Summer 2020

# Codon bias and mRNA folding stability: Two natural controls of protein expression dynamics

Anastacia Wienecke
*Western Washington University*, acia101@gmail.com

**Codon bias and mRNA folding stability:**
**Two natural controls of protein expression dynamics**

By

Anastacia Nancy Wienecke

Accepted in Partial Completion
of the Requirements for the Degree
Master of Science

ADVISORY COMMITTEE

Dr. Daniel Pollard, Chair

Dr. Suzanne Lee

Dr. Dietmar Schwarz

GRADUATE SCHOOL

David L. Patrick, Dean

**Master's Thesis**

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Western Washington University, I grant to Western Washington University the non-exclusive royalty-free right to archive, reproduce, distribute, and display the thesis in any and all forms, including electronic format, via any digital library mechanisms maintained by WWU.

I represent and warrant this is my original work, and does not infringe or violate any rights of others. I warrant that I have obtained written permissions from the owner of any third party copyrighted material included in these files.

I acknowledge that I retain ownership rights to the copyright of this work, including but not limited to the right to use all or part of this work in future works, such as articles or books.

Library users are granted permission for individual, research and non-commercial reproduction of this work for educational purposes only. Any further digital posting of this document requires specific permission from the author.

Any copying or publication of this thesis for commercial purposes, or for financial gain, is not allowed without my written permission.

Anastacia Wienecke

July 19, 2020

# Codon bias and mRNA folding stability:
# Two natural controls of protein expression dynamics

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

by
Anastacia Nancy Wienecke
July 2020

## ABSTRACT

**Introduction:** Connections between genetic variation and trait variation are complex and dynamic; the critical link between the two is gene expression variation. Though proteins are the functional products of most genes, the relative ease and throughput level of various measurement approaches has meant that gene expression is typically studied via transcript-level rather than protein-level techniques. Recent studies however, suggest that certain genetic factors act post-transcriptionally to modify rates of protein synthesis, making transcript levels imperfect indicators of protein levels. A gene's bias for 'optimal' codons (i.e., its codon bias) and a gene's mRNA folding stability appear to be two such factors, though until the present work, their individual effects on protein synthesis rates have not been systematically confirmed and quantified in a large number of genes or in a natural (non-genetically engineered) system.

**Methods:** Our study is based on sequence, mRNA and protein expression data for 1620 genes across a diverse set of 22 *Saccharomyces cerevisiae* isolates. For each gene, we model how across-isolate changes in codon bias and in mRNA folding stability relate to the across-isolate variations in the steady-state ratio of protein level to mRNA level, our approximation of protein synthesis rate.

**Results:** In general for each gene, the alleles with higher codon biases are also those with higher ratios of protein per mRNA (PPR). This relationship is especially pronounced when the gene's alleles have high mRNA folding stabilities. On a finer scale, changes in the biases of four local gene regions (protein-domain-encoding, inter-domain-encoding, 5'coding, and 3'coding) differentially contribute to the PPR of the gene. In parallel, for each gene, alleles with more stable mRNA folding are also those with higher PPR. This is particularly true if their codon biases are already high, though it does not occur through the proposed mechanisms acting on mRNA structures near the start and stop codons. Lastly, we conclude that a gene's codon bias and mRNA folding stability act synergistically to tune its PPR.

## ACKNOWLEDGEMENTS

It takes a village to raise a thesis, and with all my heart, I would like to express my sincerest appreciation and gratitude for all those who helped make this project possible: my thesis advisor, Dr. Daniel Pollard, for spending countless hours pondering my mechanistic hypotheses, discussing the literature, helping me make sense of my results, providing financial support (from the National Science Foundation Award MCB-1518314), and sharpening my initiative, my scientific writing, and my scientific thinking; my committee members, Drs. Suzanne Lee and Dietmar Schwarz, for their invaluable comments, suggestions, and words of encouragement; the WWU Biology Department for their support and their funding assistance through the 2019 Biology Summer Fellowship and teaching assistantship; the WWU Research and Sponsored Programs for so generously providing me with a more powerful computer to use for my analyses; the Yeast Resource Center at the University of Washington (particularly Drs. Maitreya Dunham and Michael MacCoss) for painstakingly collecting and processing DNA sequence, mRNA abundance, and protein abundance data for 22 yeast isolates, and then graciously sharing their data with me; the users of Stack Overflow (stackoverflow.com), of whom there are far too many to name, for sharing their expertise in Python and in R; Dr. Jose Serrano-Moreno for helping me recognize, before the beginning of this endeavor, my academic potential and for mentioning the research going on in the Pollard Lab; my fellow students in the Pollard Lab for their continual encouragement and thought-provoking questions; the Biology graduate students for lots of laughs and camaraderie; my formerly biology-layman dad, Allan, for being so curious about my project that he eventually became a trusty sounding board; my mom, Natalya, and my dear friends, Julia, Mia, Lina, and the Bellingham Baha'i Community for keeping me smiling.

## TABLE OF CONTENTS

## LIST OF FIGURES

# INTRODUCTION

Proteins play highly specialized roles in the cell. Collectively, they help cells communicate, process information, transport molecules, catalyze reactions, regulate development, and maintain structure. To sustain each moment of this elaborate symphony, cells conduct the building and breakdown of each type of protein, which in *Homo sapiens* exceeds 20,000 different types (where we define type as the set of proteins produced from a given gene). In *Mus musculus* (house mouse), *Drosophila melanogaster* (fruit fly), and *Saccharomyces cerevisiae* (baker's yeast), this number is upwards of 20,000, 14,000, and 6,000, respectively.

Instructions for protein building, i.e. genes, lie at special addresses along DNA, which itself is an ordered sequence of four nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). At any given time, both internal and external signals integrate to either promote or repress a gene's expression. If the net signal is to promote, the cell synthesizes protein in two steps: transcription and translation (see **Figure 1**). During transcription, RNA polymerase copies the requisite gene's DNA sequence into a messenger RNA (mRNA) transcript. During translation, a ribosome scans the mRNA in groups of three nucleotides called codons and synthesizes a sequence of amino acids to match. Each amino acid is brought to the ribosome by a transfer RNA (tRNA), and 'start' and 'stop' codons mark the beginning and end of translation. Based on the combined interactions between amino acids, the nascent sequence folds co-translationally into a unique shape, becoming a protein ready to perform a given function. While relatively uncommon, amino acid substitutions do occur with some frequency. Whether they be caused by mistranslation or mutation, these substitutions can change protein folding dramatically (sometimes rendering the protein useless) or not at all.

Genetic variation alters the timing, location, and extent of protein expression by acting on any of four components of protein expression dynamics: rates of mRNA transcription, mRNA decay, protein

translation, and protein decay (see **Figure 1**). Such modifications cascade to produce changes in cellular states and overall traits (Stern & Orgogozo, 2008), thereby providing fuel for adaptation and evolution (Chan et al, 2010; Skelly et al, 2009; Wang et al, 1995). Because recent sequencing advances have made mRNA abundances more straightforward to quantify than protein abundances, numerous studies have probed the effects of protein expression variation via mRNA abundance variation (see Pai et al, 2012; Skelly et al, 2009; Rockman & Kruglyak, 2006; Brem et al, 2002). An issue however, emerges with this approach: there are genetic factors that directly modify rates of protein translation and protein decay, often making mRNA levels imperfect indicators of protein levels (Pollard et al, 2016; Albert et al, 2014; Parts et al, 2014; Straub, 2011; Foss et al, 2011; Lu et al, 2007; Gygi et al, 1999). These factors can be specific to the gene sequence (*cis*-acting) or to a different part of the genome (*trans*-acting). In this inquiry, we focus on two *cis*-acting factors: codon bias and mRNA folding stability.

A gene's rate of protein synthesis is thought to be affected by changes in its bias for 'optimal' codons (i.e., its codon bias) and in its mRNA folding stability (see Hanson & Coller, 2018; Tuller et al, 2011). Until the present work however, these relationships have not been systematically confirmed and quantified in a large number of individual genes or in a natural (non-genetically-engineered) system. We base our investigation on a recently-collected dataset of DNA sequences, mRNA abundances, and protein abundances (see Methods). These data span 22 genetically-diverse yeast isolates sampled from six continents and 12 types of microenvironments (e.g. bee hairs, throat sputum, fermenting palm sap, leavening bread, and forest soil) (Skelly et al, 2013). In many ways, this is an ideal dataset for our type of study – it spans 1620 (25%) genes across isolates, and the yeast are: 1) haploid, so all mRNA and protein levels are allele-specific, 2) single-celled, so tissue-specific differences are irrelevant, 3) measured at steady-state, so the effects of time need not factor into our models, and 4) quite genetically-diverse, so we can compare the translation-related effects of single nucleotide polymorphisms (SNPs) across the alleles of each gene. Because we are examining the variation of each gene across strains (instead of just across genes), we can home in on codon bias and mRNA folding as methods of protein expression regulation. Below, we

review codon bias and mRNA folding on the global and local levels, and we review their predicted effects on protein synthesis rates.

## Global Codon Bias

The genetic code is redundant: 61 codons map to 20 amino acids. As a result, any one of a group of 'synonymous' codons specify a given amino acid during translation. For example, in the standard genetic code (which is used by yeast and humans), the amino acid tryptophan is encoded by *TGG*, while amino acid valine is equally encoded by *GTT*, *GTC*, *GTA*, or *GTG*. It therefore came as a surprise when several decades ago, the results of early DNA sequencing analyses suggested that across all domains of life and for every amino acid, synonymous codons are used in uneven frequencies (Sharp & Li, 1988) (see **Figure 2**). This pattern was named 'codon bias', and for each gene, it appears to have arisen from a unique balance of neutral mutation and natural selection; the former is biased towards A and T nucleotides in yeast (Zhu et al, 2014) and the latter concerns both the rate and the accuracy of synonymous codon translation (Trotta, 2013; Wallace at al, 2013; Plotkin & Kudla, 2010; Hershberg & Petrov, 2008).

In highly expressed genes especially, the most 'preferred' synonymous codons tend to be those with high abundances of cognate tRNAs, likely because the ribosome translates these codons most quickly and most precisely (Dana & Tuller, 2014; Sharp & Li, 1986; Ikemura, 1982); genes with this characteristic codon composition are said to have "high global codon bias" and codons with the highest cognate tRNA gene copy numbers (an approximation of cognate tRNA abundance) are said to be "quickly translated" (Tuller et al, 2010). We wonder: does a gene's protein synthesis rate increase when SNPs raise its global codon bias? We know that fluctuations in tRNA supplies occur in circadian rhythms (Frenkel-Morgenstern et al, 2012), the cell-cycle (Xu et al, 2013), cell proliferation/differentiation (Gingold et al, 2016), and stress (Torrent et al, 2018); genes whose codon biases closely match the new pool become upregulated, while genes whose biases no longer align with this pool become downregulated. Based on these and other

observations from the literature (reviewed in Quax et al, 2015), we predict that codon bias is a potent *cis*-acting form of gene expression regulation, and that patterns long observed in individual genetically-engineered cases (see Angov et al, 2008; Burgess-Brown et al, 2008; Gooch et al, 2008; Gustafsson et al, 2004; Morgan et al, 2003) also exist in all genes in the genome. From our model, we expect to see a positive association between a gene's changing global codon bias across isolates and its changing logarithm of protein per mRNA (logPPR) across isolates (protein per mRNA being our estimate of protein synthesis rate).

## Local Codon Bias

Synonymous codons are not randomly interspersed along the mRNA sequence. The distinct advantages of their individual levels of cognate tRNA abundances may explain why varying selection pressures create and maintain these localized patterns of codon bias. Generally, codons with high levels of cognate tRNAs are the most quickly translated, and a ribosome's instantaneous translation rate appears to have a role in defining the nascent protein's co-translational folding and stability, the specifics of which are critical for its proper function (Buhr et al, 2016; Yu et al, 2015; Chartier et al, 2012). Further, in certain mRNA regions, faithful translation is critical, and it can be achieved via careful codon choice: codons with high supplies of cognate tRNAs are generally the most accurately translated (Kramer et al, 2010, Zhou et al, 2009, Kramer & Farabaugh, 2007). Below, we discuss a few mechanistic ideas concerning these selection pressures, and we discuss the characteristic biases of four types of mRNA regions – protein domain encoding, protein inter-domain ("linker") encoding, 5'coding, and 3'coding (see **Figure 3**). We also review several hypotheses outlining how the changing biases of these regions might affect a gene's logPPR (see **Figure 3** for a pictorial summary).

a. Domains are independently operating subunits of a protein, and together, they define the protein's overall function. Many structurally important regions lie within domains, and an amino acid

substitution in any one of them could destabilize domain configuration, resulting in a useless, and potentially toxic, protein product (Zhou et al, 2015; Geiler-Samerotte et al, 2011; Zhou et al, 2009, Drummond & Wilke, 2009). This might explain why domain-encoding regions show particularly high densities of accurately (and quickly) translated codons (Kramer et al, 2010, Zhou et al, 2009, Kramer & Farabaugh, 2007). A secondary benefit to this high characteristic codon bias is faster overall protein translation, which from a fitness standpoint, is important for fast-growing organisms such as yeast. For each gene, certain SNPs across isolates increase this region's codon bias, and we predict that such increases are associated with alleles having higher levels of logPPR.

b. Some of the most mildly structured protein regions lie in-between domains; these are the inter-domain linkers, or just "linkers", and they are typically specified by codons with low cognate tRNA abundances (Weinberg et al, 2016). The translation of these particular codons takes the most time (at least in yeast), and this linker region slowdown could facilitate the proper co-translational folding of the preceding domains (Pechmann & Frydman, 2013; Makhoul & Trifonov, 2002; Thanaraj & Argos, 1996). Could we then expect that the general relationship between allele linker bias and allele logPPR is negative? Given our prediction for global codon bias, and given that increases in linker bias necessitate increases in global bias, the answer to the above question is not clear; if however, the negative effects of increasing linker bias are strong enough to overpower the positive effects of increasing global bias, this local-level relationship will be negative.

c. The 5'coding region (or "rare codon ramp") generally spans the first 30-50 codons of the mRNA. It is characterized by a high density of ribosomes – nearly three times that of any other mRNA region (Ingolia et al, 2009). Such a density can be explained by its low codon bias (i.e. by its frequent usage of codons with lowly abundant cognate tRNAs). The slow translation of these codons may serve to decelerate and respace ribosomes (Tuller et al, 2010), minimizing the risk for downstream collisions and jams that often lead to translation delays and early translation termination (Doma & Parker,

2006). Another hypothesis suggests that slow 5'coding region translation could be a mechanism for controlling translation initiation rates because any ribosomes located immediately downstream of the start codon preclude initiation (Chu et al, 2014). Thus, we could expect to see a negative relationship between allelic logPPR and 5'coding region bias, but since increases in 5'coding region bias necessitate increases in global codon bias, we could also expect to see a weakly positive relationship.

d. The 3'coding region is bordered by the 3'-most domain-encoding region and the translation stop codon. Compared with elsewhere in the gene, this region has the highest proportion of quickly translated codons (as measured by tRNA gene copy numbers) (Tuller et al, 2010). Such levels of bias are thought to protect against ribosome collisions and the ensuing interruptions in protein synthesis; an example of the latter is a premature translation termination, and it is especially costly (in terms of energy and resources) when the ribosome is this far past the start codon (Tuller et al, 2010; Qin et al, 2004). This hypothesis also supports the observed correlation between codon bias and gene length in *E. coli* (Plotkin & Kudla, 2010). Alleles showing higher bias in their 3'coding regions (due to SNPs across isolates) should also have higher levels of logPPR; this is what we expect to see from our models.

## Global mRNA Folding Stability

Elaborate structures form as an mRNA transcript folds onto itself; the result of this process is termed "global mRNA folding". While a vast number of folding configurations is possible, only a few are thermodynamically likely, and these comprise the mRNA's "thermodynamic ensemble". Remarkably, even a single nucleotide substitution can dramatically change the folding and the stability of ensemble structures (see **Figure 4**), shifting the speed at which ribosomes unravel them (Takyar et al, 2005). We wonder: is mRNA folding stability a mechanism for protein expression regulation?

The stable folding of mRNA is selected for (Park et al, 2013), and may be key for fast translation. Indeed, the across-gene correlation between mRNA folding stability and protein abundance is positive and strong (R = 0.68 from Zur & Tuller, 2012; Tuller et al, 2011). Yet, these observations may seem surprising because intuition could expect stronger structures to require more time and energy to unfold, and therefore lead to slower translation. Three recently proposed hypotheses attempt to reconcile this apparent contradiction, and we summarize them here. First, Zur & Tuller, 2012 propose that structured mRNAs are less prone to aggregate. Any negative effects associated with aggregation, they suggest, may well-outweigh those imparted by stable folding. Second, Lai et al, 2018 observe that strong folding maintains a short distance between 5' and 3' mRNA termini, thereby preserving favorable entropy for mRNA circularization. Such a looped arrangement mediates translation initiation (Paek et al, 2015) by facilitating ribosome re-initiation after translation termination. Third, from their theoretical model of yeast translation, Mao et al, 2014 noticed that more stably folded mRNAs have faster predicted rates of translation and higher densities of ribosomes. Their conclusions, we predict, reflect a clustering of ribosomes by strong structures: mechanistically, as the first ribosome ("$R_1$") slowly unwinds a strong structure, several ribosomes ("$R_2$", "$R_3$", . . . "$R_n$") accumulate behind it in queue. The length of this train, *n*, depends on the amount of time $R_1$ spends unfolding, which in turn depends on the structure's folding stability. Once the structure is successfully unraveled by $R_1$, $R_1$ resumes translation, and since the densely packed queue of ribosomes is so close behind $R_1$, the newly unwound structure does not have time to reform, and the whole queue translates the sequence without having to spend time and energy unwinding it. If the distance between $R_{n+1}$ and $R_n$ is sufficiently large, the newly-unfolded segment refolds before $R_{n+1}$ reaches it, in which case, $R_{n+1}$ becomes the next $R_1$, i.e. the leader of the next queue, or train, of ribosomes. In this way, only the queue-leading $R_1$ ribosomes spend a substantial amount of time unwinding the sequence. In weakly folded mRNAs, most ribosomes are $R_1$ ribosomes, but in strongly folded mRNAs, relatively few are $R_1$s. From the above observations, we predict that in general, alleles with stronger mRNA folding also show higher levels of logPPR.

## Local mRNA Folding Stability

Across genes, certain folding patterns appear at key mRNA locations. Some are conserved across all domains of life (Gebert et al, 2019) and act to modulate speeds of translation initiation, elongation, and termination. Here, we review the inherent stabilities of five mRNA regions: the coding sequence (CDS), *+1 to +10* bases from the 5'cap, *-9 to +3* bases from the start codon, *+4 to +10* bases from the start codon, and *+1 to +18* bases from the stop codon. In **Figure 5** and below, we summarize our predictions for how changes in these regions' folding stabilities should affect logPPR.

a. The coding sequence (CDS) of yeast mRNA folds more stably than either the 5' or the 3' untranslated region (UTR) (Wan et al, 2012; Kertesz et al, 2010); across all genes, this hallmark is both selected for (Katz & Burge, 2003) and well-correlated with gene expression (Zur & Tuller, 2012). Kertesz et al, 2010 propose that this quality could prevent costly translation initiation within the coding sequence, or it could aid in tRNA translocation within the ribosome. Further, stable folding may shorten the distance between translating ribosomes, and thereby decrease the probability of mRNA structures re-forming for each ribosome to unwind (Mao et al, 2014). Alleles with more stable folding in this region should also have higher levels of logPPR.

b. In human and in mouse mRNAs, strong structures near the 5' cap dramatically lower the efficiency of translation initiation (Babendur et al, 2006; Kozak, 1989). While the exact mechanism behind this action is unknown, one hypothesis suggests that this region's strong folding interferes with the binding of translation preinitiation complexes. In yeast genes however, for reasons still unknown, the opposite appears to be true: strong folding is most associated with increased protein yield when located *+1 to +10* bases from the 5'cap, as opposed to elsewhere in the 5'UTR (see Supplementary Figure S2b in Cuperus et al, 2017 and Figure 3c in Kertesz et al, 2010). If yeast engage in cap-dependent translation initiation like humans and mice do, then despite whatever pressures lead to the increased structure stabilities near their 5'caps, we would expect to see a weakly negative or a

somewhat positive relationship between logPPR and this region's folding stability; such a result would suggest that these local-level effects transcend the pressures that try to keep the entire mRNA (*+1 to +10* bases of 5'cap included) stably structured for high logPPR.

c. The region of mRNA just upstream of (or equivalently, 5' of) the translation start codon (AUG) tends to be relatively free of structure. Ever since early studies saw associations between translation inhibition and stable stem-loops located 5' of AUG, this observed pattern of weak folding has been a predicted means of facilitating AUG recognition by the preinitiation complex (Lamping et al, 2013; Vega Laso, 1993; Baim & Sherman, 1988; Kozak, 1986). Another hypothesis suggests that weaker structures could prevent steric hindrance during ribosome assembly. In eukaryotic mRNAs, folding is especially mild within *-9 to +3* bases from AUG (Wan et al, 2012; Kertesz et al, 2010; Shabalina et al, 2006). Thus, among the alleles of each gene, we predict to see a connection between logPPR and this region's folding stability. This connection should be weakly positive (but could be negative) because of the predicted positive effect of high global folding stability on logPPR.

d. Structures just downstream of (or equivalently, 3' of) the start codon can, in some cases, improve the efficiency of translation initiation, perhaps by slowing the ribosome as it approaches the start codon. This is especially prevalent in genes with suboptimal start codon contexts (Kozak, 1990). It is also the case when upon infection, the mammalian immune response inactivates an essential translation initiation factor; at this point stable stem-loops downstream of viral start codons uphold steady translation rates (Ventoso et al, 2006). In eukaryotic mRNAs, a spike in folding stability occurs within the region *+4 to +10* bases from AUG (see Figure 3c in Kertesz et al, 2010 and Figure 1a in Shabalina et al, 2006). We predict that among alleles, SNPs that cause increases in this region's folding stability create a sort of 'speed bump' for the ribosome as it nears AUG, improving the efficiency of translation initiation and increasing logPPR.

e. Strong mRNA folding surrounding the translation stop codon generally lowers translation efficiency. It may act to sterically hinder translation termination machinery (Niepel et al, 1999), or (in some cases) encourage stop codon read-through (Napthine et al, 2012). As could be expected, across yeast, mouse, and human genes, this region, like the rest of the 3'UTR, tends to be devoid of structure (see Figure 3c in Kertesz et al, 2010 and Figure 1b in Shabalina et al, 2006). Since fifteen nucleotides span the distance between the front of the ribosome and the part that holds the growing chain of amino acids (the P site), we consider each gene's relationship between logPPR and folding stability *+1 to +18* bases from the stop codon. We predict that this association will be lower in magnitude than the inferred positive association between global folding stability and logPPR, though if it is sufficiently potent, it could overpower the global-level relationship and take a negative sign.

# METHODS

## Data Collection and Processing

From Skelly et al, 2013, we download genome sequence, processed mRNA abundance, and processed peptide abundance data for each of 1636 genes in the following 22 yeast isolates: 273614N, 378604X, BC187, DBVPG1106, DBVPG1373, DBVPG6765, L_1374, NCYC361, SK1, UWOPS05_217_3, UWOPS05_227_2, UWOPS83_787_3, UWOPS87_2421, Y12, Y55, YJM975, YJM978, YJM981, YPS128, YPS606, YS2, and YS9.

For each gene in each strain, we express protein abundance as a sum of peptide levels (Michael J. MacCoss, personal communication, July 2018); we define the coding sequence (CDS) based on coordinates supplied by the Skelly et al, 2013 general feature format (.gff) file; and we define 5'UTR and 3'UTR sequences based on the UTR length specifications from Tuller et al, 2009. The whole mRNA sequence is then the concatenation of 5'UTR, CDS, and 3'UTR sequences.

## Approximating Protein Synthesis Rates

For each gene in each strain, we approximate protein synthesis rates as the steady-state ratio of the gene's protein abundance to its mRNA abundance (protein per mRNA, or PPR). Before we calculate this ratio, for each isolate, we normalize mRNA abundance and protein abundance measurements by estimates of actual cell-wide mRNA and protein molecule counts (see Miura et al, 2008; von der Haar, 2008). In this way, we correct for any variability in translation machinery among isolates. Additionally, rather than PPR being in arbitrary units, it becomes approximately in units of protein molecules per mRNA molecule. After computing PPR, we log transform it (logPPR).

It is important to note that the magnitudes of our PPR ratios also depend on protein decay rates. As such, while we seek to understand the effect of SNPs on protein synthesis rates, we also understand that

these same SNPs could affect protein decay rates (though no evidence suggests that this is through codon bias, folding stability, or a combination of the two).

## Approximating Global Codon Bias

Three classic methods of estimating a gene's codon bias are the Codon Adaptation Index (CAI), the tRNA Adaptation Index (tAI), and the normalized tRNA Adaptation Index (ntAI). Each relies on its own respective codon table, where every codon maps to one value in the range (0, 1]. A gene's CAI, tAI, or ntAI equals the geometric mean of values assigned to its comprising codons by the requisite table. Below, we briefly review each of these measures as well as one additional measure, length-normalized CAI (nlCAI). Each measure is computed with Python (version 3.7.1).

i. CAI quantifies a gene's tendency to use the synonymous codons most favored by a pre-defined training set of genes (Sharp & Li, 1987). A CAI value of 1 indicates total usage of these codons, while a CAI value approaching 0 indicates complete avoidance. Based on the limited gene expression data available at the time, 24 highly expressed genes were selected for the first CAI training set (Sharp et al, 1986). Years later, Carbone et al, 2003 proposed a more structured set-selection approach; one that seeks to capture the dominant genomic-level driver of codon bias, which in yeast and in several other fast-growing organisms, is very similar to the bias of the most highly expressed genes (Carbone et al, 2003; Sharp et al, 1988). We use the 61 genes identified by the Carbone et al, 2003 approach to calculate one CAI codon table per isolate. We then compute a single median CAI codon table across isolates. This is the table we use to measure the CAI of the coding sequence (CDS) of each gene, i.e. a gene's global codon bias.

ii. Normalized-by-length CAI (nlCAI) is our slightly modified version of CAI. Longer training set genes contribute more to the CAI codon table, and because all genes have their own intrinsic biases (see Quax

et al, 2015), these large contributions may misrepresent the overall bias of the Carbone-selected training set. Instead of computing the CAI codon table based on each gene's synonymous codon counts, we compute it based on each gene's synonymous codon percent abundances. Specifically, we calculate the fraction of codons that are codon *i* in each gene, and then add up all such fractions across genes. This gives a 61-element array, where each value matches to a sense codon. For each group of synonymous codons, we divide their corresponding array values by the maximum array value within that group. In this way, we compute a single nlCAI codon table for each isolate, and then take their median table for nlCAI calculations.

iii. tAI estimates how often a gene uses synonymous codons with high supplies of cognate/near-cognate tRNAs. A gene always using such codons has a tAI near 1, and a gene never using such codons has a tAI near 0. This measure accounts for cases in which one tRNA recognizes more than one codon (wobble) (Crick, 1966), and it approximates tRNA supply by tRNA gene copy number in the genome (dos Reis et al., 2004). The high positive correlation (r = 0.76) between tRNA gene copy number and tRNA abundance (in yeast) suggests that this is a reasonable approximation for our study (Tuller et al, 2010). Based on the dos Reis et al, 2014 approach, we compute a single tAI codon table for use in all strains.

iv. ntAI considers both the abundance of tRNAs (as measured by tRNA gene copy number in the genome) and the abundance of codons competing for them (as measured by the sum of codon translation frequencies across all mRNAs) (Pechmann & Frydman, 2012). From this view, a codon optimal for fast translation is one whose tRNA species are high in abundance and low in demand. A gene always using such synonymous codons has a ntAI value near 1, while a gene never using such values has a ntAI value near 0. We use the Pechmann & Frydman, 2012 approach to calculate an individual ntAI codon table per isolate. For each isolate, we then compute ntAI with its corresponding table.

## Approximating Local Codon Bias

We download each gene's protein domain coordinates, as predicted by Pfam, from the *Saccharomyces* Genome Database (SGD) (date of access: February, 2019). For each gene in each isolate, we concatenate the sequences encoding Pfam-defined protein domains with the sequence belonging to the 3'coding region (i.e. the region downstream of the 3'-most domain-encoding sequence and upstream of the translation stop codon). This is the "D+3'" mRNA region. For the "L+5'" mRNA region, we concatenate the sequences encoding any inter-domain linkers with the sequence of the 5' coding region (i.e. the region downstream the start codon and upstream of the 5'most domain-encoding sequence). Using Python (version 3.7.1), we then compute tAI, our chosen measure of codon bias, for D+3' and L+5' regions.

## Approximating Global mRNA Folding Stability

Throughout this study, we fold entire mRNA sequences, where "entire" is defined as the CDS flanked by the 5' and 3' UTRs. Because translating ribosomes linearize segments of the overall mRNA structure, previous studies have used sliding windows to approximate co-translational folding stability. A key question arises here: how likely are nucleotides to pair with their nearest (within ~29 nucleotides) neighbors? What about if the mRNA is circularized or if the ribosome densities are low? According to Shabalina et al, 2006, mfe structures commonly include base-pairings between nucleotides far apart along the linear mRNA sequence. For these reasons, and because sliding windows smooth any peaks and dips in local folding stabilities, we choose to fold mRNAs as single units.

Three gauges of mRNA folding stability are mean base-pair probability, minimum free energy (mfe) deltaG, and thermodynamic ensemble deltaG. All are predicted for entire mRNA transcripts (at 30°C) with the RNAfold algorithm (version 2.4.14) from the ViennaRNA Package. We summarize each measure below.

i. Mean base-pair probability is the arithmetic mean of nucleotide pairing probabilities. One such pairing probability represents the chance that a given nucleotide will base-pair with another nucleotide, given its mRNA's weighted set of thermodynamic ensemble configurations. It is calculated via the partition function (see McCaskill, 1990). A mean base-pair probability near 1 suggests that the mRNA's nucleotides are very likely to be base-paired and that its folded form is highly structured and very stable.

ii. Minimum free energy (mfe) deltaG represents the change in Gibbs free energy an mRNA experiences after folding into its most energetically stable (mfe) configuration, as predicted by RNAfold. A negative deltaG value of large magnitude indicates spontaneous formation of a highly stable structure.

iii. Ensemble deltaG is a Boltzmann-weighted sum of deltaG values; one deltaG value per mRNA structure in the mRNA's thermodynamic ensemble. Because mfe structure is a best-guess prediction and because mRNA folding is far from static (Crothers et al, 1974), ensemble deltaG is expected to be a more accurate measure of mRNA folding strength.

While RNAfold is a reasonable and widely-used approximation tool (see Zhou et al, 2018), it is worth mentioning that *in silico* folding is an incomplete representation of *in vivo* folding: it does not account for the actions of RNA binding proteins, translating ribosomes, mRNA modifications, tertiary/quaternary structures, and the specifics (excluding temperature) of the surrounding environment. We lose accuracy with this approach, but we do maintain the breadth necessary for a large-scale systematic study.

## Defining mRNA Folding Stability on a Local Scale

With the RNAeval tool from the ViennaRNA Package (version 2.4.14), we obtain a detailed thermodynamic description of each gene's mfe mRNA structure at 30°C. Specifically, the algorithm reports a deltaG approximation for all substructures that fully describe an mRNA's overall folding shape: multi

loops, external loops, interior loops, and hairpin loops. To compute the deltaG of an mRNA region (e.g. the coding sequence), we first sum the deltaGs of all substructures completely enclosed within it. Then, for any partially enclosed substructure, we 1) calculate what fraction of the substructure is built by nucleotides from our region, 2) multiply this value by the substructure's deltaG, and 3) add the result to our existing sum.

## Gene Criteria and GO Term Enrichment Analyses

Limitations in the availability of data require us to compute our various models with different sets of genes across isolates: 1620 genes (each with nonsynonymous and/or synonymous polymorphisms) and 185 genes (each with synonymous polymorphisms only) for global codon bias, 1092 genes for local codon bias, 1458 genes for global mRNA folding stability, 779 genes for local mRNA folding stability, and 1447 genes for our study probing the in-tandem actions of both bias and folding stability. Here, we explain how these gene sets were selected and we summarize the results of their Gene Ontology (GO) term enrichment analyses (see **Figure S1**).

i. Of the 1636 genes with mRNA and protein abundance data across isolates, 1620 show one or more SNPs across isolates. Of these, 185 show only synonymous SNPs. To obtain the latter information, we translate the 1636 coding sequences from each isolate via the translate tool from the SeqIO Biopython package (Cock et al, 2009). For each gene, we then align the corresponding set of amino acid sequences (one sequence from each isolate) via the MUltiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm (Edgar, 2004). Those genes with 100% amino acid identity scores and SNP(s) across isolates are used in our 185-gene analyses. When considering model results based on this smaller set of genes, we are able to discount any effects amino acid substitutions may have on translation rates.

ii. In the models pertaining to local codon bias, we consider a 1092-gene subset of the 1620 genes defined above. Each gene belonging to this subset is characterized by an absence of premature stop codons,

available protein domain location data from Pfam, and SNP(s) in both the D+3' sequence (concatenated domain-encoding and 3'coding sequence) and the L+5' sequence (concatenated linker-encoding and 5'coding sequence).

iii. We use 1458 of 1636 genes in our models of global folding stability. These genes have available length data for the 5'UTR and the 3'UTR, and they have one or more SNPs in their concatenated 5'UTR, CDS, and 3'UTR sequences.

iv. To arrive at a subset of genes suitable for local folding stability models, we filter the 1458-gene set defined above by the following criteria: genes must have variation (across isolates) in local mfe deltaG within the 5'UTR, the CDS, the 3'UTR, *+1 to +10* from the 5'cap, *-9 to +3* from translation start, *+4 to +10* of translation start, and *+1 to +18* from translation stop. Based on the sizes of these local mRNA regions, additional criteria are for 5'UTRs to be at least 19 nucleotides in length and 3'UTRs to be at least one nucleotide in length.

v. The intersection of the 1620-gene set and the 1458-gene set defines the 1447-gene set we use when analyzing the synchronous actions of codon bias and folding stability. We rank these 1447 genes by their median tAI across isolates, choose the bottom 723 genes as our 'low tAI' group, the middle 723 genes as our 'medium tAI' group, and the top 724 genes as our 'high tAI' group. This process is then repeated for mfe deltaG in place of tAI.

With few exceptions, our GO-term enrichment analyses show that genes in every set are most enriched for GO-terms related to 1) general metabolism, 2) nucleotide synthesis and metabolism (purine's especially), 3) peptide biosynthesis and metabolism, 4) amino acid synthesis and metabolism, 5) ATP metabolism, and 6) translation. This result was not unexpected since all isolates were grown in nutrient rich

broth at a steady-state temperature of 30°C, all were sampled at log-phase growth, and mass spectrometry most reliably detects highly expressed proteins.

## The Linear Mixed Effects Regression Model

The alleles of any given gene show variable codon biases and variable mRNA folding stabilities; how well does this explain the among-allele differences in logPPR? To investigate this question, we build several linear mixed effects regression models (or simply, "mixed models") that for each gene across isolates, predict logPPR as a function of a single explanatory variable; these explanatory variables include our approximations of global codon bias, local codon bias, global folding stability, and local folding stability. Each mixed model also produces a comprehensive linear relationship ("fixed effects slope") which summarizes the individual gene-specific relationships.

Rather than incorporating all explanatory variables into a single mixed model predictive of logPPR, we compute several mixed models, each with a different explanatory variable as the sole predictor of logPPR. We take this approach because our measures of global/local codon bias and global/local folding stability are calculated from the same gene sequences, and are thus not fully independent of one another. Variances differ considerably among our explanatory variables, so we caution against drawing too many conclusions when comparing fixed effects slopes.

We compute each mixed model in R (version 3.6.0) via the lmer function from the lme4 package (version 1.1.21; Bates et al, 2015). We use the default settings for lmer and we fit all models with the maximum likelihood method. "Gene" is always included as our "random" effect and we allow it to have both random slopes and random intercepts; this enables us to quantify gene-level relationships between the explanatory variable and logPPR since the random effects assign a unique slope and a unique intercept to each gene.

The statistical significance of an explanatory variable infers influence on protein synthesis rates. We assess the predictive ability of an explanatory variable via the log-likelihood ratio test; our significance threshold is $p = 0.05$ and we have 1 degree of freedom.

# RESULTS

## Modeling the Effects of *Cis*-Acting Genetic Variation on Protein Synthesis Rates

Studies in genetically-engineered systems have revealed that single nucleotide polymorphisms (SNPs) tune the rates of protein synthesis by acting on codon bias and mRNA folding stability. Our tool for exploring these patterns in a natural system is the linear mixed effects regression model. We begin our analysis by estimating each gene's codon bias and folding stability on global and local scales (see Methods). We then use the mixed effects model to capture how gene-specific variation in each estimate affects the rate at which proteins are made (as approximated by the natural log of the steady-state ratio of protein per mRNA, or logPPR). The result yields, for each gene, a slope between our measure of interest and logPPR. Their average is the fixed effects slope and it provides insight on logPPR's response to changes (due to SNPs) in the explanatory variable. Due to limitations in the availability of data, our different models use different numbers of genes across isolates: 1620 for global codon bias, 1092 for local codon bias, 1458 for global folding stability, 779 for local folding stability, and 1447 for our study probing the synchronous actions of bias and folding (see Methods).

## Selection of a Training Set for Codon Bias Reveals that Moderately GC-Rich Synonymous Codons Have the Highest Supplies of Cognate tRNAs and are the Most Preferred by Highly Expressed Genes

The original measure of codon bias, the Codon Adaptation Index (CAI), compares the synonymous codon usage of a given gene with that of a training set of genes (see Methods). The consequences of choosing this training set are not well understood, though the two major approaches involve selecting a rather arbitrary number of highly expressed genes or computing the set with the Carbone et al, 2003 algorithm. The latter method is a more structured approach; it seeks to capture the dominant genomic-level driver of codon bias, which in yeast and in several other fast-growing organisms, strongly resembles the bias of the most highly expressed genes (Carbone et al, 2003; Sharp et al, 1988). While we ultimately chose the 61 genes identified by this algorithmic approach (see Supplementary Table 9 in Carbone et al, 2003),

we initially investigated the former method by quantifying how codon preferences vary with gene expression. To do so, we first rank 6179 genes by their median transcript abundance across isolates. Then, for each isolate, we compute several CAI codon tables based on training sets containing 6179 genes and then the top $2^i$ (where $i \in [1, 12]$) genes from our ranked list. Each CAI codon table value represents a codon's prevalence among genes in the training set, and is always relative to the prevalence of its most abundant synonymous sibling. For all codon tables computed with the same number of training set genes, a median codon table is computed, yielding a total of 13 tables (see **Figure 6**).

In CAI training sets with 512 or fewer highly expressed genes, we see a distinct bias for synonymous codons inferred to be quickly translated (see **Figure 6**); this inference is based on tAI[1] codon table values, which are themselves based on cognate/near-congate tRNA gene copy numbers (Tuller et al, 2010; dos Reis et al, 2004). The above pattern is true for every amino acid except cysteine and glycine, suggesting that codon translation rates are major drivers of the unique codon bias present in the most highly expressed genes.

When the number of training set genes exceeds 2500, relative codon preferences stabilize and the level of bias for high-tAI codons wanes. Preference for AT-rich synonymous codons rises and by the 6179-gene training set, higher codon values directly relate to higher codon AT-content (for all amino acids but leucine). Further, for all amino acids with less than six synonymous codons (except proline), the order of synonymous codon preferences consistently depends on the codon's third nucleotide position; those codons ending in a thymine (T) are most-preferred, followed by those ending in an adenine (A), in a cytosine (C), and finally, in a guanine (G). It appears that the bias of more lowly expressed genes is shaped by neutral mutation, which in yeast, is biased towards A's and T's.

[1] tAI is the tRNA Adaptation Index and we discuss it in detail in later sections. We mention it here because it is a commonly used approximation for codon translation rates in yeast.

Interestingly, we find that the usage of GC-rich synonymous codons increases as more genes enter the training set. Could this alleviate some pressure on tRNA pools?, i.e. might the high-tAI codons be reserved predominantly for highly expressed genes, while the lower-tAI codons (especially if AT-rich, but even if GC-rich) be reserved for more lowly expressed genes? This hypothesis would be consistent with the **Figure 6** results.

As training sets increase in size, we also see a spike in the usage of 100%AT-rich codons. This pattern does not appear with 100%GC-rich codons, no matter how large or small their corresponding tAI values are. Usage of 100%GC-rich codons remains low across all training sets. For eleven amino acids (alanine, arginine, cytosine, glutamine, glutamic acid, glycine, leucine, methionine, proline, serine, and tryptophan), the most preferred synonymous codon remains unchanged across the various training sets.

Varying the number of highly expressed genes in our CAI training set varies codon table values, sometimes dramatically, and almost always based on inferred codon translation rates (when all training set genes are highly expressed) and on codon AT-content (when more training set genes have low to medium expression). As mentioned in the Introduction, there is a balance between neutral mutation and natural selection in the shaping of a gene's codon bias, and from **Figure 6** we see that for highly expressed genes, this balance tips in favor of natural selection, and for more lowly expressed genes, it tips in favor of neutral mutation. Based on these results, it is not clear how many genes we should hand-pick for our CAI training set. Thus we use the Carbone et al, 2003 set selection approach for our CAI calculations.

## Higher logPPR Associates with SNPs that Align Gene Codon Bias with the Dominant Genomic-Level Driver of Codon Bias

Our first mixed effects regression models show that slight shifts in CAI among alleles predict logPPR at a statistically significant level. Specifically, among the set of 1620 genes, and more importantly,

among the set of 185 genes with no nonsynonymous SNPs, the slope of CAI vs. logPPR is 0.90 (G = 73, df = 1, $p = 1.3\times10^{-17}$), and 0.47 (G = 11, df = 1, $p = 7.7\times10^{-4}$), respectively (see **Figure 7**). What does this mean in terms of protein expression regulation? The positive signs of both slopes suggest that faster speeds of translation characterize alleles with codon biases in line with the dominant genomic-level driver of codon bias. As evidenced by the 185-gene-set results, this result is not an artifact of amino acid substitutions due to nonsynonymous SNPs.

We next wanted to confirm that the Carbone et al, 2003 training set is itself unbiased in its representation of codon bias. Longer training set genes contribute more codons to the codon pool from which the CAI codon table (the foundation for all CAI calculations) is calculated, but because all genes have their own intrinsic biases (see Quax et al, 2015), these large contributions may misrepresent the dominant genomic-level driver of codon bias. As such, we length-normalize training set genes so that each contributes an equal portion to the codon pool (see Methods). We then re-calculate a new version of CAI, length-normalized CAI (nlCAI), based on this new codon table. We see that nlCAI predicts logPPR about as well as CAI does: for the 185-gene-set, the fixed effects slope for nlCAI vs. logPPR is 0.46 (G = 11, df = 1, $p = 9.7\times10^{-4}$) (see **Figure 7**). Based on these similarities, we conclude that it is not necessary to length-normalize the codon counts of Carbone et al, 2003 training set genes.

## Higher logPPR Associates with SNPs that Specifically Increase Gene Codon Bias for Synonymous Codons with High Estimated tRNA Supplies

Quick translation is a unifying feature of the synonymous codons most favored by the dominant genomic-level driver of codon bias. Through the tRNA Adaptation Index (tAI) (dos Reis et al, 2004) we can directly capture the degree of each gene's bias for codons inferred to be quickly translated; this inference is based on each codon's cognate and near-cognate tRNA gene copy numbers. How does tAI compare to CAI? Among the set of 185 genes, tAI explains more variation in logPPR; the slope of tAI vs. logPPR is 1.62 (G = 19, df = 1, $p = 1.6\times10^{-5}$) (see **Figure 7**). For the 1620-gene set, the tAI vs. logPPR

fixed effects slope is 1.82 (G = 96, df = 1, $p = 1.4\times10^{-22}$). Thus, faster translation is specifically associated with polymorphisms that increase a gene's usage of codons matching to highly abundant tRNAs. For reference, if for each gene in the 185-gene-set, we compute the standard deviation of tAI across isolates, and then take the median of these values across genes, we obtain a value 0.0010. Such a change would correspond to an estimated 0.162% increase in PPR, which in the 185-gene-set, ranges from 411 to 36,100 protein molecules per mRNA. This PPR-range is 320 to 86,500 for genes in the 1620-gene-set.

Adaptations of codons to the tRNA pool can be estimated by just tRNA supply, as mentioned above, or they can be estimated by both tRNA supply *and demand*. This brings us to a fourth and final measure of bias: the normalized tRNA Adaptation Index (ntAI). With this measure, we consider the quantity of tRNAs (as measured by tRNA gene copy number and wobble) as well as the quantity of codons competing for them (as measured by the sum of codon translation frequencies across the transcriptome) (Pechmann & Frydman, 2012). From this view, a codon optimal for fast translation is one whose tRNA species are high in abundance and low in demand. Compared to CAI and tAI, ntAI is a relatively poor predictor of logPPR; for the 185-gene-set, the ntAI vs. logPPR slope is 1.6 (G = 6, df = 1, $p = 1.3\times10^{-2}$) (see **Figure 7**). Thus, a greater adaptation of codon usage to both the supply *and demand* on the tRNA pool does not appear to be a leading factor in determining protein synthesis rates. We proceed with tAI for all further codon bias calculations.

## Higher logPPR Associates with SNPs that Increase tAI Mainly in Domain-Encoding and 3'Coding mRNA Regions

Domain-encoding and 3'coding mRNA regions show a disproportionally high composition of quickly and accurately translated synonymous codons. Do rates of protein synthesis rise when SNPs further increase their combined bias? From our mixed effects model, we see that within concatenated domain-encoding and 3'coding sequences (D+3'), higher tAI is strongly associated with higher logPPR: the fixed effects slope of D+3' tAI vs. logPPR is 1.83 (G = 64, df = 1, $p = 1.6\times10^{-15}$) (see **Figure 8**). This means that

if for each gene, we compute the standard deviation of D+3' tAI across isolates, and then take the median of these values across genes, we obtain a value of 0.0011, and this would correspond to an expected increase of 0.2% in protein per mRNA levels.

The disproportionally low appearance of quickly-translated synonymous codons in the 5'coding and linker-encoding mRNA regions suggests a strong selection pressure for their low codon bias. We therefore expected to see a relatively small rise (or even a fall) in protein synthesis rates following SNPs that increase the tAI of these concatenated (L+5') regions. While this is the relationship our mixed effects model shows, it is not at a statistically significant level. The fixed effects slope of L+5' tAI vs. logPPR is 0.173 (G = 3.3, df = 1, $p$ = 0.069) (see **Figure 8**), meaning that a 0.0017 unit increase in L+5' tAI corresponds to an expected 0.03% increase in protein per mRNA. Here, 0.0017 represents the median (across genes) standard deviation in L+5' tAI (across the 22 isolates). The locations of SNPs appear to strongly dictate their influence on translation rates via local codon bias.

## Higher logPPR Associates with SNPs that Stabilize Global mRNA Folding

Some SNPs strengthen the folding patterns of their mRNAs, and a growing body of evidence has linked this action to faster rates of translation. It has been suggested that stable mRNA structures could block mRNA aggregation, optimize the distance between translating ribosomes, and promote mRNA circularization (see Introduction). To investigate this connection as a naturally-occurring means of protein expression regulation, and to investigate the predictive ability of three commonly-used measures of mRNA folding stability (minimum free energy (mfe) deltaG, ensemble deltaG, and mean base-pair probability – see Methods), we compute three mixed effects models, each with a different measure as a predictor of logPPR.

Across 1458 genes in the genome, all three models show that added folding stability (due to SNPs) associates with increasing logPPR. Two of our three measures, mfe deltaG and ensemble deltaG (both in

units of Mcal/mol) are very strong predictors of logPPR: the mfe deltaG vs. logPPR fixed effects slope is -0.442 ($G = 45$, df = 1, $p = 2.5\times10^{-11}$) and the ensemble deltaG vs. logPPR fixed effects slope is -0.479 ($G = 52$, df = 1, $p = 6.0\times10^{-13}$) (see **Figure 9**). Note that the negative slopes here arise because increasing stability is associated with decreasing deltaG. The fixed effects slope of mean base-pair probability (calculated across the whole transcript) vs. logPPR is 0.280 ($G = 8.3$, df = 1, $p = 4.1\times10^{-3}$) (see **Figure 9**). Of all measures, ensemble deltaG explains the most variation in logPPR. To put these results into context, when for each gene, we compute the standard deviation of ensemble deltaG across the 22 isolates, and then take the median of these values across genes, we arrive at 0.0029 Mcal/mol. Such an increase in ensemble deltaG corresponds to weaker folding and is expected to decrease PPR by 0.14%.

## Changes in logPPR Associate with SNPs that Modify Local mRNA Folding Stability

Unexplained patterns of weak mRNA folding recur *+1 to +10* bases of the 5'cap, *-9 to +3* bases of translation start, and *+1 to +18* bases of translation stop. Why do these regions remain relatively unstructured? Some recent hypotheses suggest that respectively, this quality may facilitate binding of the 40S ribosomal subunit (at least in mammals), promote the successful recognition of translation start, and prevent steric hindrance during translation termination (see Introduction). Can SNPs promote rates of protein synthesis by further weakening the folded structures in these regions? Would this effect overpower the pressure for increased global folding stability?

Just as the above three regions show particularly low folding stability, the coding sequence (**CDS**) and the mRNA region ***+4 to +10*** bases of translation start show consistently high folding stabilities[2]. Hypotheses from the literature suggest that strong structures in the former shorten the distance between ribosomes, while in the latter, create a type of 'speed bump' for the ribosome as it passes over AUG (see

[2] By printing these regions' names in **bold**, we draw attention to their inherently high folding stabilities.

Introduction). Will logPPR increase when SNPs further stabilize these regions' folding? We summarize the results of our investigations in **Figure 10**.

For each of the following three regions of mRNA, *-9 to +3* bases from start, ***+4 to +10*** bases from start, and *+1 to +18* bases from stop, the fixed effects slope of deltaG vs. logPPR is positive: 0.195 (G = 0.84, df = 1, $p$ = 0.36), 0.147 (G = 4.2, df = 1, $p$ = 0.040), and 0.545 (G = 0.53, df = 1, $p$ = 0.47), respectively. From this result, we see that the weaker folding (due to SNPs) of all three regions leads to faster translation, though this relationship is only statistically significant (and counter to expectation) for the span of mRNA ***+4 to +10*** bases from start.

For the **CDS** and for the region *+1 to +10* bases from the 5'cap, the deltaG vs. logPPR fixed effects slope is negative: -0.140 (G = 4.9, df = 1, $p$ = 0.027) and -0.578 (G = 3.4, df = 1, $p$ = 0.066), respectively. Stronger folding (due to SNPs) within the **CDS** seems to accelerate translation, and notably, this is also the case with folding *+1 to +10* of the 5'cap (though not with any statistical significance).

For completion, we report the near-zero slopes of 5'UTR deltaG vs. logPPR and then 3'UTR deltaG vs. logPPR: -0.0299 (G = 0.075, df = 1, $p$ = 0.78) and -0.0116 (G = 0.0074, df = 1, $p$ = 0.93). SNP-caused changes in the folding stabilities of either region have no predicted effects on protein synthesis rates. In all cases, we have approximated regional deltaG via the thermodynamic description of the mRNA's overall mfe structure (see Methods).

## Global Codon Bias and Global mRNA Folding Stability Tune logPPR in Tandem

Mao et al, 2014 propose that as the distance between ribosomes decreases due to more stable structures, codon bias plays an ever-larger role in determining final translation elongation rates. Likewise, we predict that the relationship between tAI and logPPR is highest in genes whose alleles have the strongest

mRNA folding. This turns out to be the case: strong structures amplify the effects bias has on logPPR (see **Figure 11A**). More specifically, when we 1) rank 1447 genes by their median ensemble deltaG among the 22 isolates, 2) split this ranking into three partially overlapping groups of 723 (or 724) genes of low, medium, and high deltaG, and then 3) for each group, compute a tAI vs. logPPR model, we see that especially in genes with highly folded mRNAs (low deltaG), genetic variation leading to increased bias also leads to an increased logPPR. For instance, the **tAI vs. logPPR** slope is 2.81 (G = 86, df = 1, $p = 1.7\times10^{-20}$) for the half of genes with *low deltaG* and 1.00 (G = 20, df = 1, $p = 8.2\times10^{-6}$) for the half of genes with *high deltaG*.

For each above-defined group of low, medium, and high ensemble deltaG genes, we compute a second linear mixed effects regression model, this time with deltaG vs. logPPR (see **Figure 11C**). We find that increases in the mRNA folding stability of low deltaG (high stability) genes is mildly associated with changes in logPPR, but not in a statistically significant way: the **ensemble deltaG vs. logPPR** slope for *low deltaG* genes is -0.085 (G = 0.80, df = 1, $p = 0.37$). In genes whose mRNAs have weak or medium folding stabilities, decreases in deltaG (increases in stability) greatly magnify logPPR: the **ensemble deltaG vs. logPPR** slope for *high deltaG* genes is -0.87 (G = 23, df = 1, $p = 2.1\times10^{-6}$) (see **Figure 11C**). This would be expected under the Mao et al, 2014 model.

To investigate the reciprocal case, we repeat these tAI vs. logPPR and deltaG vs. logPPR calculations, but this time with genes ranked by their median tAI value across alleles. In the 723 genes with the lowest bias, increases in tAI have the most pronounced and positive effects on logPPR: the **tAI vs. logPPR** slope is 1.02 (G = 12, df = 1, $p = 4.9\times10^{-4}$) for the half of genes with *low tAI* and it is 0.61 (G = 6.8, df = 1, $p = 9.3\times10^{-3}$) for the half of genes with *high tAI* (see **Figure 11B**). Then, in high-tAI genes, decreases in deltaG (increases in stability) have the strongest positive effects on logPPR: the **ensemble deltaG vs. logPPR** slope is -0.22 (G = 14, df = 1, $p = 1.6\times10^{-4}$) for the half of genes with *low tAI* and it is -0.67 (G = 37, df = 1, $p = 1.5\times10^{-9}$) for the half of genes with *high tAI* (see **Figure 11D**). These results are

not due to an underlying correlation between tAI and deltaG, as shown by **Figure 11E** and examined in the Discussion.

All of the above relationships can be summarized in the following two statements: (1) codon bias or folding stability will be more strongly associated with logPPR when the *other* factor is high, and (2) codon bias or folding stability will be more strongly associated with logPPR when that factor is low (see **Figure 12** for a summary).

## DISCUSSION

A wide array of mechanisms contributes to gene expression regulation, and because gene expression forms such a critical link between genes and traits, these mechanisms have long been topics of experimental and computational studies (Stern & Orgogozo, 2008). Most of the recent work has focused on the determinants of mRNA abundances, and while these mRNA levels do explain some aspects of protein levels, inconsistencies between the two remain. Given this observation, and given that an individual's traits are largely determined by its protein content, we look to the 'post-transcriptional' layer of gene expression regulation to help bridge some of the wide gaps in the currently known web of connections between genes and traits. Specifically, we seek a better understanding of what makes mRNA levels frequently non-indicative of protein levels, how genetic variation influences post-transcriptional regulation, and importantly, whether the known post-transcriptional regulatory patterns apply to all genes in the genome.

At least in the lab setting, artificial shifts in gene codon bias and in mRNA folding stability appear to modify rates of protein synthesis; their effects have been characterized by mutagenizing genes (e.g. Cuperus et al, 2017; Kudla et al, 2009), comparing gene sequences across the genome (e.g. Kertesz et al, 2010; Shabalina et al, 2006), and perturbing unique patterns common to certain gene regions (e.g. Lamping et al, 2013; Babendur et al, 2006). Our study is the first to examine these observed relationships both systematically in over 1000 genes in the genome and in a natural (non-genetically-engineered) setting. In this way, we have begun to understand systems more fully and directly extend lab findings to the natural world. The specific question guiding our research is: do SNP-caused changes in codon bias and in folding stability explain any of the translation rate variation among 22 yeast isolates? We quantified these effects individually for each gene across isolates, and we discovered evidence that supports codon bias and folding stability as two *cis*-acting regulators of protein expression.

Our models show that a gene's codon bias and its mRNA folding stability act on a genomic scale and in tandem to tune the gene's rate of protein synthesis (which we approximate as the logarithm of protein per mRNA level (logPPR)). Specifically, an increase in mRNA folding strength across genes coupled with an increase in codon bias across isolates, produces an increase in logPPR. We see that the specific locations of SNPs strongly influence their effects on logPPR via bias, e.g. SNPs located in the protein-domain-encoding and 3'coding mRNA regions have about 10 times more influence than SNPs located elsewhere along the mRNA. Further, we discovered a second way in which bias and folding stability act in tandem, and it mirrors the first method: an increase in bias across genes coupled with an increase in folding stability across isolates associates with higher levels of logPPR. In general, SNP-caused changes in global folding stabilities strongly relate to the observed variations in logPPR, though we do see mild evidence suggesting that a gene's logPPR is also promoted by weakened folding in the regions surrounding the start codon and downstream of the stop codon, and by strengthened folding near the 5'cap (not the case in mammals) and in the CDS. All of these results confirm and quantify on a large scale and in a natural system what has previously been hypothesized from modeling work and from experimental work in genetically-engineered systems, and it suggests that codon bias and mRNA folding stability are two *cis*-acting forms of protein expression regulation.

It is worth noting that we consider only two components of the additive genetic variation. This, of course, excludes other *cis*-acting components (e.g. ORF length and uORFs: see Dana & Tuller, 2012 and Meijer & Thomas, 2002), *trans*-acting components (e.g. RNA binding proteins and microRNAs: see Moore & von Lindern, 2018 and Oliveto et al, 2017), as well as the interactions between them. Therefore, even though our two components are highly significant, they explain a small fraction of the total variation in logPPR.

## Implications of Changing Local Codon Bias at Key mRNA Locations

A SNP in one region of mRNA affects the bias of just that region, so in the interests of statistical power, for each gene, we combine linker-encoding with 5'coding sequences (L+5'), and we combine domain-encoding with 3'coding sequences (D+3'). While this does make our local tAI vs. logPPR models more reliable, it also means that we cannot separate the individual effects (on a gene's logPPR) of SNP-caused bias changes in its linker-encoding, 5'coding, domain-encoding, and 3'coding regions. We recommend this as an avenue for future research.

Because of the low and statistically insignificant slope of the L+5' tAI vs. logPPR model, we can conclude that SNP-caused perturbations in L+5' bias do not consistently associate with the observed variations in logPPR. We attempt to explain this result with two speculations: 1) the bias of linker-encoding sequences should slow the ribosomes enough to facilitate proper co-translational protein folding, but not so much that the ribosomes collide with one another, and 2) the bias of 5'coding sequences must be sufficient to re-space ribosomes, but not so low that the distance between them becomes larger than necessary (since high ribosome density is associated with faster protein synthesis). Conjectures aside, this L+5'tAI result coupled with the statistically significant and positive slope of the D+3' vs. logPPR model, suggests that SNP-caused increases in a gene's D+3' tAI affect its logPPR because of the SNPs' specific association with the D+3' region (rather than because they increase the gene's global codon bias).

## Implications of Changing Local Folding Stabilities at Key mRNA Locations

In general, we see from our model that changes (due to SNPs) in folding stability near the 5'cap do not associate with the observed variations in logPPR at a statistically significant level. What could explain this outcome? Across our 779-gene set, local folding stabilities *+1 to +10* bases of the 5'cap remain fairly weak (between -12kcal/mol and +2kcal/mol), while the synthetic hairpins introduced to this region in lab experiments show considerably more stability (between -10kcal/mol and -50kcal/mol) (Babendure et al,

2006). Perhaps in our yeast isolates, folding stabilities near the 5'caps are sufficiently mild and the perturbations in their folding stabilities sufficiently small, that translation initiation remains unhindered. The negative fixed effects slope of our 5'cap deltaG vs. logPPR model (see Results), almost suggests that a certain stability might even be required in this region. This would be consistent with recent results that show genome-wide tendencies for particularly strong structures near yeast 5'caps (Kertesz et al, 2010). We propose that structure of a certain stability is necessary, but when this structure becomes sufficiently strong, it begins to interfere with translation initiation. A future study could repeat the approach outlined by Babendure et al, 2006, but with several weakly stable (> -10kcal/mol) hairpin structures. Alternatively, we also consider the possibility that increases in the folding stability near the 5' cap contribute to increases in global folding stability, and these increases in global folding stability, according to our other models, have positive effects on logPPR. In any case, at least in our yeast, folding near the 5'cap does not appear to be a potent means for protein expression regulation on the genomic scale.

Zooming in on the region *-9 to +3* bases from the translation start codon, we find that increases in local folding stability accompany decreases in logPPR (as expected), though not at a statistically significant level. What explains the lack in significance? Our measured stabilities are certainly on par with those seen in the Lamping et al, 2013 lab study (see Introduction), so there is no issue of overly-stable artificial structures creating a signal not present in the natural setting. Perhaps the span of nucleotides incorporated into artificial vs. natural structures is responsible: in the artificial case, hairpins form from the base-pairing of a contiguous stretch of mRNA, while in the natural case, nucleotides within *-9 to +3* bases of the start codon generally pair with those further away in the 5'UTR or even more likely, in the CDS (Shabalina et al, 2006). This area of research would benefit from studies that engineer sequences so as to facilitate base-pairing between this *-9 to +3* region and several randomized locations throughout the mRNA sequence.

One of our models suggests that lower logPPR follows increases in the folding stability *+4 to +10* bases from the translation start codon. Both Kertesz et al, 2010 and Shabalina et al, 2006 see a spike in the

folding stability of this region, and this spike is conserved across the genome; could it be a way of preserving weak folding just upstream of the start codon? We considered the possibility, but found no correlation between the median deltaG *+4 to +10* of translation start and the median deltaG *-9 to +3* of translation start (where median deltaG is calculated for each gene across isolates). Higher stability is conserved in this region, though slight decreases in its strength are, for reasons yet to be discovered, weakly associated with increases in logPPR.

Like the result for 5'cap deltaG, SNP-induced changes in folding stability *+1 to +18* bases from the translation stop codon do not predict logPPR levels with any statistical significance (see Results). Here too, the absence of sufficiently stable structures may explain why our small SNP-caused changes in local folding stability show no association with the observed changes in logPPR.

Lastly, we consider how the folding stabilities of the 5'UTR, the CDS, and the 3'UTR could tune logPPR in unison. Of these three regions, the deltaG of the CDS is the only statistically significant predictor of logPPR, though its fixed effects slope of -0.140 is surprisingly small in magnitude compared to that of the global mfe deltaG vs. logPPR model: -0.34 ($G = 17$, $df = 1$, $p = 3.9\times10^{-5}$; computed on the same 779-gene set). If UTR folding stabilities really have no predicted effect on logPPR, why do we see this discrepancy in fixed effects slopes between the CDS deltaG vs. logPPR and the global mfe deltaG vs. logPPR models? To begin our inquiry, we consider how well CDS deltaG correlates with global mfe deltaG, and we find that for 555 of our 779 genes (71%), the Spearman's rank correlation coefficient (across isolates) is statistically significant (*p*-value threshold: 0.05), though not always positive (22 genes show negative correlations) or close to 1. We conclude that the folding stabilities of both UTRs must play some role. We explore this possibility with two additional correlations: for each of 293 genes in our 779-gene set (38%), the across-isolate Spearman's rank correlations between 5'UTR and CDS deltaG, as well as between 3'UTR and CDS deltaG are statistically significant (*p*-value threshold: 0.05). Most frequently, SNPs that *increase* folding stability within the CDS *decrease* folding stability in both UTRs (165 of 293 genes).

Almost never do we see SNPs have the same effects on the folding stability of all three regions (4 of 293 genes). In the second-most common case, SNPs *increase* folding stability in both the CDS and a single UTR (124 of 293 genes). Could this be evidence for some type of hidden relationship between the folding of all three regions?

## Mechanistic Speculations on the In-Tandem Effects of Codon Bias and Folding Stability

Though they appear to work together, in our set of 1447 genes, changes in global codon bias and changes in global folding stability are not related in any consistent way. For instance, the Spearman's rank correlation coefficient between tAI and ensemble deltaG is statistically significant (*p*-value threshold: 0.05) for 665 of 1447 genes (46%), and of these 665, 283 (43%) show a negative correlation while 382 (57%) show a positive correlation. These correlation coefficients are independent of both the median tAI value and the median ensemble deltaG value of each gene across isolates. Further, there is essentially no correlation between the median gene tAI and the median gene ensemble deltaG across isolates: the Spearman's rank coefficient is 0.09, $p = 7.1\times10^{-4}$. Thus, we have no reason to suspect that the effects of codon bias and folding stability confound each other, on the contrary, they act together to tune protein synthesis rates. Below, we discuss the results of **Figure 11** in detail and propose several mechanistic hypotheses that explain the associations we see.

Mao et al, 2014 propose that as the distance between ribosomes decreases due to more stable structures, codon bias plays an ever-larger role in determining final translation elongation rates. Likewise, we see that the relationship between tAI and logPPR is highest in genes whose alleles have the strongest mRNA folding stabilities (see **Figure 11A**). Mechanistically, we speculate that stable structures create clusters, or trains, of ribosomes: once the first ribosome unwinds a structure, those that have accumulated behind it merely follow and translate the newly linearized sequence. If large stretches of the mRNA become linearized in this way, stable structures are no longer the elongation-rate-limiting factors, and because of

this, tAI is able to gain a more prominent influence on the rate of protein synthesis. In the context of less stable mRNAs, ribosomes could no longer have the opportunity to form long trains. In such a scenario, each might need to unravel every folded structure it encounters. The effects of high unravelling times may overshadow the effects of any changes in tAI, explaining the small slope of tAI vs. logPPR in genes with the weakly folded mRNAs.

We see that the relationship between tAI and logPPR is highest in genes with low tAI, and lowest in genes with high tAI (see **Figure 11B**). Even though the overlap of 95% confidence intervals hints at the possibility of no meaningful difference in these fixed effect slopes, we still wonder: under what circumstances could the magnitude of a gene's existing tAI dictate the effects of its SNP-caused tAI changes on logPPR? A simple explanation is that a given variation in bias represents a larger percent difference for low tAI genes than for high tAI genes.

In parallel, with the same tAI-ranking scheme as mentioned above, we observe that the deltaG vs. logPPR association is strongest in genes with high tAI (see **Figure 11D)**. We predict that the increases in mRNA folding stabilities better cluster ribosomes into trains, at which point the effects (on logPPR) of the underlying high tAI may strengthen. Conversely, in genes with low tAI, an increase in folding stability (decrease in deltaG) is not associated with a change in logPPR. Could this mean that ribosomes still cluster into trains, but move so slowly (due to low tAI) that logPPR is unaffected by any changes in deltaG? Or perhaps, maybe the clustered ribosomes translate slowly enough that they collide with one another?

Lastly, we consider why the magnitude of the deltaG vs. logPPR fixed effects slope increases between models comprising of low, medium, and high deltaG genes (where low deltaG indicates high folding stability) (see **Figure 11C**). One explanation could involve a threshold above which additional stability (and therefore additional ribosome density) no longer affects translation rates: in mRNAs with folding stabilities above this threshold, the ribosomes could already be sufficiently clustered so that the

mRNAs stay somewhat linearized during translation. Alternatively, folding of a certain stability may already be enough to preclude mRNA aggregation or facilitate mRNA circularization. Any further increases make little to no difference. Distinguishing amongst these mechanisms will require further investigation and we present this as an exciting avenue for future research.

## Validating Our Global Codon Bias Results in the Lab

The distinct advantages of small-scale vs. large-scale and experimental vs. computational methods can be leveraged to produce a more complete answer to a common question. As such, we plan to validate our global codon bias results in the lab. Specifically, we will compute tAI for each gene in two yeast strains: S288c and YJM145. *Will genes with larger differences in tAI between strains also show larger differences in logPPR?* This would be consistent with our model. To find out, we will focus on a subset of genes with the most variable tAI between S288c and YJM145. To remove the potentially confounding effects of *trans*-acting factors, for each gene in the subset, we will measure logPPR of the S288c allele, swap the S288c allele for the YJM145 allele, and then measure logPPR of the YJM145 allele in the S288c background.

## Validating Our Global Folding Stability Results in the Lab

From our models, we see that alleles with higher global folding stabilities tend to have higher levels of logPPR. While we are currently unable to determine the exact method by which this occurs, in the near future, we plan to explore the three recently proposed mechanistic hypotheses: the Mao et al, 2014 "Ribosome Train" hypothesis, the Zur & Tuller, 2012 "mRNA Aggregation" hypothesis, and the Lai et al, 2018 "mRNA Circularization" hypothesis. Our ideas are summarized below. All three hypotheses are discussed in detail in the Introduction.

**Proposed Ribosome Train Investigation #1:** If trains of ribosomes exist, a light endonuclease digestion should reveal their presence; "light" here is a function of the digest's duration and the

endonuclease's concentration, both of which will need to be optimized. For reference, an endonuclease is an enzyme that when introduced into the cell, cleaves the bonds between mRNA nucleotides. Translating ribosomes however, protect the ~30 nucleotides of mRNA they overlap, so by digesting mRNA with this enzyme, we can detect the instantaneous location, or "footprint", of each ribosome scanning it (see Ingolia et al, 2009). An mRNA transcript (or an mRNA region) with larger numbers of these footprints is inferred to have higher ribosome density. What if, in our light digestion, we saw footprints much longer than 30 nucleotides? Such a result could be indicative of ribosome trains: cuts in-between the closely-spaced ribosomes of a train are (presumably) far less likely than cuts elsewhere along the mRNA, especially in a light digest. Rather than being a single ribosome's footprint, this longer footprint would be that of a ribosome train's.

Another question that may be asked is: do ribosome footprints tend to be larger near the stop codon rather than near the start codon? If so, it could suggest that downstream mRNA structures combine short trains of ribosomes into longer trains.

**Proposed Ribosome Train Investigation #2:** With already-available sequence and ribosome footprinting data for two closely-related yeasts: *S. cerevisiae* and *S. paradoxus* (see McManus et al, 2014), we will model how an allele's folding stability relates to its ribosome density. If dense trains of ribosomes are present, we would expect to see a positive association between a gene's folding stability, as approximated by its ensemble deltaG, and its ribosome density, as approximated by the gene's number of ribosome footprints.

**Proposed mRNA Circularization Investigation:** To determine whether highly stable mRNA structures facilitate mRNA circularization, we will first compute the ensemble deltaG of all genes in S288c and all genes in YJM145. We will then select a few genes with the largest between-strain deltaG differences for further FLIM-FRET analyses (where *FLIM* is Fluorescence Lifetime Imaging Microscopy, and *FRET* is Förster Resonance Energy Transfer; see Periasamy et al, 2015). To exclude any confounding effects from

*trans*-acting factors, we plan to study the S288c allele normally, then swap the S288c allele with the YJM145 allele, and then study the YJM145 allele in the S288c background. In accordance with the FRET method, the 5' and 3' ends of each allele's mRNA will be labelled with a FRET pair: an excited donor fluorophore and a non-excited acceptor fluorophore. When the acceptor is in close enough proximity (on the order of several nanometers) to the donor, it draws on some of the donor's stored energy, causing a measurable decrease in donor luminescence as well as a measurable increase in acceptor luminescence; this change is a function of the distance between fluorophores, though it is difficult to measure accurately (Orthaus et al, 2009). However, as soon as the passage of energy occurs, the average length of time that the donor glows (its "fluorescence lifetime") measurably decreases ("quenches"). If to compare the fluorescence lifetime of a donor on its own with that of a donor in the presence of an acceptor, we can estimate the distance between donor and acceptor fluorophores (i.e. the distance between 5' and 3' mRNA ends) (Pelet et al, 2006). Such an analysis could be done in live cells. By this approach, we will answer: is donor lifetime lower for the alleles with lower ensemble deltaG (higher folding stability)? If so, it would suggest that more stable mRNA folding allows for a shorter distance between 5' and 3' mRNA termini.

**Proposed mRNA Aggregation Investigation:** We plan to use the MS2-GFP system (see Querido & Chartrand et al, 2008) as our tool for visualizing mRNA aggregation in living cells. The coat protein of the MS2 RNA bacteriophage allows it to both enclose the phage's genome in a capsid and to halt production of its replicase protein. This halting mechanism involves the binding of coat protein dimers to an "operator" hairpin structure which overlaps the replicase-encoding gene (Peabody, 1993). In the lab, we can attach GFP (green fluorescent protein) to this MS2 coat protein, and then engineer our mRNAs such that they have an operator hairpin that binds MS2-GFP (Querido et al, 2011). The glow of the GFP tag will allow us to track mRNAs and quantify their propensity for aggregation. It will allow us to determine if mRNAs with high ensemble deltaG values aggregate the most.

## WORKS CITED


Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., & Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature, 506*(7489), 494-497.

Angov. E. Hillier, C. J., Kincaid, R. L., & Lyon J. A. (2008). Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PloS One, 3*, e2189.

Babendure, J. R., Babendure, J. L., Ding, J. H., & Tsien, R. Y. (2006). Control of mammalian translation by mRNA structure near caps. *RNA, 12*(5), 851–861.

Baim, S. B., & Sherman, F. (1988). mRNA structures influencing translation in the yeast Saccharomyces cerevisiae. Molecular and Cellular Biology, 8(4), 1591–1601.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, NY), 296*(5568), 752–755.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., & Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular Cell*, *61*(3), 341-351.

Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., & Gileadi, O. (2008). Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expression and Purification, 59*(1), 94-102.

Carbone, A., Zinovyev, A., & Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics, 19*(16), 2005-2015.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Jr, Shapiro, M. D., Brady, S. D., … Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science, 327*(5963), 302–305.

Chartier, M., Gaudreault, F., & Najmanovich, R. (2012). Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, *28*(11), 1438-45.

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M. F., & von der Haar, T. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. *The EMBO Journal, 33*(1), 21–34.

Cock, P. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422-1423.

Crick, F. H. (1966). Codon–anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology, 19*(2), 548–555.

Crothers, D. M., Cole, P. E., Hilbers, C. W., & Shulman, R. G. (1974). The molecular mechanism of thermal unfolding of Escherichia coli formylmethionine transfer RNA. *Journal of Molecular Biology, 87*(1), 63–88.

Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., Fields, S., & Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Research, 27*(12), 2015–2024.

Dana, A., & Tuller, T. (2012). Efficient manipulations of synonymous mutations for controlling translation rate: an analytical approach. *Journal of Computational Biology, 19*, 200-231.

Dana, A., & Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research, 42*(14), 9171–9181.

Doma, M. K., & Parker, R. (2006). Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature, 440*(7083), 561–564.

dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research, 32*(17), 5036-5044.

Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics, 10*(10), 715–724.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Foss E. J., Radulovic, D., Shaffer, S. A., Goodlett, D. R., Kruglyak, L., & Bedalov, A. (2011). Genetic Variation Shapes Protein Networks Mainly through Non-transcriptional Mechanisms. *PLOS Biology 9*(9): e1001144.

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y. M., & Jensen, L. J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Molecular Systems Biology, 8*, 572.

Gebert, D., Jehn, J., & Rosenkranz, D. (2019). Widespread selection for extremely high and low levels of secondary structure in coding sequences across all domains of life. *Open Biology, 9*(5), 190020.

Geiler-Samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M., Hartl, D. L., & Drummond, D. A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America, 108*(2), 680–685.

Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, N. S., Christensen, L. L., Borre, M., Sorensen, K. D., Andersen, L. D., Andersen, C. L., Hulleman, E., Wurdinger, T., Ralfkiaer, E., Helin, K., Gronbaek, K., Orntoft, T., Waszak, S., Dahan, O., Pedersen, J. S., Lund, A. H., & Pilpel, Y. (2014). A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell, 158*(6), 1281-1292.

Gooch, V. D., Mehra, A., Larrondo, L. F., Fox, J., Touroutoutoudis, M., Loros, J. J., & Dunlap, J. C. (2008). Fully codon-optimized luciferase uncovers novel temperature characteristic of the Neurospora clock. *Eukaryotic Cell, 7*(1), 28-37.

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

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology, 19*(3), 1720–1730.

Hanson, G., & Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature Review Molecular Cell Biology, 19*(1), 20-30.

Hershberg, R. and Petrov, D. A. (2008) Selection on codon bias. *Annual Review of Genetics, 42*, 287–299

Hofacker I. L., Fontana W., Stadler P. F., Bonhoeffer S., Tacker M., & Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte fur Chemie, 125*, 167–188.

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. *Journal of Molecular Biology, 158*(4), 573-597.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science, 324*(5924), 218–223.

Katz, L., & Burge, C. B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research, 13*(9), 2042–2051.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., & Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature, 467*(7311), 103–107.

Konczal J., Bower J., Gray C. H. (2019) Re-introducing non-optimal synonymous codons into codon-optimized constructs enhances soluble recovery of recombinant proteins from *Escherichia coli*. *PLoS One, 14*(4), e0215892.

Kozak M. (1986). Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences of the United States of* America, 83(9), 2850–2854.

Kozak, M. (1990). Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences of the United States of America, 87*, 8301-8305.

Kramer, E. B., & Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA, 13*(1), 87–96.

Kramer, E. B., Vallabhaneni, H., Mayer, L. M., & Farabaugh, P. J. (2010). A comprehensive analysis of translational missense errors in the yeast *Saccharomyces cerevisiae*. *RNA, 16*(9), 1797–1808.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science, 324*(5924), 255-258.

Lai, W. C., Kayedkhordeh, M., Cornell, E. V., Farah, E., Bellaousov, S., Rietmeijer, R., Salsi, E., Mathews, D. H., & Ermolenko, D. N. (2018). mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. *Nature Communications, 9*(1), 4328.

Lamping, E., Niimi, M., & Cannon, R. D. (2013). Small, synthetic, GC-rich mRNA stem-loop modules 5' proximal to the AUG start-codon predictably tune gene expression in yeast. *Microbial Cell Factories, 12*, 74.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology, 6*(26).

Lu, P., Vogel, C. Wang, R., Yao, X., & Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology, 25*, 117-124.

Makhoul, C. & Trifonov, E. N. (2002). Distribution of Rare Triplets Along mRNA and Their Relation to Protein Folding. Journal of Biomolecular Structure and Dynamics, 20(3), 413-420;

Mao, Y., Liu, H., Liu, Y., & Tao, S. (2014). Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae*. *Nucleic Acids Research, 42*(8), 4813–4822.

McCaskill, J. S. (1990). The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers, 29*, 1105–1119.

McManus, C. J., May, G. E., Spealman, P., & Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research, 24*(3), 422-430.

Meijer, H. A., & Thomas, A. A. (2002). Control of eukaryotic protein synthesis by upstream open reading frames in the 5'untralsated region of an mRNA. *The Biochemical Journal, 367*(1), 1-11.

Miura, F., Kawaguchi, N., Yoshida, M., Uematsu, C., Kito, K., Sakaki, Y., & Ito, T. (2008). Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics, 9*, 574.

Moore, K. S., & von Lindern, M. (2018). RNA Binding Proteins and Regulation of mRNA Translation in Erythropoiesis, *Frontiers in Physiology, 9*, 910.

Morgan, L. W., Greene, A. V., & Bell-Pedersen, D. (2003) Circadian and light-induced expression of luciferase in Neurospora crassa. *Fungal Genetics Biology, 38*(3), 327-332.

Napthine, S., Yek, C., Powell, M. L., Brown, T. D., & Brierley, I. (2012). Characterization of the stop codon readthrough signal of Colorado tick fever virus segment 9 RNA. *RNA, 18*(2), 241–252.

Niepel, M., Ling, J., & Gallie, D. R. (1999). Secondary structure in the 5'-leader or 3'-untranslated region reduces protein yield but does not affect the functional interaction between the 5'-cap and the poly(A) tail. *FEBS Letters, 462*(1-2), 79-84.

Oliveto, S., Mancino, M., Manfrini, N., & Biffo, S. (2017). Role of microRNAs in translation regulation and cancer. *World Journal of Biological Chemistry, 8*(1), 45–56.

Orthaus, S., Buschmann, V., Bülter, A., Fore, S., König, M., & Erdmann, R. (2009). Quantitative in vivo imaging of molecular distances using FLIM-FRET. *Application Note.* https://www.picoquant.com/images/uploads/page/files/7267/appnote_flim_fret.pdf

Paek, K. Y., Hong, K. Y., Ryu, I., Park, S. M., Keum, S. J., Kwon, O. S., & Jang, S. K. (2015). Translation initiation mediated by RNA looping. *Proceedings of the National Academy of Sciences of the United States of America, 112*(4), 1041–1046.

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J. B., … Gilad, Y. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genetics, 8*(10), e1003000.

Park, C., Chen, X., Yang, J. R., & Zhang, J. (2013). Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America, 110*(8), E678–E686.

Parts, L., Liu, Y. C., Tekkedil, M. M., Steinmetz, L. M., Caudy, A. A., Fraser, A. G., Boone, C., Andrews, B. J., & Rosebrock, A. P. (2014). Heritability and genetic basis of protein level variation in an outbred population. *Genome Research, 24*(8), 1363–1370.

Peabody, D. S. (1993). The RNA binding site of bacteriophage MS2 coat protein. *The EMBO Journal, 12*(2), 595-600.

Pechmann, S., & Frydman, J. (2012). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology, 20*(2), 237-43.

Pelet, S. B., Previte, M. J. R., & So, P. T. C. (2006). Comparing the quantification of Förster resonance energy transfer measurement accuracies based on intensity, spectral, and lifetime imaging. *Journal of Biomedical Optics, 11*(3), 034017.

Periasamy, A., Mazumder, N., Sun, Y., Christopher, K. G., & Day, R. N. (2015). FRET Microscopy: Basics, Issues and Advantages of FLIM-FRET Imaging. *Advanced Time-Correlated Single Photon Counting Applications – Springer Series in Chemical Physics, 111*, 249-279.

Pollard, D. A., Asamoto, C. K., Rahnamoun, H., Abendroth, A. S., Lee, S. R., & Rifkin, S. A. (2016). Natural Genetic Variation Modifies Gene Expression Dynamics at the Protein Level During Pheromone Response in Saccharomyces cerevisiae. bioRxiv, 090480.

Plotkin, J. B., & Kudla, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics, 12*(1), 32-42.

Python Software Foundation. Python Language Reference, version 3.7.1. Available at http://www.python.org

Quax, T. E., Claassens, N. J., Soll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell, 59*(2), 149–161.

Querido, E., & Chartrand, P. (2008). Using Fluorescent Proteins to Study mRNA Trafficking in Living Cells. *Methods in Cell Biology, 85*, 273-292.

Querido, E., Gallardo, F., Beaudoin, M., Menard, C., & Chartrand, P. (2011). Stochastic and reversible aggregation of mRNA with expanded CUG-triplet repeats. *Journal of Cell Science, 124*(10), 1703-1714,

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics, 7*(11), 862–872.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Skelly, D. A., Ronald, J., & Akey, J. M. (2009). Inherited Variation in Gene Expression. *Annual Review of Genomics and Human Genetics, 10*, 313-332.

Skelly, D. A., Merrihew, G. E., Riffle, M., Connelly, C. F., Kerr, E. O., Johansson, M., Jaschob, D., Graczyk, B., Shulman, N. J., Wakefield, J., Cooper, S. J., Fields, S., Noble, W. S., Muller, E. G., Davis, T. N., Dunham, M. J., Maccoss, M. J., & Akey, J. M. (2013). Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Research, 23*(9), 1496-504.

Shabalina, S. A., Ogurtsov, A. Y., & Spiridonov, N. A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research, 34*(8), 2428–2437.

Sharp, P. M., Tuohy, T. M., & Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research, 14*(13), 5125-5143.

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution, 24*(1-2), 28–38.

Sharp, P. M., & Li, W. H. (1987). The Codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research, 15*(3), 1281-1295.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., & Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research*, *16*(17), 8207–8211.

Stern, D. L., & Orgogozo, V. (2008). The Loci of Evolution: How Predictable is Genetic Evolution? *International Journal of Organic Evolution, 62*(9), 2155-2177.

Straub, L. (2011). Beyond the Transcripts: What Controls Protein Variation? *PLoS Biol 9*(9): e1001146.

Takyar, S., Hickerson, R. P., & Noller, H. F. (2005). mRNA Helicase Activity of the Ribosome. *Cell, 120*(1), 49-58.

Thanaraj, T. A., & Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. *Protein Science: A Publication of the Protein Society, 5*(8), 1594-612.

Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., & Babu, M. M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling, 11*, eaat6409

Tuller, T., Ruppin, E., & Kupiec, M. (2009). Properties of untranslated regions of the *S. cerevisiae* genome. *BMC Genomics, 10*, 391. https://doi.org/10.1186/1471-2164-10-391

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., & Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell, 141*(2), 344–354.

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

Trotta E. (2013). Selection on codon bias in yeast: a transcriptional hypothesis. *Nucleic Acids Research, 41*(20), 9382-95.

The UniProt Consortium, UniProt: a worldwide hub of protein knowledge (2019). *Nucleic Acids Research, 47*(D1), D506–D515.

Vega Laso, M. R., Zhu, D., Sagliocco, F., Brown, A. J., Tuite, M. F., & McCarthy, J. E. (1993). Inhibition of translational initiation in the yeast *Saccharomyces cerevisiae* as a function of the stability and position of hairpin structures in the mRNA leader. Journal of Biological Chemistry, 268(9), 6453–6462.

Ventoso, I., Sanz, M. A., Molina, S., Berlanga, J. J., Carrasco, L., & Esteban, M. (2006). Translational resistance of late alphavirus mRNA to eIF2α phosphorylation: a strategy to overcome the antiviral effect of protein kinase PKR. *Genes & Development, 20*(1), 87-100.

von der Haar T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Systems Biology, 2*, 87.

Wallace, E. W., Airoldi, E. M., & Drummond, D. A. (2013). Estimating selection on synonymous codon usage from noisy experimental data. *Molecular Biology and Evolution, 30*(6), 1438–1453.

Wang, Y., Harvay, C. B., Pratt, W.S., Sams, V., Sarner, M., Rossi, M., Auricchio, S., & Swallow, D. (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Human Molecular Genetics, 4*(4), 657-662.

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports, 14*(7), 1787–1799.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., & Johnson, C. H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature, 495*(7439), 116–120.

Yu, C. H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., & Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell, 59*(5), 744–754.

Zhou, T., Weems, M., & Wilke, C. O. (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution, 26*(7), 1571-80.

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

Zhao, Y., Wang, J., Zeng, C., & Xiao, Y. (2018). Evaluation of RNA secondary structure prediction for both base-pairing and topology. *Biophysics Reports*, *4*, 123–132.

Zhu, Y. O., Siegal, M. L., Hall, D. W., & Petrov, D. A. (2014). Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences of the United States of America, (22)*111, E2310-E2318.

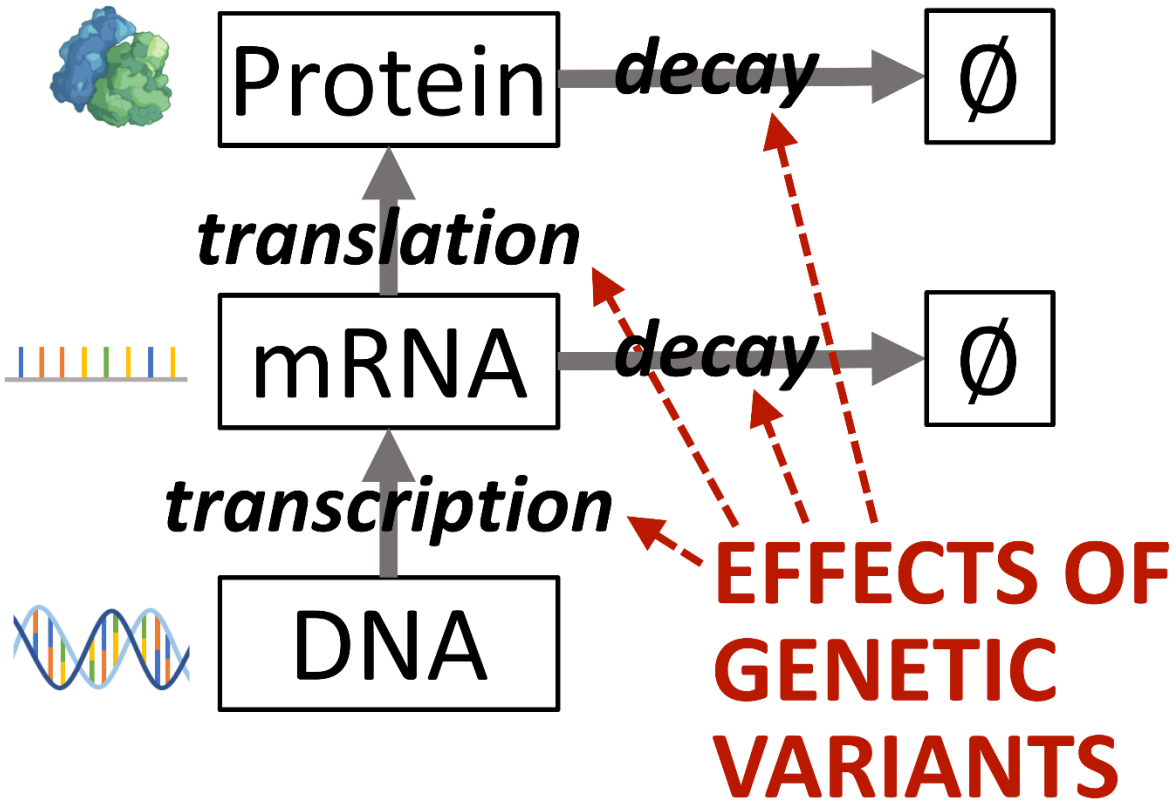



**Figure 1.** Genetic variation can affect rate of transcription, mRNA decay, translation, and protein decay, leading to variation in protein expression and ultimately, to variation in traits.

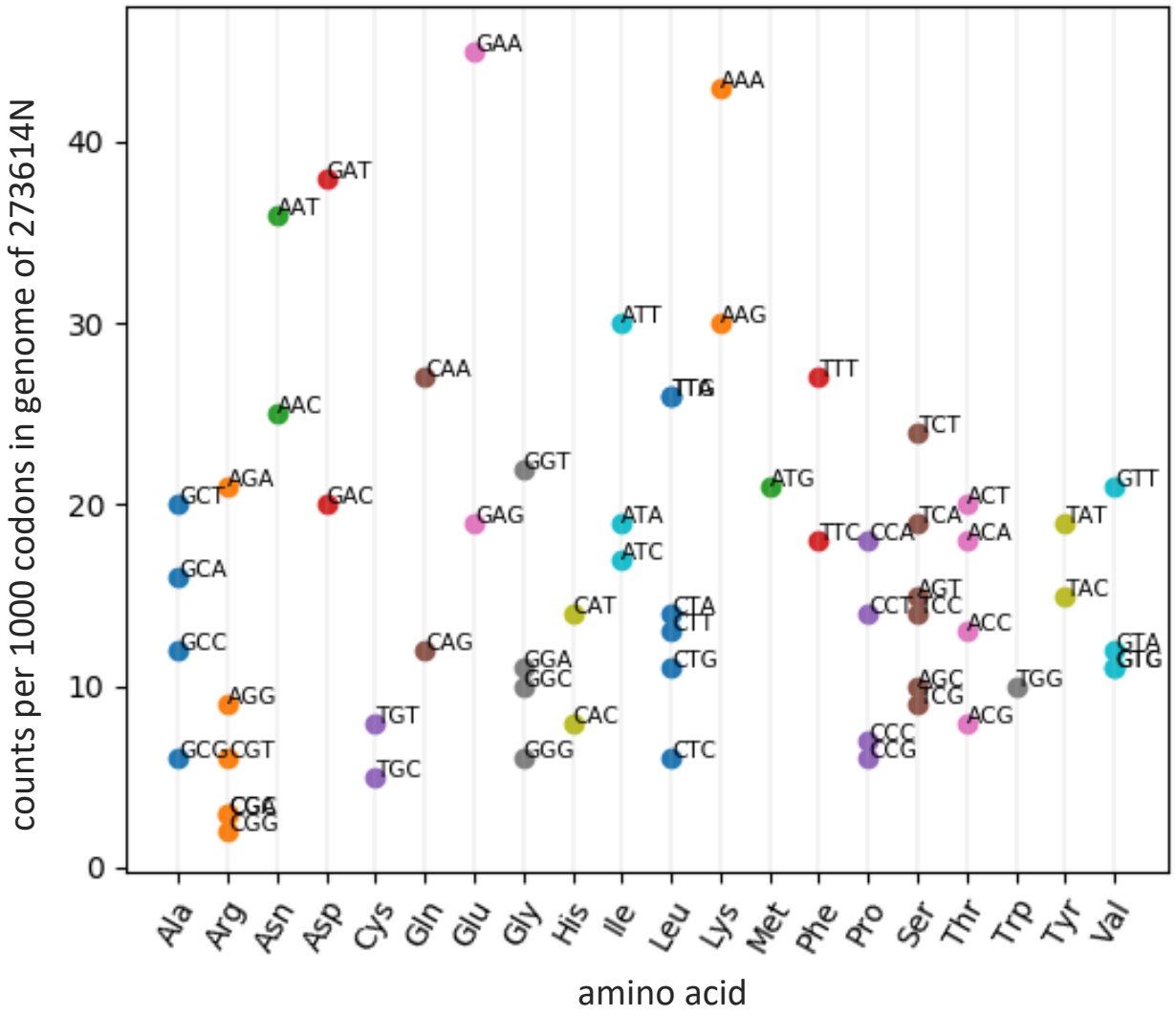



**Figure 2.** Frequency of codon usage per 1000 codons in the genome of yeast isolate 273614N. Synonymous codons encoding the same amino acid lie on the vertical gridline corresponding to that amino acid. Each gene has its own level of codon bias, and this level is shaped by natural selection and neutral mutation.

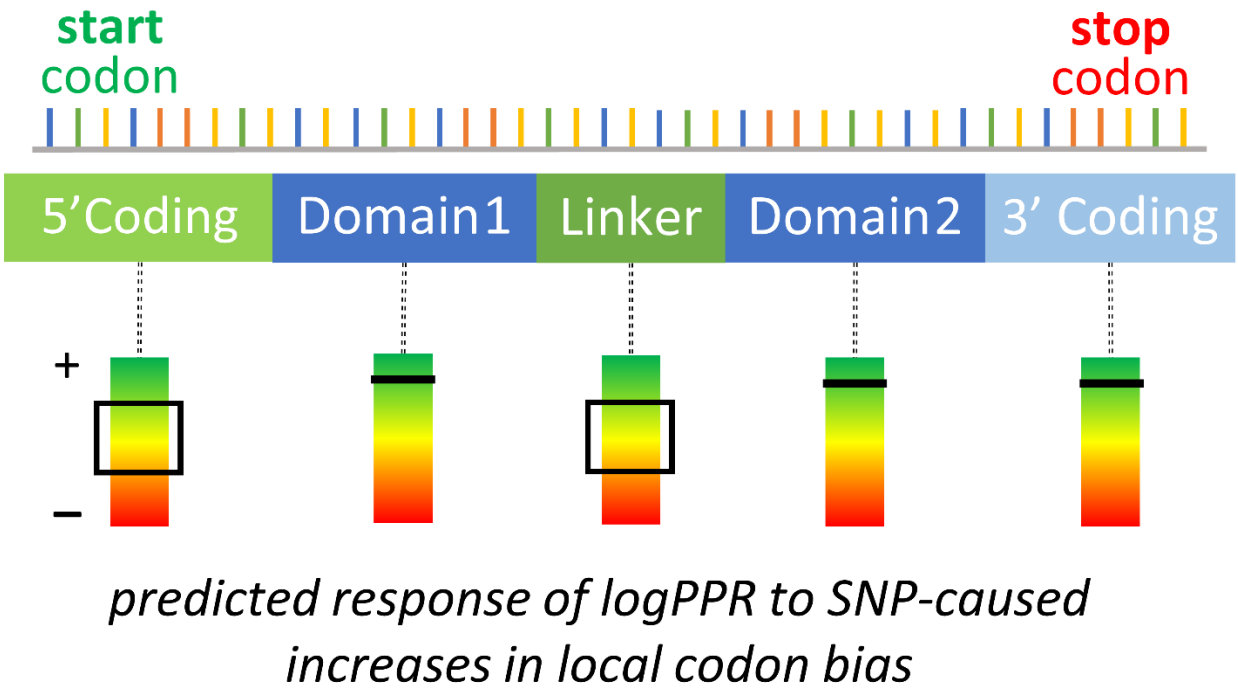



**Figure 3.** Schematic representation of the mRNA protein-encoding sequence for a two-domain protein. We predict that higher logPPR (log of protein per mRNA level) is associated with alleles whose domain-encoding and 3’coding regions show higher codon bias. We predict that the same should also be true regarding linker-encoding and 5’coding regions, though to a lesser (and potentially negative) extent since these regions require a certain degree of slowed translation, as explained in the main text.

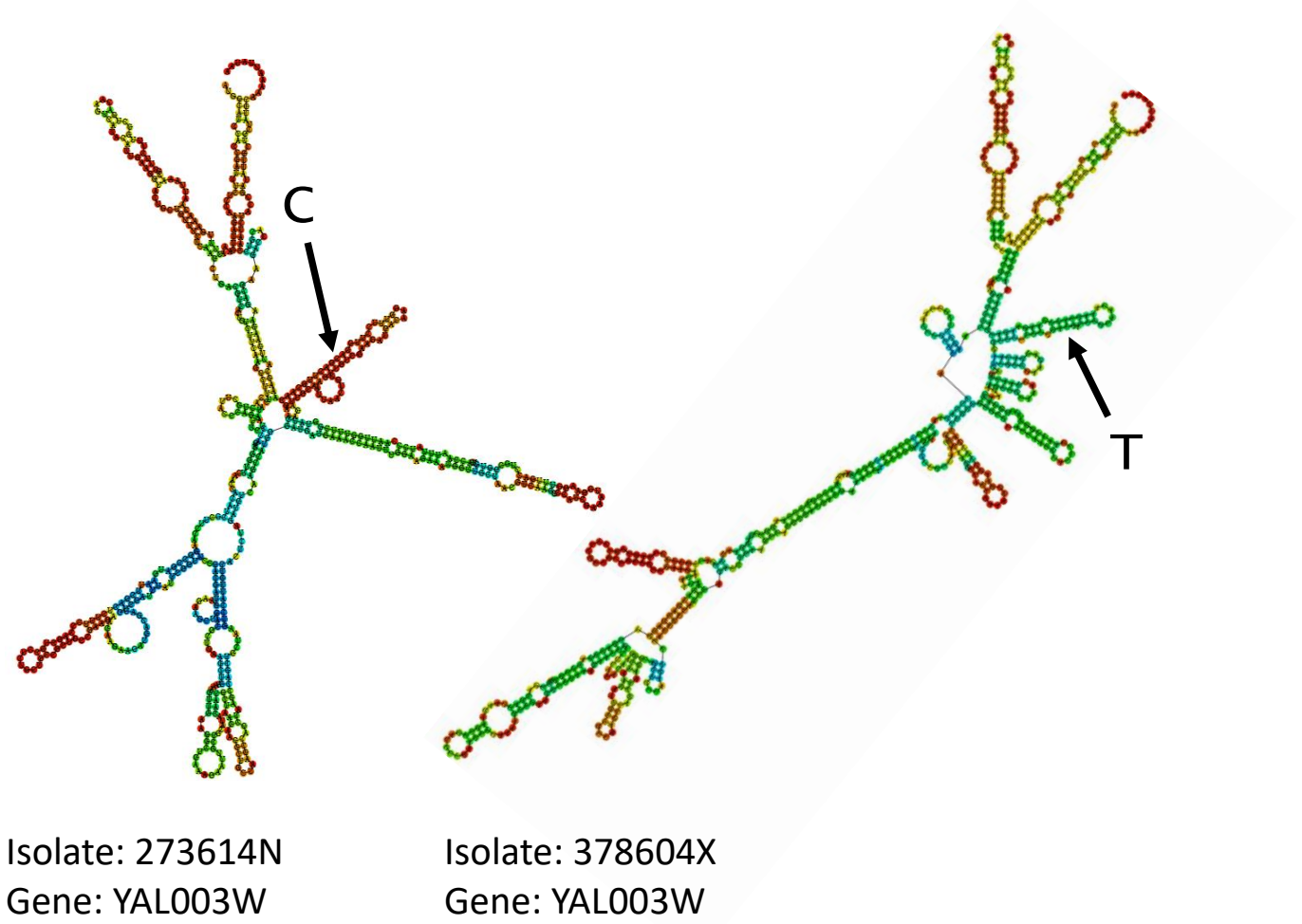



**Figure 4.** The estimated minimum free energy (most thermodynamically stable) folding structure of the YAL003W transcript in two isolates of yeast. A nucleotide's color represents its probability of being in the shown configuration (paired or unpaired) across all Boltzmann-weighted structures in the thermodynamic ensemble. A SNP at position 559 (cytosine in 273614N and thymine in 378604X) results in the structural differences shown above. This figure was generated via the ViennaRNA Package RNAfold Web Server (version 2.4.14.).

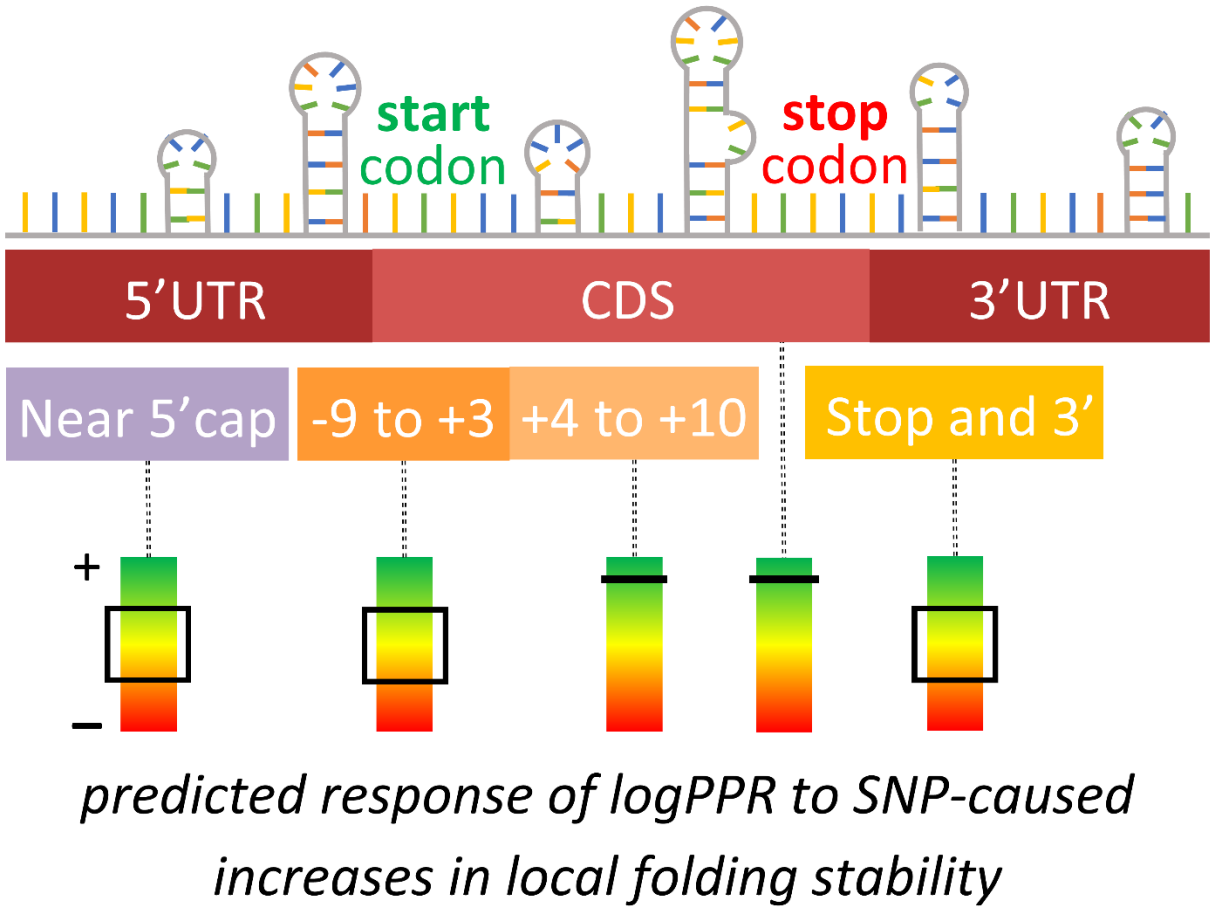
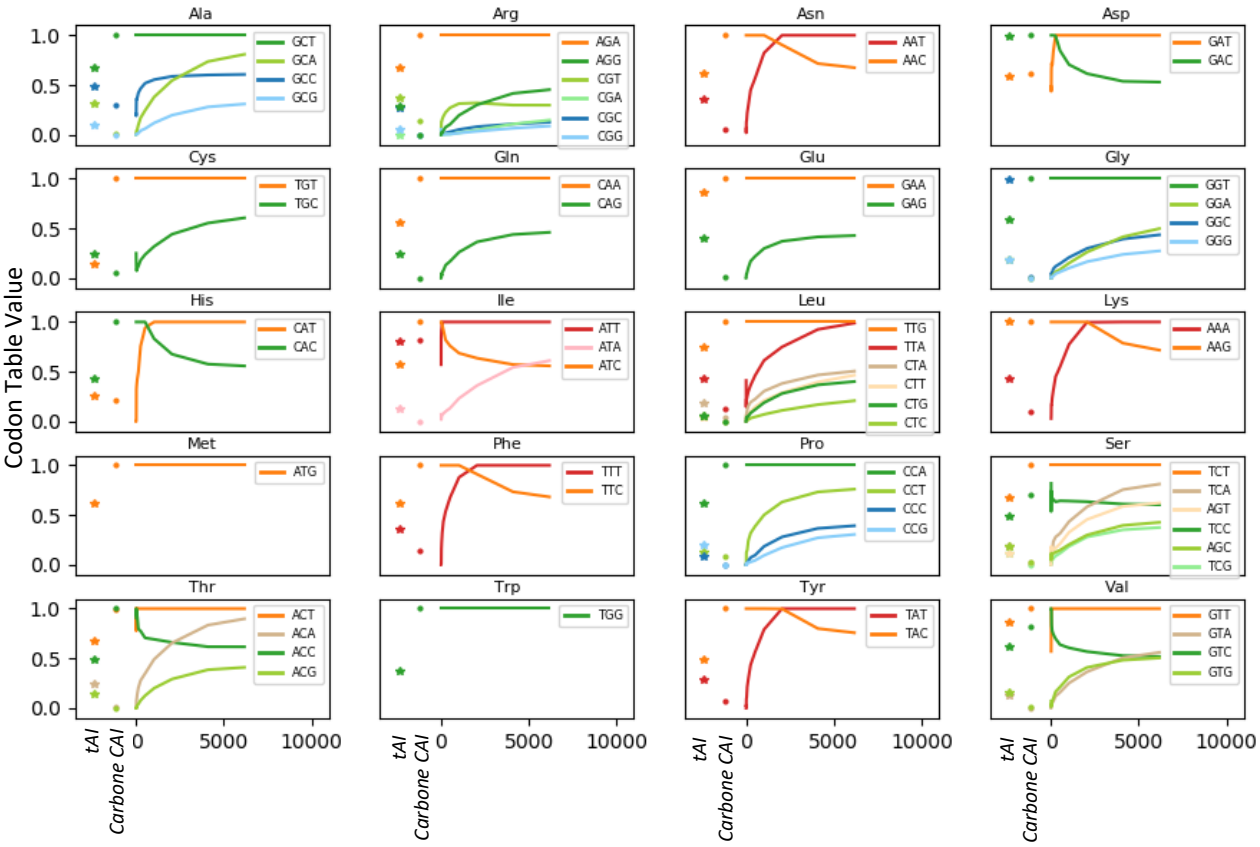



**Figure 5.** Schematic representation of a structured mRNA transcript. We predict that higher logPPR (logarithm of protein per mRNA level) is associated with alleles showing higher folding stability in the CDS and in the sequence *+4 to +10* bases from the start codon. Since we expect a positive association between logPPR and global folding stability, and since overly-strong structures in key places frequently lower the rate of protein synthesis, we expect a weakly positive (or negative) relationship between logPPR and the changing local bias of the following three regions: *+1 to +10* bases from the 5'cap, *-9 to +3* bases from the start codon, and *+1 to +18* bases from the stop codon, as explained in the main text.




In the Training Set: Number of Genes with the Highest Median Transcript Level Across our 22 Yeast Isolates

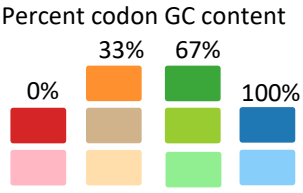



**Figure 6.** Lineplot showing how CAI codon table values change in response to the number of high expressing genes in the CAI training set. Datapoints are taken for training sets containing 6179 or $2^i$ (where $i \in [1, 12]$) of the most highly expressed genes (as ranked by median transcript abundance across our 22 yeast isolates). For each amino acid, the most frequently used synonymous codon among training set genes has a value of 1. A sibling synonymous codon appearing 60% as often would have a value of 0.6. Each subplot corresponds to an amino acid (specified by the subplot title) and its synonymous codons (color-coded inset). Codon color is based on %GC content (see legend). Synonymous codons with the same %GC content are assigned a color within the same color group – red (0%GC), orange (33%GC), green (67%GC), and blue (100%GC). For reference, we also present each codon's Carbone et al, 2003 CAI codon table value as well as each codon's tAI codon table value (this is a commonly used approximation for codon translation rates in yeast).

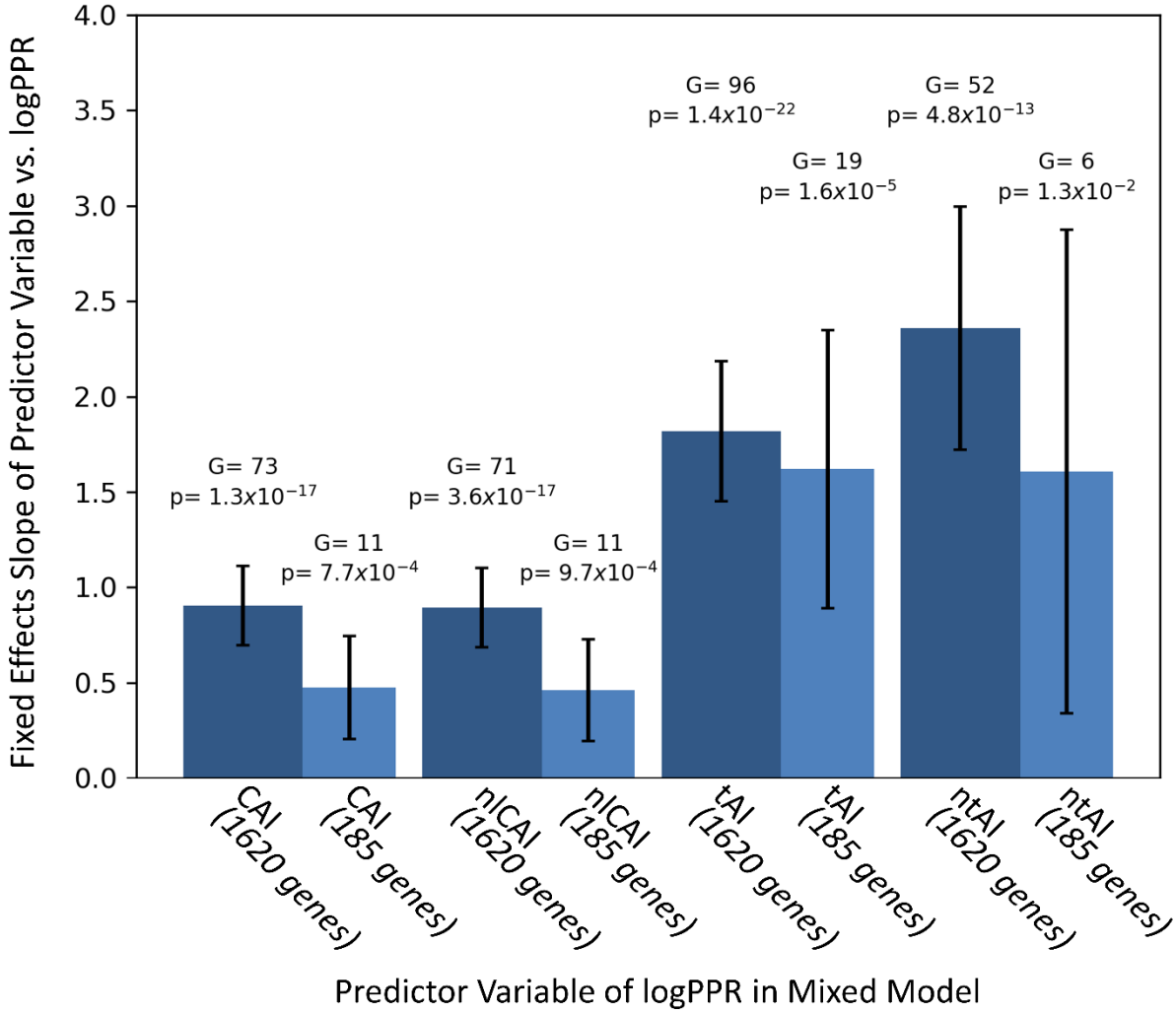



**Figure 7.** Bar graph representing the fixed effects slope of each codon bias measure (CAI, nlCAI, tAI, and ntAI) as the sole predictor of logPPR in a linear mixed effects regression model. Four models are computed based on the set of genes with SNPs across isolates and available PPR data (1620 genes). Four additional models are computed based on the subset of these genes with synonymous SNPs only (185 genes). Fixed effects slope significance is evaluated via log-likelihood ratio test with 1 degree of freedom, and the G statistic and *p*-value are reported for each model. Error bars represent 95% confidence intervals.

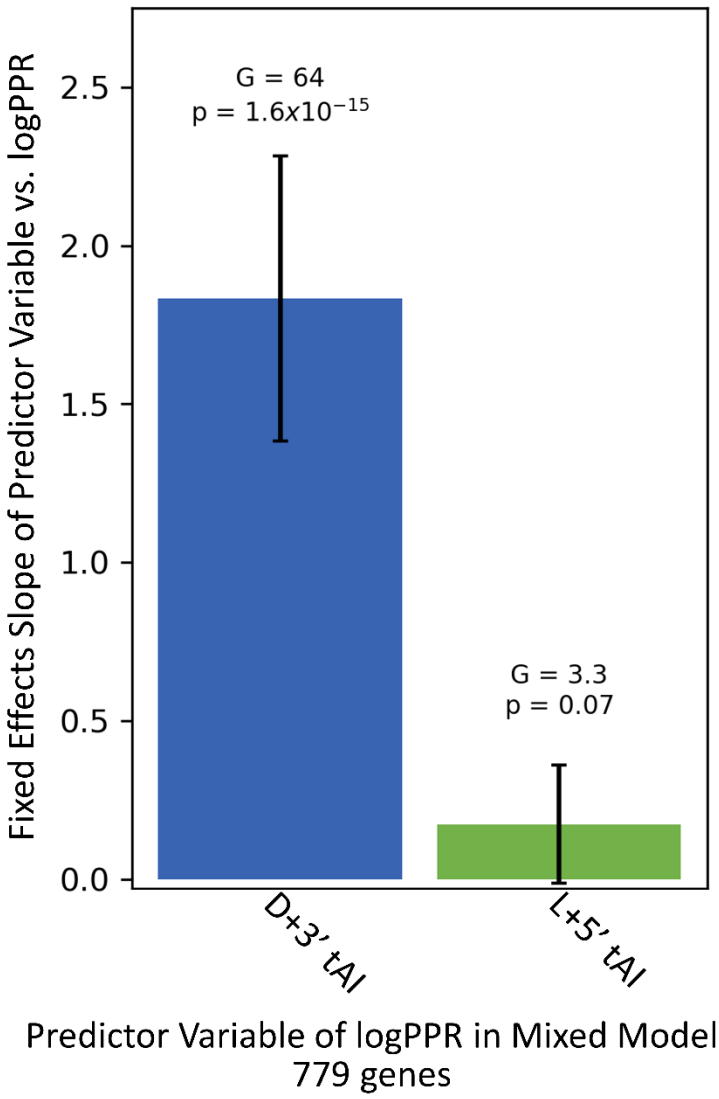



**Figure 8.** Bar graph representing the fixed effects slopes of concatenated domain-encoding and 3'coding sequence tAI (D+3' tAI) vs. logPPR, and of concatenated linker-encoding and 5'coding sequence tAI (L+5' tAI) vs. logPPR. Fixed effects slope significance was determined via log-likelihood ratio test with *df* = 1. The resulting G statistics and *p*-values are reported here, as are the error bars representing 95% confidence intervals.

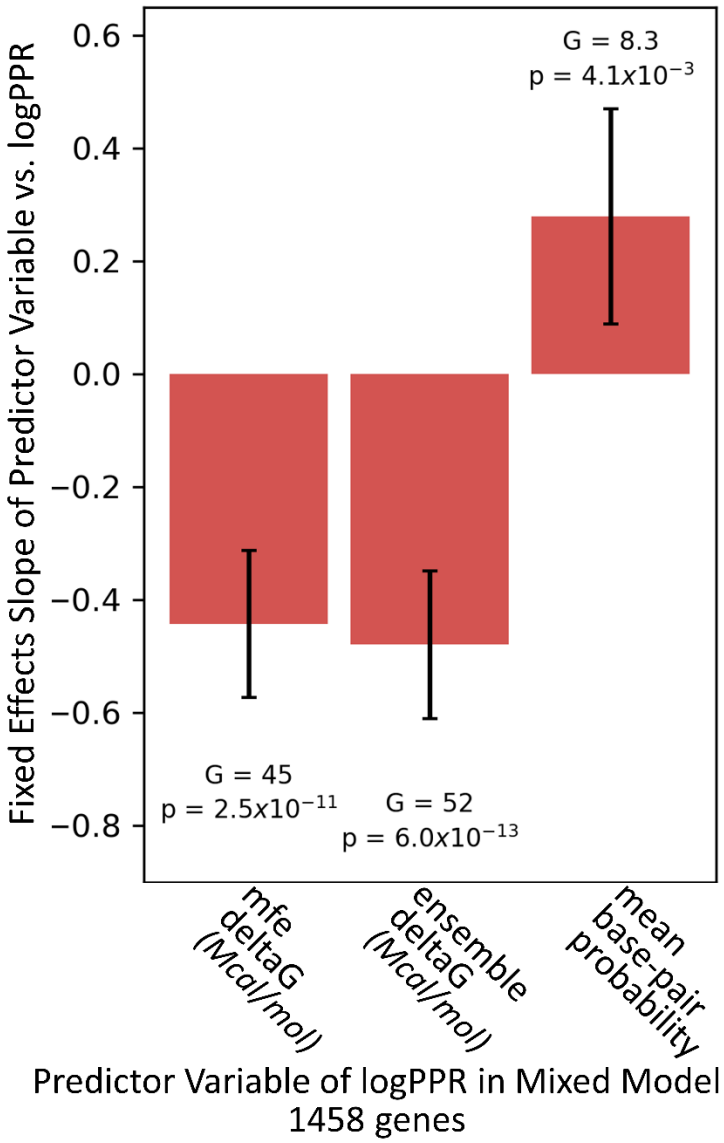



**Figure 9.** Bar graph representing the fixed effects slope of mfe deltaG, ensemble deltaG, and mean base-pair (bp) probability as individual predictors of logPPR in their own linear mixed effects regression models. The significance of each fixed effects slope is computed via log-likelihood ratio test with *df* = 1. G statistics and *p*-values are reported above their respective model's explanatory variable, and error bars represent 95% confidence intervals.

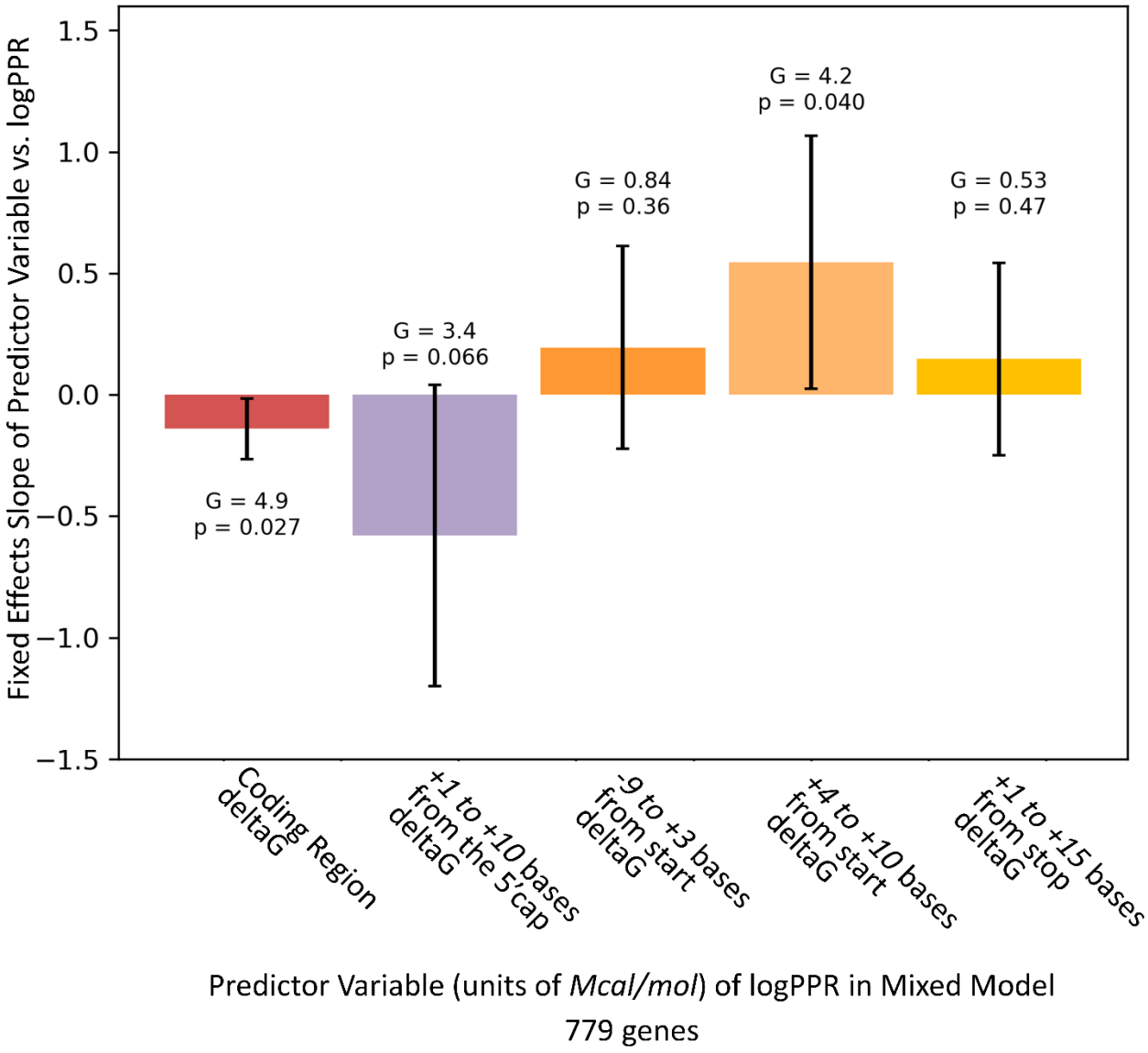



**Figure 10.** Bar graph summarizing the deltaG fixed effects slope for five mRNA regions: the CDS, +1 to +10 bases from the 5'cap, -9 to +3 bases from translation start, +4 to +10 bases from translation start, and +1 to +18 bases from translation stop. Each regional deltaG measure is computed based on an mfe loop-free-energy decomposition, and then incorporated as an individual predictor of logPPR in a linear mixed effects regression model. Fixed effects slope significance is quantified by log-likelihood ratio test with df = 1. G statistics and p-values are listed above the predictor variable of each model, and error bars signify 95% confidence intervals.

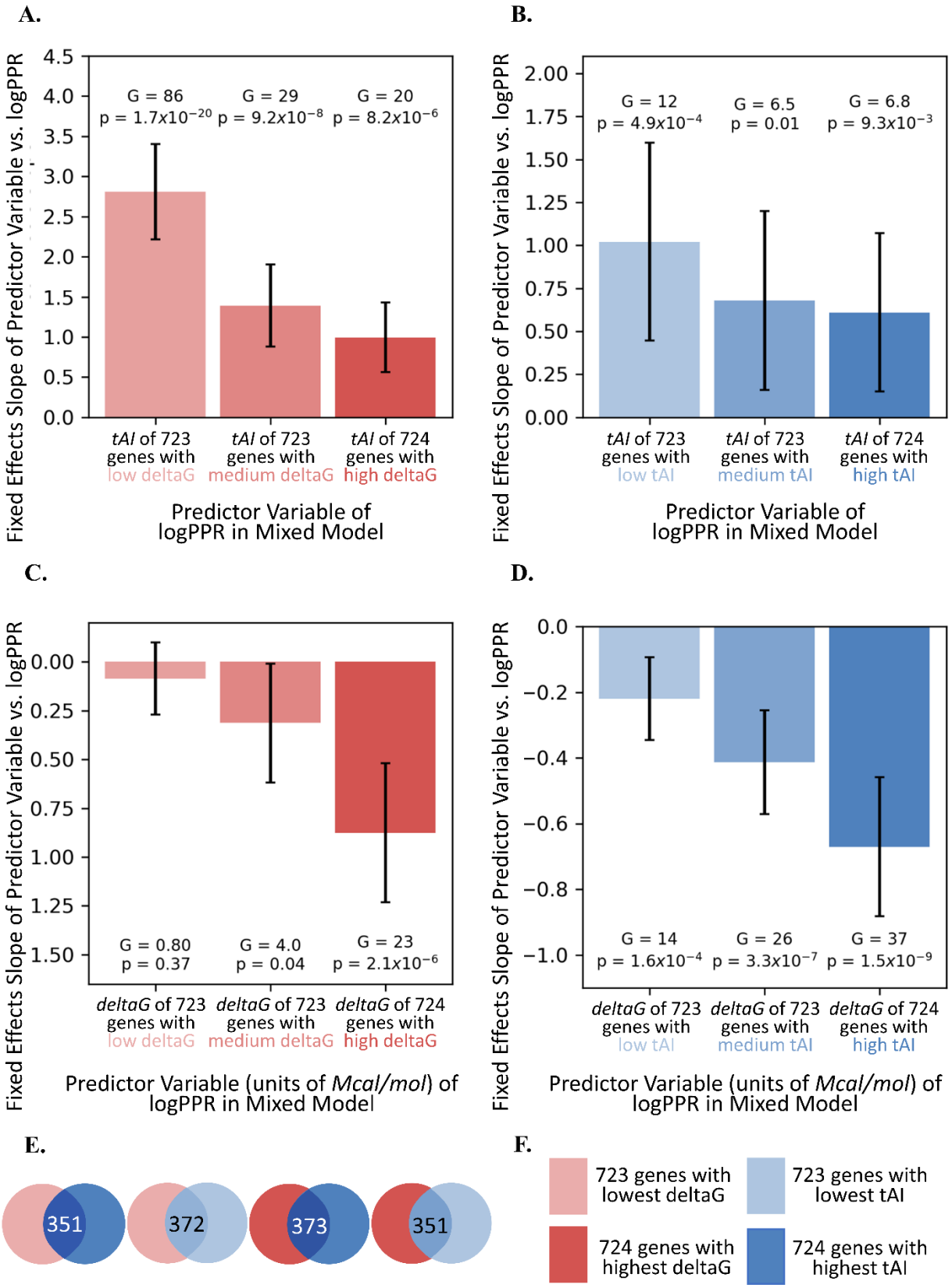

A.
B.
C.
D.
E.
F.
Fixed Effects Slope of Predictor Variable vs. logPPR
G = 86
p = 1.7x10⁻²⁰
G = 29
p = 9.2x10⁻⁸
G = 20
p = 8.2x10⁻⁶
tAI of 723 genes with low deltaG
tAI of 723 genes with medium deltaG
tAI of 724 genes with high deltaG
Predictor Variable of logPPR in Mixed Model
G = 12
p = 4.9x10⁻⁴
G = 6.5
p = 0.01
G = 6.8
p = 9.3x10⁻³
tAI of 723 genes with low tAI
tAI of 723 genes with medium tAI
tAI of 724 genes with high tAI
G = 0.80
p = 0.37
G = 4.0
p = 0.04
G = 23
p = 2.1x10⁻⁶
deltaG of 723 genes with low deltaG
deltaG of 723 genes with medium deltaG
deltaG of 724 genes with high deltaG
Predictor Variable (units of Mcal/mol) of logPPR in Mixed Model
G = 14
p = 1.6x10⁻⁴
G = 26
p = 3.3x10⁻⁷
G = 37
p = 1.5x10⁻⁹
deltaG of 723 genes with low tAI
deltaG of 723 genes with medium tAI
deltaG of 724 genes with high tAI
351
372
373
351
723 genes with lowest deltaG
723 genes with lowest tAI
724 genes with highest deltaG
724 genes with highest tAI

**Figure 11.** (A-D) Bar graphs summarizing 12 fixed effects slopes; we compute a model of ensemble deltaG vs. logPPR as well as a model of tAI vs. logPPR first for the genes characterized by low (*light red*), by medium, and by high (*bright red*) deltaG, and then again for the genes characterized by low (*light blue*), by medium, and by high (*bright blue*) tAI. This yields 12 total models. All gene groupings are based on 1447 genes ranked by their median deltaG and then by their median tAI across the 22 isolates. Each 'low' and 'medium' group contains 723 genes, while each 'high' group contain 724 genes. The significance of each fixed effects slope is evaluated by the log-likelihood ratio test with df = 1. G statistics and *p*-values are listed above the predictor variable of each model, and error bars signify 95% confidence intervals. (E) Venn diagrams showing (from left to right) the number of genes with memberships in both the low deltaG and the high tAI group, the low deltaG and the low tAI group, the high deltaG and the high tAI group, and finally, the high deltaG and the low tAI group. (F) Color key for low and high tAI, and for low and high deltaG groups of genes.

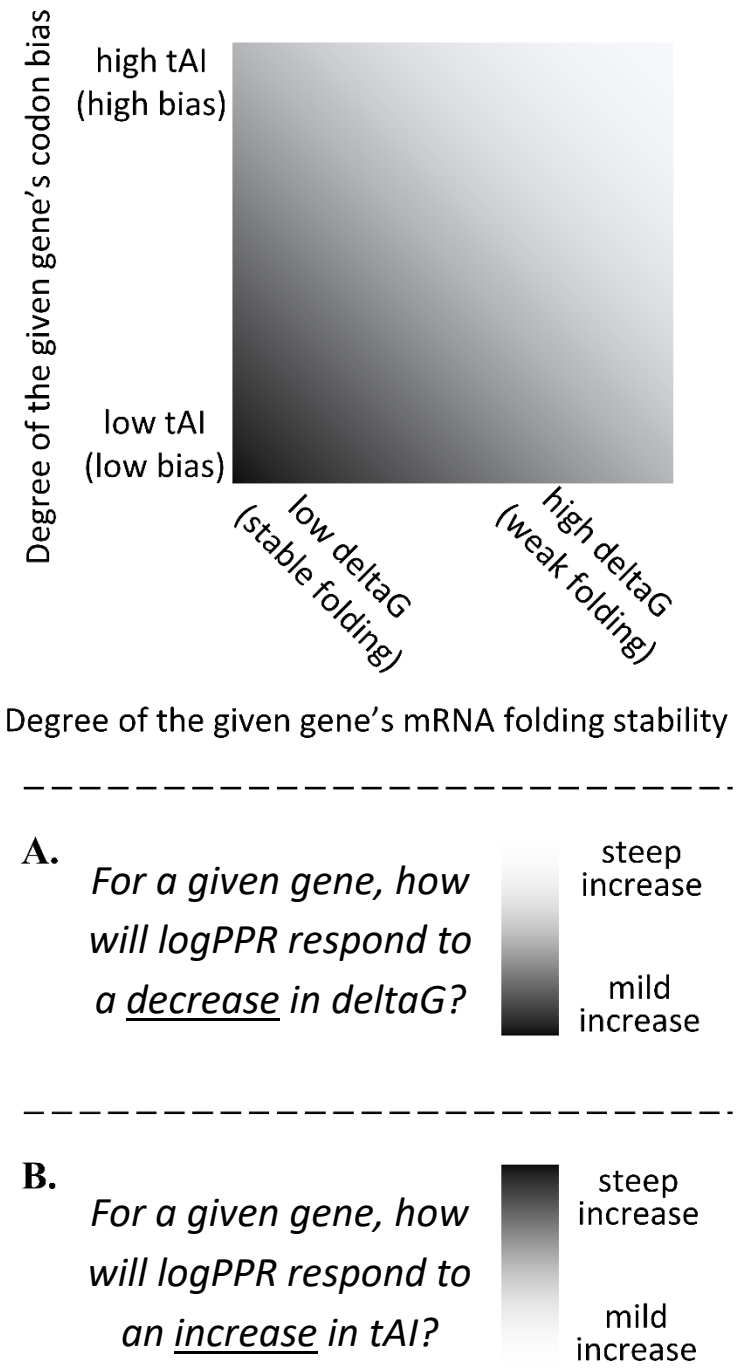



**Figure 12.** Fine-grain schematic summary of the predicted in tandem actions of codon bias and folding stability. (A) For a given gene, how will logPPR respond to a *decrease* in deltaG? (B) For a given gene, how will logPPR respond to an *increase* in tAI? The black/white color gradient is meant to be a visual answer to each question when posed separately. For instance, if the chosen gene has low deltaG and low tAI, we would expect that a decrease in its deltaG (an increase in its folding stability) leads to a steep increase in its logPPR; an increase in its tAI however, would lead to a mild increase in its logPPR.

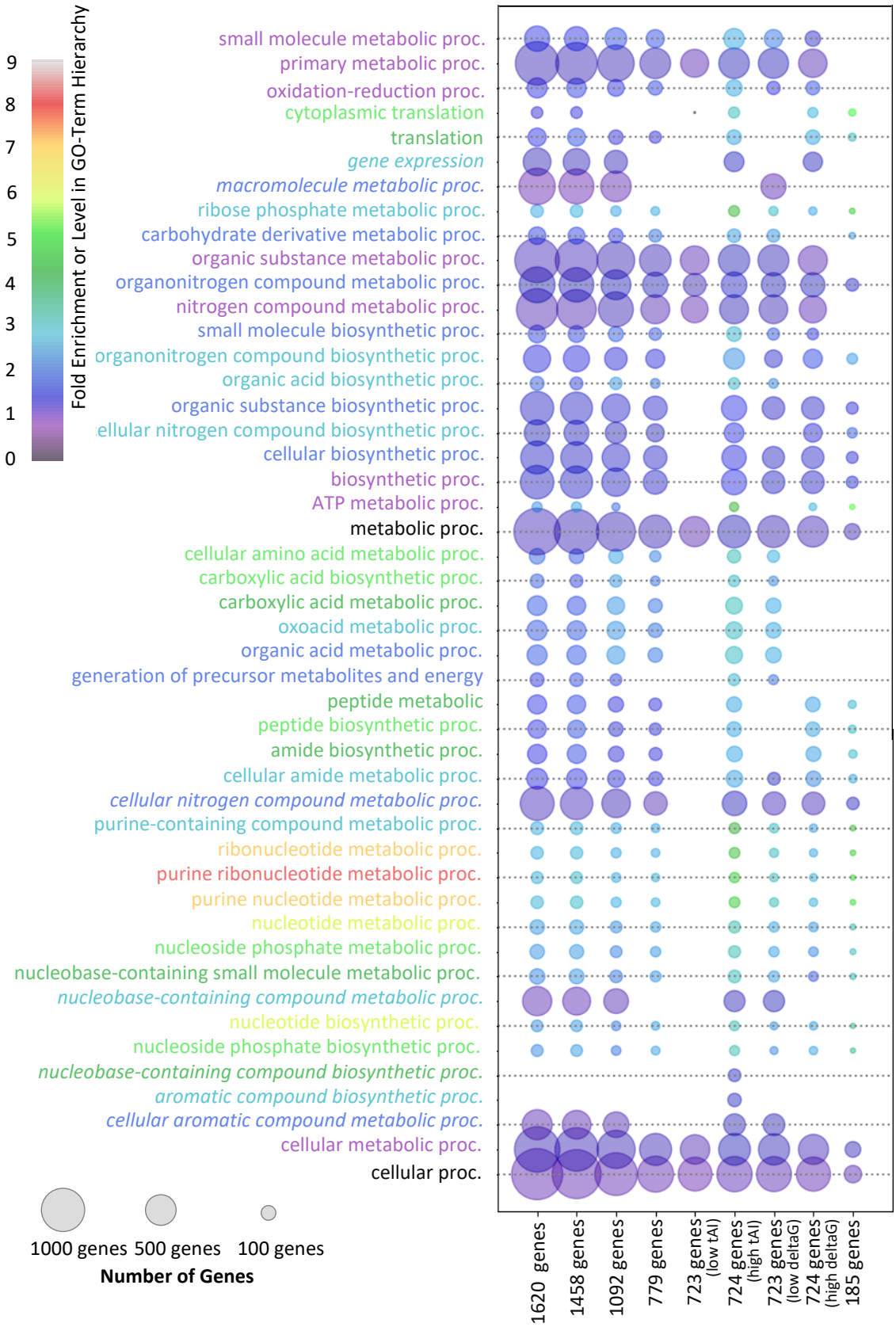

Fold Enrichment or Level in GO-Term Hierarchy
9
8
7
6
5
4
3
2
1
0
small molecule metabolic proc.
primary metabolic proc.
oxidation-reduction proc.
cytoplasmic translation
translation
gene expression
macromolecule metabolic proc.
ribose phosphate metabolic proc.
carbohydrate derivative metabolic proc.
organic substance metabolic proc.
organonitrogen compound metabolic proc.
nitrogen compound metabolic proc.
small molecule biosynthetic proc.
organonitrogen compound biosynthetic proc.
organic acid biosynthetic proc.
organic substance biosynthetic proc.
ellular nitrogen compound biosynthetic proc.
cellular biosynthetic proc.
biosynthetic proc.
ATP metabolic proc.
metabolic proc.
cellular amino acid metabolic proc.
carboxylic acid biosynthetic proc.
carboxylic acid metabolic proc.
oxoacid metabolic proc.
organic acid metabolic proc.
generation of precursor metabolites and energy
peptide metabolic
peptide biosynthetic proc.
amide biosynthetic proc.
cellular amide metabolic proc.
cellular nitrogen compound metabolic proc.
purine-containing compound metabolic proc.
ribonucleotide metabolic proc.
purine ribonucleotide metabolic proc.
purine nucleotide metabolic proc.
nucleotide metabolic proc.
nucleoside phosphate metabolic proc.
nucleobase-containing small molecule metabolic proc.
nucleobase-containing compound metabolic proc.
nucleotide biosynthetic proc.
nucleoside phosphate biosynthetic proc.
nucleobase-containing compound biosynthetic proc.
aromatic compound biosynthetic proc.
cellular aromatic compound metabolic proc.
cellular metabolic proc.
cellular proc.
1000 genes
500 genes
100 genes
Number of Genes
1620 genes
1458 genes
1092 genes
779 genes
723 genes (low tAI)
724 genes (high tAI)
723 genes (low deltaG)
724 genes (high deltaG)
185 genes

**Figure S1.** Summary of nine GO term enrichment analyses. Each analysis corresponds to one set of genes, and the sets of genes considered are those used in our models of codon bias and folding stability on the global and local scales. Circle size indicates the number of genes associated with a given GO term (see legend), while circle color indicates the fold enrichment (relative to the genome) (see color scale). The color of the GO term label indicates its level in the hierarchy of biological process GO terms, where black is level zero (broadest level term) and red is level seven (our narrowest level term) (see color scale). No circle indicates that a term's FDR q-value exceeds 0.05. Non-italicized GO terms have q-values less than $10^{-10}$ in at least one of the nine gene sets; italicized GO terms do not satisfy this criterion, but we include them because they are parent terms. Results are generated by the PANTHER Overrepresentation Test (released 2020-04-07) via the GO biological process complete annotation for *S. cerevisiae* (version 2020-03-23). Fisher's Exact test with the False Discovery Rate correction was used to obtain q-values.