proportion for a pair with the reverse codon order. Error bars indicate the lower 95% confidence interval limit.
same tRNAs. The importance of codon order to inhibition by these pairs and the other wobble-

decoded pairs further demonstrates that tRNA availability is not a primary determinant of inhibition and suggests that inhibition occurs with specific pairs in particular ribosomal site positions.

Codon Order is a Critical Factor in Elongation Speed

We also compared translation rates along yeast transcripts between the inhibitory pair and reverse order pair, using the Lareau et al. ribosome profiling dataset. Of the 12 inhibitory pairs for which a single order of codons was inhibitory in the GFP assay, 2 pairs (AGG-CGG and AUA-CGG) did not have elevated ribosome occupancies. We think that inhibition by these two pairs is likely due to alternative mechanisms. Of the remaining 10 pairs, we found that 9 had a significantly higher overall proportion of footprints when the pair occupied ribosomal site positions than when the reverse pair occupied ribosomal site positions (corrected Fisher's exact p-value $\leq 6.68 \times 10^{-4}$; **Figure 4.3C**). Since the order of codons within many inhibitory pairs is central to elevated ribosome occupancy, we conclude that slower translation of these pairs is most likely due to specific properties of the pairing.

Codon Position Matters for tRNA Suppression Outcomes

If an inhibitory codon pair impairs a single elongation cycle when the codons are located in the ribosomal P (5' codon) and A (3' codon) sites, then overproduction of aa-tRNA corresponding to the 3' codon is likely to suppress inhibition. However, we would not expect native aa-tRNA corresponding to the 5' codon to suppress, since the 5' codon would already be located within the ribosome at the time of inhibition. Thus, we evaluated the degree of suppression by native tRNA corresponding to the 5' codon and compared results to those for tRNA decoding the 3'

codon. With respect to the 3' codon, tRNA for 9 of the 11 evaluated pairs suppressed inhibition (CGA-CGA excluded; **Figure 4.4A** purple). In contrast, native 5' codon tRNA increased expression for only 2 of the 11 pairs tested (CUC-CCG and marginally for GUA-CGA) (**Figure 4.4A** orange). In one of these cases (CUC-CCG), the native, exact-matching tRNA for the 5' codon improved expression, but overproduction of a native, wobble-decoding tRNA for the same codon reduced expression. Thus, we think overproduction of the exact base-pairing tRNA for the 5' codon improved expression by enabling the exact base-pairing tRNA to out-compete native, wobble-decoding tRNA.

The failure of $tRNA^{Arg(ICG)}$ to suppress inhibition by pairs with a 5' CGA was not due to a failure to overproduce charged tRNA; charged $tRNA^{Arg(ICG)}$ increased 9-fold when the tRNA was expressed from a high copy 2μ plasmid (**Figure 4.4B**), but suppression was undetectable. Furthermore, use of higher copy *leu2-d* 2μ plasmid (Beggs, 1978) resulted in a 20-fold increase in charged $tRNA^{Arg(ICG)}$ (**Figure 4.4B**), but no detectable suppression of inhibitory pairs with a 5' CGA (**Figure 4.4C**). Thus, the 5' and 3' codons in an inhibitory pair behave differently, suggesting inhibition arises from the specific position of each codon in the ribosome.

In contrast to the native, wobble-decoding tRNA, overexpression of the non-native, exact-match tRNA ($tRNA^{Arg(UCG)*}$) suppressed all CGA tested pairs, including those with CGA in the 5' position (**Figure 4.4D**). This was also true of non-native, exact-match tRNA ($tRNA^{Leu(CAG)*}$) for the 5' codon of CUG-CCG (**Figure 4.4D**). These results are consistent with the idea that codon-anticodon interactions in the P site affect decoding in the A site.

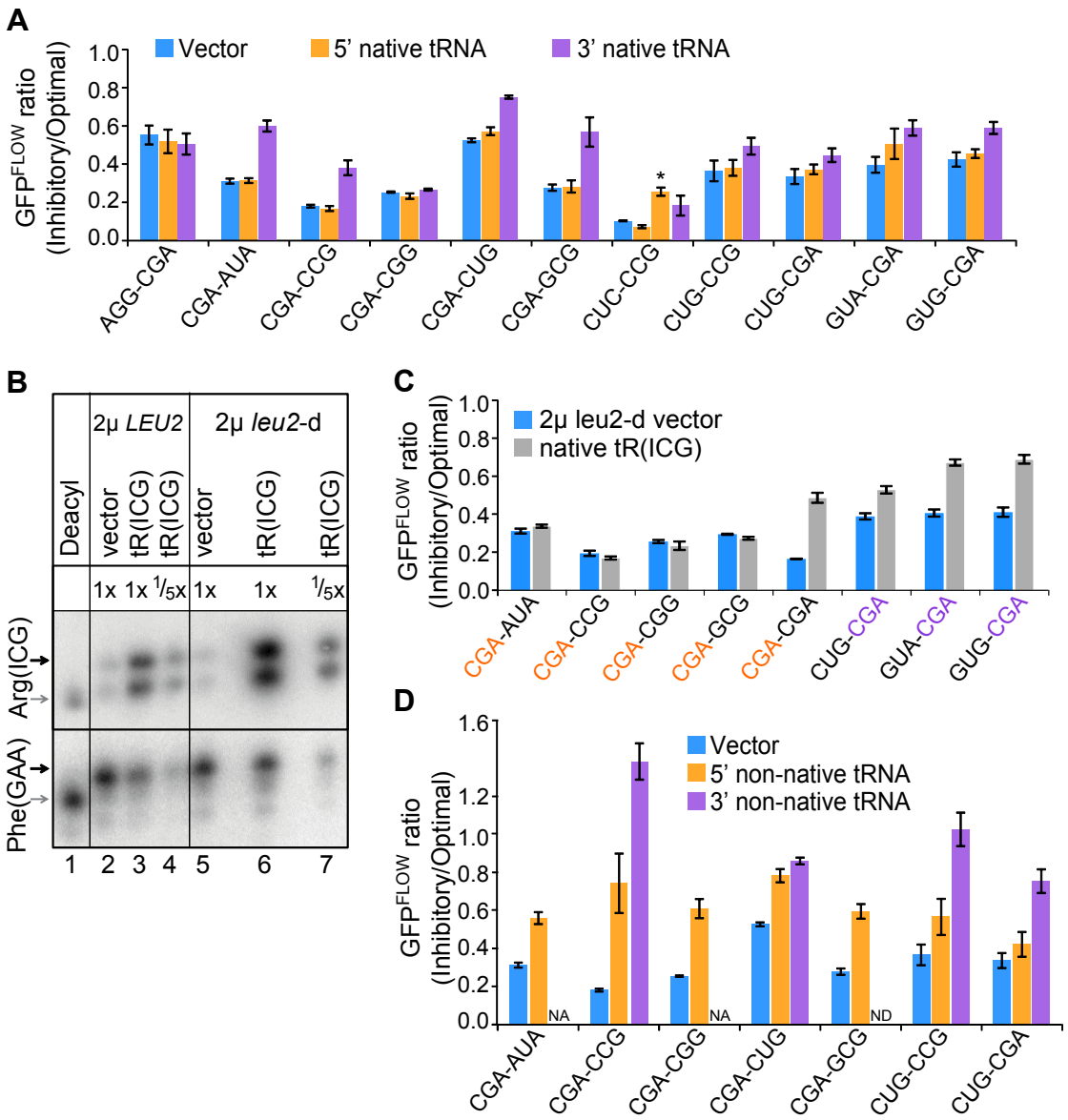

Figure 4.4 Effectiveness of tRNA Suppression Varies by Codon Position in the Pair
(A) Inhibitory:optimal GFP^{FLOW} for variants with the indicated pairs in strains with either an empty 2μ vector (blue) or native tRNA expressed from the vector (5' tRNA orange; 3' tRNA purple). For the CUC-CCG pair, a native wobble decoding tRNA (no star) and a native exact-matching tRNA (star) compete to decode the 5' CUC codon. Error bars represent ±SD.
(B) Charged tRNA levels increase when expressed from either a 2μ or 2μ *leu2*-d vector. Acidic Northern blot probed for charged (black arrow) and uncharged (gray arrow) tRNA$^{Arg(ICG)}$ and tRNA$^{Phe(GAA)}$.
(C) Effects of native tRNA$^{Arg(ICG)}$ on inhibitory:optimal GFP^{FLOW} when expressed from a 2μ *leu2*-d vector. Tested pairs have a 5' (left) or 3' (right) CGA codon. Error bars represent ±SD.
(D) Effects of exact base-pairing, non-native tRNA expressed from a 2μ vector (5' tRNA orange; 3' tRNA purple). Error bars represent ±SD.

4.3 DISCUSSION

We have found that interactions between individual codons and the tRNA anticodon contribute to inhibition by codon pairs and that codons in an inhibitory codon pair have differing effects based on their position in the pair. Both findings demonstrate that translation inhibition by synonymous codon variation encompasses factors beyond tRNA availability, and they suggest greater interplay between codons and tRNAs within the ribosome than previously understood.

Wobble decoding, rather than tRNA abundance, is a key factor for inhibition based on three observations: First, wobble codons frequently occur in the inhibitory pairs; all four codons that are decoded primarily by I:A or U:G wobble are present in the identified pairs and 15 of 17 pairs have one of these codons. Moreover, 3 of the 4 codons found only in the 5' position of inhibitory pairs are decoded by two competing tRNAs, one of which decodes by U:G wobble (Johansson et al., 2008). Second, GFP variants with inhibitory pairs had lower expression than those with synonymous pairs using an identical tRNA set. Third, wobble-decoding tRNAs were weak suppressors (when decoding 3' codons) or did not suppress (when decoding 5' CGA codons), despite increases in charged tRNA species. If inhibition were primarily a result of tRNA abundance limitations, we would expect to see stronger suppression with overproduction of an aa-tRNA.

Only I:A and U:G wobble decoding mechanisms were associated with reduced GFP expression and inhibitory codon pairs, whereas I:C and G:U wobbles were not. This observation squares relatively well with wobble-decoding weights used in the tRNA adaptation index. These weights were inferred by optimizing the correlation between codon use and protein expression in *S. cerevisiae* (Reis, 2004) and gave I:A and U:G wobbles the greatest discounts to their

adaptation value. Our findings are also consistent with biochemical observations of anticodon:codon base-pairing, since G:U base-paring forms an anticodon-codon minihelix that is almost isomorphic/isosteric to a Watson-Crick base-pair. In contrast, U:G and I:A pairs have wide geometries across the anticodon-codon minihelix (Agris et al., 2007), especially I:A pairs. Whereas a G nucleotide in the anticodon third position effectively pairs with an unmodified U in the codon sequence (G:U), for effective decoding by the reverse orientation (U:G), the anticodon U must be modified (Agris, 2004; Nasvall et al., 2007). Correct base-pairing geometry is critical to acceptance by the ribosome and a rapid GFP hydrolysis reaction. In this way, tRNAs act as allosteric effectors of the forward elongation reaction (Agris et al., 2007; Murphy et al., 2004; Ogle and Ramakrishnan, 2005) and may differ in their effectiveness.

A network of protein and RNA connections between the P-site and A-site (Demeshkina et al., 2010) may further influence the geometry and stability of A-site anticodon-codon helices. Seven codons are found in only the 3' or only the 5' position of inhibitory pairs. The idea that the ribosomal site position of each codon in an inhibitory pair impacts pair inhibition is supported by the dependence of both GFP inhibition (12 pairs) and ribosome footprint densities (9 of 12 pairs) on the order of codons in the pair. Furthermore, overexpression of tRNAs had different impacts on expression, depending in part on whether the tRNA decoded the 5' or 3' codon of an inhibitory codon pair (7 pairs). Given the importance of codon positions within the pair, we conclude that inhibition arises from a pair effect, rather than from sequential, independent codon effects. As a pairing effect, inhibition is most likely to occur primarily during a single elongation cycle. Concerted effects of adjacent codons are plausible during tRNA accommodation, formation of the hybrid state, translocation, and/or tRNA exit. Since overproducing the native tRNA for the 3' codon improved expression for 10 pairs, these pairs may impair acceptance of

the 3' codon into the A site. Previous studies have demonstrated that tRNA:codon interactions at the P site can impact A site interactions during programmed frameshifting (Atkins and Bjork, 2009) and as part of a quality control mechanism after peptide bond formation in *E. coli* (Zaher and Green, 2008). Our findings extend this understanding of ribosomal site interactions to include impacts on elongation efficiency mediated by specific codon-pairs.

4.4 METHODS

Statistical Methods

We ran one-sided Wilcoxon rank sum tests to compare syn-GFP^{SEQ} distributions. Wilcoxon rank sum tests were carried out in R. For 15 pairs with wobble-decoded codons, we compared the distribution of variants with the inhibitory pair to the distribution for a pair decoded by identical tRNAs. For 12 inhibitory pairs we compared the distribution of variants with the inhibitory pair to the distribution for variants with the codons in reverse order. In each case, we corrected the Wilcoxon p-values for 67 tests (including previous tests comparing out-of-frame and separated codon variants; **Chapter 2**) using the Holms-Bonferroni procedure.

Additional Methods

See Chapter 2.4 METHODS for GFP expression scoring and APPENDIX B: SUPPLEMENTAL MATERIALS & METHODS for strain growth and acidic northern blot.

4.5 NOTES AND CONTRIBUTIONS

This chapter is part of the publication, "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast" by Caitlin E. Gamble, Christina E. Brule (University of Rochester),

Kimberly M. Dean (University of Rochester), Stanley Fields, and Elizabeth J. Grayhack (University of Rochester); in press at *Cell* (2016).

I wrote custom software for the computational analysis; created Figures 4.1, 4.2A, and 4.3A and C; and wrote the associated manuscript together with EJG, SF, and CEB. CEB performed flow cytometry; tRNA suppression analysis; and made Figures 4.2B, 4.3B, and 4.4.

CHAPTER 5: 5'-mRNA Structure and its Impacts on Initiation Efficiency in Yeast

5.1 BACKGROUND

Due to the degeneracy of the genetic code, the same amino acid sequence can be encoded by many different mRNA sequences. In addition to changes in the use of particular codons and codon pairs, one major consequence arising from sequence variation is the generation of diverse mRNA folding preferences. Strong RNA secondary structure in the 5' end of ORFs is associated with reduced protein expression in both *E. coli* and *S. cerevisiae* (Goodman et al., 2013; Gu et al., 2010; Kudla et al., 2009; Tuller et al., 2010b). In *E. coli*, Kudla et al. measured the fluorescence of 154 GFP variants with an average of 114 synonymous changes in each variant and found that the overall degree of predicted structure in the first ~40 nucleotides explained between 44% and 59% of the variation in the 250-fold range of library fluorescence scores (Kudla et al., 2009). Similarly, Goodman et al. analyzed a library of 14,234 GFP variants and found that structure predictions centered 10 nucleotides downstream of the start explained roughly 34% of the variation (Goodman et al., 2013). These observations raise the question of what mechanism links mRNA structure in the first 30-40 nucleotides with protein translation.

In *E. coli*, mRNA conformational changes can turn translation on and off through exposure and blockage of the ribosome-binding site (RBS) (de Smit and van Duin, 1994; Kortmann and Narberhaus, 2012). Blocking of the RBS is a possible mechanism of inhibition in both the Kudla et al. (2009) and Goodman et al. (2013) studies. However, these later studies found the strongest relationships between structure and expression in windows slightly downstream of the start site, supporting the notion that reduced expression is most likely linked to start codon recognition rather than ribosome binding. In bacteria, mRNAs that do not contain

specific sites for binding ribosomes have an especially prominent lack of structure near the start codon, further supporting the notion that access and recognition of the start codon is a key factor for efficiency, especially when initiation is independent of an RBS (Mortimer et al., 2014; Scharff et al., 2011). Eukaryotes lack ribosome binding sites. Based on a scanning model of translation initiation in eukaryotes (Kozak, 2002), the small ribosomal subunit enters the 5'-end of an mRNA transcript and in complex with methionine initiator tRNA scans through the 5' untranslated region (UTR) until it encounters a start codon, recognized by the codon to anticodon base pairing with the initiator tRNA (Cigan et al., 1988). Start codon recognition is in some way facilitated by favorable flanking sequence, called the Kozak sequence in vertebrates or Kozak-like sequence in other eukaryotes. When the pause from this base pair interaction is sufficiently long, GTP hydrolysis releases initiation factors, allowing the large ribosomal subunit to join with the small unit and form a complete 80S initiation complex (Kozak, 2005). If mRNA structure were to impair expression, either by preventing recognition of the start codon during scanning or by disrupting formation of a complete initiation complex, then we might expect some forms of 5' structure to have more detrimental effects then others; base pair stems of a given strength, position, and/or shape could reduce expression levels to a greater extend than other types of stems within the same 5' region.

A detailed working knowledge of how mRNA structure parameters impact protein expression levels has practical importance for designing optimal expression constructs used for diverse applications. Yet there are currently few experimentally-based guidelines, especially outside of *E. coli*. Synthetic libraries of reporter genes, like the GFP reporters used by Kudla et al. (2009) and Goodman et al. (2013), provide a means of evaluating many different 5' UTR or

mRNA sequences driven by the same promoter and encoding an identical gene product. Both studies examined RNA structure through computational predictions.

Computational prediction of mRNA secondary structure is often achieved through free energy minimization calculations, which quantify the favorability of possible structures from base-pair bonding rules and energies in order to identify the most energetically favorable structure. In addition, partition function calculations consider many possible structures to assign the probability of specific base pairs (Reuter and Mathews, 2010; Seetin and Mathews, 2012; Zuker, 2000). However, these predictive models have limitations, especially with regard to predicting long-range interactions and pseudoknots. Experimentally, mRNA secondary structure is deciphered using nuclease enzymes or chemical probes that assess the paired/unpaired state of RNA nucleotides. Recently, the coupling of these approaches with high-throughput sequencing has enabled assessment of RNA structure across transcriptomes (Ge and Zhang, 2015; Kertesz et al., 2010; Li et al., 2012; Lucks et al., 2011; Rouskin et al., 2014; Underwood et al., 2010). However, with these high throughput experimental approaches, it remains difficult to compare open reading frames with subtle variation and high degrees of sequence similarity (as in the case of libraries composed of variants of a single reporter). Also, the ease and speed of computational prediction often makes it a preferable, initial approach for gene design.

Building on findings in *E. coli*, there is a need to characterize the impact of 5' RNA structure on translation efficiency in eukaryotic systems and to further develop practical gene design guidelines. Here, I have applied available computational prediction tools to investigate the impact of mRNA structure on translation initiation in yeast. I evaluate characteristics in predicted mRNA structure that differentiate low expression variants from high expression variants in a synthetic library of 35,811 GFP coding sequences, each with a randomized, three-codon

insertion starting at nucleotide +16 (codon position 6) (see **Chapter 2**). I learn that while the degree of 5' structure is an important factor related to expression, structural predictions limited to the first 40 nucleotides do not always provide an accurate assessment of whether structure is likely to impact expression. Identifying positions of the most likely base pairs, especially for the four nucleotides immediately downstream of the start codon and with consideration to base pairing positions over a ~100 nucleotide window, is a critical step in assessing the relationship between predicted structure and expression level outcomes.

5.2 RESULTS

The chromosomally integrated yeast reporter (Dean and Grayhack, 2012) used in our study has a bidirectional *GAL1,10* promoter that separately drives expression of both the GFP variant and RFP, allowing normalization of GFP expression to that of RFP and thus control for transcriptional effects. The construct encodes a GFP fusion protein with a site for 3C protease, an HA epitope, and His6, followed by superfolder GFP. Thus our overall GFP construct had substantial 5' sequence differences from constructs used in Kudla et al. and Goodman et al. (Goodman et al., 2013; Kudla et al., 2009). The library of GFP variants was created by inserting 3 randomized codons into amino acid positions 6-8 of the fusion protein (see **Chapter 2** for description of library construction and GFP expression scoring).

Predicted Degree of 5' Structure Weakly Correlates with Reduced Expression

To evaluate the impact of RNA structure on expression, we first examined the correlation between predicted degrees of structure and reduced expression, using both minimum free energy calculations, similar to Kudla el al. (Kudla et al., 2009), as well as partition function base pair

probability comparisons, similar to Goodman et al. (Goodman et al., 2013). In mRNA sequences the A of the AUG start codon is designated as the +1 position, while the preceding base is -1. As with previous reports (Goodman et al., 2013; Kudla et al., 2009), we found significant correlations between 5' structure and reduced expression across our 35,811 variants. Also similar to previous reports, the position of peak correlation centered on positions +5 through +7 (10 nucleotide window) and +10 through +14 (20 nucleotide window) (**Figure 5.1**). These structure correlations are based the mean probability of a nucleotide being double-stranded at positions within the window compared to the mean probability for the same window in a high-expression, synonymous variant (relative pair probability; see 5.4 Methods for more detail). Despite significant correlations, however, the degree of 5' structure was an extremely weak predictor of expression levels (Spearman's $r^2 = 0.02$ and p-value = 2.33×10^{-142} for a -4 to +38 minimum free energy window, as in Kudla et al. (2009); and Spearman's $r^2 = 0.01$ and p-value = 2.34×10^{-102} for a 20-nucleotide relative pair probability window centered at +10, as in Goodman et al. (2013); **Figure 5.1**).

Three contributing factors may explain the weak correlation: 1) A low degree of structure in the library overall. We had designed our vector to minimize local 5'-end structure, and we found that only 3% of sequences had local structure predictions that fell into the inhibitory range reported in Kudla et al. ($\Delta G \leq -10$ kcal/mol for the -4 to +38 window; (Kudla et al., 2009)). 2) Limited resolution of intermediate expression levels. We designed our cell-sorting assay primarily to sample outlier expression differences, and we did not obtain high-resolution measures for a large number of sequences with intermediate expression (variants with less than a ~25% reduction in expression), where local structure may explain more expression level variance. 3) More extreme expression effects could result from specific local structural features,

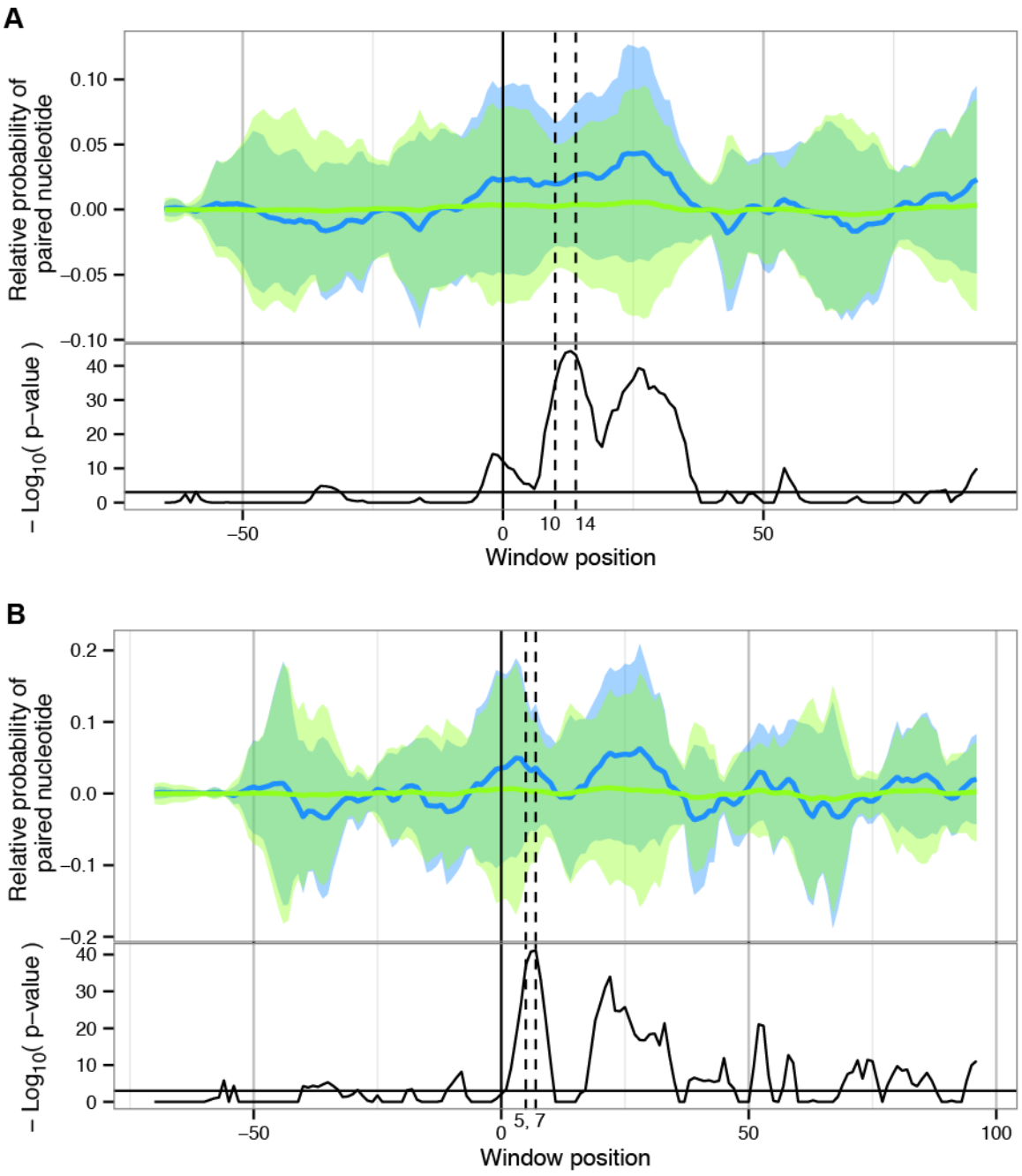

Figure 5.1 Relative Probability of Nucleotide Pairing in a Sliding Window
Predicted structure, relative to synonymous references, in a sliding window of 20-nucleotides (A) and 10-nucleotides (B). X-axis position is the window center. Solid vertical line marks the translation start. Dashed lines indicate a range of window center positions with peak expression correlations. Upper panel shows the mean (line) ± SD (shading) for low expression variants (blue; n = 1,119) and high expression variants (lime green, n = 29,565). Lower panel shows significance of each window's correlation with expression.

long-range structural interactions, or other inhibition mechanisms entirely (e.g. codon pairs or mRNA degradation).

High Probability of Base-Pairing at +3 through +6 Nucleotides Increases the Likelihood of Low Expression

We reasoned that if specific 5' structure characteristics were the primary cause of reduced expression, then these characteristics would most likely involve nucleotides in the window positions having a higher overall correlation with expression. We further reasoned that variable positions, from +16 to +24 in our library, would directly contribute to relevant base pairing differences in many variants. Thus, to investigate whether specific base pairs contributed to expression inhibition, we first examined the direct base pairing probabilities between nucleotides downstream of the translation start site and variable region positions. We asked whether variants with pairing at specific positions tended to have lower expression, and to evaluate this association we compared variants in 3 expression categories: high (≤ 1 standard deviation away from the mean; n = 29,565), intermediate (> 1 standard deviation and < 3 1 standard deviations below the mean; n = 5,148), and low (≥ 3 standard deviation below the mean; n = 1,119). For each variant and for each nucleotide position between the start site and the variable region, we took the probability of base pairing with at least one variable base. Then for each nucleotide position, we found the mean probability in high expression variants and in low expression variants. We saw that on average, high and low expression variants showed similar base pairing probability at each position. However, the mean probabilities at positions +3 through +6 suggested these position had increased base pairing in low variants relative to high variants (**Figure 5.2A**). The distribution of base pairing probabilities for these positions was highly

skewed. The vast majority of both high and low variants had very low probabilities of pairing. However, within the library there were 75 variants for which at least one of the +3 through +6 nucleotides had an extremely high (> 90%) probability of pairing with a variable region position. Of these 75 variants, more than half had below average expression (intermediate or low category), with 20% (n = 15) in the low category and 34% (n = 26) in the intermediate category. At lower base pair probabilities (between 70% and 90%), 30% (n = 166) of variants had below average expression, and intermediate variants made up a larger component of the below average expressers (25% intermediate compared to 5% low). We conclude that variants with the highest probabilities of pairing between variable positions and nucleotides in and immediately downstream of the start codon (+3 through +6) are more likely to have low expression.

In native yeast transcripts, the Ser UCU codon (which immediately follows the start codon in our GFP construct) is used at the +4 through +6 positions in roughly 50% of highly expressed yeast genes (Gingold and Pilpel, 2011; Hamilton et al., 1987). These nucleotides compose the 3' end of the yeast Kozak-like sequence context. In particular, a U at the +4 position is strongly associated with high expression. Our finding that RNA base pairing with nucleotides in the Kozak-like sequence leads to an increased likelihood of low expression suggests that base-pairing reduces the favorability of Kozak-like sequence, most likely by disrupting start codon recognition and translation initiation.

Within the same group of 75 variants, we compared the predicted degrees of structure in the 40-nucleotide 5' window (Kudla et al., 2009) between the low (20%; n = 15) and high expression variants (45%; n = 34) (**Table 5.1**). Most of the low expression variants tended to have higher degrees of structure (lower ΔG values) than the high variants (**Figure 5.2**). The 5' window prediction was a better indicator of reduced expression (Spearman's $r^2 = 0.24$, p-value =

Variable DNA Sequence	syn-GFP^{SEQ}	Expression category	+3 Prob.	+4 Prob.	+5 Prob.	+6 Prob	ΔG
TCAGTGGGC	0.48	low	99.9%	99.9%	99.9%	99.8%	-14.953
CAGTGGGCA	0.52	low	99.8%	99.8%	99.9%	99.8%	-14.622
GCGGTAGAC	0.59	low	99.9%	100.0%	100.0%	99.8%	-14.243
ACAGTAGAC	0.44	low	100.0%	100.0%	100.0%	99.8%	-13.826
CAGTGGGCG	0.55	low	98.0%	98.0%	99.6%	99.5%	-13.765
TCAGTAGAT	0.64	low	97.7%	99.8%	100.0%	99.8%	-13.76
AGTAGACGG	0.54	low	95.6%	99.5%	99.8%	99.7%	-13.344
TCGGTAGAT	1.00	high	97.7%	99.8%	100.0%	99.8%	-13.289
CAGTGGGCT	0.45	low	97.9%	97.9%	99.5%	99.4%	-12.78
GAGGTAGAC	1.00	high	99.3%	99.3%	99.3%	99.1%	-12.661
GGTAGACAG	0.99	high	99.7%	99.8%	99.8%	99.7%	-12.513
GTAGTAGGC	0.47	low	99.5%	99.4%	99.4%	99.3%	-12.235
AGGGTAGAC	1.00	high	99.1%	99.2%	99.2%	99.0%	-12.035
TCAGCAGAC	0.98	high	99.9%	100.0%	100.0%	98.8%	-11.776
CAGCGGACA	0.96	high	99.1%	99.1%	99.1%	98.4%	-11.756
CTCGGTAGA	0.98	high	0.0%	94.7%	99.8%	99.7%	-11.742
CTAGTAGGC	0.57	low	99.9%	99.9%	99.9%	99.8%	-11.551
ATAGTAGGC	0.54	low	99.8%	99.8%	99.8%	99.7%	-11.317
AGACGTTTT	0.52	low	99.9%	99.9%	99.9%	96.4%	-11.293
TCAATAGAC	0.97	high	99.9%	100.0%	100.0%	99.8%	-11.201
TCAGTTGGC	0.97	high	99.7%	99.3%	99.5%	0.3%	-11.155
CAGTTGACG	0.63	low	98.5%	99.3%	99.3%	0.0%	-11.038
TCAGTAAAC	1.00	high	97.4%	97.4%	0.0%	99.4%	-10.936
TCAGCAGGC	1.00	high	99.9%	99.8%	99.8%	98.7%	-10.705
AGTAGACTT	1.00	high	99.3%	99.9%	99.9%	99.8%	-10.585
TCAGTAGTC	0.50	low	85.9%	0.0%	98.3%	99.7%	-10.486
CGTAGACGT	1.00	high	99.6%	99.7%	99.7%	99.5%	-10.421
TCGTTAGAC	0.97	high	99.5%	99.6%	99.6%	99.3%	-10.419
GTTAGTAGA	0.99	high	0.0%	94.7%	99.7%	99.7%	-10.381
CAGCAGGCA	1.00	high	99.7%	99.6%	99.7%	98.6%	-10.374
CGGTACACA	0.99	high	99.1%	99.1%	0.0%	99.2%	-10.134
AGCAGTAGA	0.98	high	0.0%	94.7%	99.7%	99.7%	-10.047
CGCAGTAGA	0.96	high	0.0%	94.8%	99.8%	99.8%	-10.047
CTCGTAGGC	0.44	low	99.7%	99.7%	99.6%	99.6%	-10.01
GTAGACAGT	0.97	high	99.6%	99.6%	99.6%	99.5%	-9.97
CTCAGAGAT	0.99	high	96.9%	99.1%	99.2%	98.2%	-9.443
ACCAGTAGA	1.00	high	0.0%	94.6%	99.6%	99.6%	-9.435
AGTGACATT	0.99	high	99.1%	99.1%	97.6%	1.5%	-9.43
CGGCATTTT	1.00	high	99.5%	97.7%	98.1%	1.3%	-9.388
CAAGACATT	0.97	high	99.4%	99.4%	99.3%	91.0%	-9.274
AACAGTAGA	1.00	high	0.0%	94.4%	99.4%	99.4%	-9.159
AGTAGTACA	0.97	high	97.1%	96.2%	99.0%	99.1%	-8.615
AGTACGACA	0.97	high	99.1%	99.2%	99.1%	97.3%	-8.615
CAAATAGAC	1.00	high	99.3%	99.3%	99.3%	98.8%	-8.523
AAGTTAGAC	0.98	high	99.2%	99.2%	99.2%	91.7%	-8.377
AGTAGAAAT	0.97	high	2.1%	96.7%	99.2%	99.1%	-8.247
CAAGTCGAC	1.00	high	99.2%	99.2%	99.2%	0.0%	-8.218
CATAGTAGA	1.00	high	0.1%	94.2%	99.2%	99.2%	-7.721
CATTTTATT	1.00	high	99.8%	0.0%	0.0%	0.0%	-6.453

Table 5.1 Variant Sequences with Highest Probability of Direct Pairing to a +3, +4, +5 or +6 Nucleotide

Forty-nine variants ranked by predicted minimum free energy (for -4 to +38). The probability of base pairing between a given nucleotide (+3, +4, +5, or +6) and a variable region nucleotide is shown along with expression data.

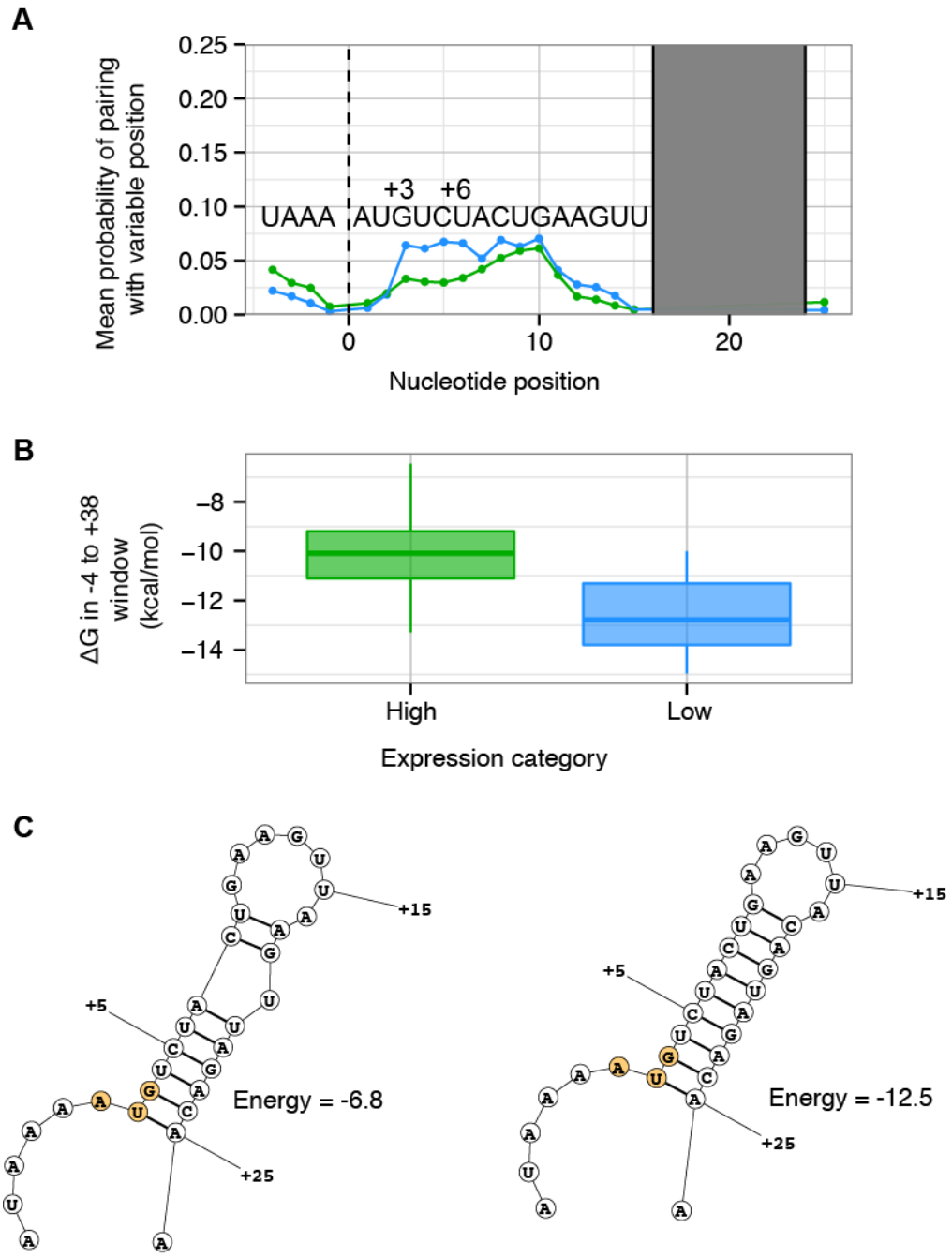

Figure 5.2 Stable Pairing between a +3 through +6 and Variable Region Nucleotide is Associated with Low Expression

(A) Mean probability by expression category (high expression in green and low in blue) of a base pair between each 5' nucleotide and a nucleotide in the variable region (gray box positions)

(B) ΔG distribution (40 nucleotide window) for 34 high (green) and 15 low (blue) variants with high base pairing probabilities between a nucleotide at positions +3 through +6 and variable region nucleotides.

(C) Examples of a high (left) and low (right) expression variant, each of which had high base pairing probabilities between a nucleotide at +3 through +6 and variable region nucleotides.

1.0 x 10^{-5}) than in the overall library (Spearman's $r^2 = 0.02$). We conclude that when there is potential for an RNA molecule to incorporate nucleotides immediately downstream of the start codon into a stem structure with neighboring bases, these variants are more likely to have substantially reduced expression; and the probability of reduced expression is governed in large part by stability of the stem structure (with an inhibitory threshold falling around -12 kcal/mol for 40 nucleotides, similar to the threshold in *E coli* (Kudla et al., 2009)).

Stabilization of a Longer-Range Hairpin Increases the Likelihood of Strong Inhibition

Since only a small proportion of the library showed a high probability of direct base pairing between a +3 through +6 nucleotide and a variable position nucleotide, we next investigated if other sequence stretches in the GFP construct were likely to pair with one of the +3 through +6 nucleotides. For each variant and at each nucleotide position over a range from the transcription start site (-74) through +101, we took the probability of base pairing with one of the +3 through +6 nucleotides. Then for each nucleotide position, we calculated the mean probability in high and low expression variants (**Figure 5.3**). We found that variable region positions had a higher probability of base pairing than most sequence regions, especially in low variants (corrected Wilcoxon p-value $\leq 2.4 \times 10^{-29}$ for the +17 position). However, by far the most probable base pairs with a +3 through +6 nucleotide were at positions +76 through +78. Both high and low expression variants had high probabilities of base pairs with these nucleotides, but the probability was highest for low expression variants (corrected Wilcoxon p-value $\leq 5.4 \times 10^{-52}$). This observation strongly suggested that base pair interactions extending beyond the first 30 nucleotides were an important factor in overall expression.

To further investigate these longer-range base pairs, we visualized the most likely base-pair interactions for low and high variants in the overall sequence. We accomplished this visualization by adopting the RNAbow tool (Aalberts and Jannen, 2013) to plot the mean probability of specific nucleotide pairs, extending from the transcription start site to +102 (**Figure 5.4**). Through this visualization we identified a hairpin, centered on +38, which tended to form a stem of base pairs that were more prevalent in low expression variants (Wilcoxon p-value = 2.64×10^{-49}). On average, the hairpin stem incorporated 6 bases on either side of the AUG start codon (including +4 through +9 nucleotides) in a longer-range interaction (including +73 through +78 nucleotides). Moreover, the hairpin had a region, from +51 through +68, where pairing with positions in and around the upstream variable region could extend the number of pairs in the overall hairpin stem (**Figure 5.4**). Since structures around the initiator AUG are likely to block translation initiation, we considered that the incorporation of positions +4 through +9 in a hairpin stem was likely to cause translation inhibition and that hybridization between variable positions and the +51 through +68 region might contribute to the structural stability of inhibitory conformations.

Based on the predicted change in change free-energy (ΔG) of hybridization between just the +51 through +68 sequence and variable region sequences, together with the probability that the +4 through +9 bases would be incorporated into an overall hairpin, we were able to identify a group of GFP variants in which 27 of 33 sequences (more than 75%) had below average expression. In the process of identifying these sequences, we first split the library into two subsets: Subset A variants had a high combined probability of forming the longer-range +4 through +9 hairpin base pairs ($P > 50\%$; $n = 8{,}763$), while Subset B variants had a lower probability ($P < 50\%$; $n = 27{,}048$). Then, within each subset we evaluated the potential degree of

base pairing at variable positions within a hairpin structure; we calculated the ΔG of hybridization between each variable region sequence (15-bases long, including the codon on each side) and the downstream region of potentially stabilizing pairs within the predicted hairpin structure (18-base sequence from +51 through +68). This hybridization ΔG provided a relative measure of pairing strength between the two regions (**Figure 5.5**). Among the variants within each subset and in bins of different hybridization strengths, we then determined the proportion of variants with below average expression (proportion of intermediate or low variants). Subset A had more sequences with strong hybridization values (lower ΔG), and at these stronger hybridization values there was a dramatic increase in the fraction of variants with below average expression (**Figure 5.6A**). We observed that within the 10% of sequences with the greatest hybridization strength from Subset A (n = 897; $\Delta G \leq -12.15$ kcal/mol; dashed line), 15% (n = 134) were low expression variants. These low variants represented 12% of all low variants in the library overall, compared to 4% (n = 203) of all intermediate variants, and 2% (n = 560) of all high variants (**Figure 5.6B** left panel). Thus, the proportion of low expression variants was enriched 5-fold compared to the proportion in the general library (Chi-square p-value = 8.96×10^{-82}). By contrast, in the Subset B, only 0.3% of the variants (n = 86) had hybridization ΔG values that fell below the same threshold ($\Delta G \leq -12.15$ kcal/mol; dashed line), and only one variant falling below this threshold had low expression (**Figure 5.6B**; right panel). Thus, we think secondary structure is the most likely cause of low expression in a group of ~134 sequences, which are likely to incorporate bases near the start codon into longer-range base pairs and an overall structure strengthened by pairing at variable positions.

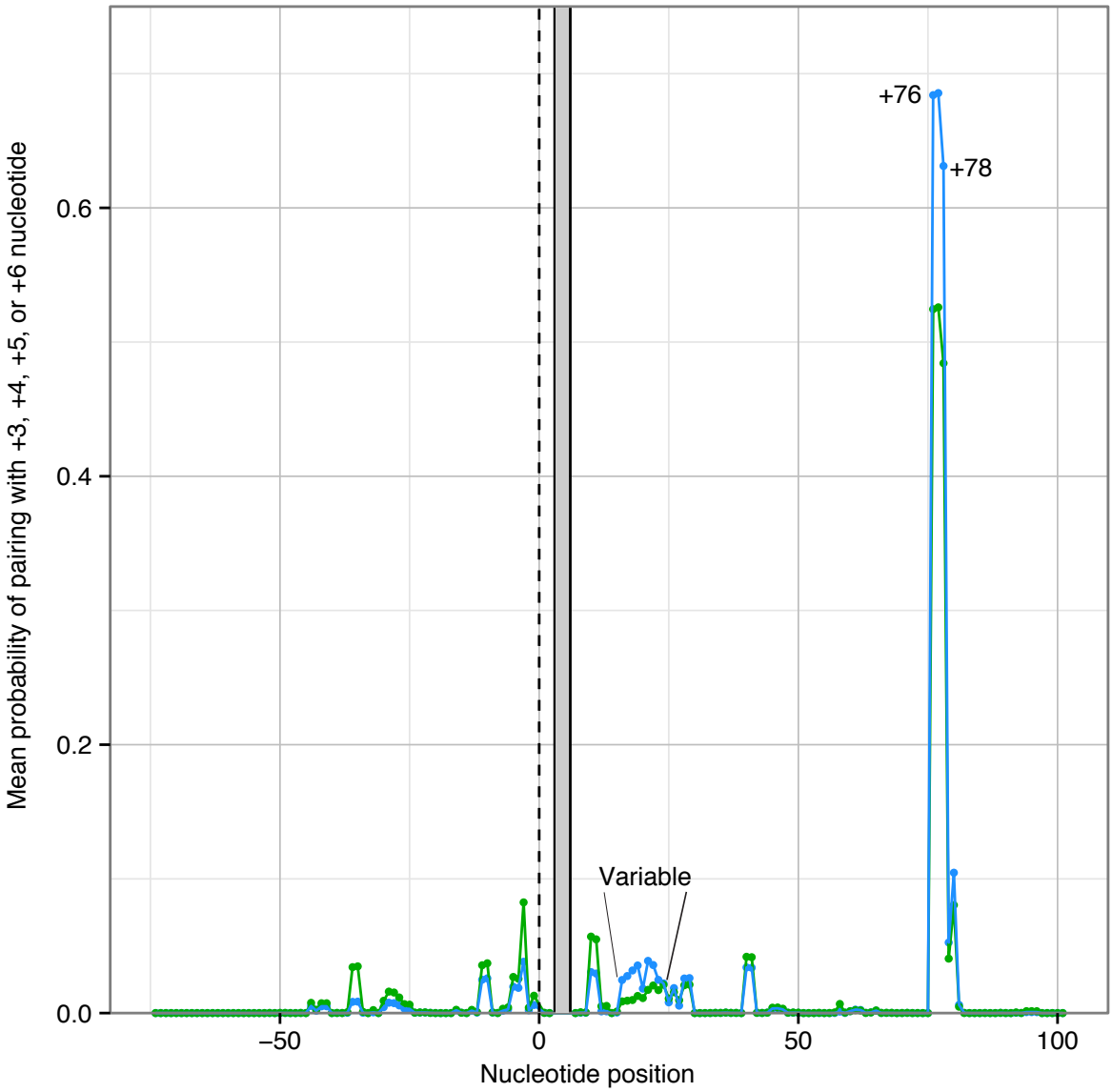

Figure 5.3 Mean Probability of Pairing with +3, +4, +5 or +6 Nucleotide
Mean probabilities for high expression (green) and low expression variants (blue) are shown for base pairing between a given position and +3, +4, +5 or +6 nucleotides (gray box).

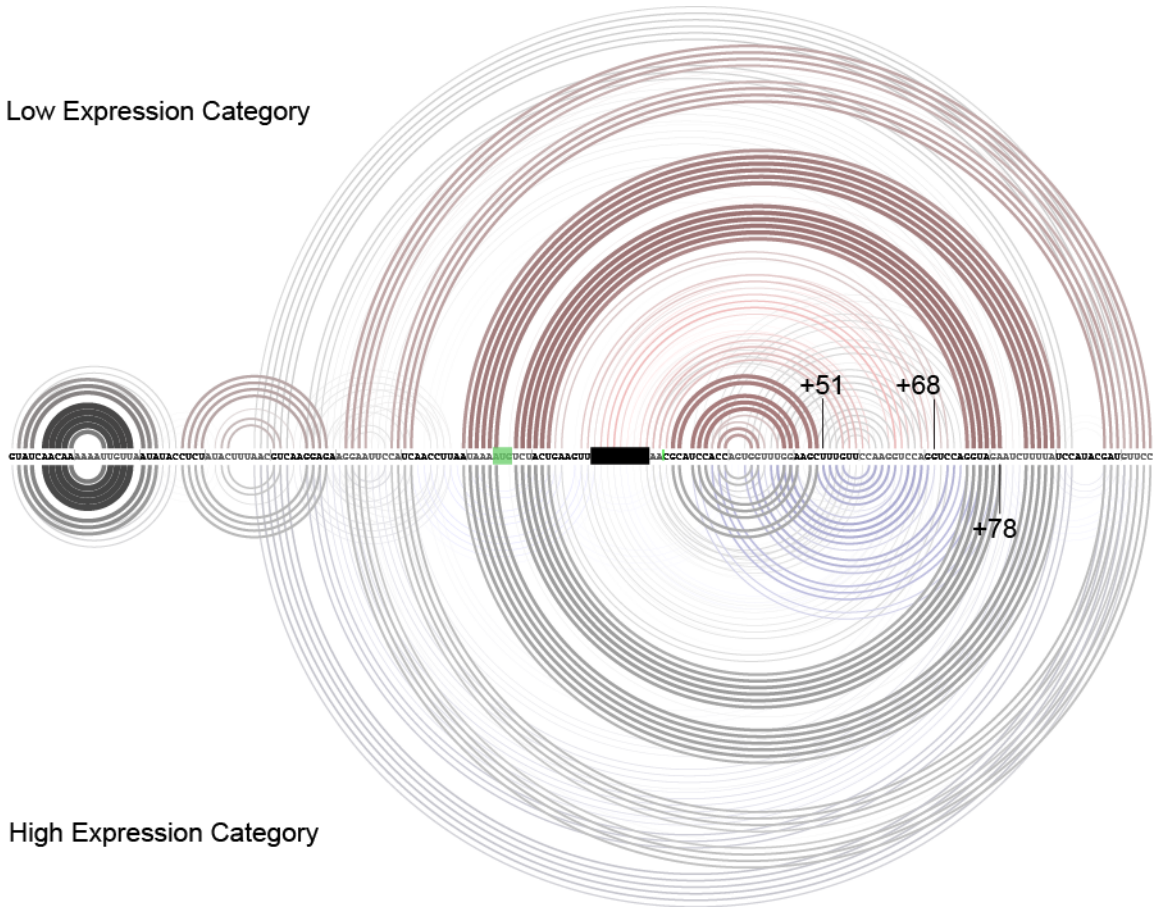

Figure 5.4 Mean Pair Probabilities by Expression Category
Mean pair probabilities by expression category. Low variants (top; n = 1,119) and high variants (bottom; n = 29,565) are plotted with custom inputs to diffRNAbow. Arc line thickness and shading are proportional to the mean pair probability, with thicker, darker lines representing higher probabilities. Arc coloration provides a measure of uniqueness to each category. It reflects strength of the relative pair probability difference in one category vs. the other. Red arcs have a greater relative mean probability in low variants, while blue arcs have a greater relative mean probability in high variants. A green box identifies the start codon and a black box identifies our library's variable base positions.

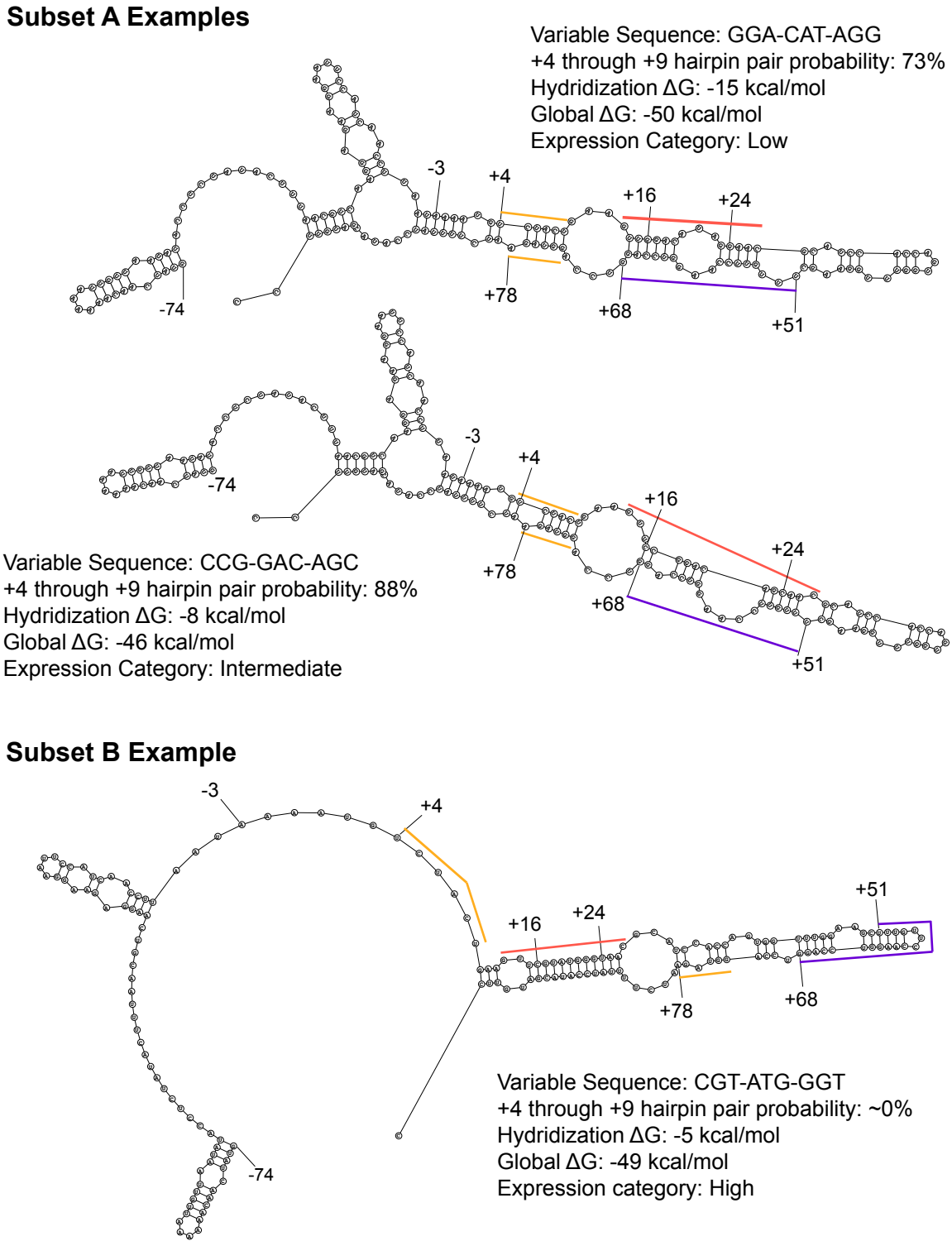

Figure 5.5 Example Minimum Free Energy Structure from Each Subset
Subset A (top examples) has a high probability of pairing between +4 through +9 and +73 through +78 (yellow), while Subset B has a low probability. The top Subset A example has a stronger hybridization potential between +51 through +68 (purple) and the nucleotides in and around the variable region (red).

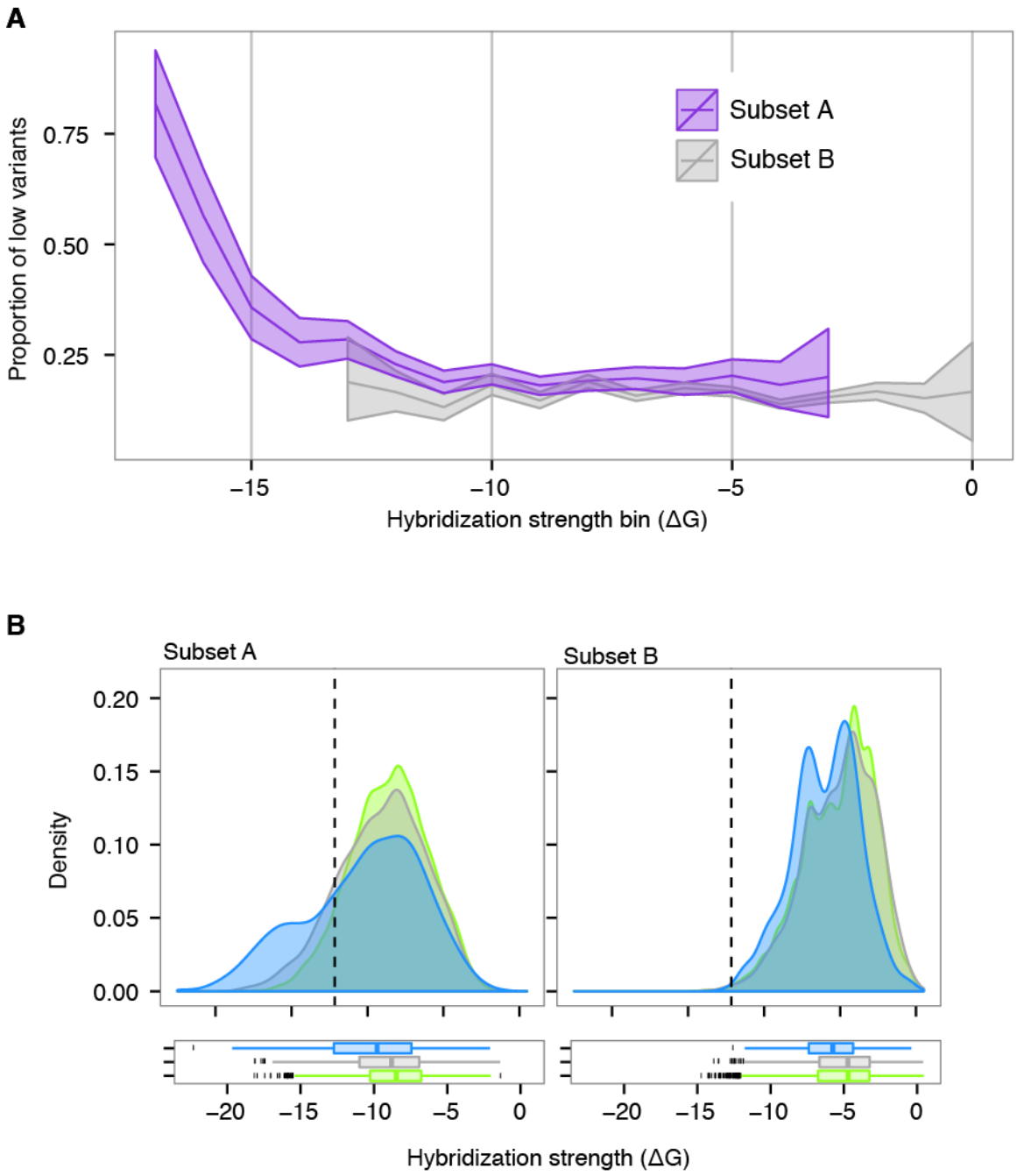

Figure 5.6 Hybridization Strength of Key Regions by Subset and Expression Category
Hybridization strength (Δ G) between the variable sequence region (from +13 through +27, including the codon on each side of variable positions) and nucleotides +51 through +68.
(A) The proportion of variants with below average expression by Δ G bin and subset. Bins are 1 kcal/mol in size and have at least 30 sequences. Subset A (n = 8,763; purple) is compared to Subset B (n = 27,048; gray).
(B) The density distribution of hybridization strengths (Δ G) by subset and expression category. Subset A (n = 8,763; left panel) is compared with Subset B (n = 27,048; right panel). High (green), intermediate (gray), and low (blue) expression variants.

5.3 DISCUSSION

We have investigated mRNA structure characteristics influencing gene expression levels in a eukaryote. Similar to previous reports in *E. coli* (Goodman et al., 2013; Kudla et al., 2009), we found that the degree of mRNA structure in a window downstream of the start site is associated with low expression in yeast. Unexpectedly, longer-range base pairs were an important factor. Within our reporter construct, mRNA structure across more than a 75-base range, which had a high probability of incorporating bases around the start of translation, may account for approximately 12% of extremely low expression variants in our library, which had reductions in expression of at least 35%. Furthermore, we observed that regions of sequence variation may stabilize inhibitory base pair stems, creating an increased likelihood of low expression.

Due in part to the high probability of longer-range interactions, we initially observed that structure predictions within windows of 40 nucleotides downstream of AUG were weakly predictive of large expression level differences in our overall library. However, taking into account the position of likely base pair partners for nucleotides immediately downstream of the start and including these nucleotide positions in the structure prediction window strengthened associations between the degree of predicted structure and expression. For a group of 75 variants, which had very high probability of base pairing between nucleotides in the first 24 bases, we found that predicted structure in the 5' window could account for about 24% of the variance in expression. This number is comparable to that found in previous studies in *E. coli* (Kudla et al. r^2 = 0.44**;** Goodman et al., r^2 = 0.34**)**. Furthermore, the approximate threshold for inhibition (ΔG = -10 to -12 kcal/mol) was very similar to the threshold as reported in *E. coli* (Kudla et al., 2009).

Whether we examined base pairs within 30 nucleotides or more than 75 nucleotides, the effects of these base pairs on expression were not consistent across all variants that were

predicted to incorporate key nucleotides downstream of the start into stem structures. Some of this variation related to the stability of stems, as more contributing base pairs lead to greater overall stability of stems and a higher probability of low expression. However, even in the highly stabilized, longer-range interaction group with many low expression variants, as many as 25% of the variants maintained above average expression. We conclude that much of this variation is likely due to inaccuracies in predicting structure and the challenge of predicting *in vivo* structure that is dynamic in nature.

RNA base pairing is likely to change in response to ribosomes translating across the mRNA molecule. Based on a model in which mRNA structure inhibits expression by hampering initiation, structural inhibition could be reduced or removed once the first ribosome has moved through initiation and entered into elongation cycles, as the base pairs key to inhibition may not have the opportunity to re-form prior to the arrival of subsequent ribosomes. Thus, ultimate expression levels may partially reflect the probability with which this initial barrier is overcome. As evidence of this possibility, we have seen intermediate (rather than low) expression variants make up a larger portion of all the variants with below average expression, when the base pair probabilities with nucleotides +4 through +6 were lower (70-90% rather than > 90%). Furthermore, such a model predicts that the degree of inhibition by mRNA structure will depend in part on the rate with which ribosomes enter the 5'-end of a transcript and proceed toward the start codon, such that shorter spacing between ribosomes is more likely to limit RNA structure effects. Future studies will need to examine the validity of this model.

Based on findings from the Kudla et al (2009) and Goodman et al. studies (2013) we had not anticipated the importance of base pairs outside the first 30-40 nucleotides in the association of 5' structure with reduced expression. Although using a different set of prediction tools, others

have reported that a window size of 100 to 150 nucleotides is more accurate than smaller windows and that more than 80% of the base pair spans within known structures in yeast fall into this range (Lange et al., 2012). In bacteria, there are reports of large thermometer structures blocking the ribosome binding site (Kortmann and Narberhaus, 2012). Much longer-range pairing between bases in the *mok* initiation site and those in the 3′ end of the same mRNA transcript block translation of the co-transcribed, toxic *hok* gene (Moller-Jensen et al., 2001). However, in general there are limited specific reports of larger mRNA structures blocking translation initiation, perhaps because these structures are both hard to predict and challenging to validate. Having many GFP variant examples from a concentrated region near the 5' end has lent confidence to our observations of likely structural effects, but these variants do not inform us of how general or rare longer-range structural inhibition may be in either synthetic or naturally occurring eukaryotic genes. Future studies will need to examine the extent to which longer range pairing may play a role in translation regulation.

Examining specific genes for pairing outside of the first 40 nucleotides will need to entail a combination of both minimum free energy and partition function base-pair probability predictions. RNA structure is an important consideration in gene design for many biotechnology applications. A greater focus on pair probabilities, the stability of structures that incorporate +1 to +10 nucleotides, and evaluation of potential interactions over a longer range (~100 to 150 nt) will help to fine-tune and improve these efforts in obtaining high protein yields for therapeutics and vaccines.

5.4 METHODS

Library RNA Structure Prediction and Pair Probability Analysis

To assess RNA structure across all 35,811 sequence variants of our library, we used both UNAFOLD and NUPACK software packages, and we ran both Spearman rank correlation and Wilcoxon tests in R. In making comparisons to the Kudla et al. (2009) dataset, we ran minimum free energy (mfe) calculations with UNAFOLD's hybrid-ss-min (Markham and Zuker, 2008) with a temperature setting of 30ºC. We also found the global free energy (ΔG) for each variant and compared local, paired nucleotide probabilities between synonymous sequences in a manner similar to Goodman et al. (Goodman et al., 2013). For these calculations we used NUPACK pairs to make ΔG and paired nucleotide predictions (Zadeh et al., 2010). We ran these calculations at 30 ºC for a 175 base window, from the transcription start site (-74) through the 101st base in the open reading frame. NUPACK computes the pair probability between each nucleotide pair combination across the length of the sequence window. Then it estimates the probability that a given position will be unpaired. Taking NUPACK's unpaired estimate for each position in a variant, we calculated the mean probability that a position would be unpaired within sliding windows of 5, 10, 20, and 40 bases. For each unpaired score we subtracted from 1 to arrive at mean paired nucleotide probability for each window.

We then identified windows where the mean paired nucleotide probabilities differed between synonymous sequences. To identify these windows, for each window position and each variant, we took the ratio of probabilities between the window in the variant sequence and the same window position in a synonymous reference sequence. This ratio yielded the relative paired nucleotide probability for each window, similar to Goodman et al. (Goodman et al., 2013). To evaluate whether structure may contribute to synonymous expression differences, for each

window position we found the Spearman rank correlation between relative paired nucleotide probabilities and syn-GFP^{SEQ} (Figure 5.1). Here we have reported the relative values for the 10- and 20-nucleotide windows. Significantly correlated window positions were consistent with and without the inclusion of sequences that had high degrees of global structure.

To predict the lowest free energy structure for examples with individual variants, we used the RNAstructure Fold server (Reuter and Mathews, 2010) with default settings.

Visualization of Mean Pair Probabilities by Position and Expression Category

To visualize the strongest mean pair probabilities across positions, we provided diffRNAbow (Aalberts and Jannen, 2013) with custom pair file inputs that included all possible pair relationships with a ≥ 5% mean probability.

RNA Hybridization Free Energy Between Two Sequence Stretches

To estimate the relative strength of potential interactions between the variable positions (plus a codon on each side) and an 18-base, downstream region from +51 to +69 (5'-UUUGUUCCAAGGTCCAGG-3'), we computed the ΔG of hybridization, as though these sequence regions were separate strands. We performed this analysis using UNAFOLD melt.pl at a temperature setting of 30ºC and strand concentration (Ct) of 0.00001 M. Then within hybridization ΔG bins of 1 kcal/mol (and bins with at least 30 sequences), we found the proportion of sequences made up of either low or intermediate variants.

5.5 NOTES AND CONTRIBUTIONS

Figure 5.1B is contained in the supplementary material of the publication, "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast" by Caitlin E. Gamble, Christina E. Brule (University of Rochester), Kimberly M. Dean (University of Rochester), Stanley Fields, and Elizabeth J. Grayhack (University of Rochester); in press at *Cell* (2016).

The work in this chapter is my own. I thank Stan Fields, Elizabeth Grayhack, and Christina Brule for their discussion and commentary.

CHAPTER 6: Summary and Future Directions

Many diverse factors contribute to the quantity and quality of protein produced from a coding sequence. In the previous chapters, my collaborators and I have established that codon pairs impact translation elongation and efficiency in yeast, most likely through cross talk between codons and tRNA in two ribosomal site positions. In addition we have shown how synonymous variation within the coding sequence might impair translation initiation by the increased tendency of an RNA molecule to form base pairs with nucleotides immediately downstream of the translation start site. Future work will be needed to further develop our mechanistic understanding of these inhibitory effects. Characterizing the full spectrum of possible codon pair effects on translation efficiency might include identifying other codon pairs with more subtle inhibitory effects, elucidating how the translational properties of individual codons pairs and their tRNAs exert their effects within the translation machinery, and assessing whether environmental conditions and cell type characteristics influence inhibitory properties (such as through tRNA modifications to wobble decoding capabilities). The ribosome is highly conserved in both sequence and function; thus, mechanistic insights from one species will likely apply to others. Identification of genes in which inhibitory mRNA structure conformations or codon pairs may play functional roles in protein synthesis would facilitate characterization of inhibitory properties. These genes could also provide key insights for applying structure and codon pair knowledge to the analysis of disease-associated genetic variation and synthetic gene design.

General Parameters under which RNA Structure is Disruptive to Initiation

Gene sequences are often re-designed or placed in alternative contexts for many biotechnology applications. Identifying the positions and energy parameters under which RNA structure may impair expression of a designed construct would provide useful guidelines for construct design. Knowledge of these parameters could provide insight into regulatory functions of mRNA structure in naturally occurring coding sequence. From our analysis, we have found that when nucleotides from the +3 through +6 positions of an mRNA transcript are likely to base pair, the consequence is that reduced protein expression from the transcript is more likely (**Chapter 5**). Furthermore, our analysis suggests that the probability of below average expression substantially increases once the base pairing results in a change in free energy of about -10 to -12 kcal per mol. Future studies will be needed to clarify how generally applicable these position and energy parameters may be with respect to other gene contexts and species.

Mechanisms of Codon Pair Inhibition: Wobble within the Ribosome

U:G and I:A wobbles are a central component of inhibition by a number of the inhibitory codon pairs identified in our assay. Based on differences between expression of tRNAs that decode the 5' and 3' codons of inhibitory pairs, our data further suggest that for some codon pairs, wobble base pairing may contribute to inhibition even after the wobble-decoded codon and associated tRNA have entered the ribosomal P-site (**Chapters 3 and 4**). tRNA gene copy numbers provide evidence that the use of some wobble-decoding mechanisms as the sole means of decoding particular codons is species-specific. Yet some decoding inequities, such as a reliance on inefficient I:A wobble decoding of the CGA codon, are found in both yeast and many species of bacteria (Chan and Lowe, 2015).

Since wobble base pairing does not conform to the typical geometry of Watson:Crick base pairing, tRNA modifications are often needed to facilitate wobble decoding (Agris et al., 2007). Anticodon modifications to U34 are known for having particularly rich chemistries (Agris et al., 2007). In bacteria, cmo^5U modifications to U34 expand the decoding repertoire of the anticodon U to include U- or C-ending codons. Surprisingly, it was found in Salmonella that the cmo^5U modification of U34 is also necessary for U:G wobble decoding of Pro, Ala, and Val codons (Nasvall et al., 2007). Three codons (CUG, CCG, and CGC) are present in many of the identified inhibitory codon pairs that rely on a U:G wobble mechanism in yeast. Differential expression and activity of tRNA modification enzymes may influence how codon pairs behave in certain environments and cell types. For example, conditions of stress increase activity of the anticodon modification enzyme, Trm9, and lead to differential translation of yeast mRNAs enriched in AGA codons (Begley et al., 2007).

Several other codons (GUG, AGG, and CUC) that occur in the inhibitory codon pairs are decoded by competing tRNA anticodon species. When more than one tRNA species can decode a given codon, the outcome of this tRNA competition will determine whether a wobble decoding mechanism is utilized. Since, wobble decoding contributes to reduced translation efficiency of codon pairs, the outcomes of tRNA competitions will likely have implications for elongation efficiency and could affect the function or availability of the associated protein products.

Cell Type and Environmental Influences on tRNA Modification and Competition

For some genes, the selective pressures acting on codon usage may not be readily apparent under standard laboratory growth conditions. Evolution takes place in a dynamic environment with frequent changes in temperature, nutrient availability, and other stresses. Additionally, in

multicellular organisms, cells undergo considerable differentiation and are defined to a large extent by the composition of their proteome. Cells undergo dramatic changes in gene expression depending on stress, state of growth, and cellular identity. These global changes alter metabolic activity and codon usage demands. In yeast, tRNA concentrations vary during the cell cycle, and sets of genes expressed during stages of the cell cycle have similar preferences with regard to non-optimal and wobble decoding use (Frenkel-Morgenstern et al., 2012). In mammalian cells codon use as well as tRNA isoacceptor expression differ substantially between proliferating cells and differentiated tissues (Gingold et al., 2014). There are many potential points of tRNA regulation, including: expression, localization, modification enzyme activity, and amino acid charging. Future studies will be needed to investigate how changing environmental and cellular conditions impact tRNA pools and the translational efficiency of codon pairs.

Alternative Mechanisms of Pair Inhibition

In addition to the importance of wobble-decoding mechanisms in translation inhibition, our data suggest that crosstalk between ribosomal sites influences translation efficiency, especially when there are codons decoded by wobble. Suppression of inhibition by overexpression of native tRNA for the 3' codon of 9 tested codon pairs suggests inhibition occurs when codons of the pair are positioned in the P and A-sites of the ribosome (**Chapter 5**). Future studies will be needed to determine the precise mechanisms of crosstalk and inhibition arising from properties of the codons and tRNAs.

Pair-mediation inhibition may not, however, be limited to P and A site positions. For example, the arginine-arginine codon pair, AGG-CGA, is most likely to inhibit in an alternative manner. We found AGG-CGA had higher ribosome occupancy, but only when CGA was in a P- or E-site position. Consistent with inhibition occurring away from a ribosomal A-site, inhibition

by AGG-CGA was not suppressed by either native or exact-match tRNA for the 3' CGA codon. A similar arginine-arginine pair, AGG-CGG, was also identified as an inhibitory codon pair in the GFP assay, but we found no ribosome profiling evidence that it slows down the ribosome, and it has not been tested for suppression by tRNA. The 3 codons in these two pairs are rarely used in highly expressed genes. Based on the G rich nucleotide content of the pairs, it's tempting to speculate that inhibition by these pairs may result from mRNA structure conformations. However, within our library low expression variants with these pairs had both local and global structure predictions that fell within a standard deviation of most high expression variants. Others have also reported inhibition of translation by AGG and CGG codons near the start codon in *E. coli* (Gonzalez de Valdivia, 2004). Gonzalez de Valdivia and Isaksson (2004) concluded that inhibition by AGG and CGG codons could not simply be attributed to augmented secondary structure, since shifting the reading frame removed inhibition (Gonzalez de Valdivia, 2004; Kozak, 2005). Thus, separate constructs in *E. coli* and in *S. cerevisiae* provide data that the presence of G-rich codons around the start site is deleterious, but the reasons from this effect are not known.

Beyond direct mechanisms by which codon pairs inhibit the translation machinery, suboptimal translation efficiency of codon pairs may also lead to secondary effects that contribute to a reduction in protein levels. For instance, studies have suggested that differences in decoding rates along transcripts may affect mRNA decay (Presnyak et al., 2015), feedback on translation initiation rates (Chu et al., 2013; Hersch et al., 2014) or cause recruitment of quality control systems (Letzring et al., 2013). With regard to secondary effects such as these, the context of codon pairs within the transcript as a whole is likely to be an important factor.

Prediction and Design: Toward a Richer Understanding of Translation Efficiency

Many inhibitory pairs are inhibitory only with the two codons in a specific order. Together with weak suppression by tRNA for the 5' codon, our data support the conclusion that tRNA availability is not the primary determinant of large expression differences mediated by codon pairs (**Chapter 4**). On the other hand, increasing the availability of 3' or exact base-pairing tRNA species helps to suppress pair inhibition, and tRNA availability could influence pair inhibition based on the competition between wobble decoding and non-wobble decoding tRNA. Overall, the data call into question a model whereby synonymous codon usage influences translation efficiency primarily through the small, additive effects of individual codons (Pechmann and Frydman, 2012; Reis, 2004; Sharp and Li, 1987; Tuller et al., 2010a), since in our data individual codon effects would be subtle relative to codon pair inhibition.

For heterologous protein expression, "codon optimization" strategies based on individual codon frequencies in highly expressed genes (Sharp and Li, 1987) or their approximate tRNA abundances (Pechmann and Frydman, 2012; Reis, 2004; Tuller et al., 2010a) have yielded mixed results (Gustafsson et al., 2004). Shifting away from optimization of each codon individually and toward an expanded focus on identifying translation bottlenecks (Navon and Pilpel, 2011; Shah et al., 2013), accounting for key context factors (e.g. codon pairing), and assessing tRNA abundance with respect to the decoding properties of competing tRNA, will ultimately improve yield predictions and reduce potentially detrimental effects arising from unintended alterations to RNA and protein structure conformations (discussed in (Mauro and Chappell, 2014)).

Bottlenecks are a key determinant of efficiency as with any synthesis process. For translation, bottlenecks could take a variety of forms, including: infrequent translation initiation (e.g. due to limited ribosome availability), obstruction of translation initiation (e.g. through

protein-binding or RNA conformational masking) to stalling of the ribosome at difficult to translate sequences (e.g. a strong inhibitory codon pair or an ulta-stable mRNA hairpin). We have discussed two likely forms of bottlenecks in our GFP construct library: inhibitory codon pairs and mRNA structure that incorporates bases in and immediately downstream of the start codon. Stable mRNA structure greatly increases the probability of dramatic reductions in expression, most likely by blocking translation initiation. Ribosome footprints typically cover ~30 nucleotides. Given the positioning of the variable region in our GFP library construct at nucleotide +16, the magnitude of observed codon pair bottlenecks may result not only from a pause in elongation but also from the obstruction by the paused ribosome of initiation by other ribosomes.

In addition to limiting the overall rate of production, a bottleneck has important consequences for the selective pressures acting upon the rest of the coding sequence. If sub-optimal regions of a gene do not exceed the rate limitation established by a bottleneck, then there may be little selective pressure to improve the efficiency of these other sub-optimal steps. If there is selective pressure to improve their efficiency, it is likely a diffuse pressure, selecting for efficient allocation of limited translation machinery resources throughout the cell. On the other hand, acute selective pressures to facilitate proper protein folding (Zhang and Ignatova, 2009) may favor "sub-optimal," slowly translated sequence. Thus, identifying the likely bottlenecks in a coding sequence and the relative magnitudes of bottlenecks will be important steps for interpreting functional consequences of coding sequence variation throughout a gene, and an ability to do so will ultimately improve both translation efficiency predictions and *de novo* gene design.

Applications in Biotechnology and Medical Genetics

Gene design is a critical step related to the cost and effectiveness of protein engineering products. Products from the field of protein engineering are projected to reach a market size $168 billion in 2017 (Liszewski, 2015). These products include protein-based reagents, diagnostics, therapeutics, and vaccines for which the accuracy and yield of a protein are critical to ultimate success of the product. Here, awareness and management of mRNA structure, codon context, and translational bottlenecks could dramatically improve gene design results. More directly, as reductions in translation speed have been shown to facilitate proper protein folding (Zhang and Ignatova, 2009; Zhou et al., 2013) as well as to facilitate modification of key residues (Zhang et al., 2010), it's possible the inhibitory codon pairs we've identified in yeast could be applied to solving folding and modification problems in a protein engineering context. For instance, desirable end products could be achieved through the placement of a moderately inhibitory pair between protein folding domains or at an appropriate distance downstream of residues needing modification, thereby allowing more time for folding or modification to occur before additional residues are introduced.

Identification of therapeutic protein targets often comes from genetic data and genetic variation influences disease predisposition. Some polymorphisms near the initiator AUG codon have been associated with disease pathologies in humans and mice (Kozak, 2002). Many of these mutations are believed to negatively impact the Kozak sequence at positions -3 and +4, but mRNA structure effects need to be ruled out (Kozak, 2005). Also, in many disease association studies, synonymous polymorphisms have been overlooked as potential contributors to disease risk in large part due to lack of means to identify when synonymous variation is likely impact protein quantity or quality. In 2014, Supek et al. implicated synonymous mutations in oncogenes

as drivers of cancer (Supek et al., 2014). Many of these synonymous changes are likely to impact alternative splicing mechanisms. However, for roughly half of synonymous drivers, the authors could provide no mechanistic explanation. While it remains to be seen to what extent inhibitory codon pairs may influence translation dynamics throughout the yeast genome and in other species, improving our overall mechanistic understanding of how codon context influences translation efficiency will aid in interpreting the likely consequences of genetic variation and assessing disease risk.

In summary, the dramatic effects of inhibitory codon pairs that we have demonstrated highlight the importance of both tRNA decoding properties and sequence context considerations when evaluating synonymous variation. These considerations have future application in interpreting phenotypes that arise from natural genetic variation and in designing genes for heterologous protein expression. Future investigation into the broader role of codon pairs, wobble decoding, and tRNA regulation will contribute to a richer understanding of translation, one of the most central processes in the life of a cell.

Appendix A: GFP^{FLOW} of Individual Constructs

This is the work of Christina E. Brule, who made the constructs and obtained flow cytometry measures. GFP^{FLOW} (GFP/RFP) is the mean of 3 or 4 independent isolates ±SD.

Inhibitory Codon Pair	tRNA vector	Inhibitory pair variant	GFP^{FLOW}	Optimal pair variant	GFP^{FLOW}
AGG-CGA	none	AGGCGAAAT	0.58 ±0.02	AGAAGAAAT	1.38 ±0.04
	ev LEU2 2u	AGGCGAATG	0.46 ±0.01	AGAAGAATG	0.83 ±0.07
	tR(CCU)		0.42 ±0.05		0.81 ±0.01
	tR(UCU)		0.50 ±0.02		0.86 ±0.02
	tR(ICG)		0.45 ±0.05		0.90 ±0.02
	tR(UCG)		0.47 ±0.03		0.71 ±0.04
AGG-CGG	none	AGGCGGCAC	0.58 ±0.03	AGAAGACAC	1.12 ±0.04
AUA-CGA (Candidate)	none	ATACGAGAT	0.56 ±0.01	ATTAGAGAT	1.43 ±0.03
AUA-CGG (Candidate)	none	ATACGGACG	0.57 ±0.00	ATTAGAACG	0.89 ±0.03
CGA-AUA	none	CGAATACAT	0.38 ±0.02	AGAATTCAT	1.13 ±0.01
	ev LEU2 2u		0.29 ±0.01		0.94 ±0.02
	tR(ICG)		0.30 ±0.01		0.92 ±0.01
	tR(UCG)*		0.44 ±0.02		0.79 ±0.02
	tI(UAU)		0.52 ±0.01		0.87 ±0.04
CGA-CCG	none	CGACCGATG	0.16 ±0.00	AGACCAATG	0.86 ±0.01
	ev LEU2 2u	CGACCGAAG	0.14 ±0.05	AGACCAAAG	0.76 ±0.05
	tR(ICG)		0.13 ±0.00		0.76 ±0.09
	tR(UCG)*		0.51 ±0.09		0.69 ±0.07
	tP(UGG)		0.26 ±0.00		0.68 ±0.07
	tP(CGG)*		0.88 ±0.02		0.63 ±0.04
CGA-CGA	none	CGACGAACT	0.25 ±0.01	AGAAGAACT	1.27 ±0.04
	ev LEU2 2u		0.18 ±0.02		0.93 ±0.07
	tR(ICG)		0.37 ±0.02		0.97 ±0.03
	tR(UCG)*		0.78 ±0.07		0.87 ±0.03
	tR(CCU)		0.18 ±0.05		0.91 ±0.03
	tR(UCU)		0.18 ±0.06		1.00 ±0.03
	tR(CCG)		0.16 ±0.02		0.99 ±0.01
CGA-CGG	none	CGACGGAGC	0.42 ±0.01	AGAAGAAGC	1.22 ±0.02
	ev LEU2 2u		0.28 ±0.04		1.09 ±0.00
	tR(ICG)		0.25 ±0.02		1.09 ±0.00
	tR(UCG)*		0.58 ±0.05		0.95 ±0.00
CGA-CUG	none	AACCGACTG	0.72 ±0.01	AACAGATTG	1.55 ±0.04
	ev LEU2 2u		0.52 ±0.01		1.45 ±0.05
	tR(UCG)		0.94 ±0.07		1.22 ±0.06

	tL(UAG)		1.08 ±0.02		1.35 ±0.13
	tL(CAG)*		1.08 ±0.03		1.22 ±0.01
CGA-GCG	none	CGAGCGAGT	0.35 ±0.01	AGAGCTAGT	1.36 ±0.03
	ev LEU2 2u	CGAGCGAAT	0.33 ±0.02	AGAGCTAAT	1.19 ±0.05
	tR(ICG)		0.33 ±0.04		1.17 ±0.03
	tR(UCG)*		0.65 ±0.03		110 ±0.05
	tA(UGC)		0.67 ±0.09		1.19 ±0.05
CUC-CCG	none	CTCCCGACT	0.18 ±0.01	TTGCCAACT	1.27 ±0.01
	ev LEU2 2u		0.14 ±0.00		1.37 ±0.03
	tL(UAG)		0.10 ±0.01		1.40 ±0.01
	tL(GAG)		0.34 ±0.03		1.33 ±0.05
	tP(UGG)		0.24 ±0.07		1.34 ±0.10
	tP(CGG)*		0.91 ±0.14		1.37 ±0.10
CUG-AUA (Candidate)	none	CTGATAATG	0.58 ±0.03	TTGATTATG	0.94 ±0.08
CUG-CCG	none	CTGCCGACC	0.50 ±0.00	TTGCCAACC	1.28 ±0.02
	ev LEU2 2u		0.35 ±0.04		0.97 ±0.08
	tL(UAG)		0.41 ±0.03		1.07 ±0.08
	tL(CAG)*		0.54 ±0.05		0.95 ±0.13
	tP(UGG)		0.50 ±0.04		1.02 ±0.03
	tP(CGG)*		0.96 ±0.07		0.94 ±0.04
CUG-CGA	none	CTGCGAAGT	0.46 ±0.01	TTGAGAAGT	1.24 ±0.01
	ev LEU2 2u		0.32 ±0.02		0.96 ±0.08
	tL(UAG)		0.36 ±0.02		0.96 ±0.03
	tL(CAG)		0.39 ±0.05		0.92 ±0.08
	tR(ICG)		0.42 ±0.01		0.94 ±0.07
	tR(UCG)		0.61 ±0.03		0.80 ±0.05
CUG-CUG (Candidate)	none	CTGCTGACA	0.62 ±0.03	TTGTTGACA	0.82 ±0.08
CUU-CUG (Candidate)	none	CTTCTGACG	0.65 ±0.04	TTGTTGACG	0.99 ±0.06
GUA-CCG (Candidate)	none	GTACCGAGT	0.60 ±0.02	GTTAGAAGT	1.42 ±0.08
GUA-CGA	none	GTACGACAA	0.36 ±0.03	GTTAGACAA	0.91 ±0.02
	ev LEU2 2u		0.26 ±0.07		0.65 ±0.02
	tR(ICG)		0.38 ±0.02		0.64 ±0.03
	tR(UCG)*		0.48 ±0.01		0.56 ±0.01
GUG-CGA	none	GTGCGAACT	0.50 ±0.01	GTTAGAACT	1.16 ±0.02
	ev LEU2 2u		0.37 ±0.03		0.88 ±0.00
	tV(CAC)		0.38 ±0.01		0.83 ±0.03
	tR(ICG)		0.52 ±0.10		0.88 ±0.04
	tR(UCG)		0.57 ±0.10		0.77 ±0.10
	tI(UAU)		0.34 ±0.00		0.83 ±0.02

Appendix B: Supplemental Material & Methods

Methods in this section primarily reflect work by Christina E. Brule. They are part of a manuscript submitted for publication as, "Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast" by Caitlin E. Gamble, Christina E. Brule, Kimberly M. Dean, Stanley Fields, and Elizabeth J. Grayhack.

Strains, Plasmids, and Oligonucleotides

The yeast strain BY4741 (MATa *his3Δ1, leu2Δ0, met15Δ0, ura3Δ0*) was the parent strain for all linear transformations, except for constructs with the *Renilla* luciferase-codon insert-GFP fusion reporter where the parent strain was AW765 (BY4741, *nmd2Δ::kanMX*) (Wolf and Grayhack, 2015). The plasmid vector pEAW315, in which P_{GAL1} controls transcription of a *GLN4* NTD-GFP fusion, is a variation on plasmids previously described (Wolf and Grayhack, 2015). The $GLN4_{(1\text{-}99)}$ fragment was obtained by PCR amplification of pJE1012a (Grant et al., 2012), and then it was inserted into the Pac1 site of pEKD1024 (Dean and Grayhack, 2012) using ligation independent cloning (LIC), which regenerated the Pac1 site as well as a LIC site. Thus, all insertions of codon sequences into the *GLN4*-GFP fusion protein were performed via LIC of two annealed oligos into pEAW315.

Yeast GFP Library Construction

DNA was obtained from *E. coli* libraries (1 ml frozen aliquots) grown at 37 °C to saturation in 100 ml LB+amp, using a QiaFilter kit (Qiagen). For the $(NNN)_3$ and $(VNN)_3$ libraries, 1900 ng and 1950 ng of StuI cut and gel-purified linear DNA were transformed into 10 ml BY4741 yeast cells, which were plated four times on selective media (SD-met) to obtain 73,068 [$(NNN)_3$ Library 1], 50,532 [$(NNN)_3$ Library 2] and 77,200 [$(VNN)_3$ Library] yeast transformants, as described (Guy et al., 2014). For each transformation, transformants from four plates were pooled by scraping the plates into YP + 2% raffinose + 8% DMSO and saved at -80 °C. Aliquots were grown for 3.5 generations in selective media (S-met + 2% raffinose) at 30 °C. Then cells were diluted into 25 ml YP + 80 mg/L Ade + 2% raffinose + 2% galactose media at a starting OD_{600} of 0.08 and grown for 3.5 generations at 30 °C, diluted into another 50 ml at a starting OD_{600} of 0.05, grown overnight for 5 generations, and diluted into 10 ml at a starting OD_{600} of 0.3, followed by another 4h of growth at 30 °C.

Analysis of Individual Variants by Flow Cytometry

Variant sequences in GFP (at amino acids 6 to 8), in $GLN4_{(1\text{-}99)}$-GFP (beginning at amino acid 100), and in *Renilla* luciferase-GFP (beginning at amino acid 318) were inserted using LIC cloning as previously described (Dean and Grayhack, 2012); (Wolf and Grayhack, 2015). Synonymous optimal codons were chosen based on CAI. If two synonymous codons had a similar CAI, the more A-U rich codon was chosen. Strains to be analyzed by flow cytometry were grown overnight in YP + 80 mg/L Ade + 2% raffinose + 2% galactose media at 30 °C, followed by dilution to OD_{600} between 0.1-0.2 and grown for 4-6 hours in the same media to OD_{600} of ~0.8. Analytical flow cytometry and downstream analysis were performed for 4 independent isolates of each strain (sometimes 3) as previously described (Dean and Grayhack, 2012). To examine the effects of expressing various tRNA genes from 2μ vectors, strains were grown using the same protocol in S minimal media lacking leucine + 2% raffinose + 2% galactose media (Sherman, 1986).

Strain Growth for *leu2*-d Selection in tRNA Suppression Studies
Strains transformed with the 2μ *leu2-d* vectors, pECB1118 and pECB1406, were grown for ~18 hours at 30 °C in 5 ml S-ura + 2% raffinose + 2% galactose + 80 mg/L Ade media, followed by dilution to OD_{600} of 0.01 in 5 ml S-ura-leu + 2% raffinose + 2% galactose + 80 mg/L Ade media and grown overnight at 30 °C. Approximately 4 hours before flow cytometry analysis, the strains were diluted to OD_{600} of 0.25 in 5 ml S-ura-leu + 2% raffinose + 2% galactose + 80 mg/L Ade media (Whipple et al., 2011).

Acidic Northern Blot
Bulk RNA, prepared from ~3 OD pellets, was resolved on 6.5% acrylamide gels at pH 5 as described (Alexandrov et al., 2006).

References

Aalberts, D.P., and Jannen, W.K. (2013). Visualizing RNA base-pairing probabilities with RNAbow diagrams. Rna *19*, 475–478.

Agris, P.F. (2004). Decoding the genome: a modified view. Nucleic Acids Res *32*, 223–238.

Agris, P.F., Vendeix, F.A.P., and Graham, W.D. (2007). tRNA's wobble decoding of the genome: 40 years of modification. J Mol Biol *366*, 1–13.

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

Ames, B., and Hartmann, P. (1963). The histidine operon. Cold Spring Harbor Symposium Quantitative Biology *28*, 349–356.

Artieri, C.G., and Fraser, H.B. (2014). Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Research *24*, 2011–2021.

Atkins, J.F., and Bjork, G.R. (2009). A Gripping Tale of Ribosomal Frameshifting: Extragenic Suppressors of Frameshift Mutations Spotlight P-Site Realignment. Microbiology and Molecular Biology Reviews *73*, 178–210.

Beggs, J.D. (1978). Transformation of yeast by a replicating hybrid plasmid. Nature *275*, 104–109.

Begley, U., Dyavaiah, M., Patil, A., Rooney, J.P., DiRenzo, D., Young, C.M., Conklin, D.S., Zitomer, R.S., and Begley, T.J. (2007). Trm9-Catalyzed tRNA Modifications Link Translation to the DNA Damage Response. Mol Cell *28*, 860–870.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. J. Biol. Chem. *257*, 3026–3031.

Bonetti, B., Fu, L., Moon, J., and Bedwell, D.M. (1995). The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. J Mol Biol *251*, 334–345.

Bossi, L. (1983). Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J Mol Biol *164*, 73–87.

Boycheva, S., Chkodrov, G., and Ivanov, I. (2003). Codon pairs in the genome of Escherichia coli. Bioinformatics *19*, 987–998.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2011). High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. Science *335*, 552–557.

Brockmann, R., Beyer, A., Heinisch, J.J., and Wilhelm, T. (2007). Posttranscriptional Expression Regulation: What Determines Translation Rates? PLoS Comput Biol *3*, e57.

Buchan, J.R. (2006). tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res *34*, 1015–1027.

Cannarozzi, G., Cannarrozzi, G., Schraudolph, N.N., Faty, M., Rohr, von, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. Cell *141*, 355–367.

Chan, P.P., and Lowe, T.M. (2015). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res gkv1309.

Charneski, C.A., and Hurst, L.D. (2013). Positively charged residues are the major determinants of ribosomal velocity. Plos Biol *11*, e1001508.

Chevance, F.F.V., Le Guyon, S., and Hughes, K.T. (2014). The Effects of Codon Context on In Vivo Translation Speed. PLoS Genet *10*, e1004392.

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M.F., and Haar, von der, T. (2013). Translation elongation can control translation initiation on eukaryotic mRNAs. Embo J. *33*, 21–34.

Cigan, A.M., Feng, L., and Donahue, T.F. (1988). tRNAi(met) functions in directing the scanning ribosome to the start site of translation. Science *242*, 93–97.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. Science *320*, 1784–1787.

Crick, F.H.C. (1966). Codon—anticodon pairing: The wobble hypothesis. J Mol Biol *19*, 548–555.

Curran, J.F., and Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol *209*, 65–77.

Dana, A., and Tuller, T. (2014a). Mean of the Typical Decoding Rates: A New Translation Efficiency Index Based on the Analysis of Ribosome Profiling Data. G3&Amp;#58; Genes|Genomes|Genetics.

Dana, A., and Tuller, T. (2014b). The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res *42*, 9171–9181.

de Smit, M.H., and van Duin, J. (1994). Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. J Mol Biol *235*, 173–184.

Dean, K.M., and Grayhack, E.J. (2012). RNA-ID, a highly sensitive and robust method to identify cis-regulatory sequences using superfolder GFP and a fluorescence-based assay. Rna *18*, 2335–2344.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., and Yusupova, G. (2012). A new understanding of the decoding principle on the ribosome. Nature *484*, 256–259.

Demeshkina, N., Jenner, L., Yusupova, G., and Yusupov, M. (2010). Interactions of the ribosome with mRNA and tRNA. Current Opinion in Structural Biology *20*, 325–332.

Dinman, J.D. (2012). Mechanisms and implications of programmed translational frameshifting. WIREs RNA *3*, 661–673.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

Elf, J. (2003). Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage. Science *300*, 1718–1722.

Fedorov, A., Saxonov, S., and Gilbert, W. (2002). Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res *30*, 1192–1197.

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y.-M., and Jensen, L.J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Molecular Systems Biology *8*, 1–10.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. eLife *3*.

Ge, P., and Zhang, S. (2015). Computational analysis of RNA structures with chemical probing data. Methods *79-80*, 60–66.

Gerashchenko, M.V., and Gladyshev, V.N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res *42*, e134–e134.

Gerashchenko, M.V., Lobanov, A.V., and Gladyshev, V.N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proceedings of the National Academy of Sciences *109*, 17394–17399.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gingold, H., and Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. Molecular Systems Biology *7*, 1–13.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sørensen, K.D., et al. (2014). A Dual Programfor Translation Regulationin Cellular Proliferation and Differentiation. Cell *158*, 1281–1292.

Gonzalez de Valdivia, E.I. (2004). A codon window in mRNA downstream of the initiation

codon where NGG codons give strongly reduced gene expression in Escherichia coli. Nucleic Acids Res *32*, 5198–5205.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science *342*, 475–479.

Graille, M., and Séraphin, B. (2012). PERSPECTIVES. Nature Publishing Group *13*, 727–735.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res *9*, r43–r74.

Grosjean, H., de Crécy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: Coevolution with specific tRNA modification enzymes. FEBS Lett *584*, 252–264.

Gu, W., Zhou, T., and Wilke, C.O. (2010). A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLoS Comput Biol *6*, e1000664.

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. Trends Biotechnol. *22*, 346–353.

Gutman, G.A., and Hatfield, G.W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. Proc Natl Acad Sci USA *86*, 3699–3703.

Hamilton, R., Watanabe, C.K., and de Boer, H.A. (1987). Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. Nucleic Acids Res *15*, 3581–3593.

Hersch, S.J., Elgamal, S., Katz, A., Ibba, M., and Navarre, W.W. (2014). Translation Initiation Rate Determines the Impact of Ribosome Stalling on Bacterial Protein Synthesis. Journal of Biological Chemistry *289*, 28160–28171.

Hiraoka, Y., Kawamata, K., Haraguchi, T., and Chikashige, Y. (2009). Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes to Cells *14*, 499–509.

Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S., and Press, W.H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLoS Genet *11*, e1005732.

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

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

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009a). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science *324*, 218–223.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009b). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Irwin, B., Heck, J.D., and Hatfield, G.W. (1995). Codon pair utilization biases influence translational elongation step times. J. Biol. Chem. *270*, 22801–22806.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics *9*, 102.

Johansson, M.J.O., Esberg, A., Huang, B., Bjork, G.R., and Bystrom, A.S. (2008). Eukaryotic Wobble Uridine Modifications Promote a Functionally Redundant Decoding System. Mol. Cell. Biol. *28*, 3301–3312.

Kane, J.F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr Opin Biotechnol *6*, 494–500.

Kato, M., Nishikawa, K., Uritani, M., Miyazaki, M., and Takemura, S. (1990). The difference in the type of codon-anticodon base pairing at the ribosomal P-site is one of the determinants of the translational rate. J. Biochem. *107*, 242–247.

Keppler-Ross, S., Noffz, C., and Dean, N. (2008). A New Purple Fluorescent Color Marker for Genetic Studies in Saccharomyces cerevisiae and Candida albicans. Genetics *179*, 705–710.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007). A “Silent” Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science *315*, 525–528.

Kortmann, J., and Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. Nat Rev Microbiol *10*, 255–265.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene *361*, 13–37.

Kozak, M. (2002). Emerging links between initiation of translation and human diseases. Mamm. Genome *13*, 401–410.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science *324*, 255–258.

Lange, S.J., Maticzka, D., Mohl, M., Gagnon, J.N., Brown, C.M., and Backofen, R. (2012).

Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Res *40*, 5215–5226.

Lareau, L.F., Hite, D.H., Hogan, G.J., and Brown, P.O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife *3*, e01257.

Letzring, D.P., Dean, K.M., and Grayhack, E.J. (2010). Control of translation efficiency in yeast by codon-anticodon interactions. Rna *16*, 2516–2528.

Letzring, D.P., Wolf, A.S., Brule, C.E., and Grayhack, E.J. (2013). Translation of CGA codon repeats in yeast involves quality control components and ribosomal protein L1. Rna *19*, 1208–1217.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012). Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. The Plant Cell *24*, 4346–4359.

Liszewski, K. (2015). Speeding Up the Protein Assembly Line. GEN: Genetic Engineering and Biotechnology News *35*, 1.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proceedings of the National Academy of Sciences *108*, 11063–11068.

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet *39*, 415–421.

Markham, N.R., and Zuker, M. (2008). UNAFold: Software. 1–33.

Mauro, V.P., and Chappell, S.A. (2014). A critical analysis of codon optimization in human therapeutics. Trends Mol Med *20*, 604–613.

McManus, C.J., May, G.E., Spealman, P., and Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Research *24*, 422–430.

Moller-Jensen, J., Franch, T., and Gerdes, K. (2001). Temporal Translational Control by a Metastable RNA Structure. Journal of Biological Chemistry *276*, 35707–35713.

Mortimer, S.A., Kidwell, M.A., and Doudna, J.A. (2014). Insights into RNA structure and function from genome-wide studies. Nat Rev Genet *15*, 469–479.

Moura, G., Pinheiro, M., Silva, R., Miranda, I., Afreixo, V., Dias, G., Freitas, A., Oliveira, J.L., and Santos, M.A.S. (2005). Comparative context analysis of codon pairs on an ORFeome scale. Genome Biol *6*, R28.

Murgola, E.J., Pagel, F.T., and Hijazi, K.A. (1984). Codon context effects in missense

suppression. J Mol Biol *175*, 19–27.

Murphy, F.V., Ramakrishnan, V., Malkiewicz, A., and Agris, P.F. (2004). The role of modifications in codon discrimination by tRNALysUUU. Nat Struct Mol Biol *11*, 1186–1191.

Nasvall, S.J., Chen, P., and Bjork, G.R. (2007). The wobble hypothesis revisited: Uridine-5-oxyacetic acid is critical for reading of G-ending codons. Rna *13*, 2151–2164.

Navon, S., and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biol *12*, R12.

Nedialkova, D.D., and Leidel, S.A. (2015). Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. Cell *161*, 1606–1618.

Novoa, E.M., and de Pouplana, L.R. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. Trends in Genetics 1–8.

Ogle, J.M., and Ramakrishnan, V. (2005). STRUCTURAL INSIGHTS INTO TRANSLATIONAL FIDELITY. Annu. Rev. Biochem. *74*, 129–177.

Pechmann, S., and Frydman, J. (2012). nsmb.2466. Nat Struct Mol Biol *20*, 237–243.

Pelechano, V., Wei, W., and Steinmetz, L.M. (2015). Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. Cell *161*, 1400–1412.

Percudani, R., Pavesi, A., and Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol *268*, 322–330.

Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2005). Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol *24*, 79–88.

Phillips-Jones, M.K., Watson, F.J., and Martin, R. (1993). The 3' codon context effect on UAG suppressor tRNA is different in Escherichia coli and human cells. J Mol Biol *233*, 1–6.

Plotkin, J.B. (2011). Cell biology. The lives of proteins. Science *331*, 683–684.

Plotkin, J.B., and Kudla, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. Nature Publishing Group *12*, 32–42.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular Systems Biology *10*, 770–770.

Precup, J., and Parker, J. (1987). Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. *262*, 11351–11355.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., et al. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. Cell *160*, 1111–1124.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet *8*, e1002603.

Quartley, E., Alexandrov, A., Mikucki, M., Buckner, F.S., Hol, W.G., DeTitta, G.T., Phizicky, E.M., and Grayhack, E.J. (2008). Heterologous expression of L. major proteins in S. cerevisiae: a test of solubility, purity, and gene recoding. Journal of Molecular Evolution *66*, 210–223.

Quax, T.E.F., Wolf, Y.I., Koehorst, J.J., Wurtzel, O., van der Oost, R., Ran, W., Blombach, F., Makarova, K.S., Brouns, S.J.J., Forster, A.C., et al. (2013). Differential Translation Tunes Uneven Production of Operon-Encoded Proteins. CellReports *4*, 938–944.

Reis, dos, M. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res *32*, 5036–5044.

Reuter, J.S., and Mathews, D.H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*, 129.

Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M., and Humphreys, G. (1984). Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res *12*, 6663–6671.

Rosenberg, A.H., Goldman, E., Dunn, J.J., Studier, F.W., and Zubay, G. (1993). Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. J. Bacteriol. *175*, 716–722.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M., and Yusupova, G. (2015). Structural insights into the translational infidelity mechanism. Nature Communications *6*, 1–9.

Sander, I.M., Chaney, J.L., and Clark, P.L. (2014). Expanding Anfinsen's Principle: Contributions of Synonymous Codon Selection to Rational Protein Design. J. Am. Chem. Soc. *136*, 858–861.

Satchidanandam, V., and Shivashankar, Y. (1997). Availability of a second upstream AUG can completely overcome inhibition of protein synthesis initiation engendered by mRNA secondary structure encompassing the start codon. Gene *196*, 231–237.

Scharff, L.B., Childs, L., Walther, D., and Bock, R. (2011). Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites. PLoS Genet *7*, e1002155.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Seetin, M.G., and Mathews, D.H. (2012). RNA Structure Prediction: An Overview of Methods.

(Totowa, NJ: Humana Press), pp. 99–122.

Shabalina, S.A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res *34*, 2428–2437.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J.B. (2013). Rate-Limiting Steps in Yeast Protein Translation. Cell *153*, 1589–1601.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res *15*, 1281–1295.

Stadler, M., and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. Rna *17*, 2063–2073.

Stoebel, D.M., Dean, A.M., and Dykhuizen, D.E. (2008). The Cost of Expression of Escherichia coli lac Operon Proteins Is in the Process, Not in the Products. Genetics *178*, 1653–1660.

Stoletzki, N., and Eyre-Walker, A. (2006). Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Molecular Biology and Evolution *24*, 374–381.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Ben Lehner (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324–1335.

Takyar, S., Hickerson, R.P., and Noller, H.F. (2005). mRNA Helicase Activity of the Ribosome. Cell *120*, 49–58.

Thanaraj, T.A., and Argos, P. (1996a). Ribosome-mediated translational pause and protein domain organization. Protein Sci. *5*, 1594–1612.

Thanaraj, T.A., and Argos, P. (1996b). Protein secondary structural types are differentially coded on messenger RNA. Protein Sci. *5*, 1973–1983.

Thomas, L.K., Dix, D.B., and Thompson, R.C. (1988). Codon choice and gene expression: synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes in vitro. Proc Natl Acad Sci USA *85*, 4242–4246.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell *141*, 344–354.

Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proceedings of the National Academy of Sciences *107*, 3645–3650.

Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods *7*, 995–1001.

Varenne, S., and Lazdunski, C. (1986). Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. Journal of Theoretical Biology *120*, 99–110.

Varenne, S., Buc, J., Lloubes, R., and 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*, 549–576.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D.L., Nutter, R.C., Segal, E., and Chang, H.Y. (2012). Genome-wide Measurement of RNA Folding Energies. Mol Cell *48*, 169–181.

Weatheritt, R.J., and Babu, M.M. (2013). The Hidden Codes That Shape Protein Evolution. Science *342*, 1325–1326.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2015). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009). Design Parameters to Control Synthetic Gene Expression in Escherichia coli. PLoS ONE *4*, e7002.

Wolf, A.S., and Grayhack, E.J. (2015). Asc1, homolog of human RACK1, prevents frameshifting in yeast by ribosomes stalled at CGA codon repeats. Rna *21*, 935–945.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., and Johnson, C.H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature *495*, 116–120.

Yarus, M., and Folley, L.S. (1985). Sense codons are found in specific contexts. J Mol Biol *182*, 529–540.

Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M., and Pierce, N.A. (2010). NUPACK: Analysis and design of nucleic acid systems. J. Comput. Chem. *32*, 170–173.

Zaher, H.S., and Green, R. (2008). Quality control by the ribosome following peptide bond formation. Nature *457*, 161–168.

Zhang, F., Saha, S., Shabalina, S.A., and Kashina, A. (2010). Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. Science *329*, 1534–1537.

Zhang, G., and Ignatova, Z. (2009). Generic Algorithm to Predict the Speed of Translational Elongation: Implications for Protein Biogenesis. PLoS ONE *4*, e5036.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol *16*, 274–280.

Zhou, J.-H., You, Y.-N., Chen, H.-T., Zhang, J., Ma, L.-N., Ding, Y.-Z., Pejsak, Z., and Liu, Y.-S. (2013). The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus. Infection, Genetics and Evolution *16*, 270–274.

Zhou, T., Weems, M., and Wilke, C.O. (2009). Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. Molecular Biology and Evolution *26*, 1571–1580.

Zhou, Z., Schnake, P., Xiao, L., and Lal, A.A. (2004). Enhanced expression of a recombinant malaria candidate vaccine in Escherichia coli by codon optimization. Protein Expression and Purification *34*, 87–94.

Zinshteyn, B., and Gilbert, W.V. (2013). Loss of a conserved tRNA anticodon modification perturbs cellular signaling. PLoS Genet *9*, e1003675.

Zuker, M. (2000). Calculating nucleic acid secondary structure. Current Opinion in Structural Biology *10*, 303–310.