Evolutionary principles underpinning codon usage bias: patterns, functions, and mechanisms

Alexander L. Cope[1,2], Michael A. Gilchrist[3], Premal Shah[1], and Edward W.J. Wallace[4,*]

[1]Department of Genetics, Rutgers University, Piscataway, NJ, United States

[2]Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, United States

[3]Department of Ecology and Evolution Biology, University of Tennessee Knoxville, Knoxville, TN, United States

[4]Institute for Cell Biology and Centre for Engineering Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

[*]Corresponding Authors: edward.wallace@ed.ac.uk

Thursday 30th May, 2024

Keywords: Codon usage bias, translation, selection-mutation-drift equilibrium, evolutionary spandrels

Abstract

Synonymous codons are used unevenly despite coding for the same amino acid. Recent work has provided critical insights into the functions, mechanisms, and fitness consequences of codon usage bias and synonymous mutations. However, experiments aimed at understanding the role of synonymous mutations often involve only a small number of reporter genes. How do these observations generalize across genomes, where confounding factors include gene expression and GC content? We propose the following principles for making inferences about the functions, mechanisms, and evolution of codon usage. First, use additive selection-uniform mutation-drift equilibrium as the null model. This evolutionary model explains how codon usage in low-expressed genes is driven by mutation bias and, in high-expressed genes, is driven by selection. It performs well enough to serve as a sensible default for understanding the evolution of codon usage patterns. Second, analyses of codon usage should control for gene expression. The effect of a synonymous change on mRNA translation scales with a gene's total protein production rate such that evolutionary selection on codon usage is strongest in highly-translated genes. Because protein production rate correlates with many other gene features, researchers must control for its effects. Third, researchers must consider mechanistically how codon usage affects biological processes. While correlations between codon usage and other molecular measurements are valuable, proposed mechanistic roles of codon usage must be consistent with established biological mechanisms. In conclusion, the underlying architecture of molecular evolution should be considered before invoking other superficially plausible explanations of codon usage.

Introduction

While the existence and prevalence of synonymous codon usage bias is non-controversial, the biological causes of this bias are controversial. As seasoned researchers in the field, we believe that controversies in the codon usage literature are the result of multiple factors. First, researchers have no agreed-upon best way to quantify codon usage. There are a staggering number of metrics for quantifying and identifying the "optimality" of a codon, as well as quantifying the overall codon adaptation of a gene (Roth *et al.*, 2012). Second, there is no agreed-upon null distribution

for the expected frequencies of codon usage, to compare to alternative hypotheses. Synonymous codons are used unequally in every known organism, making the null model of equal codon usage unsupportable. Third, codon usage metrics can be confounded by many factors, particularly gene expression and amino acid usage biases.

We believe that most of these issues can be addressed through the integration of molecular models of the processes of protein translation with evolutionary models of allele fixation. This approach naturally leads to a cogent and powerful null model that includes the effects of selection, mutation bias, and genetic drift. We begin our argument by providing background context and a basic rationale for our claim. We next lay out a well-supported yet simple and mechanistic model, the additive selection-uniform mutation-drfit equilibrium (ASUMDE), that we argue should be adopted as a default null model for codon usage bias in a genome. We present examples where the lack of appropriate null models in published analyses has led to spurious conclusions about codon usage bias. Finally, we end with a discussion of the limitations of the ASUMDE model and call for more nuanced models.

A brief history of codon usage bias research

The genetic code is degenerate, as most amino acids are coded for by more than one codon (Crick *et al.*, 1957). As gene sequences became available in the 1970s, it became clear that codons were not used at equal frequencies within a species (Clarke, 1970; Fitch, 1976; Grantham *et al.*, 1980). This non-uniform usage of synonymous codons, or codon usage bias, has been the subject of intense study over the last 40 years.

In the 1980s, a clear relationship between codon usage and the tRNA pool emerged: codon frequencies tended to correlate with tRNA abundances, and high-expressed genes were biased towards codons corresponding to more abundant tRNA (Gouy and Gautier, 1982; Ikemura, 1981, 1982, 1985). All other things being equal, codons with higher concentrations of cognate tRNA are translated both faster, i.e. higher elongation rate, and also more accurately, i.e. lower missense error rate. These results suggested synonymous mutations were neither irrelevant nor neutral, but could be under natural selection to promote translation, and coevolving with the tRNA pool

(Bulmer, 1987). During this early period of codon usage research, two major hypotheses emerged to explain this observation: (1) the regulatory hypothesis (Grosjean and Fiers, 1982; Konigsberg and Godson, 1983; Walker *et al.*, 1984; Hinds and Blake, 1985; Burns and Beachamn, 1985) and (2) selection-mutation-drift equilibrium models, also referred to as the Li-Bulmer model (Li, 1987; Bulmer, 1991).

The regulatory hypothesis posited that codon usage regulated protein production, such that using slower codons would produce fewer proteins. However, the regulatory hypothesis overestimates the general impact of codons on protein production. Both theoretical and experimental evidence suggest that total protein production per mRNA is primarily limited by translation initiation, while codons primarily determine translation elongation (Andersson and Kurland, 1990; Bulmer, 1991; Arava *et al.*, 2003; Salis *et al.*, 2009; Kosuri *et al.*, 2013; Erdmann-Pham *et al.*, 2020). Thus, synonymous mutations have a smaller effect on protein production, on average, than mutations to regions near the start codon that determine translation initiation, a conclusion supported by reporter gene studies (Kudla *et al.*, 2009), omics-scale measurements of ribosome positioning (Arava *et al.*, 2003), and theoretical studies of mRNA translation dynamics (Shah *et al.*, 2013; Subramaniam *et al.*, 2014; Erdmann-Pham *et al.*, 2020). It is simpler for evolution – or genetic engineers – to change protein production by altering a few regulatory elements near the start codon than by altering 100s of codons across the gene.

In contrast, selection-mutation-drift equilibrium models posit that genome-wide codon usage frequencies are at an equilibrium between natural selection favoring "optimal" synonymous codons, while "non-optimal" codons are introduced via mutation and fixed (i.e., found in all members of a population) via genetic drift. The most successful version is the additive selection-uniform mutation-drift- equilibrium (ASUMDE), which models selection per protein produced, and mutation as uniform across the proteome. Unlike the regulatory hypothesis, the ASUMDE hypothesis assumed that natural selection on codon usage related to mRNA translation acts additively depending only on the total protein production rate, rather than to optimize protein-specific regulation. The major interpretation is that additive selection acts via the pool of free ribosomes in the cell, such that slow codons cause ribosomes to pause on transcripts, resulting in a reduction to the pool of ribosomes available to initiate translation. As the evidence indicates mRNA translation is

initiation-limited, a reduction to the pool of free ribosomes would negatively impact cellular-wide mRNA translation dynamics, consistent with both theoretical and empirical analysis (Shah *et al.*, 2013; Subramaniam *et al.*, 2014; Frumkin *et al.*, 2018; Ballard *et al.*, 2019; Erdmann-Pham *et al.*, 2020). However, the additive selection model is also interpretable as selection acting against less accurate codons due to a fitness cost for each mistranslated protein (Wallace *et al.*, 2013).

The emergence of omics-scale technologies and advancements in molecular biology, genetics, and bioinformatics led to vast improvements in our understanding of the mechanisms and functions of codon usage and synonymous mutations. Genome-wide correlations between synonymous codon usage and gene expression were observed in multiple species spanning the tree of life (Drummond *et al.*, 2005; Drummond and Wilke, 2008; Hiraoka *et al.*, 2009). Aside from elongation speed or efficiency, codon usage has been implicated in translation accuracy (Kurland, 1992; Akashi, 1994; Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Drummond and Wilke, 2008; Mordret *et al.*, 2019), mRNA secondary structure (Chamary and Hurst, 2005; Stoletzki, 2008), translation initiation (Kudla *et al.*, 2009; Hockenberry *et al.*, 2014), cotranslational protein folding (Komar *et al.*, 1999; Kimchi-Sarfaty *et al.*, 2007; Tsai *et al.*, 2008; Buhr *et al.*, 2016; Walsh *et al.*, 2020), protein secretion (Fluman *et al.*, 2014; Zalucki *et al.*, 2009), and mRNA decay (Presnyak *et al.*, 2015; Wu *et al.*, 2019; Forrest *et al.*, 2020) (for a comprehensive overview, see reviews by Chaney and Clark (2015); Hanson and Coller (2018); Nieuwkoop *et al.* (2020); Wu and Bazzini (2023)). Codon usage has also been implicated in non-translation processes, such as transcription (Zhou *et al.*, 2016; Zhao *et al.*, 2021). As a result, there is a resurgence in the idea that codon usage can play a regulatory role in protein production. Work with reporter genes implicates synonymous mutations in many of these functions and mechanisms. However, a key challenge is understanding the general role of codon usage and synonymous mutations in these mechanisms and functions on a genome-wide scale. To this end, numerous bioinformatics studies attempted to extrapolate observations made from a relatively small number of genes to genome-wide trends by looking for associations with codon usage. The results of these studies often conflicted, creating numerous controversies.

Before returning to these controversies, we explain the ASUMDE model and argue why it is useful in resolving them.

The additive selection-uniform mutation-drift equilibrium (ASUMDE) model

The ASUMDE model quantifies the relative contributions of mutation bias and natural selection to shaping codon frequencies. Generally, mutation bias drives codon usage in low-translated genes while selection drives codon usage in high-translated genes (Figure 1). Because the ASUMDE model involves selection specifically on total protein production rate per gene(Shah and Gilchrist, 2011; Wallace *et al.*, 2013; Gilchrist *et al.*, 2015), we shall carefully distinguish protein production rate from the more ambiguous term "gene expression".

Precisely, ASUMDE is a population genetics model of codon usage, incorporating selection that scales additively with the gene's protein production rate, mutation with a uniform bias across the genome, and genetic drift that limits the impact of selection based on the (effective) population size. The probability of observing codon c, in gene g, is:

$$p_{c,g} \propto \exp\left(M_c + N_e P_g S_c\right) \tag{1}$$

where:

M_c	Mutation bias towards codon c
N_e	Effective size of population
P_g	Protein production rate of gene g
S_c	Selection coefficient towards codon c.

Note that the prediction is constant across a gene, i.e., every position in a gene that encodes the same amino acid has the same probabilities for codon usage.

We approach the ASUMDE model from three perspectives. First, ASUMDE is a regression of codon counts on protein production rate, i.e., the simplest statistical model for the dependence of codon usage on protein production rate. Technically, ASUMDE is a logistic regression, a common statistical model within the generalized linear model family, such that the parameters can be

estimated using standard statistical methods (Agresti, 2002). The same framework extends to 3-, 4- and 6- codon families as a multinomial logistic regression, again with well-developed fitting algorithms that are widely implemented and that quantify uncertainty in the parameter estimates.

Second, ASUMDE is a mechanistic model of evolution that quantifies codon usage in terms of underlying biological causes. The mechanistic motivation for the model is that first, mutations happen at rates that depend only on the codon sequence, e.g. AAA to AAG. Next, these mutations are fixed in a population at rates depending on selection for speed or accuracy of translation, and drift. Selection scales with a codon-specific selection coefficient that is the same for all genes, the protein production rate that is different for each gene, and the effective population size from population genetics that determines the strength of selection compared to drift (Berg *et al.*, 2004; Sella and Hirsh, 2005; McCandlish *et al.*, 2015). Mechanistic selection-mutation-drift equilibrium models exist within a larger framework of origin-fixation models with a 50-year history as successful population genetics tools (see McCandlish and Stoltzfus (2014) for a review of these models and the relevant population genetics theory). The selection coefficient is in the sense of population genetics, where $P_g(S_c - S_{c'})$ is interpreted as change in average number of offspring from having codon c rather than codon c' in gene g. The mutation bias M_c is proportional to the log mutation rate such that $\exp(M_c - M_{c'})$ is the ratio of mutation rates between codon c and codon c'. The ASUMDE model is separate for each amino acid, and so also implicitly relies on a separation of scales where synonymous codon substitutions have on average, smaller selection coefficients than nonsynonymous substitutions.

Third, ASUMDE is the steady state equilibrium of a Markov chain, similar to other stochastic dynamical models. This equilibrium approximation is inaccurate because evolution is by definition, a departure from a steady state, and limitations of the equilibrium assumption are discussed by McCandlish and Stoltzfus (2014). However, the success of ASUMDE's statistical predictions of *average* codon usage show that the model is explanatory.

This situation - a standard statistical model that is also a mechanistic model with interpretable parameters - is very helpful. Standard statistical models can be fit to data with a low risk of errors in model specification and parameter estimation and, therefore, a low risk of spurious conclusions. Interpretable parameters - selection coefficients and mutation bias - can be meaningfully compared

with other biological data.

Development of the ASUMDE model: incorporating protein production rate per gene

The earliest selection-mutation-drift equilibrium models for codon usage proposed a selection coefficient reflecting the overall strength of selection on codon usage (Bulmer, 1991). This model did not explicitly incorporate protein production rate per gene, which has since been shown to dominate per-gene evolutionary rate (Drummond *et al.*, 2006). As quantitative gene expression data became available, researchers began to account for different codon usage in high and low expression genes (Sharp *et al.*, 2005; Harrison and Charlesworth, 2011; Pechmann and Frydman, 2013; de Oliveira *et al.*, 2021). Such approaches for quantifying adaptive codon usage trace back to the Codon Adaptation Index (CAI) of Sharp and Li (1987), which identifies "optimal" codons based on a set of high-expressed genes. However, protein production rate and other gene expression measures are continuous variables, not simply "high" or "low".

Shah and Gilchrist (2011) developed the ASUMDE model to quantify the effect of protein production rate on codon usage precisely. They introduced equation 1, with a uniform mutation bias term and a selection coefficient that is multiplied by per-gene protein production rate estimates derived from high-throughput data. This separated the effects of mutation and natural selection to provide codon-specific estimates of mutation bias and selection coefficients. Importantly, these selection coefficients were well-correlated with expected waiting times estimated from tRNA abundances, suggesting that selection related to elongation speed or efficiency is a major driver of adaptive codon usage bias. Overall, the ASUMDE model developed by Shah and Gilchrist (2011) was able to explain 92% of the variation in codon counts across the *Saccharomyces cerevisiae* yeast genome.

A limitation of the Shah and Gilchrist (2011) ASUMDE model was that it failed to account for the noise present in empirical gene expression data. By "noise", we mean both that estimates from any one study are randomly inaccurate or biased by growth conditions, and also that measuring

gene expression by RNA abundance (for example) is not a perfectly accurate measure of protein production rate. Thus, Wallace *et al.* (2013) incorporated the ability to account for noise in gene expression data using a more complex Bayesian statistical approach. The success of Wallace *et al.* (2013) exploited their observation that the ASUMDE model is the same as a (multinomial) logistic regression on codon frequencies, allowing for parameter estimation with standard statistical methods. To test the predictions of the model with independently obtained data, Wallace *et al.* (2013) showed that codon-specific estimates of mutation bias correlated well with mutation biases estimated from mutation accumulation experiments. Thus, the ASUMDE model precisely quantifies the observation that codon usage in low-translated genes is primarily driven by mutation bias, and in high-translated genes can be driven by selection.

To extend the ASUMDE model to species lacking empirical gene expression data, Gilchrist *et al.* (2015) developed a Bayesian framework to estimate an evolutionary-average protein production rate per gene simultaneously with estimating per-codon mutation and selection coefficients. This model, termed the Ribosomal Overhead Cost version of the Stochastic Evolutionary Model of Protein Production Rates (ROC-SEMPPR), can be applied to any species with an annotated genome (i.e., a FASTA file containing protein-coding sequences). ROC-SEMPPR estimates of protein production rate per gene are often well-correlated with empirical gene expression data (Cope *et al.*, 2018; Landerer *et al.*, 2020; Cope and Shah, 2022). Indeed, Gilchrist *et al.* (2015) showed that incorporating empirical gene expression data had little impact on model performance, demonstrating that codon usage itself is sufficient to accurately estimate protein production rates per gene, mutation biases, and selection coefficients. The ROC-SEMPPR framework is implemented in the **AnaCoDa** R package (Landerer *et al.*, 2018), for wider use.

A key lesson is that codon "optimality" should not be determined by codon frequencies in a set of high-translated genes, but by the continuous changes in synonymous codon frequencies as protein production rate varies. This is because if mutation bias is very strong or natural selection is very weak, the selectively favored codon may not be the most frequent even in high-translated genes. A similar idea was proposed by Hershberg and Petrov (2012), who argued that the "optimal" synonymous codon should be defined as the codon whose gene-specific frequencies correlated best with gene-level estimates of codon bias. The ASUMDE model makes this argument precise.

Molecular spandrels and codon usage

In their 1979 essay "The Spandrels of San Marco and the Panglossian Paradigm", evolutionary biologists Stephen Jay Gould and Richard Lewontin argued that biologists had become enamored with natural selection and adaptation, too often attempting to explain biology with no consideration to developmental constraints or neutral evolutionary processes such as genetic drift (Gould and Lewontin, 1979). They likened the adaptive explanations to the presence of spandrels in St. Mark's cathedral. At first glance, the beautiful artwork painted within the spandrels may lead to the conclusion that the building was designed to accommodate such artwork; however, spandrels are merely the result of stacking a dome on arches, with the artwork made to fit the available space (Figure 3). Analogously, a correlation between codon usage and gene-level traits is often interpreted in the light of adaptive evolution. As efforts attempt to unlock signals of selection on codon usage related to various processes and mechanisms, it is important to ensure that the observed bias is not more simply explained by generic models such as ASUMDE (Figure 3). Failing to properly control for factors such as gene expression and amino acid biases could lead to spurious conclusions regarding the nature or direction of natural selection on codon usage.

A common observation is the apparent enrichment of slow codons at the 5'-ends of coding regions, spawning both adaptationist and non-adaptationist explanations. One hypothesis argued that slow codons are selected for at the 5' end forming a "ramp", an adaptation to prevent downstream ribosome queueing and promote overall efficient translation (Tuller *et al.*, 2010; Sejour *et al.*, 2023). In contrast, we found that selection on codon usage is positively correlated between the 5'-end and the remainder of the gene (Cope *et al.*, 2018). Our results show that the same codons are generally favored at the 5'-end as the remainder of the coding region, but the strength of selection on codon usage is generally weaker at the 5'-end. Several other studies argue that selection at the 5'-end may be quantitatively different from the rest of coding regions, possibly due to conflicting selection pressures related to mRNA secondary structure (Kudla *et al.*, 2009; Bentele *et al.*, 2013; Goodman *et al.*, 2013; Hockenberry *et al.*, 2014) or weaker selection against premature termination errors that are more likely to occur at slower codons (Eyre-Walker, 1996; Qin *et al.*, 2004; Gilchrist, 2007; Gilchrist *et al.*, 2009; Yang *et al.*, 2019). Direct experimental testing of 5' ends also found

that the "ramp" hypothesis is a "spandrel", not supported by evidence: substituting faster codons at 5' ends of genes in budding improves expression (Sejour *et al.*, 2023). So, 5'-end selection still generally conforms to the assumptions of the ASUMDE: codon usage is well-described by a balance between additive selection (for speed or accuracy), uniform mutation, and genetic drift.

Numerous studies have focused on the role of codon usage in regulating protein biogenesis (e.g., protein folding, protein secretion), often looking for regions of slow codons that are thought to be connected to these processes (Chaney and Clark, 2015). Results differ drastically across studies due to differences in how codon usage bias was quantified. Furthermore, some of these studies failed to test their hypotheses relative to the ASUMDE expectation explicitly and so failed to account for the effects of gene expression and amino acid biases (Figure 2).

Other previous work found an enrichment of slow codons in signal peptides – N-terminal determinants of protein secretion – in *E. coli*, which was hypothesized to be due to increased selection for inefficient codons to modulate protein secretion (Burns and Beachamn, 1985; Power *et al.*, 2004; Zalucki *et al.*, 2009). Empirical studies indicate that synonymous mutations in signal peptides can impact protein secretion in specific cases (Zalucki *et al.*, 2007; Zalucki and Jennings, 2007; Zalucki *et al.*, 2008, 2010), but does the enrichment of slow codons in signal peptides reflect a true evolutionary adaptation? We concluded that the enrichment of slow codons in the signal peptides of *E. coli* relative to the 5'-ends of non-secreted proteins is consistent with the ASUMDE model(Cope *et al.*, 2018). We simulated coding sequences using ASUMDE as a null model: assuming no differences in selection on codon usage in the 5'-ends encoding signal peptides and non-secreted proteins, we found that signal peptides always had a lower average Codon Adaptation Index (CAI) (Cope *et al.*, 2018). This difference did not reflect selection but instead was driven partly by differences in amino acid composition of signal peptides, because CAI normalizes codon-specific coefficients separately for each amino acid. After controlling for both gene expression effects and amino acid biases, there was no enrichment for slow codons in signal peptides: the effect disappeared.

Another possible molecular spandrel is the proposal that slow codons are selected based on their effect on protein folding. Much like protein secretion, empirical evidence indicates that altering synonymous codon usage can affect protein folding (Purvis *et al.*, 1987; Krasheninnikov *et al.*, 1991; Kimchi-Sarfaty *et al.*, 2007; Holtkamp *et al.*, 2015; Buhr *et al.*, 2016; Walsh *et al.*, 2020). In

well-supported cases, codon usage appears to modulate cotranslational protein folding by slowing down translation during key parts of the folding process. As a result, numerous studies attempted to connect differences in codon usage with protein structure by looking for patterns across larger sets of protein-coding sequences.

Zhou *et al.* (2015) investigated codon usage within intrinsically disordered protein regions, finding a negative correlation between CAI and the "disorderedness" of a region within a protein across many species. This led them to conclude that disordered regions had a "preference" for slow codons, supposedly to assist upstream structured or ordered regions fold co-translationally. However, this work did not account for the fact that disordered regions are generally avoided in high-expression genes (Singh and Dash, 2008; Gsponer *et al.*, 2008; Dubreuil *et al.*, 2019) and have distinct amino acid biases (Singh, 2015). We used ROC-SEMPPR to test for differences in natural selection on codon usage between structured and disordered regions of proteins in *E. coli* and *S. cerevisiae* (Cope and Gilchrist, 2022). In contrast to the findings of Zhou *et al.* (2015), we found that additive selection on protein production rate was the predominant selective force driving codon usage in both structured and disordered regions. Much like with 5'-ends, selection was weaker in disordered regions, but this does not indicate a change in codon "preference". Indeed, such results could also be explained by reduced selection against missense errors as concluded in a similar study of disordered region codon usage (Homma *et al.*, 2016). Based on simulations, Cope and Gilchrist (2022) concluded that if selection for slow translation does occur in disordered regions, it likely affects less than 1% of codon sites. This means that the ASUMDE is generally a good description of codon usage within disordered regions. Similarly, Cope and Gilchrist (2022) also used simulations under the ROC-SEMPPR model to show that the apparent enrichment of "non-optimal" codons at the second and third positions of α-helices in yeasts (Pechmann and Frydman, 2013) was perfectly consistent with expectations under ASUMDE. In other words, structured regions and IDRs preferentially used the same codons, as do different positions in α-helices, in both cases consistent with an ASUMDE model and refuting arguments for specific selection based on structure.

These and other examples (Akeju and Cope, 2024) show that failing to use a null model that appropriately accounts for confounding factors when testing for selection on codon usage can lead to spurious conclusions. The examples have two important implications. First, amino acid biases

can impact codon usage metrics that only consider relative codon usage, because the strength of selection on codon usage often varies between amino acids. Second, differences in metrics such as CAI or tAI should not be interpreted as reflecting differences in selection on codon usage without carefully controlling for other factors affecting coding sequence evolution. In particular, the ASUMDE model calculates selection coefficients on the same scale for every amino acid, avoiding biases found in metrics that normalize coefficients for each amino acid in a way that is not theoretically grounded.

A tangled web: the relationship between gene expression, codon usage, and other biological mechanisms

Not accounting for gene expression when explaining codon usage patterns can also lead to spurious conclusions. Protein production rate is correlated with many other processes involved in gene expression. If unaccounted for, these correlations could give the false impression that codon usage plays a mechanistic role in another process.

Problems arising from shared correlations with gene expression extend beyond the study of codon usage. For example, previous studies concluded that the evolutionary rate of a protein – often measured as the ratio of the nonsynonymous to synonymous substitutions across species – correlated with gene dispensability (Hirsh and Fraser, 2001) and properties of the protein-protein interaction network (Fraser *et al.*, 2002; Han *et al.*, 2004; Fraser *et al.*, 2004). However, these correlations largely disappeared after controlling for gene expression (Pál *et al.*, 2003; Batada *et al.*, 2007; Bloom and Adami, 2003, 2004; Wang and Zhang, 2009).

A recent hypothesis is the role of codon usage in modulating mRNA decay. The Codon Stabilization Coefficient (CSC) intends to reflect a codon's contribution to mRNA stability - i.e., longer lifetime - by correlating how a codon's frequency changes as a function of mRNA half-life (Presnyak *et al.*, 2015). However, recent work identified synonymous codon variants of transcripts that increase mRNA secondary structure and thus mRNA half-life (Zhang *et al.*, 2023). This sequence space has been largely unexplored by previous design algorithms and evolution, as few natural sequences fall

within this space. Selection for extensive RNA secondary structure requires base-pairing between distant regions of the same mRNA including non-adjacent codons.

How then do we explain the high correlation between CSC and mRNA half-life? Simulating under ROC-SEMPPR using protein production rate estimates from ribosome profiling data (Weinberg *et al.*, 2016), we find that predicted CSC estimates for simulated genes agree with the CSC estimates from real protein-coding sequences 4. Thus, if mRNA decay is primarily determined by translation dynamics (Chan *et al.*, 2018), and protein production rates are translation-initiation-limited, then a correlation between codon usage and mRNA decay is expected even if the former has relatively little to no mechanistic role in the latter on a genome-wide scale. Our results do not invalidate a mechanistic role for codon usage on mRNA decay on a genome-wide scale, but correlations between these various gene-level traits (e.g., protein production rate, mRNA half-life) make it difficult to distinguish selection on mechanistically distinct processes.

Other recent work has hypothesized a role for codon usage in transcription. Work proposing a mechanistic role for codon usage in transcription must overcome two key challenges. First, RNA polymerase only recognizes nucleotides, not codons. Second, transcription and translation are highly correlated. Zhao *et al.* (2021) proposed that the correlation between codon usage and mRNA abundances in the nucleus (where translation does not occur) of the fungi *N. crassa* serves as evidence that codon usage impacts transcription. However, this ignores the fact that whole-cell mRNA and nuclear mRNA abundances are well-correlated, necessitating the use of partial correlations. When using partial correlations, we find that codon usage negatively correlates with nuclear mRNA abundances (Table 1).

Limitations and extensions of the ASUMDE model

As said by George Box, “All models are wrong, but some are useful,” (Box, 1979). All current versions of the ASUMDE model assume each codon within a sequence evolves independently of other codons and ignores the effects of recombination, i.e., linkage-related effects are absent. Furthermore, ASUMDE models for codon usage fall into a class of ”mutation-limited” models in molecular evolution known as origin-fixation models in which a mutation is either fixed or purged from a

population before the arrival of the next mutation (McCandlish and Stoltzfus, 2014). As a result, the ASUMDE models ignore polymorphism within populations.

The ASUMDE model does not fit codon usage bias well in humans, where by contrast variations in GC-richness and dinucleotide biases are dominant (Radrizzani *et al.*, 2024). This is thought to be due to humans' small effective population size, which both limits the impact of translation selection and also allows more scope for junk DNA and mobile genetic elements (Radrizzani *et al.*, 2024). By contrast, the ASUMDE model is particularly effective in quantifying codon usage in fast-growing microbial species with strong selection.

We have encountered other species where the ROC-SEMPPR implementation of ASUMDE did a poor job of explaining codon usage patterns. In some cases, this appears to be due to intragenomic variation in non-adaptive nucleotide biases, which can result from biased gene conversion (Duret and Galtier, 2009; Galtier *et al.*, 2018), context-dependent mutation rates, strand-specific mutation biases, or lateral gene transfer events such as introgressions. Landerer *et al.* (2020) found that ROC-SEMPPR performed poorly on a budding yeast, *Lachancea kluyveri*, which is noted for having a large introgressed region (approximately 450 genes) with a higher GC% content than the rest of the genome (Payen *et al.*, 2009). By allowing the codon-specific mutation bias and selection coefficient parameters to vary between the ancestral and introgressed genes, ROC-SEMPPR obtained much better predictions of protein production rates in *L. kluyveri* (Landerer *et al.*, 2020).

GC-biased gene conversion (gBGC) has become a prevalent hypothesis for explaining variation in non-adaptive nucleotide biases in species ranging from budding yeasts to humans (Duret and Galtier, 2009) (but see Liu *et al.* (2017)). By using empirically determined recombination rates, it is possible to control for the effects of gBGC when estimating selection on codon usage using the SMDE model (Harrison and Charlesworth, 2011); however, such data is generally unavailable for non-model species. Cope and Shah (2022) showed that unsupervised machine learning approaches can help deal with intragenomic variation in non-adaptive nucleotide biases, but better insight can be gained from more nuanced models that explicitly incorporate evolutionary processes such as gBGC.

Finally, although previous work has used the ASUMDE implementation ROC-SEMPPR to test for differences in natural selection within genes (Cope *et al.*, 2018; Cope and Gilchrist, 2022),

ROC-SEMPPR does not explicitly allow for differences in the direction of selection. As noted previously, selection on codon usage is hypothesized to be related to selection for elongation speed, translation accuracy, and mRNA secondary structure, among others. Some evidence suggests codons favored by one selective pressure need not be favored by another (i.e., the fastest codon need not be the most accurate) (Stoletzki, 2008; Shah and Gilchrist, 2010). ROC-SEMPPR and similar frameworks currently average over these processes, such that selection coefficients will reflect the dominant selective pressure. Models that are able to explicitly separate these selective pressures would greatly improve our understanding of the evolution of codon usage.

Concluding Remarks

The evolutionary biologist Theodosius Dobzhansky famously said (Dobzhansky, 1973) "Nothing in biology makes sense except in the light of evolution." Michael Lynch took this idea a step further, arguing that (Lynch, 2007) "Nothing in evolution makes sense except in the light of population genetics,": in essence, evolutionary outcomes are the result of microevolutionary processes. Population genetics thus provides null models against which to evaluate adaptive hypotheses (Bromham, 2009; Koonin, 2016). We agree with Dobzhansky and Lynch: codon usage bias does not make sense without the population genetics-based ASUMDE model and its extensions. Despite its limitations, the ASUMDE model is a sensible default null model on which to build more detailed models.

Even in non-evolutionary studies of the functions and mechanisms of codon usage, researchers must be cautious that many gene-level traits and processes are correlated with protein production rate, such that naive correlations may suggest a mechanistic or functional role for codon usage where none exists. In such cases, researchers must use more advanced statistical analyses, such as partial correlations. Researchers must also be careful not to over-interpret their results and be mindful of mechanisms, noting that codons are only "seen" as codons when translated by ribosomes. In addition to numerous technical advances, the field of codon usage will benefit from models that more realistically model coding sequence evolution.

In conclusion, we propose the following principles for making inferences about the functions, mechanisms, and evolution of codon usage:

1. Use additive selection-uniform mutation-drift equilibrium as the null model.

2. Control for gene expression.

3. Consider mechanistically how codon usage affects biological processes, starting with translation.

Acknowledgments

The authors would like to thank Elizabeth Ballou for comments and feedback on the manuscript. We thank the organizers and participants in the 2022 EMBO Workshop on Codon Usage, for giving us the opportunity to develop this piece. This work was supported by the Biotechnology and Biological Sciences Research Council [BB/S018506/1 to E.W.J.W.]; the National Science Foundation [DBI 1936046 to P.S.]; the National Institutes of Health [R35 GM124976 to P.S., NIH-IRACDA Fellowship to Rutgers University to A.R.C.].

Table 1: Partial and Semi-Partial Pearson correlation coefficients between nuclear mRNA abundances and CBI when accounting for the total mRNA abundance across three strains.

Strain	Partial Pearson	P-value
FGSC4200 (wild-type)	-0.12	5.93e-27
FKH1	-0.18	1.69e-58
set_2	-0.18	3.73e-61
Strain	Semi-Partial Pearson	P-value
FGSC4200 (wild-type)	-0.10	1.65e-20
FKH1	-0.16	2.55e-46
set_2	-0.16	2.07e-48

References

Agresti, A. 2002. *Categorical Data Analysis*. Wiley.

Akashi, H. 1994. Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136.

Akeju, O. J. and Cope, A. L. 2024. Re-examining correlations between synonymous codon usage and protein bond angles in e. coli. *Genome Biology and Evolution*, 16: evae080.

Andersson, S. G. and Kurland, C. G. 1990. Codon preferences in free-living microorganisms. *Microbiological Reviews*, 54: 198–210.

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. 2003. Genome-wide analysis of mrna translation profiles in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100: 3889–3894.

Ballard, A., Bieniek, S., and Carlini, D. B. 2019. The fitness consequences of synonymous mutations in escherichia coli: Experimental evidence for a pleiotropic effect of translational selection. *Gene*, 694: 111–120.

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D., and Tyers, M. 2007. Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLOS Biology*, 5: e154.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. 2013. Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology*, 9: 675.

Berg, J., Willmann, S., and Lässig, M. 2004. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4: –42.

Bloom, J. D. and Adami, C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evolutionary Biology*, 3: 21.

Bloom, J. D. and Adami, C. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. *BMC Evolutionary Biology*, 4: 14.

Box, G. 1979. *Robustness in the Strategy of Scientific Model Building*, pages 201–236.

Bromham, L. 2009. Does nothing in evolution make sense except in the light of population genetics? *Biology and Philosophy*, 24.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., and Komar, A. A. 2016. Synonymous codons direct cotranslational folding toward different protein conformations. *Molecular Cell*, 61: 341–351.

Bulmer, M. 1987. Coevolution of codon usage and transfer rna abundance. *Nature*, 325: 728–730.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129: 897–907.

Burns, D. M. and Beachamn, I. R. 1985. Rare codons in e. coli and s. typhimurium signal sequences. *FEBS Letters*, 189: 318–324.

Chamary, J. V. and Hurst, L. D. 2005. Evidence for selection on synonymous mutations affecting stability of mrna secondary structure in mammals. *Genome biology*, 6: 1–12.

Chan, L. Y., Mugler, C. F., Heinrich, S., Vallotton, P., and Weis, K. 2018. Non-invasive measurement of mrna decay reveals translation initiation as the major determinant of mrna stability. *eLife*, 7.

Chaney, J. L. and Clark, P. L. 2015. Roles for synonymous codon usage in protein biogenesis. *Annu. Rev. Biophysics*, 44: 143–166.

Clarke, B. 1970. Darwinian evolution of proteins. *Science*, 168: 1009–1011.

Cope, A. L. and Gilchrist, M. A. 2022. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. *BMC Genomics*, 23: 408.

Cope, A. L. and Shah, P. 2022. Intragenomic variation in non-adaptive nucleotide biases causes underestimation of selection on synonymous codon usage. *PLOS Genetics*, 18: e1010256.

Cope, A. L., Hettich, R. L., and Gilchrist, M. A. 2018. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochimica et Biophysica Acta - Biomembranes*, 1860: 2479–2485.

Crick, F. H. C., Griffith, J. S., and Orgel, L. E. 1957. Codes without commas. *Proceedings of the National Academy of Sciences*, 43: 416–421.

de Oliveira, J. L., Morales, A. C., Hurst, L. D., Urrutia, A. O., Thompson, C. R. L., and Wolf, J. B. 2021. Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. *Molecular Biology and Evolution*, 38: 3247–3266.

Dobzhansky, T. 1973. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35.

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

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102: 14338–14343.

Drummond, D. A., Raval, A., and Wilke, C. O. 2006. A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23: 327–337.

Dubreuil, B., Matalon, O., and Levy, E. D. 2019. Protein abundance biases the amino acid composition of disordered regions to minimize non-functional interactions. *Journal of Molecular Biology*, 431: 4978–4992.

Duret, L. and Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10: 285–311.

Erdmann-Pham, D. D., Duc, K. D., and Song, Y. S. 2020. The key parameters that govern translation efficiency. *Cell Systems*, 10: 183–192.

Eyre-Walker, A. 1996. Synonymous codon bias is related to gene length in escherichia coli: Selection for translational accuracy? *Mol. Biol. Evol.*, 13: 864–872.

Fitch, W. M. 1976. Is there selection against wobble in codon-anticodon pairing? *Science*, 194: 1173–1174.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. 2014. mrna-programmed translation pauses in the targeting of e. coli membrane proteins. *eLife*, 3: e03440.

Forrest, M. E., Pinkard, O., Martin, S., Sweet, T. J., Hanson, G., and Coller, J. 2020. Codon and amino acid content are associated with mrna stability in mammalian cells. *PLOS ONE*, 15: e0228730.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. 2002. Evolutionary rate in the protein interaction network. *Science*, 296: 750–752.

Fraser, H. B., Hirsh, A. E., Wall, D. P., and Eisen, M. B. 2004. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101: 9033–9038.

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., and Pilpel, Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 115: E4940–E4949.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., and Duret, L. 2018. Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and gc-biased gene conversion. *Molecular Biology and Evolution*, 35: 1092–1103.

Gilchrist, M. A. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24: 2362–2372.

Gilchrist, M. A. and Wagner, A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recylcing. *Journal of Theoretical Biology*, 239: 417–434.

Gilchrist, M. A., Shah, P., and Zaretzki, R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183: 1493–1505.

Gilchrist, M. A., Chen, W. C., Shah, P., Landerer, C. L., and Zaretzki, R. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7: 1559–1579.

Goodman, D. B., Church, G. M., and Kosuri, S. 2013. Causes and effects of n-terminal codon bias in bacterial genes. *Science*, 342: 475–479.

Gould, S. J. and Lewontin, R. C. 1979. The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London*, 205: 581–598.

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, 10: 7055–7074.

Grantham, R., Gautier, C., and Gouy, M. 1980. Codon frequencies in 1 9 individual genes confrm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8: 1893–1912.

Grosjean, H. and Fiers, W. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18: 199–209.

Gsponer, J., Futschik, M. E., Teichmann, S. A., and Babu, M. M. 2008. Tight regulation of unstructured proteins: From transcript synthesis to protein degradation. *Science*, 322: 1365–1368.

Han, J. D. J., Berlin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., and Vidal, M. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430: 88–93.

559 Hanson, G. and Coller, J. 2018. Codon optimality, bias and usage in translation and mrna decay.
560 *Nature Reviews Molecular Cell Biology*, 19: 20–30.

561 Harrison, R. J. and Charlesworth, B. 2011. Biased gene conversion affects patterns of codon usage
562 and amino acid usage in the saccharomyces sensu stricto group of yeasts. *Molecular Biology and*
563 *Evolution*, 28: 117–129.

564 Hershberg, R. and Petrov, D. A. 2012. On the limitations of using ribosomal genes as references
565 for the study of codon usage: A rebuttal. *PLOS ONE*, 7: e49060.

566 Hinds, P. W. and Blake, R. D. 1985. Delineation of coding areas in dna sequences through assign-
567 ment of codon probabilities. *Journal of Biomolecular Structure and Dynamics*, 3: 543–549.

568 Hiraoka, Y., Kawamata, K., Haraguchi, T., and Chikashige, Y. 2009. Codon usage bias is correlated
569 with gene expression levels in the fission yeast schizosaccharomyces pombe. *Genes to Cells*, 14:
570 499–509.

571 Hirsh, A. E. and Fraser, H. B. 2001. Protein dispensability and rate of evolution. *Nature*, 411:
572 1046–1049.

573 Hockenberry, A. J., Sirer, M. I., Amaral, L. A. N., and Jewett, M. C. 2014. Quantifying position-
574 dependent codon usage bias. *Molecular Biology and Evolution*, 31: 1880–1893.

575 Holtkamp, W., Kokie, G., Jager, M., Mittelstaet, J., Komar, A. A., and Rodnina, M. V. 2015.
576 Cotranslational protein folding on the ribosome monitored in real time. *Science*, 350: 1104–
577 1107.

578 Homma, K., Noguchi, T., and Fukuchi, S. 2016. Codon usage is less optimized in eukaryotic gene
579 segments encoding intrinsically disordered regions than in those encoding structural domains.
580 *Nucleic Acids Research*, 44: 10051–10061.

581 Ikemura, T. 1981. Correlation between the abundance of escherichia coli transfer rnas and the
582 occurrence of the respective codons in its protein genes: A proposal for a synonymous codon
583 choice that is optimal for the e. coli translational system. *Journal of Molecular Biology*, 151:
584 389–409.

585 Ikemura, T. 1982. Correlation between the abundance of yeast transfer rnas and the occurrence of
586 the respective codons in protein genes. differences in synonymous codon choice patterns of yeast
587 and escherichia coli with reference to the abundance of isoaccepting transfer rnas. *Journal of*
588 *Molecular Biology*, 158: 573–597.

589 Ikemura, T. 1985. Codon usage and trna content in unicellular and multicellular organisms. *Molec-*
590 *ular Biology and Evolution*, 2: 13–34.

591 Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and
592 Gottesman, M. M. 2007. A "silent" polymorphism in the mdr1 gene changes substrate specificity.
593 *Science*, 315: 525–528.

594 Komar, A. A., Lesnik, T., and Reiss, C. 1999. Synonymous codon substitutions affect ribosome
595 traffic and protein folding during in vitro translation. *FEBS Letters*, 462: 387–391.

596 Konigsberg, W. and Godson, G. N. 1983. Evidence for use of rare codons in the dnag gene and
597 other regulatory genes of escherichia coli. *Proceedings of the National Academy of Sciences of*
598 *the United States of America*, 80: 687–691.

599 Koonin, E. V. 2016. Splendor and misery of adaptation, or the importance of neutral null for
600 understanding evolution. *BMC Biology*, 14: 114.

601 Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D.,
602 and Church, G. M. 2013. Composability of regulatory sequences controlling transcription and
603 translation in escherichia coli. *Proceedings of the National Academy of Sciences of the United*
604 *States of America*, 110: 14024–14029.

605 Krasheninnikov, I. A., Komar, A. A., and Adzhubei, I. A. 1991. Nonuniform size distribution
606 of nascent globin peptides, evidence for pause localization sites, and a cotranslational protein-
607 folding model. *Journal of Protein Chemistry*, 10: 445–453.

608 Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. 2009. Coding-sequence determinants
609 of expression in escherichia coli. *Science*, 324: 255–258.

610 Kurland, C. G. 1992. Translational accuracy and the fitness of bacteria. *Annual Review of Genetics*,
611 26: 29–50.

612 Landerer, C., Cope, A., Zaretzki, R., and Gilchrist, M. A. 2018. Anacoda: analyzing codon data
613 with bayesian mixture models. *Bioinformatics*, 34: bty138.

614 Landerer, C., O'Meara, B. C., Zaretzki, R., and Gilchrist, M. A. 2020. Unlocking a signal of intro-
615 gression from codons in lachancea kluyveri using a mutation-selection model. *BMC Evolutionary*
616 *Biology*, 20: 109.

617 Li, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom
618 usage of synonymous codons. *Journal of Molecular Evolution*, 24: 337–345.

619 Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., Liti, G., Tian, D., Hurst, L. D., and Yang, S.
620 2017. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no
621 gc bias. *Nature Ecology and Evolution*, 2: 164–173.

622 Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity.
623 *Proceedings of the National Academy of Sciences of the United States of America*, 104: 8597–
624 8604.

625 McCandlish, D. M. and Stoltzfus, A. 2014. Modeling evolution using the probability of fixation:
626 History and implications. *The Quarterly Review of Biology*, 89: 225–252.

627 McCandlish, D. M., Epstein, C. L., and Plotkin, J. B. 2015. Formal properties of the probability of
628 fixation: Identities, inequalities and approximations. *Theoretical Population Biology*, 99: 98–113.

629 Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G. D., Cox, J., Geiger,
630 T., Lindner, A. B., and Pilpel, Y. 2019. Systematic detection of amino acid substitutions in
631 proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity.
632 *Molecular Cell*, 75: 427–441.

633 Nieuwkoop, T., Finger-Bou, M., van der Oost, J., and Claassens, N. J. 2020. The ongoing quest to
634 crack the genetic code for protein production. *Molecular Cell*, 80: 193–209.

635 Payen, C., Fischer, G., Marck, C., Proux, C., Sherman, D. J., Coppée, J.-Y., Johnston, M., Dujon,
636 B., and Neuvéglise, C. 2009. Unusual composition of a yeast chromosome arm is associated with
637 its delayed replication. *Genome Research*, 19: 1710–1721.

638 Pechmann, S. and Frydman, J. 2013. Evolutionary conservation of codon optimality reveals hidden
639 signatures of cotranslational folding. *Nature Structural and Molecular Biology*, 20: 237–243.

640 Power, P. M., Jones, R. A., Beacham, I. R., Bucholtz, C., and Jennings, M. P. 2004. Whole
641 genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of
642 escherichia coli. *Biochemical and Biophysical Research Communications*, 322: 1038–1044.

643 Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg,
644 D., Baker, K. E., Graveley, B. R., and Coller, J. 2015. Codon optimality is a major determinant
645 of mrna stability. *Cell*, 160: 1111–1124.

646 Purvis, I. J., Bettany, A. J., Santiago, T. C., Coggins, J. R., Duncan, K., Eason, R., and Brown,
647 A. J. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation
648 in vivo. a hypothesis. *Journal of Molecular Biology*, 193: 413–417.

649 Pál, C., Papp, B., and Hurst, L. D. 2003. Rate of evolution and gene dispensability. *Nature*, 421:
650 496–497.

651 Qin, H., Wu, W. B., Kreitman, J. M. C. M., and Li, W. 2004. Intragenic spatial patterns of codon
652 usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168: 2245–2260.

653 Radrizzani, S., Kudla, G., Izsvák, Z., and Hurst, L. D. 2024. Selection on synonymous sites: the
654 unwanted transcript hypothesis. *Nature Reviews Genetics*, 25: 431–448.

655 Roth, A., Anisimova, M., and Cannarozzi, G. M. 2012. *Measuring codon usage bias*, pages 189–217.
656 Oxford University Press.

657 Salis, H. M., Mirsky, E. A., and Voigt, C. A. 2009. Automated design of synthetic ribosome binding
658 sites to control protein expression. *Nature Biotechnology*, 27: 946–950.

659 Sejour, R., Leatherwood, J., Yurovsky, A., and Futcher, B. 2023. No ramp needed: Spandrels,
660 statistics, and a slippery slope. *eLife*, 12.

661 Sella, G. and Hirsh, A. E. 2005. The application of statistical physics to evolutionary biology.
662 *Proceedings of the National Academy of Sciences of the United States of America*, 102: 9541–
663 9546.

664 Shah, P. and Gilchrist, M. 2011. Explaining complex codon usage patterns with selection for
665 translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy*
666 *of Sciences of the United States of America*, 108: 10231–10236.

667 Shah, P. and Gilchrist, M. A. 2010. Effect of correlated trna abundances on translation errors and
668 evolution of codon usage bias. *PLoS Genetics*, 6: 1–9.

669 Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. B. 2013. Rate-limiting steps in yeast
670 protein translation. *Cell*, 153: 1589–1601.

671 Sharp, P. M. and Li, W. 1987. The codon adaptation index - a measure of directional synonymous
672 codon usage bias, and its potential applications. *Nucleic Acids Research*, 15: 1281–1295.

673 Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., and Sockett, R. E. 2005. Variation in the
674 strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33: 1141–1153.

675 Singh, G. P. 2015. Association between intrinsic disorder and serine/threonine phosphorylation in
676 mycobacterium tuberculosis. *PeerJ*, 2015.

677 Singh, G. P. and Dash, D. 2008. How expression level influences the disorderness of proteins.
678 *Biochemical and Biophysical Research Communications*, 371: 401–404.

679 Stoletzki, N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest
680 selection on mrna secondary structures. *BMC Evolutionary Biology*, 8: 1–9.

681 Subramaniam, A. R., Zid, B. M., and O'Shea, E. K. 2014. An integrated approach reveals regulatory
682 controls on bacterial translation elongation. *Cell*, 159: 1200–1211.

683 Tsai, C.-J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., and Nussinov,
684 R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and
685 distinct minima. *Journal of Molecular Biology*, 383: 281–291.

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

689 Walker, J. E., Saraste, M., and Gay, N. J. 1984. The unc operon nucleotide sequence, regulation
690 and structure of atp-synthase. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*,
691 768: 164–200.

692 Wallace, E. W. J., Airoldi, E. M., and Drummond, D. A. 2013. Estimating selection on synonymus
693 codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30: 1438–1453.

694 Walsh, I. M., Bowman, M. A., Santarriaga, I. F. S., Rodriguez, A., and Clark, P. L. 2020. Synony-
695 mous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness.
696 *Proceedings of the National Academy of Sciences of the United States of America*, 117: 3528–
697 3534.

698 Wang, Z. and Zhang, J. 2009. Why is the correlation between gene importance and gene evolu-
699 tionary rate so weak? *PLOS Genetics*, 5: e1000329.

700 Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P.
701 2016. Improved ribosome-footprint and mrna measurements provide insights into dynamics and
702 regulation of yeast translation. *Cell Reports*, 14: 1787–1799.

703 Wu, Q. and Bazzini, A. A. 2023. Translation and mrna stability control. *Annual Review of*
704 *Biochemistry*, 92: 227–245.

705 Wu, Q., Medina, S. G., Kushawah, G., Devore, M. L., Castellano, L. A., Hand, J. M., Wright,
706 M., and Bazzini, A. A. 2019. Translation affects mrna stability in a codon-dependent manner in
707 human cells. *eLife*, 8.

Yang, Q., Yu, C.-H., Zhao, F., Dang, Y., Wu, C., Xie, P., Sachs, M. S., and Liu, Y. 2019. erf1 mediates codon usage effects on mrna translation efficiency through premature termination at rare codons. *Nucleic Acids Research*, 47: 9243–9258.

Zalucki, Y. M. and Jennings, M. P. 2007. Experimental confirmation of a key role for non-optimal codons in protein export. *Biochemical and Biophysical Research Communications*, 355: 143–148.

Zalucki, Y. M., Power, P. M., and Jennings, M. P. 2007. Selection for efficient translation initiation biases codon usage at the second amino acid position in secretory proteins. *Nucleic Acids Research*, 35: 5748–5754.

Zalucki, Y. M., Gittins, K. L., and Jennings, M. P. 2008. Secretory signal sequence non-optimal codons are required for expression and export of β-lactamase. *Biochemical and Biophysical Research Communications*, 366: 135–141.

Zalucki, Y. M., Beacham, I. R., and Jennings, M. P. 2009. Biased codon usage in signal peptides: a role in protein export. *Trends in Microbiology*, 17: 146–150.

Zalucki, Y. M., Jones, C. E., Ng, P. S. K., Schulz, B. L., and Jennings, M. P. 2010. Signal sequence non-optimal codons are required for the correct folding of mature maltose binding protein. *Biochimica et Biophysica Acta*, 1798: 1244–1249.

Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., Liu, B., Ma, X., Zhao, F., Jiang, H., Chen, C., Shen, H., Li, H., Mathews, D. H., Zhang, Y., and Huang, L. 2023. Algorithm for optimized mrna design improves stability and immunogenicity. *Nature 2023*, 621: 396–403.

Zhao, F., Zhou, Z., Dang, Y., Na, H., Adam, C., Lipzen, A., Ng, V., Grigoriev, I. V., and Liu, Y. 2021. Genome-wide role of codon usage on transcription and identification of potential regulators. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

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

Zhou, Z., Danga, Y., Zhou, M., Li, L., Yu, C. H., Fu, J., Chen, S., and Liu, Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription.

Proceedings of the National Academy of Sciences of the United States of America, 113: E6117–E6125.

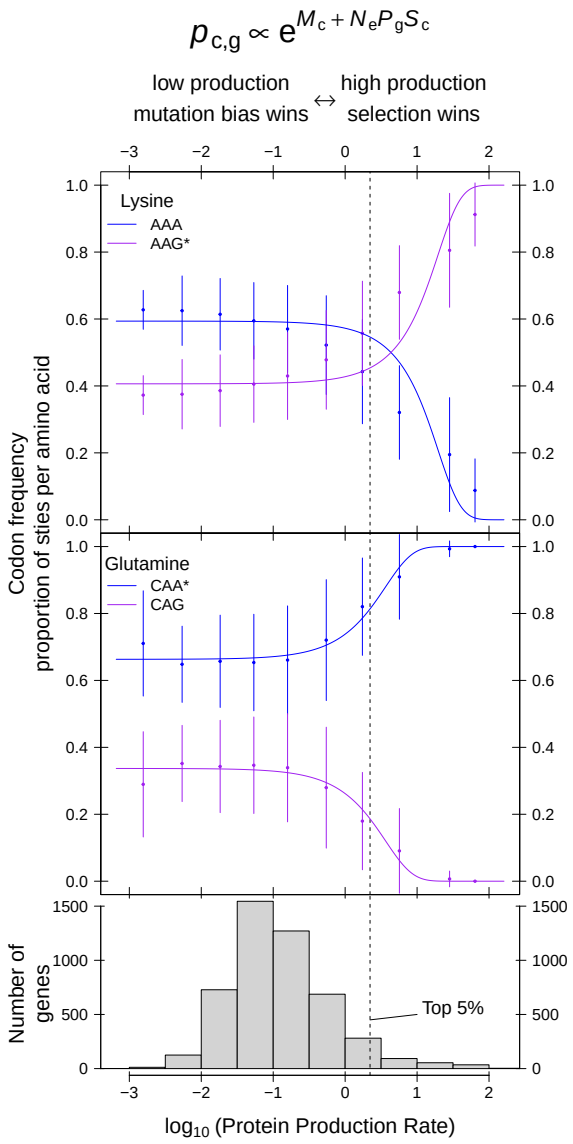

Figure 1: How codon frequencies change as a function of per-gene protein production rate in *Saccharomyces cerevisiae* yeast. Individual points and error bars represent the mean (± 1 std. dev) observed codon frequencies in genes binned based on empirical protein production rates taken from ribosome profiling data (Weinberg *et al.*, 2016). Solid lines represent the expected codon frequencies based on the additive selection-uniform mutation equilibrium. The dashed black line represents the 95^{th} percentile of empirical protein production rate values. The * indicates the codon favored by natural selection. In some cases, such as the amino acid lysine, the mutation and selection are biased towards opposite codons. As a result, the selectively-favored codon has lower frequencies in low-translation genes, but higher frequencies in high-translation genes. This contrasts with amino acids such as glutamine where mutation and selection are biased towards the same codon. In this case, the selectively favored codon is almost always used more frequently in low and high-translation genes, but this discrepancy grows for the latter genes. Natural selection has a major impact on only a small percentage of highly translated genes.

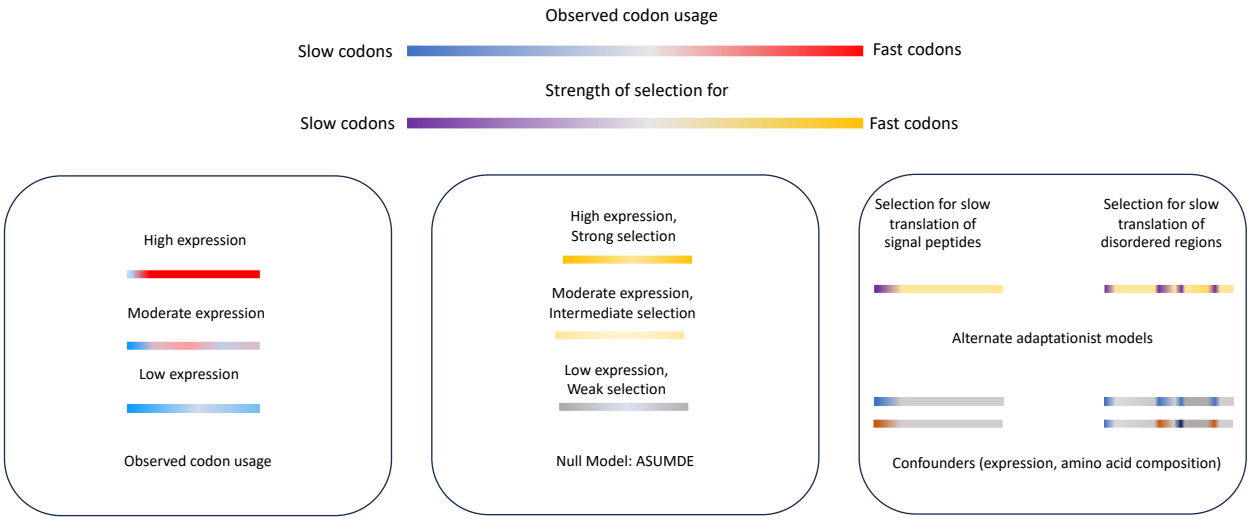

Figure 2: Models of selection on codon usage. The ASUMDE model of codon usage has selection on a coding region proportional to its protein production rate, and independent of position on the gene. Alternate models of position-dependent selection can be confounded by amino acid composition and gene expression.

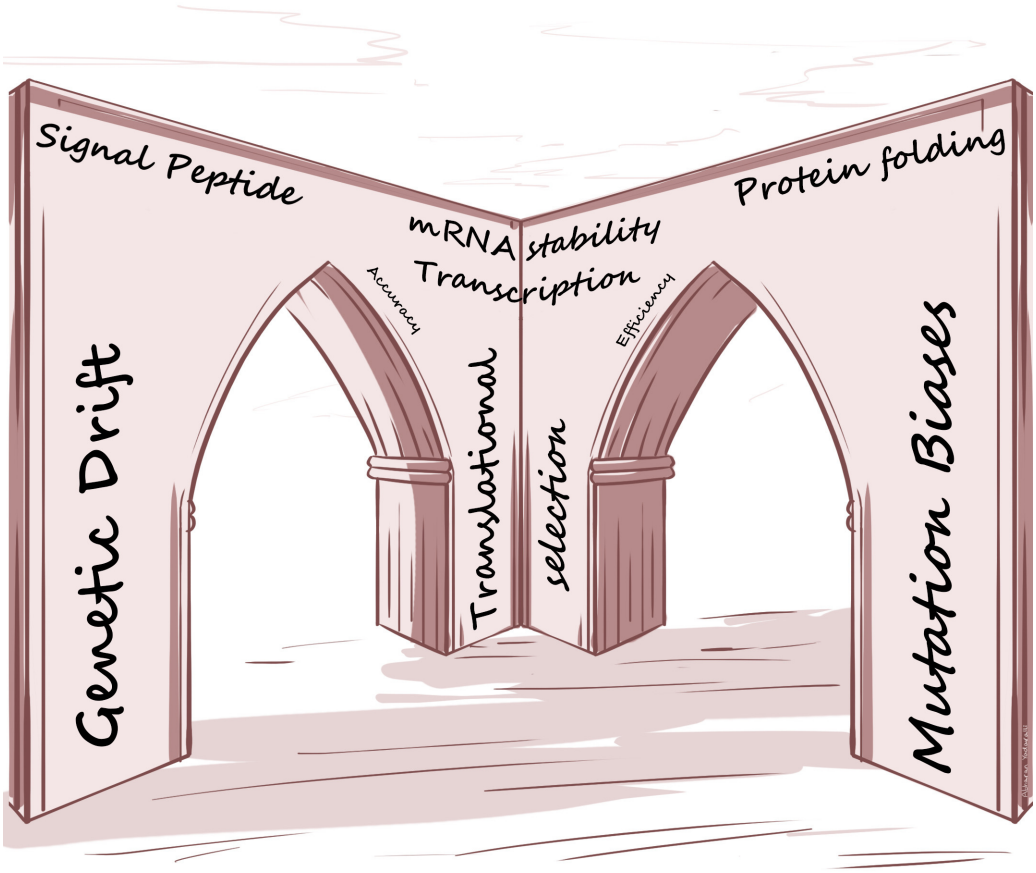

Figure 3: Molecular spandrels in codon usage bias. The major forces driving codon usage bias are translational selection, mutation biases, and genetic drift as quantified by the ASUMDE model. Other patterns detectable in codon usage may be consequences of the ASUMDE model, so this structural explanation should be excluded before getting overly excited about an alternative model.

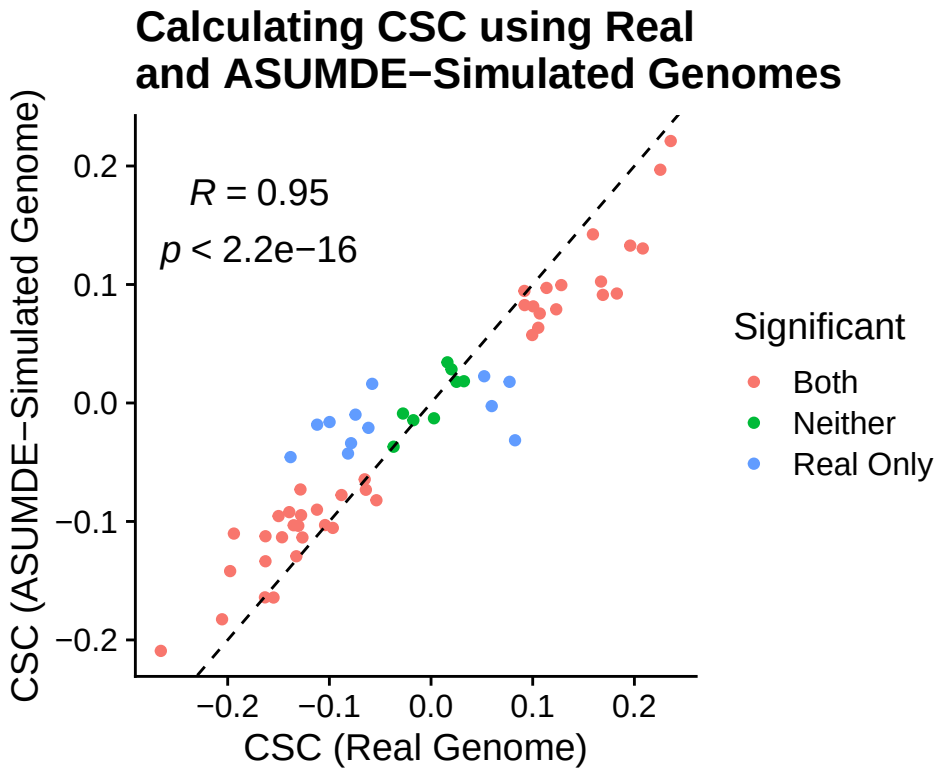

Figure 4: Comparing Codon Stabilization Coefficient (CSC) estimated for true and simulated protein-coding sequences in *S. cerevisiae*. Simulated protein-coding sequences used publicly-available ribosome profiling data from Weinberg *et al.* (2016). Colors indicate if the CSC value was significantly different from 0 in both, either, or neither of the real and simulated protein-coding sequences.