NORTHWESTERN UNIVERSITY

Sequence Determinants of Translation Efficiency

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Interdisciplinary Biological Sciences

By

Adam J. Hockenberry

EVANSTON, ILLINOIS

June 2017

ABSTRACT

Sequence Determinants of Translation Efficiency

Adam J. Hockenberry

Bacterial gene expression requires numerous steps that are energetically costly and tightly regulated. Following transcription of messenger-RNA, the translation of mRNA into protein is further regulated by a variety of sequence features both within and upstream of coding sequences. Collectively, these features contribute to the control of translation initiation, elongation, and termination rates that modulate protein abundances. Despite the near universality of the genetic code, species specific control sequences can help to impede or facilitate horizontal gene transfer across species boundaries and recombinant gene expression. A better understanding of the identity and effects of features that control protein translation can both enhance genetic engineering efforts and provide insight into the evolutionary pressures that have shaped bacterial genomes.

In the following work, I report on several efforts to increase our understanding of the link between mRNA sequences and protein translation rates. I develop several novel metrics to better quantify the effects of translation initiation motifs, as well as

synonymous codon usage biases, and find that these metrics can enhance predictions of genome-wide translation rates and protein abundances. In addition to quantifying the extent of the variation that can be explained by existing knowledge, these results provide insight into the constraints that features of translation initiation place onto the evolution of coding sequences and vice-versa. I further apply comparative genomic methods to show how that genome-wide variation in translation initiation and elongation related features are largely governed by the environments and growth strategies of different organisms.

Acknowledgements

Nothing that follows from this point onwards would have been possible without the mentorship of my two extraordinarily wonderful advisors: Michael Jewett and Luís Amaral. You both continually challenge me to be a better scientist, and provided endless support, flexibility, and inspiration that allowed me to succeed. So to Mike, Luís, and all past and present members of the Jewett and Amaral labs who made my time spent working here so enjoyable: thank you all from the bottom of my heart.

Several collaborators and co-authors were particularly integral to the work that I produced, so I want to specifically thank Irmak Sirer, Adam Pah, Aaron Stern, and Chuyue Yang for work that appears here. Additionally, Sophia Liu, Andrea Lancichinetti, Sarah Quillin, and Hank Seifert were all critical collaborators in ongoing and forthcoming work.

I want to thank my past-and-present committee members for valuable feedback: Neil Kelleher, Eric Weiss, Erik Andersen, and Erik Sontheimer. In addition, Jason Brickner and Curt Horvath have both provided advice and mentorship throughout my time here that has helped to make me a better scientist. An abundance of thanks also goes to Cathy Prullage and Justyna Gutowska for all of their hard work behind-the-scenes.

If it weren't for the overwhelmingly positive experiences I had working as a research technician, I likely never would have pursued this route, which has given me so much happiness and joy. So I want to specifically thank David Meaney and the entire Meaney Lab for turning me into a biologist so many years ago.

To the dear friends who have stood by me, if I began listing any of you now I'd inevitably leave someone cherished out and be forever regretful. So you know who you are. Some of you have been at my side for decades and are more family than friends, while others have been a daily part of my life here at Northwestern.

It is rather difficult to express just how much any words of praise and gratitude will inevitably fail to do justice to what I owe to my parents John and Cheryl, and my brother Jason. Thank you for everything that you have done for me throughout each-and-every leg of each-and-every journey that has brought me here. A lot of people have made me a better scientist. You all made, and continue to make, me a better person.

Finally, I'll never be able to look back on my time here at Northwestern without my thoughts inevitably drifting in one direction. So this thesis is dedicated to the memory of the one girl who never would have been able to actually read it in the first place, but who would have adored taking it off the bookshelf and chewing it to pieces. With endless love Dakota, this thesis is as much yours as it is mine.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Translation of messenger-RNA molecules into proteins is one of the most energetically expensive cellular processes, and is essential for the maintenance, growth, and reproduction of organisms[1–4]. While active proteins are frequently the molecule of interest to researchers studying cellular processes, the ability to measure mRNA levels has historically out-paced protein level measurements[5–8]. However, investigation into a variety of species have shown that the relationship between cellular mRNA and protein abundances is modest, with frequently cited correlations on the order of 0.5[9–14]. These comparatively low correlations were the motivation behind much of the work that follows in this thesis. Why something so fundamental—the basic formulation of the central dogma of molecular biology: DNA-to-mRNA-to-protein—does such a poor job of explaining protein abundance variation was surprising to me. In hindsight, some of this surprise was mis-guided. The *expectation* of a perfect correlation between mRNA and protein abundances ignores many physical and evolutionary constraints as I will discuss throughout this introductory chapter. The remainder of this thesis will focus on asking how knowledge of mRNA sequences can be incorporated alongside knowledge of mRNA abundances to improve predictions of protein abundances. Much is wrapped up in the statement "knowledge of mRNA

sequences", and various pieces of this puzzle form the basis of individual Chapters 2-5 where I demonstrate my own attempts to encapsulate basic biological mechanisms into a predictive framework to better explain the sequence-based regulation of mRNA translation.

1.1. Control-points in bacterial gene expression

Precisely why mRNA and protein abundances correlate so poorly within individual cells is a matter of on-going debate. In particular, several studies have questioned these low correlations, suggesting that a combination of measurement error and time-delays between mRNA and protein abundance changes are partially responsible[15–18]. In addition, different proteins, as well as their corresponding mRNA transcripts, have different half-lives and this will ultimately affect steady state protein abundances even if the rates of transcription and translation are identical between different genes[14,19–22]. Further, multiple lines of research suggest that differential translation efficiency—i.e. the number of protein molecules produced from a given transcript over a given time-period—is highly variable between different genes and likely a dominant factor affecting the relationship between steady-state mRNA and protein levels[10,14].

Ultimately, the translation efficiency of a given gene is subject to physical and biological constraints. At one end of the spectrum, physical processes place an apparent limit on the maximum achievable translation efficiency, for instance: sequential movement along an mRNA relies on diffusion limited tRNA binding[23,24]. At the other end of the spectrum, resource utilization places a floor on the *expected* levels

of translation efficiency. The production of mRNA transcripts is both energetically costly and mutagenic; having a large number of inefficiently translated mRNAs will waste cellular resources on transcription when compared to a small number of well-translated mRNAs[4,25–27]. Between these two extremes, there are a number of reasons to expect that different genes will evolve according to different *optimal* levels of translation efficiencies[3]. Some genes are highly variable in their expression, being required only in particular instances[28,29]. Other genes—so-called 'house-keeping' genes—are expressed at near constant levels throughout the life of a cell. Among the variably expressed genes, it is sometimes advantageous to have rapid change in available active protein molecules whereas this is less important for others[30]. Additionally, absolute numbers matter in terms of expression noise. While a single, highly translating transcript may be sufficient to ensure the required protein production for a particular gene, transcriptional noise resulting from the poisson-like process of transcript production may make this an inviable strategy at the population level[10,31–33].

It is important to emphasize that the rates of transcription, translation, and degradation are not the result of an engineers optimal plan. Rather, the observed rates for each individual step in this process for a given gene are the result of a steady accumulation of fixed mutations through evolutionary processes. These processes *may* include natural selection of advantageous mutations that increase overall translation efficiency for a particular gene, but this assumes that the link between efficient protein production and organism fitness is sufficiently large[34–36]. Just as

plausibly, the observed rates associated with transcription, translation, and degradation may be largely shaped by the random process of genetic drift, which dominates when the selective advantage of an individual mutation is below a threshold related to the effective population size of the species[37–39]. Further, organisms are simultaneously evolving these rates for all genes at the same time and are thus subject to clonal interference whereby advantageous mutations resulting in more energetically efficient production can still be rejected owing to the simultaneous accumulation of different advantageous mutations in competing lineages[40,41]. These issues are further complicated by tissue-specific expression, as well as a variety of other specific regulatory mechanisms affecting temporal and spatial control of gene expression in multi-cellular eukaryotes[15,17,18]. The remainder of this thesis and the results presented herein will largely focus on microbial species, particularly bacteria.

From the discussion thus far, several critical facts have already arisen that are worthy of reiteration and synthesis. First, the rates of transcription, translation, and degradation can all be manipulated to regulate and maintain the production of a given protein at a target level of abundance. Second, there are certain limits that constrain these bounds owing to cellular energetics and the reliance on diffusion. Third, any given 'strategy' may have advantages including the robustness of ultimate protein abundances to noise arising from stochastic production of transcripts and the dynamic flexibility to alter levels according to environmental change. Finally, even if an ideal combination of rates could be envisioned for a particular gene, evolution

is a sequential and partially random process that will not always produce optimal results.

By illustrating all of these issues, I hope to make it clear to the reader why a one-size fits all approach is a mis-guided expectation, and that a strong correlation between mRNA and protein abundances would be perhaps more surprising than the relatively weak correlation that we observe for most species. The critical questions that we are left with are: if we look at the genomes of extant species, and were we able to perfectly measure the transcription, translation, mRNA and protein degradation rates for every single gene, what interesting *knowledge* about this organisms evolutionary history, biosynthetic capabilities, and role within the ecosystem could we glean from this *information*? At a more practical level, how might we leverage that knowledge to better *design* genes for biotechnology purposes?

1.2. The genetic code and synonymous diversity

Thus far, I've focused on the fact that rates of the various steps required for gene expression can vary from gene to gene. This section will elaborate on the molecular source of this variation, and describe how individual rates may be tuned for the *same* protein. The genetic code is redundant such that 64 possible nucleotide triplets (codons) code for just 20 amino acids (and stop signals). How this code evolved, and its possible selective advantage according to various biochemical and information coding properties has been the subject of much debate[42–45]. Nevertheless the consequences of this redundant code are widely known: there are as many as 6 *synonymous* codons that may code for certain amino acids. An abundance of research starting

with some of the very first sequenced genes has noted that individual organisms have what appear to be preferences for and against the use of particular synonymous codons[23,46]. These apparent preferences vary both from organism-to-organism, and between genes within the same organism[47,48].

On average, individual amino acids have 3.05 synonymous codons:

$$\frac{(3_{aas} * 6_{cods}) + (5_{aas} * 4_{cods}) + (1_{aa} * 3_{cods}) + (9_{aas} * 2_{cods}) + (2_{aas} * 1_{cod})}{20_{aas}} \tag{1.1}$$

where $aa(s)$ refer to the number of amino acids within a particular class and $cod(s)$ refer to the number of synonymous codons used in that class. A back-of-the-envelope calculation shows that even a comparatively small 100 amino acid protein will thus have an astronomical 3.05^{100} ($\approx 2.69 * 10^{48}$) unique nucleotide sequences capable of coding for it. Each of these hypothetical mRNA sequences may be translated and degraded at very different rates, and these rates may depend on organism-specific factors. In a seminal paper, Kudla *et al.* (2009) showed that wide-spread variation in protein production between synonymous gene constructs does in fact exist[49]. The researchers made 154 different synonymous constructs coding for Green Fluorescent Protein, cloned them into plasmids, and expressed them in *Escherichia coli*. Despite the fact that each synonymous gene was placed in the same plasmid backbone, with the same promoter and surrounding regulatory sequences, protein production varied by several orders of magnitude. A variety of studies have pointed to multiple possible explanations for why synonymous transcripts produce different amounts of

protein, including variation in: i) cognate-tRNA levels corresponding to individual codons that alter elongation patterns[23,50], ii) binding efficiency of tRNA molecules via wobble pairing[51,52], iii) tRNA physical interactions on the ribosome and di-codon biases[53,54], iv) differences in the tRNA supply and demand[55–57], v) RNA structures precluding translation initiation and/or disrupting proper elongation[49,58,59], vi) start codon efficiency[60], and vii) transcript degradation via accessibility to RNases[61,62].

While the above explanations all relate to the overall rate of protein production, differences in *active* protein production introduce further complexity. Research has shown that some synonymous codons are more likely to result in mis-incorporation of erroneous amino acids (that possibly disrupt proper protein folding/function) or premature termination[63–65]. Differential translational accuracy may explain a large portion of the synonymous codon usage bias observed in bacterial genomes, particularly with regard to highly expressed genes where the burden of mis-folded protein molecules can have toxic effects[66]. Additionally, protein folding can occur co-translationally. Even within a single protein, it is highly possible—and evidence suggests—that the translation rate of different regions is tuned not to ensure maximum speed, but rather proper protein folding, which may require purposefully slow translation around complex folds[67–69].

1.3. The interplay of translation initiation and elongation processes

Any discussion of codon usage bias and the expression differences arising from synonymous gene constructs is complicated by one rather important fact: increasing the rate at which the ribosome can translate mRNA may have absolutely no effect on the

amount of protein produced if translation initiation is a rate-limiting step in protein translation[70]. A given mRNA generally contains both upstream and downstream non-coding regions that contain important signals for regulating initiation, termination, and transcript degradation[71]. Such 'signals' can be the presence of particular nucleotide motifs that bind various proteins and mRNAs, or these motifs may be related not to the nucleotides themselves but to the physical structures formed by mRNA folding[72]. In the critical experiment of Kudla *et al.* (2009), these signals were held constant between different constructs—as the researchers only altered the protein coding region. But this viewpoint fails to account for important boundary effects. The same 5′ untranslated region (5′ UTR) will fold into a very different structure according to the sequence identity downstream of the start codon, which can have important effects[49].

In short, biology is not entirely modular: the same sequence in different sequence contexts may behave very differently. Indeed, the dominant conclusion arising from the study of Kudla *et al.* (2009) was that variance in predicted mRNA structure surrounding the start codon between different synonymous constructs was the dominant factor influencing protein production, with little or no measurable role for codon usage biases. However, there is still outstanding debate with regard to this conclusion[73,74]. Subsequently, a variety of research has confirmed the large degree of context dependency that must be accounted for when attempting to isolate variables associated with different regulatory processes[75,76].

1.4. Novel data, novel opportunities for understanding

The synonymous codon usage of endogenous genes for particular organisms often show distinct preferences for and against individual codons. These preferences vary from species to species with possible implications for recombinant protein production, as well as horizontal gene transfer and viral infectivity[77–79]. Methods to measure codon usage biases have shown that the most highly biased genes within a given organism tend to be expressed at the highest levels such that codon usage biases are highly predictive of protein abundances in microbial species (these effects tend to be much smaller for multi-cellular eukaryotes)[80]. Interpreting this correlation as causation is a classic statistical trap, but it is nevertheless indicative of interesting biology. Further, a larger set of research is beginning to show that certain synonymous codons do seem to translate at different rates from one another. Perhaps most notably, the novel experimental technique of ribosome profiling allows researchers to map the locations of ribosomes at the genome-scale and is shedding light on to the process of translation. While initial studies failed to show any association between the speed of individual codon translation and tRNA abundance patterns or apparent codon preferences, more statistically refined techniques have shown that ribosome profiling does support this long-held hypothesis[81–84]. At the same time, several other orthogonal experimental and systems-level techniques have also lended support to a relationship between codon usage preferences, tRNA abundances, and measurable differences in translation speed and/or accuracy[85–87].

The era of recombinant protein production and biotechnology has steadily shifted as a matter of degree into the era of synthetic biology. Whether it is transferring individual genes from one organism to the next, or *de novo* design of novel genes, the ability to produce a target protein at desired levels in an organism of interest is critical for countless applications[71]. There is thus a pressing need to streamline this process, and a better understanding of *how* individual organisms code their native genes has the potential to shed important light on our ability to rationally encode genes for predictable expression. Further, library-based synthetic biology approaches are generating vast amounts of data that may prove useful in answering many of these questions[75,86,88,89].

While the functional goal of being able to precisely dictate gene expression levels through DNA sequence manipulation is of interest to engineers and basic biologists alike, there is a second reason as to why a better understanding of the process of gene encoding is important: signatures of translational selection can provide important insight into the roles of genes within an organism and between organisms within a ecological community[80,90,91]. Meta-genomic sequencing efforts continue to expand our knowledge of species abundance and dispersal patterns in the natural world[92]. Despite notable progress in the ability to culture and manipulate increasingly diverse species, it is nevertheless difficult to imagine a near-future where transcriptomic, proteomic, metabolomic, etc. measurements are available for even a fraction of the known species in realistic growth conditions. However, as researchers uncover and refine robust relationships between sequence-based proxies (such as codon usage bias)

and protein abundances in well-studied species, we can gain confidence in predictions about the most highly expressed genes from comparatively less well-studied species based solely on their genome sequences.

The ability to 'read' a genome as more than a sequence of 'As', 'Ts', 'Gs', and 'Cs' is a critical goal for biologists. Given knowledge of a species DNA sequence, we may one day be able to glean important information about its functional capacity, its evolutionary history, and its role within the ecosystem. Increasingly accurate tools for functional annotation of DNA sequences will allow biologists to translate DNA sequence information into knowledge of biological processes, and signatures of translation efficiency have thus far played an important role in this endeavor[93].

The following work has been enabled by technological advances in systems and synthetic biology, and the enormous amount of data that these advances have generated. Namely, the growing number of genome-wide RNA- and protein-abundance datasets, the development and deployment of ribosome profiling as a method to measure the translation rates of individual genes and codons, and the vast increases in the number of fully sequenced genomes together provide a rich source of data that can be utilized to advance our knowledge of translation efficiency.

My initial investigation into the constraints governing how organisms encode genes via synonymous codon usage bias, and how knowledge of these constraints can be used to better predict translation efficiencies of native genes forms the basis of Chapter 2. The conclusions resulting from this work showed that a large and

under-appreciated constraint on synonymous codon usage biases is the need for efficient translation initiation. My research thus turned to how novel datasets such as ribosome profiling can be leveraged to better understand the relationship between DNA sequence features relating to translation initiation and measured translation efficiency. In Chapter 3, I describe novel methods to use ribosome profiling data in order to support long-standing models of translation initiation and to fill in previous gaps in our understanding surrounding the functioning of an important sequence motif—the Shine-Dalgarno sequence. This work focused on understanding variation in translation initiation efficiencies between genes from the same organism. In Chapter 4, I extend this work by looking across the bacterial domain, and investigating the factors governing the evolution of efficient translation initiation mechanisms and the Shine-Dalgarno sequence between diverse organisms. In Chapter 5, I return to the subject of coding sequence evolution, this time to show how translation initiation mechanisms constrain the usage of Shine-Dalgarno-like motifs within coding sequences and the consequence this has for translation efficiency. Finally, in Chapter 6 I discuss the implications of this work, and illustrate the challenges and opportunities that big-data continue to provide with regards to our understanding of mRNA translation.

CHAPTER 2

Quantifying position-dependent codon usage bias

This work was published with M. Irmak Sirer (co-first author), Luís AN Amaral, and Michael C Jewett in Molecular Biology and Evolution, 2014.

2.1. Abstract

While the mapping of codon to amino acid is conserved across nearly all species, the frequency at which synonymous codons are used varies both between organisms and between genes from the same organism. This variation affects diverse cellular processes including protein expression, regulation, and folding. Here, we mathematically model an additional layer of complexity and show that individual codon usage biases follow a position-dependent exponential decay model with unique parameter fits for each codon. We use this methodology to perform an in-depth analysis on codon usage bias in the model organism *Escherichia coli*. Our methodology shows that lowly and highly expressed genes are more similar in their codon usage patterns in the 5′ gene regions, but that these preferences diverge at distal sites resulting in greater positional-dependency for highly expressed genes. We show that position-dependent codon usage bias is partially explained by the structural requirements of mRNAs that results in increased usage of A/T rich codons shortly after the gene start. However, we also show that the positional-dependency of 4- and 6-fold degenerate codons is

partially related to the gene copy number of cognate-tRNAs supporting existing hypotheses that posit benefits to a region of slow translation in the beginning of coding sequences. Lastly, we demonstrate that viewing codon usage bias through a position-dependent framework has practical utility by improving accuracy of gene expression prediction when incorporating positional-dependencies into the Codon Adaptation Index model.

2.2. Introduction

The initial investigations into the usage of synonymous codons occurred nearly 40 years ago[23,46]. Since then, a large body of work has shown that bias in codon usage is widespread across diverse taxa[48] and related to a variety of factors including genomic base composition[94], mutational bias[95,96], and selection for or against particular sequence motifs that are used as control elements to differentially degrade or traffic mRNAs to particular areas of the cell[45,97]. Additionally, different species of tRNA vary in their genetic copy number, overall expression level, and affinities for their target codons[23,50,51]. Under the assumption that elongation rates may be diffusion limited in at least some cases, it has long been speculated that codon usage bias may impact both the speed and accuracy of translation. This, however, remains a controversial topic with experimental support on both sides[97–100].

The consequences of codon usage bias are equally as diverse as their origins. Computational studies have shown that codon usage bias may play a role in gene transfer between species[79] and protein folding[56]. Additionally there is experimental support showing that an understanding of codon usage bias is important for viral

defense and vaccination[77,78,101], resistance to environmental fluctuations in amino acid levels[102,103], temporal or cyclic control of gene expression[104–106], co-translational protein folding[67], and recombinant protein production[49,73]. While the vast majority of studies assume that codon usage bias is uniform along the length of genes, several reports dating back to the 1980s showed that codon usage bias in particular gene regions is distinct from others[88,107–114] including clusters of 'sub-optimal' or 'rare' codons at the beginning of genes[56,115,116].

There have been several proposed mechanisms as to why rare codons are enriched in the 5′ region of genes, with one positing that a region of slow translation (a 'translational ramp' or 'bottle-neck') at the beginning of genes helps to keep ribosomes evenly spaced and avoid collisions[116,117]. In parallel to this line of research, several computational and experimental reports in recent years have also highlighted the importance for reduced secondary structure surrounding the start codon[49,52,118–120], particularly for prokaryotic gene expression. More recently, researchers have drawn a critical link between codon usage and mRNA secondary structure and showed that the choice of synonymous codons can influence secondary structure[88,114,121] and that codon usage bias in the 5′ region of genes may modulate translation initiation in addition to elongation.

However, most studies to date have analyzed aggregate measures of codon usage (Codon Adaptation Index, tRNA adaptation index, etc.) that mask the potentially important contributions of individual codons. To illustrate why this may be problematic, we note that the decreased 'translational efficiency' (for which codon and/or

tRNA adaptation indices are often a proxy) in the beginning of gene sequences may simply be the result of one or two amino acids having inverted preferences in this region as opposed to a global phenomenon whereby all amino acids select 'slow' codons to modulate translation rate. This distinction could be critical for testing mechanistic hypotheses about evolutionary/mutational origins of codon usage bias as well as in designing recombinant proteins for optimal expression.

Further, most published studies also rely on bins of codons or an unnatural delineation between gene regions (i.e. the first 10 codons versus the rest of the gene) whose physical basis or statistical rationale is rarely discussed. Lastly, although researchers have known about the positional dependence of codon usage bias for years, to our knowledge all statistical models of codon usage bias fail to account for this effect. Thus, there is a disconnect between this knowledge in principle and its usage in practice.

To address these gaps, we sought to investigate position-dependent codon usage bias through a rigorous quantitative framework with a focus on the model organism *Escherichia coli*. We validate previous observations about heterogeneous codon usage with regard to position and expand on the established link between base composition, codon usage and mRNA structure. Further, we use model selection to determine a functional form to individual codon usage biases and observe an unexpected heterogeneity of parameters that should serve as a crucial test for any proposed mechanistic explanations relating to the origins of codon usage bias. We demonstrate that our revised understanding of codon usage bias, viewed through a position-dependent

framework, can be simply incorporated into existing codon usage models and used to increase predictability in gene expression. Finally, we show preliminary support that our results are likely not unique to *E. coli* by demonstrating that the position-dependent exponential decay model more accurately describes codon usage biases in a variety of organisms.

2.3. Results

2.3.1. Codon usage bias is not uniform with regard to position

To test whether there is position-dependent bias in codon usage preferences at the genome-scale, we performed a χ^2 test on 4,139 protein-coding genes from *E. coli* (NCBI/Genbank: NC_00913.2). Briefly, we aligned all the coding sequences at their start codon and partitioned the codons into 10 position-dependent gene regions such that each bin contained approximately 130,000 total codons (Fig. 2.1A, see *Materials and Methods*). To account for uneven gene lengths and maintain a similar number of codons per bin, as illustrated in Fig. 2.1A, bin width is progressively wider at distal sites. Within these bins, we counted the occurrences of individual codons and compared those counts to the expected mean and standard deviation calculated from a null model derived by using a synonymous codon shuffling algorithm. This method preserves overall codon usage and amino acid structure within each gene allowing us to quantify codon usage bias at all positions rather than simply codon usage. We then calculated the χ^2 statistic and determined the statistical significance of the observed values.

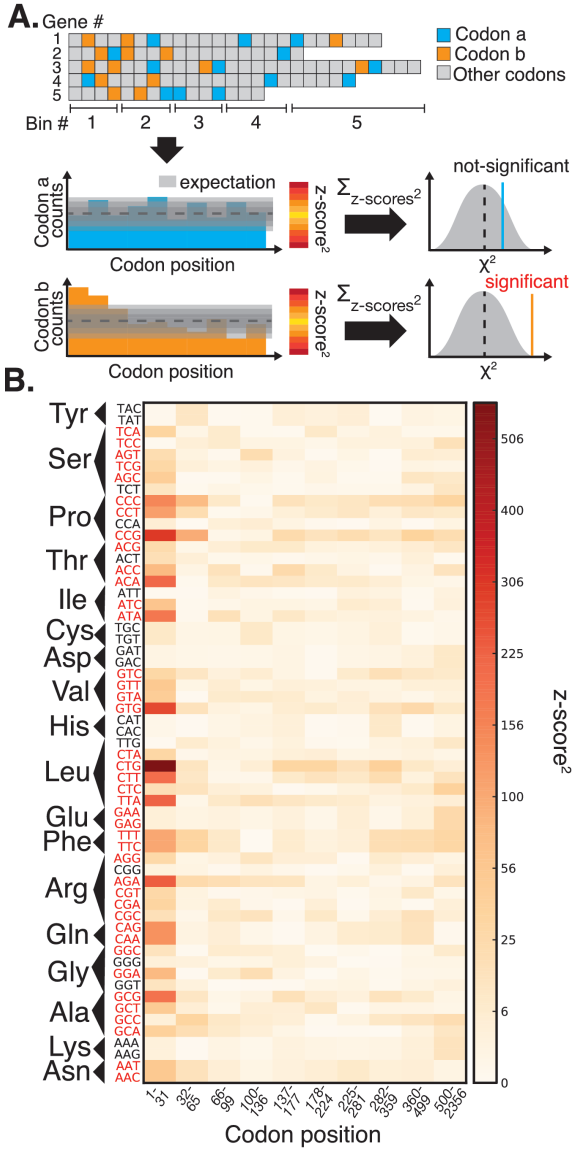

Figure 2.1. **Codon usage bias is not uniform with regard to intragenic position.** *A*, This cartoon schematic shows one codon that is used evenly throughout the toy gene-set (codon a, blue) and one codon that is not (codon b, orange). To statistically verify this, we align all genes at the 5′ region, group each codon into position-dependent bins, compare codon usage in each bin to random expectation, and sum the deviations over all bins. *B*, Squared z-scores of codon usage for *E. coli* as a function of position. Codons on the y-axis are grouped according to the amino acid they code for, and are labeled red if their usage bias is significantly non-uniform ($p < 0.00017$). Results for each bin are depicted according to the quadratically scaled color bar and the 10 bins are arranged from 5′ to 3′.

For 41 out of a possible 59 redundant codons, we found statistically significant ($p < 0.00017$) heterogeneous codon usage bias using this method (Fig. 2.1B, red codons). Further, visual analysis of the squared z-scores for each bin reveals that the observed deviations from uniformity are predominantly occurring in the $5'$ region of genes whereas there appears to be comparatively little heterogeneity in codon usage bias at distal sites. To make sure that these findings are robust and do not rely on a particular statistical test or binning scheme, we tried two different binning schemes (50 and 100 bins, 41 and 38 significant codons respectively, Supplementary Fig. A.5) and we performed 3 separate statistical tests for individual codons (all of which were compared against a synonymous shuffling null model): the position of median codon occurrence, the area under the curve of the cumulative distribution of codon usage with regard to position, and the size of the largest deviation from expectation in the cumulative distributions (see *Materials and Methods*). Using these tests, neither of which require data binning, we found that 24 out of 59 codons had significantly non-uniform codon usage bias in at least three out of the four tests and that 19 codons were significantly non-uniform in all four of our tests (see *Materials and Methods* and Supplementary Fig. A.1).

2.3.2. An exponential decay model most accurately describes patterns of codon usage bias

We extended the observation of non-uniform codon usage bias by testing the hypothesis that codon usage probability follows a specific functional form: uniform

(which assumes that codon usage bias does not vary with regard to position), linear, step-function (which would imply a distinct region of 5′ codon usage bias), and exponential decay. For each model, we used maximum likelihood estimation to determine the best-fitting parameters to the conditional codon probability data (the occurrences of the codon of interest divided by the occurrences of the amino acid of interest for all x values where x is the codon position inside of genes). We then used model selection based on Akaike Information Criterion (AIC)[122], which penalizes models with higher numbers of parameters, to determine which of the underlying models best describes *all* of the codon data in the *E. coli* genome (see *Materials and Methods*). We found strong evidence (odds ratio $\sim 10^{2263}$ relative to uniform) that an exponential decay model:

$$P_{a.a._j}(codon_i|x) = a_{i,j} \ \exp\left(-\frac{x}{\tau_{i,j}} \right) + c_{i,j} \tag{2.1}$$

provides the best description of codon usage in the *E. coli* genome where $codon_i$ refers to the ith codon that codes for the jth amino acid, $a.a._j$. Each parameter is specific to the individual codon and amino acid, hence the parameter subscripts i, j. For clarity, however, we will simply refer to these parameters in the general sense as a, c, and τ. The model parameters have straightforward interpretations: $a + c$ represents codon probability at the start codon, c is the asymptotic value that codon probability approaches, and τ is a measure of the distance over which the decay occurs. In Fig. 2.2A, we show example fits comparing the goodness of fit of the exponential decay and uniform models for the two phenylalanine codons.

Further, in Fig. 2.2B we show fits in the 5′ region (first 100 codons) for aspartic acid, phenylalanine, and glutamine to illustrate the heterogeneity of data and the best fitting forms for several 2-fold redundant amino acids (see Supplementary Figs. A.6-A.22 and Supplementary Figs. A.2 & A.3 for log-likelihood and AIC values for each codon). While aspartic acid exerts no positional-dependency, glutamine deviates sharply within a relatively short region of the gene-sequences while phenylalanine codons show a much slower decay with regard to position. The observation for aspartic acid and other amino acids such as histidine (Supplementary Fig. A.9) may be explained by the fact that the dominant codon at the genome-scale ends in a T[114] and is therefore unlikely to be further enriched in the beginning of gene sequences. In both of these cases, aspartic acid and histidine, the dominant codon in highly expressed genes is also in contrast to the dominant codon in the genome, but since highly expressed genes are relatively few in number the impact of this may be may be masked by genome-scale aggregation.

From this data, we also wish to make two further notes. First, if we restricted our analysis to a set of codon positions (e.g. the first 20, 50, 100), we would possibly miss valuable information: while 20 codons may be sufficient to encapsulate the positional heterogeneity for glutamine, it would be insufficient to faithfully evaluate phenylalanine. Second, this figure makes clear that the exponential decay model is not likely to be the *simplest* model to describe all amino acids. However, the exponential decay model is able to fit equally well to uniform and linear data-types (e.g. aspartic acid by having extremely large τ values), albeit with one or two possibly

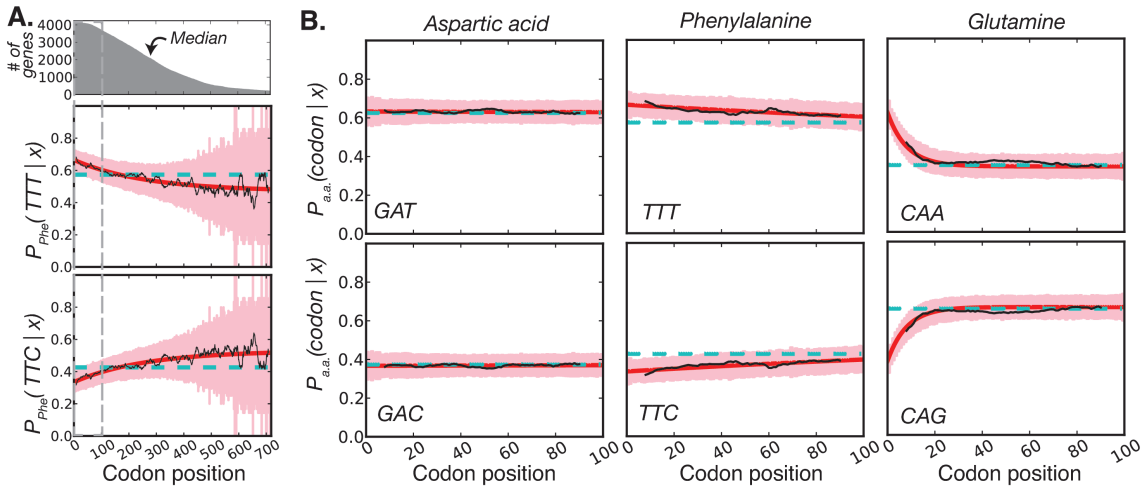

Figure 2.2. **The functional form of codon usage bias.** *A*, For the amino acid phenylalanine, we show the conditional probability of observing a codon as a function of position (black line, smoothed with a sliding window of 8 codons). We also show the best fitting exponential model (red) with corresponding 95% confidence intervals (pink) and the uniform model (cyan, confidence intervals not shown for clarity). The survival curve of *E. coli* gene lengths is highlighted at the top to illustrate the basis for increasingly wide-confidence intervals due to data sparseness at distal sites. *B*, Data for three different 2-fold redundant amino acids as in *A* but with the x-axis extending only to 100 codons to highlight heterogeneity in the 5′ region.

unnecessary parameters, whereas the uniform and linear models are simply unable to fit certain data (e.g. glutamine). For our model selection, we ask which model fits best for the set of *all* codons and arrive at the exponential decay model even though the fit is not necessarily the best/simplest for each individual codon (though it is the best for the overwhelming majority, see Supplementary Figs. A.2 & A.3). In fact, the heterogeneity of parameters that we observe between amino acids was striking and unanticipated.

2.3.3. Intragenic heterogeneity of codon usage bias is more pronounced in highly expressed genes

Most studies of intragenic codon usage bias have looked at the entire genomes of organisms. Since overall codon usage bias varies between genes from the same organism, certain *E. coli* genes may be contributing to the variation in intragenic codon usage bias more than others[108]. To test this hypothesis, we used a dataset of single molecule quantification of fluorescently tagged protein measurements collected under steady state growth conditions in rich medium at 30°C[10] to categorize low and high abundance proteins based on the top and bottom quartile of expression (see *Materials and Methods*, Supplementary Fig. A.23 for expression distribution). This delineation allows for sufficient separation of proteins such that there should be no overlap between these two bins and each bin still encompasses enough genes such that we have high confidence in the fits. Although this dataset only contains measurements for $\frac{1}{4}$ of the *E. coli* proteome, it is the largest proteome level dataset

for *E. coli* that we are aware of and covers a wide distribution of expression levels. We separately calculated the best fitting exponential parameters for each codon in each gene set. Using the three parameters of Eq. (1), we define a single metric — herein referred to as 'positional-dependency (pD)' — that encapsulates the degree and magnitude of the heterogeneity in usage bias for a $codon_i$ into a single number (Fig. 2.3A):

$$pD_{codon_i} = \int_0^L dx \Big[P_{a.a._j}(codon_i|x) - P_{a.a._j}(codon_i|L) \Big] \tag{2.2}$$

where L is the median gene length in the genome and $P(codon_i|x)$ obeys Eq. (1) with the parameter values obtained through maximum likelihood fits (see *Materials and Methods*). Essentially, our pD metric is an integral of the exponential function that is bounded by the median gene length, a limitation that we impose so as to have a high degree of confidence in the codon probability data, which gets increasingly noisy at distal sites. Positive values of pD correspond to codons used more frequently in the beginning of gene sequences, and negative values of pD correspond to codons used less frequently in the beginning relative to the end of genes.

We compared the absolute values of pD for all codons in lowly and highly expressed genes and saw that highly expressed genes have significantly greater positional-dependency in their codon usage bias compared to lowly expressed genes (Wilcoxon signed-rank test, $p < 0.0001$). Further, within both low and high expressing genes, we divided codons into two sets, which we term as 'rare' and 'abundant', according to their usage within a reference set of highly expressed genes[47]. By this definition,

'rare' codons are those whose frequency is less than random expectation in the reference set and 'abundant' codons are used at a frequency greater than expectation. We found a highly significant difference in pD values between these two codon sets within highly and lowly expressed genes (Wilcoxon rank-sum test, $p < 0.0001$ & $p = 0.0007$; Fig. 2.3B, top). Namely, the rare codons have positive values of pD and thus are enriched in the beginning of genes. This difference also persists when we use other metrics, such as the tRNA adaptation index[51], to classify codons (Supplementary Fig. A.24) and other delineations of lowly and highly abundant proteins such as the bottom and top 50% of protein abundances (Supplementary Figs. A.23 & A.25).

We also split codons into sets according to the identity of the third position base: A/T or G/C (Fig. 2.3B, bottom). Again, the difference between these sets was significant for both low and high abundance protein sets ($p < 0.0001$ and $p = 0.0006$, respectively) suggesting that the base composition of codons may play a role in determining the positional-dependency of codons and that this phenomenon is equally important in lowly and highly expressed genes.

For each codon, we have probability values as a function of position in both the low abundance and high abundance protein sets. This allows us to compute the difference between these gene sets for a given codon at two positions, the beginning of gene sequences and a distal site for which we use the median length *E. coli* gene:

$$\Delta_{beginning} = |P_{a.a._j}(codon_i|x = 1)_{high} - P_{a.a._j}(codon_i|x = 1)_{low}| \tag{2.3}$$

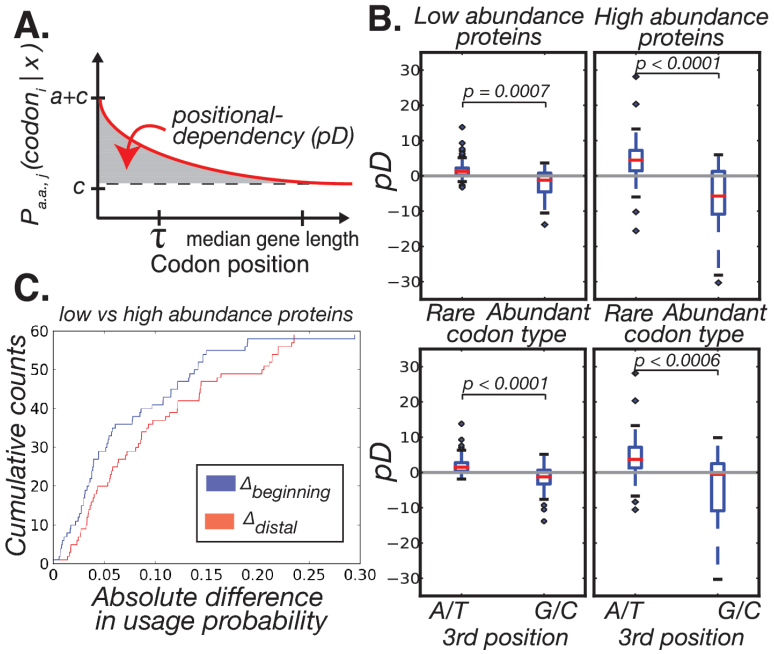

Figure 2.3. **The effect of gene expression on position-dependent codon usage bias.** *A*, Illustration of the 'positional-dependency (pD)' metric and exponential parameters. *B*, pD of codons in the genes of low and high abundance proteins split according to codon prevalence (top) and 3rd position base (bottom). We observe a significant difference in absolute pD of the codons between the two gene sets as well as differences within each gene set according to rare and abundant codons. Within gene sets, we also observed significant differences in pD between codons that end in A/T versus those that end in G/C. *C*, For each codon, we took the absolute difference in codon probabilities between the low and high abundance protein data sets and did so at two different points, the beginning of sequences and the median. Shown are the cumulative distributions of these differences.

$$\Delta_{distal} = |P_{a.a._j}(codon_i|x = 281)_{high} - P_{a.a._j}(codon_i|x = 281)_{low}| \tag{2.4}$$

In Fig. 2.3C we show that the cumulative distribution of these absolute differences. We observe that differences at the 5′ end ($\Delta_{beginning}$) are smaller in magnitude as compared to the absolute differences in codon probabilities at a distal site (Δ_{distal}) (Wilcoxon signed-rank test, $p = 0.0115$). Thus, in *E. coli* lowly and highly expressed genes are more similar in their codon usage biases at the beginning of gene sequences than at distal sites. Assessing the generality of this finding will require high quality proteome level datasets for other organisms that can be used to replicate this analysis.

2.3.4. Codon usage directly affects mRNA structure

To investigate the mechanistic basis for our findings, we next considered the effect of codon usage on mRNA structure. Several recent studies have illustrated that minimal secondary structure surrounding the start codon is important for translation initiation[49,88,114,118]. Throughout the rest of the mRNA sequence, this constraint does not exist and, in fact, strong mRNA structure may be important for regulating mRNA half lives[123]. Given that structural demands are position dependent, we sought to determine whether codon choice affects structure and thus whether this constraint may be a factor promoting position-dependent codon usage bias[88,114,121].

We therefore investigated the base pairing probability for each nucleotide in each gene within the high abundance protein set (calculated from the Boltzmann ensemble of structures, see *Materials and Methods*). We show that compared to synonymously shuffled null-model counterparts, actual genes have significantly less structure in the

5′ region (Wilcoxon rank-sum test on positions +5 to +15, $p < 0.0001$, Fig. 2.4A). Additionally, we developed a synonymous shuffling method that preserves positional frequencies of codons (and thus GC content at each position) within the gene set (see *Materials and Methods*) and saw that this method also leads to significantly less pairing probability in this region ($p < 0.0001$) compared to the null model but still higher probability compared to actual genes ($p < 0.0001$). This method suggests that the codons enriched in the 5′ region of genes are less likely to participate in strong structural interactions.

Since evolution is an iterative process, we sought to understand changes to structure in response different types of mutations. We thus looked at the effect of all possible single synonymous substitutions in the first 12 codons on the folding energy of the -36 to +36 region of mRNAs from the highly abundant proteins (see *Materials and Methods*). In Fig. 2.4B (left), we show that random mutations in this region are likely to increase structure, again verifying a selective bias for minimal mRNA structure around the start codon. As we expected, single synonymous substitutions from G/C -> A/T ending codons are more likely to decrease or maintain the structural properties of mRNA compared to 3rd position A/T -> G/C substitutions, which result in increased structure (Wilcoxon rank-sum test, $p < 0.0001$)[124]. Interestingly, we also find that synonymous mutations from abundant -> rare codons are less likely to introduce structure in the 5′ region compared to mutations from rare -> abundant codons (Wilcoxon rank-sum test, $p < 0.0001$) suggesting that the usage of rare

codons helps to maintain minimal secondary structure in this region, likely a result of their base composition which supports recent findings[88,114].

We repeated the above mutation simulation for a region distal to the initiation codon (+36 to +108, mutating the 12 codons from +72 to +108 region for direct comparison to our findings in the initiation region). At these distal sites, we confirmed that random mutations tend to decrease structure (Fig. 2.4B, right). In contrast to the 5′ region, distal gene regions are more likely to tolerate substitutions which preserve their strong structure (i.e. substitution to G/C rich and/or abundant codons). This analysis supports our hypothesis that synonymous codon choice affects mRNA structure and that requirements for reduced structure in the 5′ region of transcripts may result in selection for a unique codon set. As opposed to previous studies[121] that investigated structural robustness with regard to transcriptional fidelity, we show that robustness of the gene sequences to different substitutions depends on the position along a gene as well as the type of substitution. This likely has a mechanistic basis in translation initiation where mRNA structure around the start codon is potentially a rate-limiting barrier. Since most RNA structure is the result of local interactions, this effect should be applicable within a narrow window of codons/nucleotides that surround the start codon.

2.3.5. Position dependent bias in tRNA usage

The previous results, along with several recent studies[88,114], lend clear support for the hypothesis that mRNA structural constraints play an important role in shaping

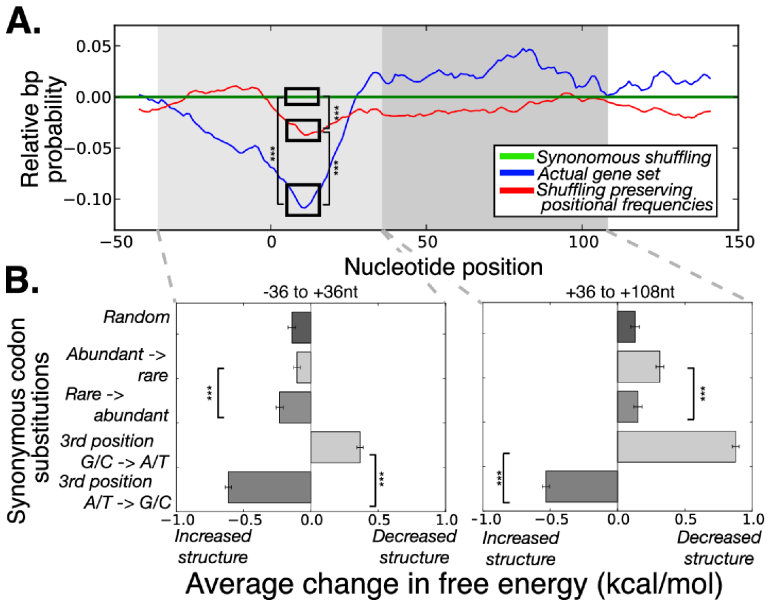

Figure 2.4. **The link between codon usage bias and mRNA structure.** *A*, We folded a 200mer (-50 to +150nt, relative to the start codon) region for each gene in the high abundance protein set and extracted the individual base pair probabilities. For clarity we illustrate median pair probabilities relative to the null model created by synonymous shuffling within genes (green). Actual genes (blue) and an alternative gene-set created by shuffling synonymous codons *between* genes in a manner that preserves positional biases (red) have significantly less structure in the $5'$ region (Wilcoxon rank-sum test on raw data, $p < 0.0001$ for all cases illustrated). *B*, We calculated the effect on folding energy of single synonymous codon substitutions in the genes of high abundance proteins. Left: the effect of substitutions in the $5'$ region (-36 to +36nt, relative to the start codon) is variable depending on the nature of the codon. Right: the same analysis for a region distal to the start codon (+36 to 108nt). For all cases illustrated, error-bars represent standard-error of the mean and $p < 0.0001$ according to Wilcoxon rank-sum test.

codon usage patterns. However, the parameter heterogeneity observed in Fig. 2.2, and in particular the large τ values — the length that it takes codon usage bias to reach its asymptotic value — that we found for some codons, suggests that mRNA structure alone is likely insufficient to explain all of the observed positional-dependencies.

In most cases, the 2-fold redundant amino acids are read by one tRNA species via wobble-rule base pairing so the results presented in Fig. 2.2 essentially represent variation in codon usage given a particular tRNA. Interestingly, we note that in *E. coli K12* the only 2-fold redundant amino acid to have two different tRNA anti-codons is glutamine, the amino acid with the sharpest positional-dependency in Fig. 2.2. To test for the possibility of a translational-ramp or bottleneck consisting of slowly translated codons at the $5'$ end, we turned to 4-fold degenerate amino acids, which are frequently read by at least two different tRNA species (one that predominantly reads purines [A & G] and another that reads pyrimidines [T & C] according to wobble-base pairing). Positional-dependency in these groups of codons would represent *between* tRNA variation in codon usage as opposed to the *within* tRNA variation that we previously observed for 2-fold redundant amino acids.

If AT/GC content variation is the main driver of codon usage patterns with regard to position, we expected that grouping the purines and the pyrimidines separately would lead to relatively uniform usage patterns with regard to position for these separate 'tRNA-classes', though we expect the class of codons read by rarer tRNAs to be less frequent overall as has been previously observed[23]. However, since the tRNAs

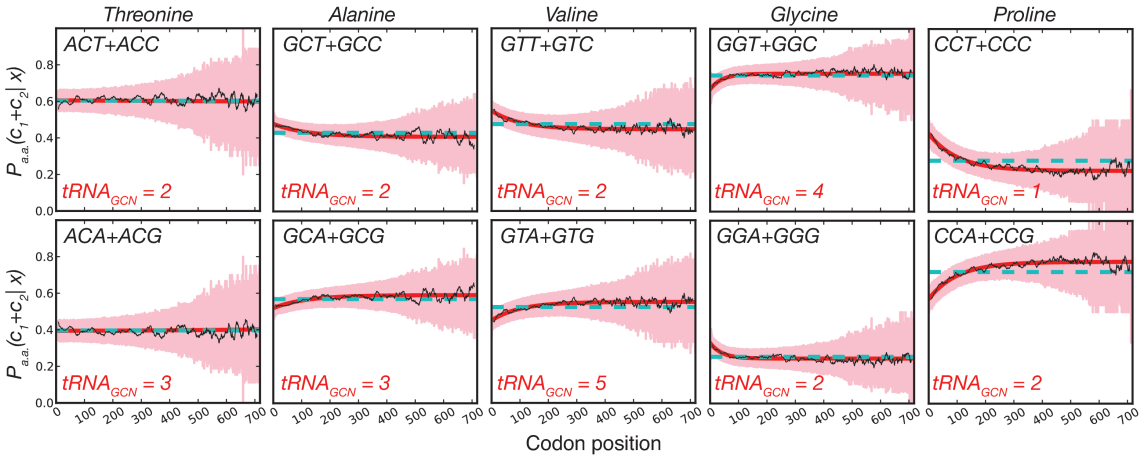

Figure 2.5. **Positional-dependency in codon groups and its association with cognate-tRNA gene copy number.** *A*, For all 4-fold redundant amino acids, we group codons into separate sets under the assumption that single tRNA species are more likely to read codons within these groupings according to wobble-base pairing than between groupings. We illustrate conditional probabilities as in Fig. 2.2 and highlight the gene copy number of the cognate tRNAs for each group ($tRNA_{GCN}$) to show that codons read by the rarer tRNAs are enriched in the 5′ region.

that read these two groups of codons are often present at different concentrations, if there is a benefit to slow translation in the 5′ region we would expect codons that are predominantly read by the less abundant tRNAs to be enriched in this region. What we observe for nearly all cases is that the rarer tRNA group (quantified by the cumulative gene copy number of the cognate tRNAs ($tRNA_{GCN}$) is indeed enriched in the beginning of coding sequences (Fig. 2.5). Further, the position-dependent usage of codons read by different tRNA species occurs over a relatively long range and not the narrow window that would be expected to influence mRNA secondary structure around the start codon. We repeated the above analysis for 6-fold redundant amino acids and reach the same conclusion (Supplementary Fig. A.26). While we did not observe any instance of codon groups read by abundant tRNAs being enriched at the 5′ end, there are several cases, such as for the amino acids threonine and serine, where we do not observe either enrichment or depletion of codon groups even though tRNA gene copy numbers are heterogeneous. While further investigation might resolve some of these differences, these data nevertheless suggest that in addition to structural requirements, codons read by rare tRNAs are enriched at the 5′ end of genes.

2.3.6. Intragenic codon usage bias can be used to more accurately predict gene expression

Our findings support a new understanding of codon usage bias: that codon preferences vary with regard to intragenic position, that this variation is partially but not

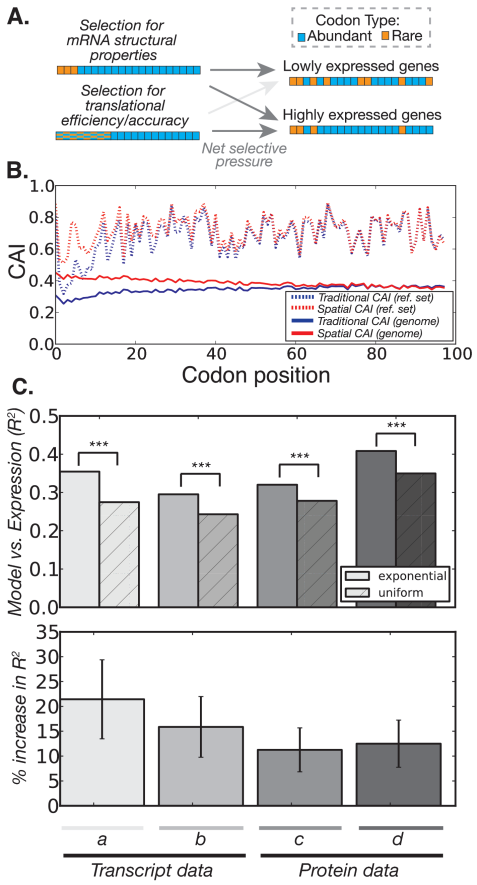

Figure 2.6. **Accounting for position-dependent codon usage leads to superior estimates of gene expression levels.** *A*, Our model posits that selection for reduced mRNA structure around the start codon acts strongly on all sequences relative to disruptive processes such as genetic drift and mutational biases. However, preference for accurate and efficient translation, is a second and weaker effect that is largely apparent in highly expressed genes and becomes stronger distal sites. *B*, Rather than calculate the CAI for each gene, we aligned genes at the start codon and calculated the CAI score for each *position* in either the reference set or genome. The dip in adaptedness after the start codon for both data sets (blue) is corrected by using exponential fits to the codon usage in the reference set (red). *C*, For two datasets of transcript abundances[10,127] and two datasets of protein abundances[10,13], we show that the R^2 correlation coefficient between the CAI and gene expression data is increased when using exponential fits to calculate the CAI as opposed to the traditional uniform assumption. Top, raw values; bottom, % increase. Error bars show standard deviation from 10,000 boot-strap re-sampled sets (paired t-test, $p < 0.0001$ for all cases).

entirely based on the structural requirements of mRNA, and that intragenic variation is particularly pronounced in highly expressed genes. Others have noted[118], and our analysis corroborates, that there are differences in both codon usage and mRNA structure between lowly and highly expressed genes at the beginning of coding sequences. Here, however, we have shown that the magnitude of codon usage bias differences between lowly and highly expressed genes at the beginning of genes is smaller than the equivalent differences at distal sites — suggesting that the pressure for minimal mRNA structure in the region surrounding the start codon is relatively stronger than the need for efficient or accurate translation of individual codons. In lowly expressed genes, selection for accurate or efficient translation may be dwarfed by other evolutionary processes such as biased mutation and genetic drift. However, in highly expressed genes, the balance of these forces may be tipped in favor of selection for individual codons (Fig. 2.6A). If this is indeed the case, accounting for heterogeneity in codon usage preferences should improve the accuracy of existing codon usage bias models.

There are many strategies to identify and quantify codon usage bias[51,57]; here we attempt to incorporate these positional dependencies into one of the most popular methods: the Codon Adaptation Index (CAI)[47]. The CAI relies on a reference set of highly expressed[47] or highly biased[125,126] genes to determine a coefficient for each codon that is based on the frequency of codon usage in the reference set. The coefficient takes a single value for each codon in the classical approach corresponding to the uniform assumption of codon usage bias. In contrast, we fit our exponential

decay model, Eq. (1), to the same reference gene set and use these position dependent functions in place of the single value approach (see *Materials and Methods*).

First, we observed that the reference set of genes has highly skewed codon usage biases (Supplementary Fig. A.27), and show that calculating the CAI at each position within the reference set (rather than for each gene) leads to a noticeable dip in CAI shortly after the start codon (Fig. 2.6B, blue dashed line). To understand why this result is slightly paradoxical, it is important to note the rationale behind the CAI: the model is a distance metric that calculates how well the codon usage patterns of a given gene match the codon usage patterns of a reference set of genes that are known to be highly expressed. However, we have shown here that the codon usage patterns of the reference set are inadequately described by a single number for each codon, and therefore we hypothesize that the distance metric should account for position-dependent codon usage. This hypothesis makes a strong prediction: if the position-dependent codon usage biases are of physiological relevance, accounting for this should lead to more accurate predictions of gene expression. However, if the position-dependent codon usage biases that we observe in the reference set are over-fitting to noise or are simply of no consequence, we would expect our predictions of genome-wide transcript and protein abundances to be worse.

We thus utilize our exponential fits to the reference (training) set to come up with a position-dependent array of coefficients for each codon, termed the position-dependent CAI (pdCAI) model. One caveat with this methodology is that we limit our analysis to the final codon-position of the longest gene in the reference set, as

we are unable to say how codon preferences in our reference set of genes might extrapolate past this point. Thus, for a given test-set gene, we only include the codons up to position 705 in our calculation of the pdCAI (though we note this cutoff encompasses the entirety of $> 90\%$ of endogenous *E. coli* genes (Fig. 2.2A, top)). Otherwise, we follow the same mathematics and logic behind the original CAI and show that, as expected, our pdCAI model corrects the dip in codon adaptedness for both the reference set and the whole genome when the calculation is performed in a way that treats all codons of a given position as a gene (Fig. 2.6B, red lines).

It is still unclear whether our correction leads to superior estimates of physiologically interesting properties. Namely, the usage of a rare codon early in a gene sequence will boost the genes overall CAI score in our model while this usage will be penalized by the standard CAI. In Fig. 2.6C, we show that in two distinct datasets of *E. coli* transcript abundances[10,127] as well as two distinct datasets of protein abundances[10,13], our pdCAI model makes more accurate predictions than the traditional approach with percent increases in the range of 10-25% (bootstrap re-sampling followed by paired t-test, for all cases $p < 0.0001$). Further, in addition to providing robust improvements in predictive power across several datasets, this increase in predictive power is also robust to an entirely different choice of reference set[125] (see Supplementary Fig. A.28).

2.4. Discussion

The pervasive understanding of codon usage bias assumes that rare codons are 'suboptimal', and their usage is thus minimized in coding sequences, particularly those

of highly expressed genes. Our work suggests that this notion of globally 'optimal' or 'sub-optimal' codons is misguided and that observed codon preferences are actually the result of contrasting forces, the magnitude of which varies significantly with distance from the start codon. A codon may at once be 'optimal' with regards to translational efficiency and/or accuracy, but 'sub-optimal' with regard to secondary structure, all of which makes a blanket term of 'optimality' problematic in light of ours and other recent results[88,113,114].

By modeling individual codon probabilities, we uncover a unifying functional form to codon usage bias. We find an unexpected heterogeneity in the easy-to-interpret parameters for the exponential decay function for different codons within *E. coli*. These results question the utility and validity of defining the $5'$ region by an arbitrary window of codons surrounding the start codon and treating this region as 'distinct'.

We draw a link between codon usage and mRNA structure and support previous findings by showing that the conflicting demands for and against mRNA structure at different positions likely contributes to synonymous codon selection[88,114,121]. By itself, this is a rather unsurprising fact since RNA secondary structure is the result of base pairing interactions and synonymous codons are composed of different bases. However, statistical investigations to support this assertion have until very recently been lacking. Our methodology is distinct and complementary to several recent studies that have investigated this link, and we draw largely similar conclusions: codon choice has a clear impact on secondary structure and empirical codon usage

biases reflect competing demands for and against secondary structure at different gene positions[88,114,121].

In contrast to these recent studies that focus on the prominent role of mRNA structure in shaping 5′ codon usage biases, we also show that mRNA structural constraints are likely inadequate to account for the heterogeneity in codon usage biases that we observe. Nucleotides distal to the start codon are unlikely to participate in secondary structure around the initiation region, which made the observation that several codons vary in their usage at relatively distal sites seem paradoxical. However, we show that codons read by less abundant tRNAs are also enriched in the 5′ of coding sequences. This finding could be interpreted as support for the translational bottle-neck hypothesis whereby enrichment of rarely used codons in the beginning of coding sequences could serve as a mechanism to space out ribosomes during translation so as to avoid collision. Another possible mechanism for the observed positional-dependencies stems from the fact that different tRNAs vary in their misreading rates[63]. Errors in translation are likely to be more costly at sites distal to the start codon, and this could lead to stronger selection with increasing gene length[35]. Teasing apart these two possibilities will require further investigations.

We note that although position-dependent codon usage bias had previously been observed, the vast majority of literature on codon usage bias has either ignored this fact or treated it as relatively inconsequential[128,129]. Our framework allows position-dependent codon usage biases to be incorporated into existing models, which we demonstrate here by re-defining the popular Codon Adaptation Index. Our aim here

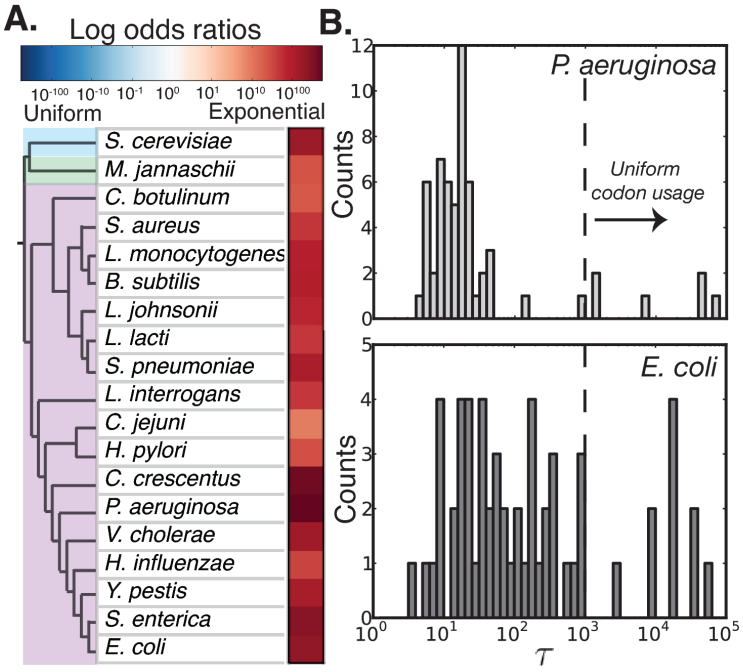

Figure 2.7. **Position-dependent codon usage bias in multiple organisms.** *A*, The observed log odds ratios for the exponential decay model fits relative to uniform model for different organisms. *B*, The distribution of τ values for *E. coli* and *P. aeruginosa* highlights potential differences in the evolutionary forces that have shaped the respective genomes.

is not to develop a model to predict protein abundances with maximal accuracy. Rather, we aim to show that the increased accuracy that we see is supportive of the fact that the 5′ usage of rare codons in the reference set and in the genome at large is likely beneficial in some regard. Additionally, this result allows us to show that positional-dependencies are far from inconsequential, and that they can be accounted for with relatively simple changes to existing models. We anticipate that more thoroughly investigating the pdCAI model with regard to different reference set choices and possible perturbations regarding how to most efficiently treat the decreasing confidence of our reference set fits at distal codon positions (which is particularly problematic for small gene-sets) may result in further improvements. All of the predictive improvements that we report are of endogenous transcript/protein levels, but the CAI model is frequently used in evaluating and designing recombinant proteins. The improvements that we demonstrate may therefore have utility in this field as well, though proper evaluation will require careful experimental controls to account for confounding issues such as mRNA structure around the start codon.

Additionally, although we focused here on intra-organism codon usage biases, the findings presented are likely not unique to *E. coli*. Towards this end, we repeated our model selection analysis on eighteen other microbial genomes randomly chosen to sample diverse taxa and found that the exponential decay model of codon usage bias is systematically selected as a better fit to the data than the uniform model (Fig. 2.7A, Supplementary Fig. A.4). While this fact alone may be unsurprising given known differences in AT/GC skews at the 5′ end of genes, further investigation of how

positional-dependency varies with organismal GC content, genome size, average gene lengths, etc. may reveal unexpected patterns. Additionally, we note as one example that the distribution of τ values for *E. coli* vary over a much larger range than equivalent values from *Pseudomonas aeruginosa* (Fig. 2.7B). Values of τ on the order of 10-100 are most likely indicators of the structural importance of mRNA, and since *P. aeruginosa* is a relatively G/C rich organism, we hypothesize that enrichment of A/T rich codons in the beginning of genes could conceivably account for the majority of positional-dependency that is observed for this organism. Conversely, large τ values (e.g. greater than 10^3) are the result of codons with little or no positional-dependency, of which there are far more in *E. coli* than *P. aeruginosa*. The generality of our method, and the ease of parameter interpretation suggest that comparative genomics investigations into inter-species parameter heterogeneity may yield novel insight into the forces that shape and constrain microbial genome evolution.

The effect of specific sequence features on a given gene's expression level is highly context dependent[71] and a multitude of factors shape the usage of codons within genes — many of which are undoubtedly particular to individual regulatory contexts or protein specific constraints[56,75,76,104–106]. However, we have uncovered a clear global pattern of codon usage within genes that is dependent on location and is partially related to differential requirements for mRNA structure. We anticipate that our results will be highly relevant in the field of synthetic biology and in genome engineering applications for which organism-specific sequence design is an important consideration. Further, the quantitative description of codon usage biases that we

have outlined here can help to serve as a testing ground for evolutionary investigations into the complex origins of codon usage bias within and between species.

2.5. Materials and Methods

2.5.1. χ^2 test of significance for uniformity in codon usage bias

After filtering out coding sequences that did not have recognizable start and stop codons, contained internal stop codons or non-standard bases, whose length was not a multiple of three, or was annotated as a pseudogene, we aligned genes at the start codon, removed the start and stop codons, and for each subsequent codon, calculated the χ^2 value:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O-E)^2}{\sigma^2} = \sum_{i=1}^{n} z^2 \tag{2.5}$$

where O is the observed counts per bin, E is the expected counts per bin, σ is the standard deviation of the expected distribution per bin, n is the number of bins, and z is the z-score per bin. We then compared this value to a chi-square distribution with degrees of freedom equal to $n-1$. A codon was deemed significant if the probability of observing that value: $p < 0.00017$ according to Bonferroni multiple-testing correction which is calculated from the number of tests (59) at a significant p-value $= 0.01$.

2.5.2. Binning schema

First, we lined genes up at the start codon and searched for an initial bin width that would contain approximately 130,000 codons (the entire *E. coli* genome contains approximately 1,300,000 codons thus 10 equal-sized bins required approximately 130,000 codons per bin). The algorithm starts with codon position one of all genes

and if there are $< 130{,}000$ codons then we add position two, etc. Once we found a bin width that contains $> 130{,}000$ codons we compare the bin with the previous width and choose the bin size that is closest to the target number (in this example: 130,000). We then start our next bin at the next position and iterate until the entire genome is partitioned with each codon position occurring in one and only one bin. In Fig. 2.1B, the first bin encompassed positions 1-34 of all genes and bins were progressively wider at distal regions to account for fewer genes and thus less data at these sites. It should be noted that one potential limitation of the χ^2 arises when bins contain fewer than 5 counts. In our published bin scheme, however, every bin for every codon has far more than actual 10 observations. Further, not content with selecting a bin scheme arbitrarily, we investigated a variety of other target bin numbers and sizes and found that these did not affect the results of Fig. 2.1 (Supplementary Fig. A.5).

2.5.3. Scrambling genes to determine expectation

For each gene, we followed a commonly used synonymous codon shuffling algorithm where codons that code for the same amino acid were randomly shuffled within genes. Thus, in a scrambled genome, each gene codes for the same amino acids and does so using the same frequency of each codon. This procedure allows us to preserve possible selection for or against particular codons or GC content within particular genes and to isolate the variable of interest, which in our case is the

deviation of spatial uniformity in codon usage bias. The expected counts in Fig. 2.1 were calculated from 200 scrambled genomes.

We also developed a novel synonymous codon shuffling method that we use to interrogate mRNA structure in Fig. 2.4: rather than shuffle synonymous codons within a gene, we allow for shuffling of codons between genes as long as the codons occur at the same position. This method preserves the amino acid structure of each gene, but not codon usage within genes. Rather, the method preserves positional codon frequencies of the gene *set* while introducing a similar number of codon changes per gene. Thus, were we to conduct the analysis in Fig. 2.1 using this as a null model, counts per bin for each codon would be identical between all shuffled genomes as well as in the actual genome.

2.5.4. Other statistical tests of codon usage bias

We performed 3 other statistical tests to determine whether any given codon was significantly non-uniform in its usage bias. All of the following required lining the genes up at the start codon as before, but neither require binning of codons which was necessary for the χ^2 test.

In the median test, we simply asked (for each codon) at which codon position the median codon in the genome occurs at. Thus if a codon appears 1000 times, we wanted to know at what position the 500th codon falls. We did this for the 200 scrambled genomes and found a discrete uniform distribution which allowed us to measure the deviation from the mean of this distribution that was observed in the

actual genome. A median closer to the start than expected would imply that this codon occurs more frequently in the beginning of genes than random expectation. The significance of this deviation was calculated via a two-tailed significance test.

In the area under the curve (*auc*) and the d-value tests, we relied on a cumulative distribution function (CDF) of codon counts where the x-axis is the absolute codon position re-scaled to 1 and the y-axis is cumulative counts of the codon of interest re-scaled to 1. If a codon occurred equally throughout the genome, and all genes were of equal length then perfect uniformity in usage would result in a diagonal line in the CDF and the *auc* would equal 0.5. A codon occurring more in the beginning would have an $auc > 0.5$ whereas a codon occurring at the end of genes would have an $auc < 0.5$. However, since genes are not of equal length, the *auc* was far greater than 0.5 due to the fact that few genes are represented at distant codon positions. However, we again assessed the significance of the actual genome findings by comparing against the *auc* for 200 scrambled genomes which resulted in a normal distribution of values to test our observed value against.

Lastly, using the CDF of scrambled genomes, we determined the 'average' CDF and found the absolute value of the largest deviation from this average CDF when plotted against the actual genome (the largest y-axis deviation regardless of where it occurred). Unlike the median and *auc* tests, the distribution of the randomized genomes was not normal since they were absolute values but a one-tailed test allowed us to determine the significance of the actual genome compared to the expectation

from 200 scrambled genomes. Crucially, we observed a large degree of overlap between these tests with the chi-square test being the most conservative estimate and the d-value the least (Supplementary Fig. A.1).

2.5.5. Maximum likelihood estimation of model parameters

When amino acid j (aa_j) is encountered at location x, the probability of codon i ($codon_i$) is defined by $P(codon_i|x, aa_j)$. We considered uniform, linear, step function, and exponential models for this codon usage probabilities. These models, each consisting of i functions (one for each codon i) of model parameters $\boldsymbol{\theta_i}$ and location x, are defined as:

$$\text{uniform: } P_i(\boldsymbol{\theta_i}, x) = \theta_{i1} \tag{2.6}$$

$$\text{linear: } P_i(\boldsymbol{\theta_i}, x) = \theta_{i1}x + \theta_{i2} \tag{2.7}$$

$$\begin{aligned} \text{step function: } P_i(\boldsymbol{\theta_i}, x) &= \theta_{i1} \text{ if } x < \theta_{i3}, \\ P_i(\boldsymbol{\theta_i}, x) &= \theta_{i2} \text{ if } x \geq \theta_{i3} \end{aligned} \tag{2.8}$$

$$\text{exponential: } P_i(\boldsymbol{\theta_i}, x) = \theta_{i1} \exp\left(\frac{-x}{\theta_{i2}}\right) + \theta_{i3}. \tag{2.9}$$

Note that for ease of following, in case of the exponential model we refer to θ_{i1} as a, θ_{i2} as τ and θ_{i3} as c in the main text.

We defined n_{ik} as the number of times we observe codon i at location $x = k$ among N_{jk} genes with amino acid j at $x = k$. The observed fractions y_{ik} of codon i usage at location k for amino acid j are obtained directly from these values: they are the ratios of n_{ik} (the number of observations of codon i at k) to N_{jk} (the number of possibilities to use codon i at k). Each n_{ik} is binomially distributed with the probability $P(codon_i|x = k, aa_j)$, giving rise to the probability density function (PDF):

$$f(n_{ij}|N_{kj}, P_i(\boldsymbol{\theta_i}\ , x)) = \binom{N_{jk}}{n_{ik}} P_i(\boldsymbol{\theta_i}\ , k)^{n_{ik}}(1 - P_i(\boldsymbol{\theta_i}\ , k))^{N_{jk}-n_{ik}} \tag{2.10}$$

Assuming that the n_{ik} values are statistically independent from each other, the log-likelihood function for the model parameters is:

$$\ln \mathcal{L}(\boldsymbol{\theta_i}\ |\mathbf{n_i}, \mathbf{N_j}) = \sum_i \ln f(n_{ik}|N_{jk}, P_i(\boldsymbol{\theta_i}\ , k)) \tag{2.11}$$

where $\mathbf{n_i}$ and $\mathbf{N_j}$ are the vectors comprised of n_{ik} and N_{jk} for all codon locations k.

For each codon, we estimated the parameters $\hat{\boldsymbol{\theta_i}}$ for each of the four models by finding the parameter set that maximizes this log-likelihood:

$$\hat{\boldsymbol{\theta_i}} = \text{argmax} \ln \mathcal{L}(\boldsymbol{\theta_i}\ |\mathbf{n_i}, \mathbf{N_j}) \tag{2.12}$$

For optimization, we used the *fmin* function of the *SciPy* scientific package for *Python* programming language, which utilizes a downhill simplex algorithm. To ensure that

the algorithm does not get stuck at local maxima, we performed each optimization 5 times, starting from different initial points.

2.5.6. Model selection

We used maximum likelihood estimation to determine the likelihood that our model fits individual codon data. To correct for the possibility of over-fitting, we used Akaike Information Criterion[122], a measure of goodness of fit for a statistical model that is grounded in information theory. It is defined as:

$$AIC = 2k - 2\ln\mathcal{L} \tag{2.13}$$

where k is the number of free parameters in the model, and $\mathcal{L}$ is the maximized likelihood for the estimated model. AIC is a relative measure of information loss caused by using the model to describe reality. The model with the minimum AIC value is the most likely model to minimize information loss compared to the underlying true process[130]. The relative probability $p_{M,AIC}$ of model M minimizing the information loss is given by:

$$p_{M,AIC} = \exp\left(\frac{AIC_{min} - AIC_M}{2}\right) \tag{2.14}$$

where AIC_{min} is the minimum AIC among all models, and AIC_M is the AIC of model M.

For all datasets of all organisms we investigated, we calculated the AIC value for each of the four tested models. First, we fit the codon usage probability function

using MLE to each codon. We obtained the log-likelihood for the entire model by summing the log-likelihoods of the individual fits. The total number of free parameters is the number of codons times the parameters in the model for a single codon. After calculating the Akaike Information Criterion values in this manner, we also calculated relative odds of each model to minimize information loss according to Eq. (2.14).

2.5.7. mRNA structural calculations

All free energy calculations were calculated using the RNAfold method of ViennaRNA[131] with default parameters. To extract the Boltzmann distribution of sequences we used RNAsubopt[131] and the -p 1000 flag.

For mutation studies, we used the transcript sequences of the 500 highly expressed genes. For each gene, we iterated through the codons within a region of interest (either 0 to +36nt, or +72 to +108nt), and if it matched the identified criteria (i.e. in the rare set) we swapped it to a synonymous counterpart with the desired criteria (i.e. in the abundant set). With 1 swap per gene, we re-folded and calculated the MFE of the structure and subtracted this from the original MFE for that sequence to determine the change in free energy from this substitution. We repeated this process for all applicable codons within the entire gene set to arrive at the distributions in Fig. 2.4.

For pair probability calculations, we created 5 separate scrambled genomes and aggregated the results in order to compare the actual pair probabilities to those

calculated from scrambled sequences. We fold each gene (-50 to +150 relative to start) and for each base calculate the number of sequences out of 1000 which that the base is paired. For each position we thus have a distribution of values (1 value representing the pair probability from each gene for that position) that we compare to the distribution created using synonymous shuffling algorithms.

2.5.8. Protein and transcript expression data

We downloaded the publicly available datasets of protein abundances in *E. coli* calculated from single molecule fluorescence counting Taniguchi *et al.* (2010)[10] and mass-spectrometry Lu *et al.* (2007)[13] and used the former dataset to classify proteins as low and high abundance due to the greater size of the dataset. After mapping genes back to the genome, we were left with a dataset of 1001 protein abundances that we split according to either the quartiles or median expression. Additionally, data of transcript abundances were downloaded from Taniguchi *et al.* (2010)[10] and Shiroguchi *et al.* (2012)[127] and again filtered for genes that we were able to map back to the genome. All of these datasets encompass only a sub-population of the transcriptome/proteome, but since each experimental technique has unique biases and limitations that restrict the sub-populations that they can measure, it is not safe to assume that genes which could not be quantified are either lowly or un-expressed. We thus only include genes for which measurements from the dataset in question exist.

2.5.9. Calculation of the Codon Adaptation Index

We make a slight alteration to the traditional calculation of the Codon Adaptation Index by replacing the frequency of codon i that codes for amino acid j ($X_{i,j}$) with a position dependent function $P_{i,j}(x)$. The function used here is the maximum likelihood estimation of the exponential decay function for each codon. The Relative Synonymous Codon Usage (RSCU) in our pdCAI is then:

$$RSCU_{i,j}(x) = \frac{P_{i,j}(x)}{\frac{1}{n}\sum_{i=1}^{n_j} P_{i,j}(x)} \tag{2.15}$$

which makes the weight of codon i also dependent on position:

$$w_{i,j}(x) = \frac{RSCU_{i,j}(x)}{RSCU_{i,max}(x)} \tag{2.16}$$

and the pdCAI:

$$pdCAI_{gene} = (\prod_{x=1}^{L} w_{i,j}(x))^{\frac{1}{L}} \tag{2.17}$$

When using the maximum likelihood fits of a uniform function, this result is analytically equivalent to the traditional CAI.

Additionally, based on the original formulation of the CAI, we use the RSCU of codons in the reference set to determine rare and abundant codons[47]. RSCU values less than 1 are categorized as rare and greater than 1 are categorized as abundant.

2.5.10. Calculation of the tRNA Adaptation Index

There are a number of ways to classify 'sub-optimal' and 'optimal' codons. We use the codon usage in a reference set of highly expressed genes to do so and adapt the nomenclature of 'rare' and 'abundant'[47]. However to demonstrate the robustness of this finding we also classify codons according to their tRNA adaptation index weights:

$$W_i = \sum_{i=1}^{n_j} (1 - s_{ij}) tGCN_{ij} \tag{2.18}$$

$$w_i = \frac{W_i}{W_{max}} \tag{2.19}$$

where n is the number of different tRNA species that read $codon_i$, $tGCN_{ij}$ is the gene copy number of the tRNA, and s_{ij} is a scaling factor to account for wobble interactions in anti-codon recognition[51]. Under this scheme, for each amino acid, we consider the lowest weight codon as 'sub-optimal' and the highest weight codon as 'optimal'.

CHAPTER 3

Leveraging genome-wide datasets to quantify the functional role of the anti-Shine-Dalgarno sequence in regulating translation efficiency

This work was published with Adam R Pah, Michael C Jewett, and Luís AN Amaral in Open Biology, 2017.

3.1. Abstract

Studies dating back to the 1970s established that sequence complementarity between the anti-Shine-Dalgarno (aSD) sequence on prokaryotic ribosomes and the 5′ untranslated region (UTR) of mRNAs helps to facilitate translation initiation. The optimal location of aSD sequence binding relative to the start codon, the full extents of the aSD sequence, and the functional form of the relationship between aSD sequence complementarity and translation efficiency have not been fully resolved. Here, we investigate these relationships by leveraging the sequence diversity of endogenous genes and recently available genome-wide estimates of translation efficiency. We show that—after accounting for predicted mRNA structure—aSD sequence complementarity increases the translation of endogenous mRNAs on the order of 50%. Further, we observe that this relationship is non-linear, with translation efficiency maximized for mRNAs with intermediate levels of aSD sequence complementarity.

The mechanistic insights that we observe are highly robust: we find nearly identical results in multiple datasets spanning 3 distantly related bacteria. Further, we verify our main conclusions by re-analyzing a controlled experimental dataset.

3.2. Introduction

The abundance of different protein species within a single cell can vary by several orders of magnitude, and multiple points of control are critical for tuning the expression of individual proteins over such a wide-range[1,10,49,71]. Transcription of the gene of interest is a necessary first step in the pathway of gene expression but, by itself, transcription is insufficient to ensure protein expression; studies in a variety of organisms have shown that mRNA abundances only modestly predict protein abundances[10–14,132]. The magnitude of these correlations remains open to debate, and part of the lack of a strong relationship between mRNA and protein abundances is likely a result of differential protein degradation rates and noisy measurements of both quantities[16]. It is, however, clear that the rate at which different mRNA species are translated into their protein product is variable and may be a significant source of variation in protein abundance and a point of regulation[71,133].

In studies dating back to the 1970s, researchers noted that a thermodynamic interaction between the 16S ribosomal-RNA and the 5′ untranslated region (UTR) of mRNAs is important for overall translation efficiency—defined here as the number of protein molecules made per mRNA per unit time—by enhancing translation initiation in prokaryotes[72]. The strength, optimal distance to the start codon, and structural accessibility of this anti-Shine-Dalgarno::Shine-Dalgarno (aSD::SD) sequence

interaction all play a crucial role in modulating the rates of translation initiation and thus protein abundances[134–138]. More recently, multiple studies have reinforced this paradigm and continue to elucidate the finer details about the importance of translation initiation signals, highlighting the fact that surrounding nucleotides may constrain SD sequence evolution due to mRNA structural constraints[75,76,88,114,139,140].

Much of our understanding about the process of translation initiation has come from experimental researchers expressing multiple genetic constructs with slightly varying 5′ UTRs placed upstream of a heterologous gene whose output is easy to quantify. However, most studies have looked at a relatively small number of such easily quantifiable genes that have been expressed in a small subset of experimentally tractable species, often at high-levels. Experimental studies present a well-controlled system to interrogate these mechanisms, but the degree to which these findings can be extrapolated more broadly to different genes, species, and expression levels remains largely unknown. Nevertheless, researchers ability to predict translation rates of heterologous genes have continually improved as more and more detailed experimental data is generated and incorporated into biophysical models[71,139].

In parallel, a number of different studies have analyzed various facets of translation initiation sequence variation across bacteria using bioinformatic or computational means, but definitions about which genes to consider as 'SD genes' vary broadly[71,141–147]. The main differences frequently concern where to look upstream of the start codon for a putative SD sequence and what bases of the 16S rRNA sequence to consider as the aSD sequence when assessing sequence complementarity to the 5′

A

	Assumed aSD sequence	5' UTR nts explored
Sakai *et al.* (2001)[1]	CUCCUGA	-23 ... -7
Ma *et al.* (2002)	AUCACCUCCUUA	-26 ... +5
Starmer *et al.* (2006)	GAUCACCUCCUUA	-28 ... +3
Chang *et al.* (2006)[1]	CCUCC	-20 ... -1
Salis *et al.* (2009)[2]	ACCUCCUUA	-35 ... -1
Nakagawa *et al.* (2010)	GAUCACCUCCUUA	-20 ... -5
Na *et al.* (2010)	UCACCUCCUU	-30 ... -1
Zheng *et al.* (2011)	UCACCUCCUU	-20 ... -1

B

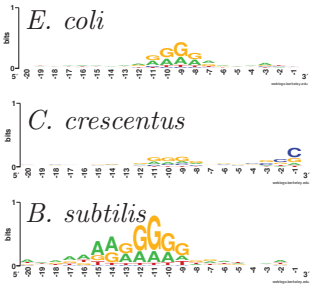

Figure 3.1. **SD sequence usage is variably defined in the literature and differs between genomes.** *A*, Several studies report a range of relevant parameters used to identify the aSD::SD sequence interaction. ([1]denotes studies that implicitly derive aSD sequences by extrapolating from over-represented UTR motifs; [2]denotes studies that explicitly penalize for non-optimal distances to the start codon). *B*, Sequence logos demonstrate that 5′ UTRs are highly non-random within a given species, largely a result of significant purine enrichment. However, the magnitude of this enrichment, and the spacing relative to the start codon varies between species despite widespread conservation in the 3′ end of the 16S rDNA.

UTR of mRNAs (Fig. 3.1A). Despite their differences, bioinformatic investigations have consistently shown that SD sequences occur much more frequently than random expectation in the 5′ UTRs of most species, further suggesting a large role for aSD sequence complementarity in regulating translation initiation (Fig. 3.1B).

Finally, as genome-scale and high-throughput sequencing technologies have come of age, a third route of investigation has become possible. By measuring the translational status of thousands of different genes within a single experiment, ribosome profiling (Ribo-seq) and RNA sequencing (RNA-seq) technologies can be combined to allow researchers to determine translation efficiencies across the genome[148]. Application of this technique to multiple organisms has already enhanced our understanding of translational regulation, stoichiometric protein production, determinants of elongation speed, and genome annotation[97,133,148,149]. However, in the context of bacterial translation initiation, several studies have suggested that the aSD binding strength shows no discernible relationship with the measured translation efficiency of endogenous genes at the genome-scale[133,149,150]. The negative results of these studies may be due to a variety of non-mutually exclusive factors, including: (i) noisy or inaccurate estimates of translation efficiency from these data, (ii) sub-optimal parameters associated with assessing the aSD sequence relationship, (iii) difficulty accounting for the effect of mRNA structures surrounding the start codon through computational means, (iv) the fact that many endogenous mRNAs are translationally regulated or present in operons, and finally (v) the lack of a relationship in these

data may be, quizzically, real—requiring researchers to re-think our understanding of the mechanisms governing translation initiation in bacteria.

Here, we investigate whether the sequence diversity of endogenous genes can be leveraged along with ribosome profiling-based estimates of translation efficiencies to precisely define the relevant parameters associated with aSD::SD sequence interaction. Rather than attempt to develop a comprehensive model to explain as much of the variation in translation efficiencies as possible, we instead propose a simpler question: can empirically measured translation efficiencies help us to better understand the particular phenomenon of aSD sequence complementarity and its role in regulating translation efficiencies? Our data-driven analysis yields definitions for the optimal distance between predicted aSD sequence binding and the start codon and the extents of the aSD sequence itself. We further highlight a highly conserved non-linear relationship between aSD sequence complementarity and translation efficiency of endogenous genes whereby intermediate complementarity maximizes translation efficiency downstream genes. We confirm these findings in multiple independent genome-scale and experimental datasets, and in doing so highlight the robustness of our conclusions while validating that the size of this effect is greatly enhanced as experimental steps are taken to reduce error in translation efficiency measurements.

3.3. Results

3.3.1. Deriving translation efficiency measurements from Ribo- and RNA-seq

For a given mRNA, ribosome density maps derived from ribosome profiling can be used to illustrate regions of relatively fast and slow translation. When used in conjunction with RNA-seq to estimate mRNA abundances, this groundbreaking technology allows researchers to roughly quantify relative translation efficiency (RTE) on a per gene basis for thousands of genes in a single experiment. However, it is important to note that estimates of RNA abundances and ribosome occupancies are both error-prone due to biological noise as well as the numerous steps in the experimental process that may introduce systemic bias[151–155]. Thus, RTE is a particularly noisy approximation because error is compounded when dividing two error-prone values. We therefore established several quality controls for gene inclusion that are stricter than those previously used in the literature (see Materials and Methods). Following on the previous work of others[133,149], we then calculated relative translation efficiency (RTE) per gene as:

$$RTE_i = \frac{RPKM_{Ribo-prof,i}}{RPKM_{RNA-seq,i}} \tag{3.1}$$

where $RPKM_{Ribo-seq}$ and $RPKM_{RNA-seq}$ are Reads Per Kilobase per Million mapped reads (RPKM) for a gene, i, obtained through ribosome profiling and RNA-seq, respectively. Using the original Ribo- and RNA-seq mappings provided by three separate studies in rich media for *Escherichia coli*, *Caulobacter crescentus*, and *Bacillus subtilis* we derived measurements of translation efficiency for 2910, 1833, and 2385 genes, respectively (Supplementary Fig. B.1)[133,149,156]. While this metric relies on some crucial assumptions, such as equivalent elongation rates between genes, prior work has shown that these assumptions are generally valid[133]; a noise-free RTE metric calculated in this manner should be highly correlated with 'true' translation efficiencies as we have defined it. We note that we investigated several variations in the above metric such as excluding the beginning and the end of genes, Winsorizing to limit extreme values, removing the lowest mRNA expression decile, etc. but none of these variations lead to distinguishably different results so for the purposes of this manuscript we opt for the simplicity of Eq. (3.1) moving forward.

As others have noted, mRNA structure surrounding the start codon is known to influence translation initiation, perhaps playing a dominant role in determining translation efficiency[49,118,133,137,140]. We confirmed this finding by showing that log-transformed translation efficiencies in all three organisms showed highly significant correlations with the predicted degree of mRNA secondary structure ($\Delta G_{folding}$) in the initiation region (defined here as -30 to +30 nucleotides relative to the first base of the start codon[which was labeled +1]) (R^2 = 0.13, 0.10, and 0.08 for *E. coli*, *C. crescentus*, and *B. subtilis*, $p < 10^{-42}$ for all cases). Given the strength of this

correlation (Supplementary Fig. B.2), we analyze the residuals from this predictive model (in units of log-scaled translation efficiency) in order to determine what role, if any, aSD sequence complementarity has in modulating translation efficiency:

$$r_i = RTE_i - \widehat{RTE_i} \tag{3.2}$$

where RTE_i is the relative translation efficiency of gene i, and $\widehat{RTE_i}$ is the estimate of RTE for gene i derived from the regression on $\Delta G_{folding}$ for each dataset. Put more simply, the residual RTE value for a gene is the difference in observed RTE minus the predicted RTE where our prediction is based off of the mRNA structure. We include this step to alleviate the source of biological variation associated with *cis*-structure, but note that these computational predictions also introduce error due to the—at best—modest correlation between computationally predicted structures and their counterparts as they exist *in vivo*[124]. Later, we show that all of our primary results remain significant, albeit with decreased magnitude when we skip this step and instead investigate RTE values directly.

3.3.2. Defining the optimal distance to the start codon and species specific aSD sequences

Using the residual RTE values described in Eq. (3.2), we took a systematic approach in order to determine where to look, in an unbiased manner relative to the start codon, for the statistical signal of aSD sequence complementarity under the assumption that the true value of this parameter should show the strongest correlation

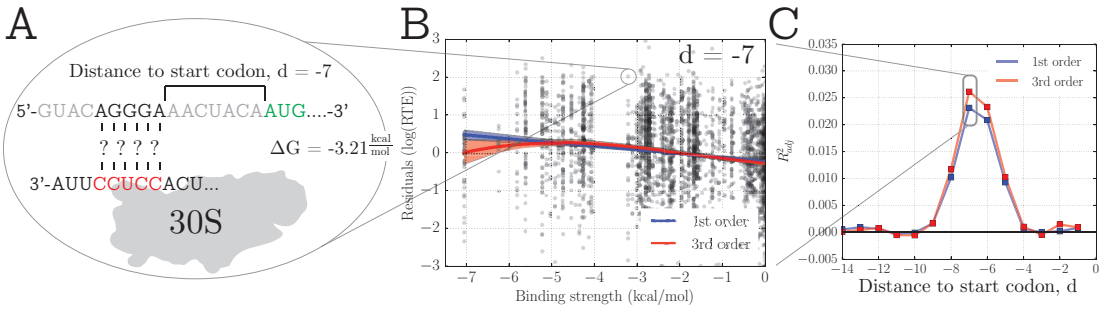

Figure 3.2. **Determining the optimal distance to the start codon.** *A*, Illustration of the method used in this study for determining the predicted Gibbs free-energy ($\Delta G_{binding}$) of the hybridization of the putative aSD sequence (highlighted in red) to the 5-nucleotide sequence at a distance of 7 nucleotides upstream from the start codon. *B*, The strength of aSD binding for each gene at a distance of -7 is correlated against the model residuals in units of $log(RTE)$. Shown are 1st and 3rd order polynomials ($R^2_{adj} = 0.023$ and 0.026 respectively, $p < 10^{-16}$ for both). *C*, We performed the same correlation analysis as in (B) for each putative distance to the start codon in the *E. coli* dataset for the given aSD sequence. Shown are the R^2_{adj} values for the relevant models with a maximum peak for d=-7.

between aSD sequence complementarity and residual RTE values. For each gene, we calculated the predicted hybridization energy of the core aSD sequence (5′-CCUCC-3′) to each sequential 5-mer upstream of the start codon (Fig. 3.2A). Hereafter, we will refer directly to the location (relative to the start codon) as the number of bases between the fragment analyzed and the start codon (this metric of distance corresponds to the aligned spacing presented by Chen *et al.*[135]). We asked how well the aSD sequence complementarity at a particular location for all genes performed at predicting residual RTE values via both linear and 3rd order polynomial regression.

In Fig. 3.2B we show example data for a distance to the start codon of -7 nucleotides (assessing complementarity of nucleotides -12 through -8 for each gene). We show both the 1st and 3rd order fits for the residual RTE data from *E. coli*, and find that both correlations are small yet nevertheless highly significant (F-test, $p < 10^{-16}$). Further, in Fig. 3.2C, we show the adjusted-R^2 (R^2_{adj}) resulting from repeating the correlations shown in Fig. 3.2B for each indicated distance relative to the start codon. We utilize the R^2_{adj} metric hereafter because unlike R^2 this adjusted metric penalizes for increasing parameter numbers associated with more complex 3rd order polynomial models and thus helps guard against over-fitting to the data. Despite the relatively small R^2_{adj} values, the sharpness of this peak shows that there is a clear and highly significant relationship between aSD sequence complementarity in the 5′ UTR of mRNAs and translation efficiency. The 3rd order polynomial model

was slightly more predictive at this stage, so we present our data in the form of 3rd order polynomial regressions hereafter except where otherwise noted.

Our choice of 5′-CCUCC-3′ as the aSD sequence in Fig. 3.2 was simply to illustrate our methodology by using the most conserved region of the 16S rRNA tail. In practice, it is not clear precisely which 16S bases belong to the aSD sequence although the 3′ tail of *E. coli* has been experimentally determined to end with 5′-...CCUCCUUA-3′. In order to see if the data would allow us to recover the expected aSD sequence, we repeated the above analysis for different putative aSD sequences extending in the 5′ and 3′ directions at different binding locations and observed increasing R^2_{adj} values and a slight re-positioning of the optimal distance to the start codon (Fig. 3.3A). It should be noted, however, that this change in the optimal distance is partially an artifact of our numbering scheme. As we include more 5′ bases in the definition of the aSD sequence, even if the location of optimal binding for a given mRNA does not change, the 'distance' will change based on the fact that it is calculated relative to the 5′ end of the putative aSD sequence (Supplementary Fig. B.3). In this analysis, we extend past the known rRNA sequence tail as a control that will allow us to test the accuracy of our method by determining whether it is able to uncover the known 3′ terminus.

We finally explored a range of variants that include extensions on both ends to determine the optimally predictive aSD sequence and distance parameters for the given dataset (Fig. 3.3B). Several of these putative aSD sequences produced similar results so we selected the shortest sequence among these candidates (5′-ACCUCCUUA-3′)

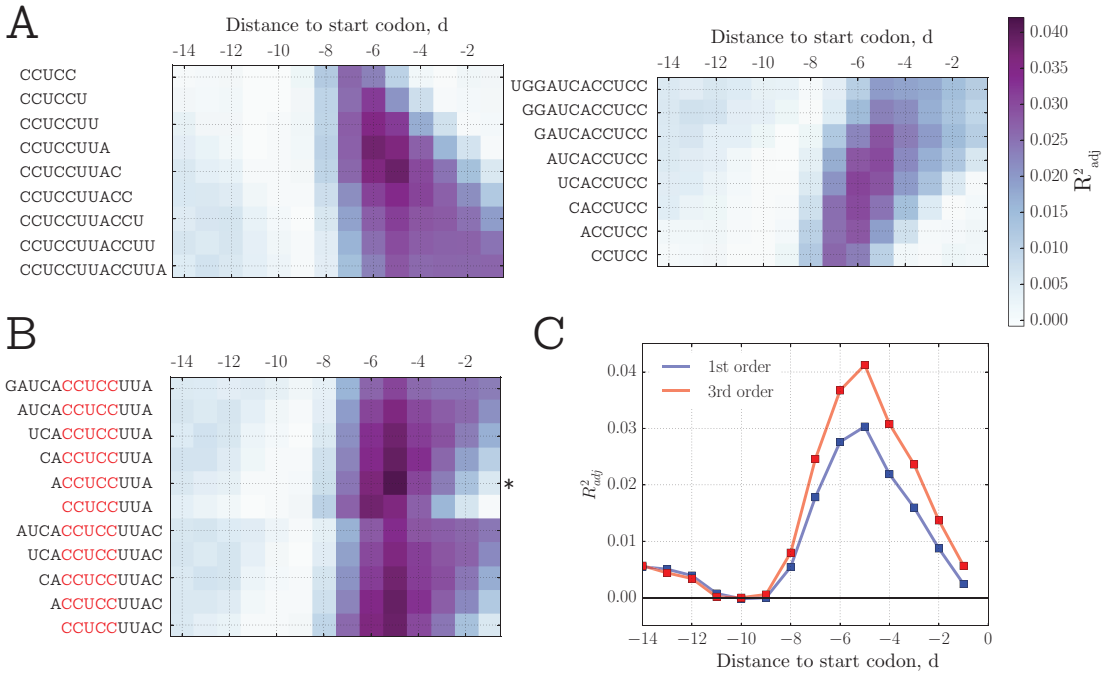

Figure 3.3. **Parameter fitting landscape to determine optimal aSD and distance values.** *A*, R^2_{adj} from the 3rd order model at different distances to the start codon and various 3′ and 5′ extensions to the core aSD for *E. coli*. *B*, Combination of best fitting putative aSDs from (A) to determine the optimal aSD sequence and distance parameters based on their fit to the residual RTE data (* denotes the selected best fitting aSD sequence). *C*, Comparison of R^2_{adj} between the 1st and 3rd order polynomial models from the best performing aSD sequence from (B).

but stress that our methodology can likely not discriminate these boundaries precisely given the small differences in R^2_{adj} values between putative aSDs with single base additions/deletions. While the overall correlation coefficient in this best fit model is still modest ($R^2_{adj} = 0.041$), the significance of this finding is extremely high ($p < 10^{-26}$) indicating that despite the potentially large error in RTE estimates, we are nevertheless able to observe a highly significant underlying relationship. These data further show that although complementarity to the core aSD sequence shows a roughly linear relationship with RTE (the 3rd order model in Fig. 3.2C performs only slightly better), the inclusion of flanking sequences results in both increasing predictive power as well as increasing non-linearity in the underlying relationship. Finally, as a further indicator of the accuracy of this method, it resulted in a frequently cited aSD sequence of 5′-ACCUCCUUA-3′, thus uncovering the experimentally determined 3′ terminus.

3.3.3. The relationship between aSD binding and translation efficiency

In order to test the generality of our findings for *E. coli*, we next tested whether our methodology could produce comparable results for *B. subtilis* and *C. crescentus*. We found that the 5′ extensions are similar for the different organisms studied with *B. subtilis* showing preference for a slightly longer 5′ aSD extension, a finding that is consistent with prior observations that the canonical SD sequence in *B. subtilis* 5′ UTRs appears shifted further upstream of the start codon (Fig. 3.1B).

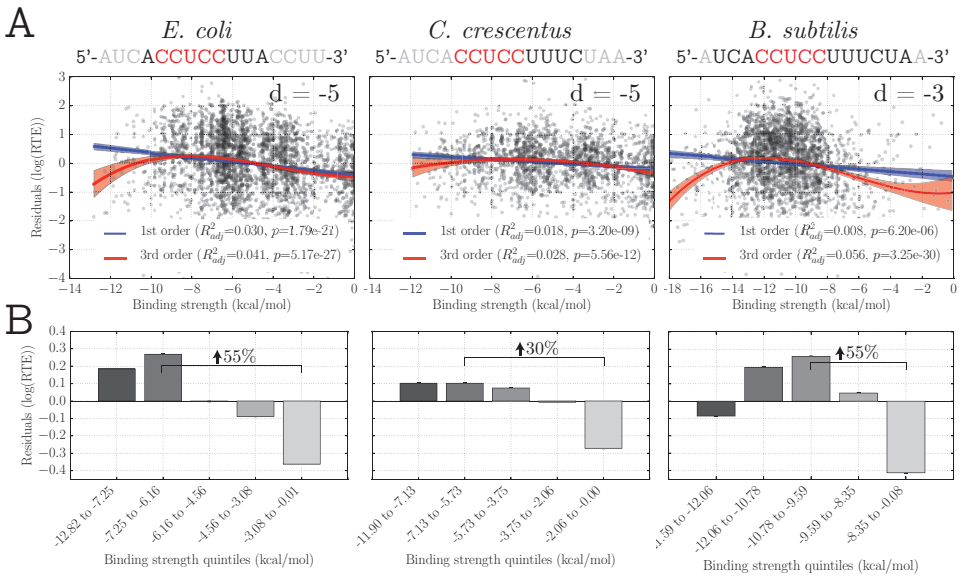

Figure 3.4. **Summary of findings for three independent organisms using ribosome profiling based data.** *A*, Scatter plot of residual RTE values after accounting for the effect of mRNA structure versus aSD sequence complementarity for the species-specific optimal aSD sequence (shown above in black) and optimal distance to the start codon (inset). *B*, Data from (A) depicted as equally sized quintile bins to illustrate the magnitude of the effect. Bars denote the mean within each bin, while error bars show standard error of the mean. Percent increase highlights the average increase in translation efficiency expected for a gene with aSD sequence complementarity at the optimal distance compared to a gene with weak aSD sequence complementarity.

We further found that species-specific 3′ extensions to the 16S rRNA result in enhanced correlations and thus are likely present in the processed 16S rRNA (to our knowledge, the precise 3′ 16S rRNA terminus for these species is unknown) and participate in message discrimination for these two organisms (Supplementary Figs. B.4 & B.5). For *C. crescentus* the aSD sequence that we obtained from our data-driven model is 5′-CCUCCUUUC-3′ while for *B. subtilis* the corresponding sequence is the 5′ extended 5′-UCACCUCCUUUCUA-3′. However, as with *E. coli*, it is difficult to discern whether single base additions/deletions to the ends of these putative aSD sequences are functional.

Despite the vast evolutionary distance between these species, the functional form of the best fitting models was highly similar for all three, showing the highest residual RTE values for intermediate binding strengths with similar predictive powers in the 3rd order model ($R^2_{adj} = 0.041, 0.028$ and 0.056, for all cases $p < 10^{-11}$) (Fig. 3.4A). We further verified that non-linear models provide a superior fit to the data—even though R^2_{adj} explicitly punishes models with more parameters—via the Akaike Information Criterion (AIC), a stringent model selection metric used to judge the relative quality of model fits while explicitly penalizing for parameter number (Supplementary Fig. B.6).

In order to more clearly show the magnitude of the observed effect—and for strictly illustrative purposes—we split the data for each organism into equally sized quintile bins (i.e. the 20% of genes with the highest aSD sequence complementarity, through to the 20% with the lowest). Notably, treating the data this way involves no

model fitting and in doing so, we observe that: (i) the average gene which binds the aSD sequence at the intermediate-to-strong binding strength level shows a 30-50% increase in translation efficiency compared to an average gene that binds the aSD very weakly (Fig. 3.4B) and (ii) the strongest binding quintile of genes exhibits either decreased or equivalent translation efficiency compared to the bin with intermediate-to-strong aSD binding strength. This suggests that mRNAs that contain sequences that bind too strongly to the aSD sequence may actually show reduced translation efficiency, a point that has support from several prior studies in the literature working with experimental systems[157,158]. We note, however, that the optimal sequence complementarity bin for *B. subtilis* is larger than the optimal bin for *E. coli* and *C. crescentus*. This variation may be a result of true underlying differences between the translation initiation mechanisms between these distantly related species, or a function of the fact that the *B. subtilis* aSD sequence is much longer resulting in a broader range of sequence complementarity values than is observed for the other species.

To test the robustness of the above findings to some of our previous assumptions, we repeated the analysis from Figs. 3.3 & 3.4 by interrogating log-transformed RTE values directly. Although *cis*-mRNA structure is thought to be an important regulator of translation initiation, we are faced with the reality that our computational predictions of structural stability are rough approximations of *in vivo* structures and therefore may introduce further error and biases into our measurements. Nevertheless, we observed very similar results for all 3 organisms in terms of the optimal aSD

sequence and distance (Supplementary Fig. B.7) as well as the functional form of the best fitting model (Supplementary Fig. B.8). The fact that the significance of our results are improved when removing the effect of mRNA structure provides further evidence that the *true* magnitude of the aSD sequence complementarity effect may be even further enhanced were we able to more accurately predict—and control for—the structural component of this relationship.

Given recent concerns in the literature about the possibility of biases arising from the size selection step of prokaryotic ribosome profiling studies, we analyzed two further *E. coli* datasets ($n = 1278$ and 1321) from an independent lab that were generated in such a way as to purportedly minimize potential sources of error[153]. After accounting for mRNA structure as before ($R^2 = 0.11$, $p < 10^{-33}$ for both datasets), we observed nearly identical results to the previous *E. coli* dataset (Fig. 3.5, Supplementary Fig. B.6). For both replicates, the 5′-ACCUCCUUA-3′ aSD sequence at a distance of -5 provided the best fit to the data with corresponding R^2_{adj} values of 0.06 and 0.07 for the best fitting 3rd order polynomial and effect sizes of 45% and 50%. While illustrating the robustness of our results for a given organism across multiple independent datasets, this analysis also highlights the sensitivity of R^2_{adj} to measurement noise. Although we observed generally low, albeit highly significant, R^2_{adj} values in the previous analyses, we saw a 50% increase in predictive power using the same modeling approach when applied to these new data while the effect size remains relatively insensitive to this scatter. Indeed, in these data the correlation between aSD sequence complementarity and residual RTE is nearly as large as the

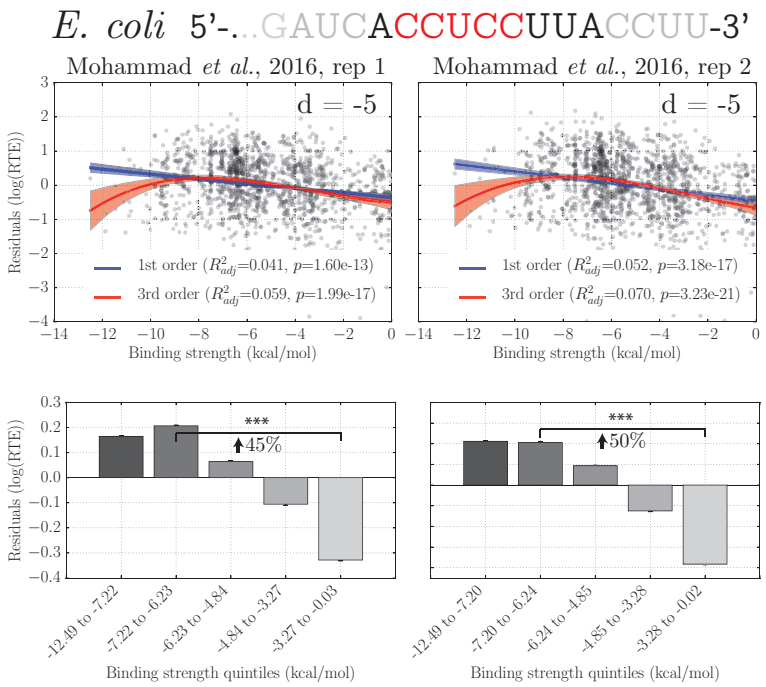

Figure 3.5. **Validation of findings in independent *E. coli* ribosome profiling datasets.** Scatter plot and quintile analysis for independent *E. coli* datasets as in Fig. 3.4. For both replicates, the optimal fitting aSD sequence and distance to the start codon were the same as shown in Fig. 3.4 for *E. coli* with largely similar trends and stronger correlation, presumably due to a reduction in measurement error.

correlation between mRNA structure and RTE supporting previous observations of a strong role for the aSD sequence in enhancing translation initiation.

Finally, given the propensity of prokaryotic genes to occur in operons, we repeated our analysis for all 5 datasets (using the previously discovered organism specific aSD and distance parameters) by splitting genes up according to whether they are predicted to be first in a transcription unit or in the middle/end (see Materials and Methods). Our results were variable for the different organisms with our model fitting procedure resulting in substantially increased predictive power for genes in the middle/end of operons for the *E. coli* datasets, while the opposite phenomenon was evident in the *C. crescentus* and *B. subtilis* data (Supplementary Fig. B.9). Nevertheless, all correlations were highly significant and the 3rd order polynomial model—having a maximum value for intermediate aSD sequence complementary—resulted in larger R^2_{adj} values compared to linear models for all datasets, further illustrating the robustness of this finding.

3.3.4. Translation efficiency in other data sets

To make sure that our observations are not a result of unknown systemic bias in the ribosome profiling based method of calculating RTE, we turned to two separate data sets. First, we utilized an independent data set from Taniguchi *et al.* (2010) who estimated protein production per mRNA from the green fluorescent protein (GFP)-tagged single-cell protein distributions for 1018 *E. coli* genes (see Materials and Methods for our quality control procedures)[10]. Using their data, we performed

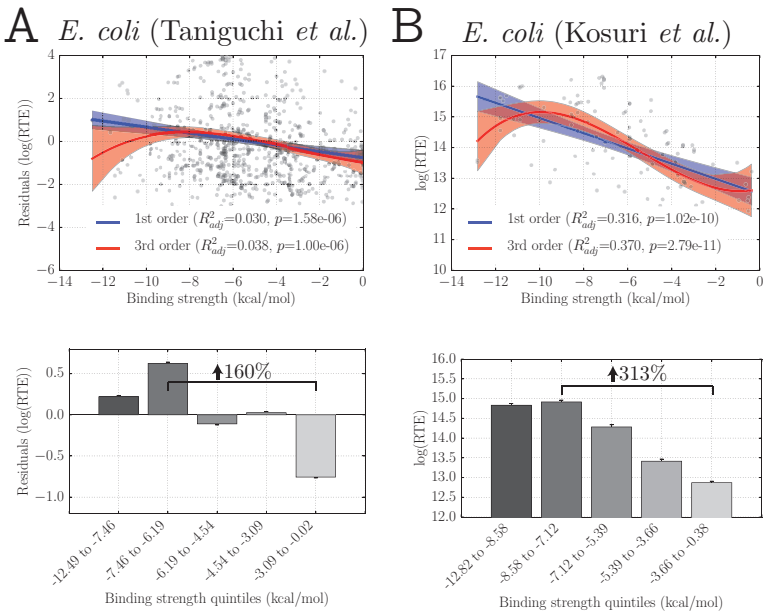

Figure 3.6. **Validation of principal findings in non-ribosomal profiling based datasets.** *A*, Genome-wide data from Taniguchi *et al.* (2010) show a significant relationship between aSD binding strength and residual RTE values. Quintile analysis shows a 160% increase in RTE between genes with weak and intermediate-to-strong aSD sequence complementarity. *B*, Experimental data from Kosuri *et al.* show the same trend as in Fig. 3.4 ($R^2_{adj} = 0.32$ and 0.37 for 1st and 3rd order models, $p < 10^{-10}$ for both cases). Quintile analysis shows a large effect size as well as a plateau / slight-decrease for the quintile with the largest degree of aSD sequence complementarity.

the same analysis as above and again observed nearly identical results to those seen in Fig. 3.4 for *E. coli*. In other words, the data exhibit a maximum at intermediate-to-strong aSD sequence complementarity (Fig. 3.6A, Supplementary Fig. B.10). When we limit our analysis of this dataset to genes with the highest signal-to-error ratio (specifically, the top 50% as calculated by Taniguchi *et al.* (2010)), the magnitude of the R^2_{adj} gets larger with 5′-ACCUCCUUA-3′ sequence complementarity at a spacing of -5 predicting residual RTE with an R^2_{adj} of 0.075 ($p < 10^{-6}$) (Supplementary Fig. B.10).

Finally, although our interest here is in the relationship between aSD sequence complementarity and the translation efficiency of endogenous genes, we further verified our main conclusions using a controlled experimental dataset[75]. Kosuri *et al.* (2013) measured the strength of 111 ribosome binding sites (RBS) by creating synthetic constructs whereby RBS/promoter combinations drove expression of a downstream GFP reporter (see Materials and Methods). For each RBS, the protein produced per mRNA, averaged across the different promoter constructs, is an indicator that we will again refer to as RTE for simplicity. For these data, we did not remove the effect of mRNA structure since each RBS data point represents an average across multiple independent mRNA species (derived from different upstream promoter sequences), and because the coding sequence remains unchanged. Alterations in 5′ structure between these different constructs are still possible, but the effect is likely diminished compared to the other studies and difficult to reliably assess computationally. We nevertheless observed that a 3rd order polynomial model again

provided a better fit to the data than a 1st order linear model (R^2_{adj} = 0.37 and 0.316, respectively, $p < 10^{-10}$ in both cases)(Fig. 3.6B, Supplementary Fig. B.10). We also observed that the intermediate binding quintile produced RTE values 85% higher than the weakest binding quintile and observed a plateau or slight decrease in RTE for the strongest binding quintile of RBS sequences. This provides further support for our conclusion that translation efficiency is maximized at intermediate levels of aSD sequence complementarity and serves as an independent validation of our genome-scale findings. The large R^2_{adj} values that we observed also provide strong empirical support for the hypothesis that some combination of error-prone mRNA structure prediction and error in the calculated RTE values strongly limit the observed R^2_{adj} values in the genome-wide analyses while the general trends and conclusions remain robust and are supported by this experimental dataset.

3.4. Discussion

Our work illustrates that there is a strong relationship between aSD sequence complementarity to the 5′ UTR of mRNAs and the translation of downstream endogenous genes. Specifically, we demonstrate that after accounting for the effects of mRNA structure: (i) aSD sequence complementarity to mRNA is predictive of translation efficiencies for endogenous genes within a relatively narrow window relative to the start codon, which can be empirically determined on a per-organism basis, (ii) slight changes in the putative aSD sequence significantly alter the statistical conclusions allowing us to determine a data-driven definition of the optimal aSD sequence for each

species, and (iii) intermediate aSD sequence complementarity maximizes the translation efficiency of downstream genes in all data sets that we encountered including well controlled experimental data.

Our study complements and extends the experimental study of Vimberg *et al.* (2007) who showed similar patterns of decreasing translation efficiency for experimentally manipulated genes with extended aSD sequence complementarity[158]. While it is possible that native sequences do not typically have strong sequence complementarity and that this effect would thus only apply to a small range of artificial gene constructs, we show here that a substantial number of genes from each genome actually fall within the regime decreased translation efficiency due to the strength of their aSD sequence complementarity. Overall translation efficiency appears to be maximized at intermediate levels of complementarity between the aSD sequence and mRNA, possibly as a result of competing processes governing the efficiency of initiation complex assembly and the transition to translation elongation (Fig. 3.7)—as originally articulated by Komarova *et al.* (2002)[132,157–159]. Alternatively, rapid loading of ribosomes on a single mRNA may cause ribosomal queuing, and potentially result in premature termination or frame-shifting as ribosomes unproductively stall—thus decreasing overall ribosomal throughput on a given message[160]. More accurate experimental and computational protocols that limit sources of error and allow for more precise mapping of ribosome locations may fully resolve these and other issues.

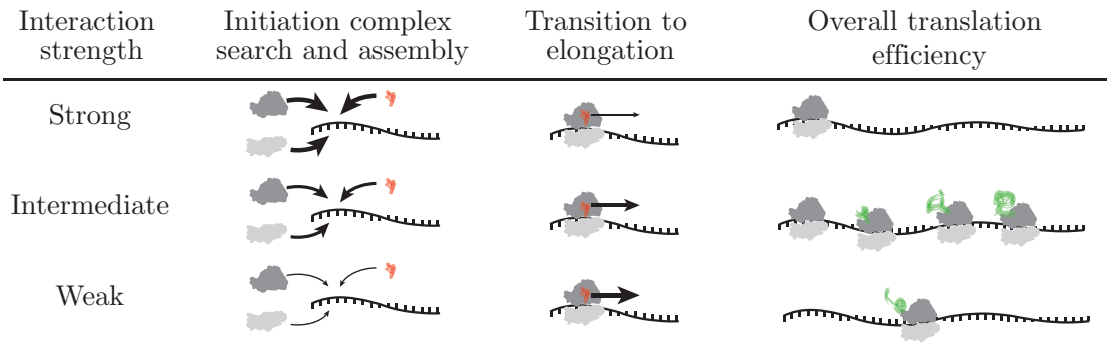

Figure 3.7. **Model explaining why translation efficiency may be maximized for mRNAs with intermediate aSD sequence complementarity.** The competing processes of initiation complex assembly and transition into elongation select for and against, respectively, strong aSD binding to mRNAs resulting in maximal translation efficiency for sequence with intermediate binding strength.

Many previous bioinformatic and experimental studies either implicitly or explicitly assume a continual increase in translation efficiency with increasing aSD sequence complementarity[71,133,143]. One possible reason for this discrepancy is that many experiments may not observe a drop-off in efficiency at high levels of aSD sequence complementarity because they fail to access the full range of sequence diversity capable of binding to the 16S tail. We show here that mRNAs with perfect sequence complementarity to the core aSD sequence appear to translate just fine (Fig. 3.2B, linear fit). However, when considering the fact that sequence binding beyond the core aSD sequence appears to occur in all of these species, perfect complementary becomes detrimental as it begins to include base pairing to these flanking sequences.

Our goal here has not been to develop a comprehensive model to predict translation efficiencies measured by ribosome profiling, but rather to ask whether the sequence diversity and translation efficiency measurements for thousands of native genes can provide insight into the basic mechanisms of initiation. It is nevertheless surprising that the predictive power of the aSD::SD relationship is so low given that the aSD sequence is so highly conserved across nearly all bacterial species, and experimental investigations have seen large changes in protein output when modulating $5'$ UTR sequence binding to the aSD sequence[71]. However, as we have stressed throughout, we note here again that our findings likely represent a lower bound on the predictive power of this interaction for several reasons. Genome-scale metrics are subject to both technical and biological noise, and translation efficiency as a metric will particularly suffer from this noise due to error-propagation. Further, mRNA

folding around the start codon is known to exert a large effect on translation efficiencies and computationally predicted structures are rough approximations of the true mRNA structure[124]. It is thus reasonable to assume that these sources of noise contribute to lowering the expected 'perfect' correlations far below 1.0 as has been observed for other systems[16]. Despite these concerns, the underlying relationship that we observe is strong enough to show robust, statistically significant correlations in all datasets that we investigated. In the most controlled dataset that we analyzed, a 3rd order model of aSD sequence complementarity explained roughly 40% of the observed variance in translation efficiency within an experimental system where the structure surrounding the start codon should be *relatively* similar across different constructs on account of the same coding sequence being expressed.

In addition to measurement noise and other caveats listed above, predicting the translation efficiency of endogenous genes poses a number of other unique challenges that contribute to low correlations. The location of transcription start sites relative to the start codon is variable and experimentally measured $5'$ UTRs are often shorter than 30 bases and sometimes far longer. Further, a number of important genes such as ribosomal proteins are known to be regulated at the level of translation by various mechanisms that obscure statistical signal and which act in addition to the general patterns that we are trying to study. On top of all these limitations, we are also aware that translation efficiency may be modulated by differential elongation and termination in a non-trivial manner and that even within the realm of translation initiation other mechanisms such as the binding of ribosomal protein S1 may further

modulate initiation efficiencies. Investigating the full-range of possible contributions from each of these effects is far-beyond the scope of our study, but doing so in the future will be valuable for better understanding translational regulation.

A better understanding of the rules governing translation initiation and translation efficiency stemming from this systems-biology approach has the practical potential to enhance our ability to design and engineer optimal protein expression systems for a host of biotechnological purposes. Particularly, orthogonal ribosome systems consisting of 30S subunits with altered aSD sequences (and corresponding mRNA sequence preferences) are an increasingly utilized tool in the synthetic biology community[161,162]. The effect that expression of these ribosomes has on endogenous genes governs their orthogonality, and predicting these effects based on the results that we show here may form an important part of rationally designing optimal systems that balance orthogonality against native genes and high expression of target genes[97].

Continued development and application of the ribosome profiling technique and associated technologies to diverse organisms will be critical for clarifying a number of outstanding questions in the field of translation and advancing our understanding of less-well understood species. While detailed experimental studies that systematically express and measure heterologous constructs remain the gold standard for studying sequence-based control of gene expression, we show here that genome-scale approaches combining RNA sequencing and ribosome profiling of native genes can provide valuable insight into these same mechanisms—making this approach particularly attractive for species with less established experimental protocols. Studying

the sequence effects on translation in endogenous genes thus provides a valuable and complementary approach to long standing experimental and bioinformatic investigations.

3.5. Materials and Methods

3.5.1. The data and relative translation efficiency

We downloaded ribosome profiling reads and corresponding RNA-sequencing reads for *Escherichia coli*, *Caulobacter crescentus*, and *Bacillus subtilis*[133,149,156]. We used the original researchers mapping of sequence reads to the respective genomes (.wig files) and removed genes with coverage below 25% in either the RNA-seq or ribosome profiling data sets in order to enrich for high confidence measurements. We also removed any gene shorter than 30 codons as well as potentially mis-annotated genes with zero ribosome profiling reads to the first 10 nucleotides. For all remaining genes, we calculated translation efficiency for each gene as the RPKM in the Ribo-seq dataset divided by the RPKM in the RNA-sequencing dataset. We separately compiled 2 further datasets for *E. coli*, subjecting them to the same pipeline as above[153]. We settled on this approach as it is far more strict in data inclusion criteria than previous studies (which should partially limit noise in RTE measurements) while still providing reasonably large numbers of genes for analysis.

We further utilize two experimental datasets to independently validate our conclusions. The first from Taniguchi *et al.* (2010) utilized single-cell distributions of

protein counts to estimate the proteins produced per mRNA from fitted gamma-distributions of single-cell expression[10]. From the original dataset of 1018 genes we remove 4 from our analysis for quality control, *i.e.* coding sequences which are not a multiple of 3, do not have a 'product' annotation, contain internal stop codons, etc. Since estimates for translation efficiency in this dataset were based off of model fitting under the assumption of gamma-distributed protein concentrations, we analyzed the subset of proteins ($n = 717$) for whom the probability of gamma fit was >95%. For clarity we maintain the label of relative translation efficiency (RTE) to describe these data but stress that their derivation is un-related to ribosome profiling based estimates of translation efficiency and that RTE in this context has a slightly different interpretation[10].

We also downloaded experimental data from recombinant gene expression in *E. coli*[75]. Each of 110 different ribosomal binding sites (RBSs) were characterized using FLOW-Seq (a method that combines fluorescence-activated cell sorting and high-throughput DNA sequencing) and can be described by their average protein levels across different promoters divided by the average mRNA levels (roughly equivalent to RTE when calculated for the same protein) (from their initial data we exclude the 'Dead-RBS' construct because its short length is prohibitive to our analysis). Here we analyze this 'mean.xlat' data (as described in their supporting tables[75]) as a measure of relative translation efficiency. As before, for ease of language we continue to refer to this as relative translation efficiency (RTE) but note the slight differences in interpretation. Rather than subtracting out the effect of mRNA structure as in

the previous datasets, we simply provide regressions on this raw data here since: (i) the downstream gene is the same and thus structure is mostly preserved between constructs and (ii) each promoter will introduce slightly different sequences upstream of the RBS but their structural effects of this introduction should be accounted for in the averaging process.

3.5.2. Gene classification and quantification of aSD binding strength

All calculations of RNA folding were performed using the RNAfold method from ViennaRNA with default parameters[163]. Estimations of *cis*-structure were based on calculated folding energies for the -30 to +30 nt region relative to the start codon ('A', 'T', 'G' are bases +1, +2, and +3, respectively). RNA::RNA hybridizations were performed using the RNAcofold method with default parameters. For each gene, we iterated through all x-mers (where x is the length of the putative aSD sequence) upstream of the start codon in order to capture 14 hybridization events.

3.5.3. Operon predictions

We utilized predicted operons from the Database of Prokaryotic Operons[164]. From these data tables, we classified each gene according to whether it is predicted to occur first within a transcription unit or whether another gene precedes it within a transcription unit, regardless of the distance.

3.5.4. Statistics and code sharing

All code used to perform translation efficiency measurements, as well as all statistics were written using custom scripts in Python that are included in the supplementary information. All regression models and statistics (including R^2, R^2_{adj} and Akaike Information Criteria (AIC)) were performed using the statsmodels package from Python; reported p-values in all regressions are based on the the F-test. Code and necessary data to re-create figures are available at `https://github.com/adamhockenberry/OpenBiology_2016`

CHAPTER 4

Growth demands shape variation in translation initiation mechanisms across bacterial species

This work was performed with Aaron J Stern, Michael C Jewett, and Luís AN Amaral.

4.1. Abstract

The Shine-Dalgarno (SD) sequence is often found upstream of protein coding genes across the bacterial kingdom—enhancing start codon recognition via hybridization to the anti-SD (aSD) sequence on the small ribosomal subunit. Despite wide-spread conservation of the aSD sequence, the proportion of SD-led genes within a genome varies widely across species and the evolutionary pressures shaping this variation remain largely unknown. Here, we use phylogenetic comparative methods to show that species capable of rapid growth have a significantly higher proportion of SD-led genes in their genome, suggesting a role for SD sequences in meeting the protein production demands of rapidly growing species. Further, in a larger dataset we show that utilization of the SD sequence mechanism co-varies with: i) genomic traits that are indicative of efficient translation, and ii) optimal growth temperatures. In contrast to prior surveys, our results demonstrate that variation in translation initiation mechanisms across genomes is largely predictable after accounting for phylogenetic effects,

and that SD sequence utilization is part of a larger suite of translation-associated traits whose variation is driven by the differential life-history strategies of individual species.

4.2. Introduction

Translation of a given messenger-RNA (mRNA) into a functional protein is contingent on the ability of the translational apparatus to recognize the proper start codon. To discriminate between potential start codons, bacteria have evolved several distinct mechanisms[144]. Most-utilized is the so-called Shine-Dalgarno mechanism—named for a purine rich sequence commonly defined as 5′-AGGAGG-3′—whereby mRNAs hybridize with a complementary anti-SD (aSD) sequence on the 16S rRNA of the small ribosomal subunit (5′-...ACCUCCUU...-3′)[72]. Varieties of the canonical SD sequence are enriched upstream of known start codons across nearly the entire bacterial kingdom, and the aSD sequence is highly conserved (though notable exceptions exist, see: Lim *et al.* 2012)[135,142,145,146,165–167]. For a given gene within an organism, researchers have shown that a number of context-dependent factors including the structural accessibility of the SD sequence, the thermodynamic binding potential between the sequence and the aSD sequence, and the exact positioning of the SD sequence relative to the start codon are all factors that modulate the translation initiation rate of downstream genes[71,75,134,136,139,143,158,168–171].

The importance of the SD sequence is further supported by the fact that these sequences are under-represented in the coding sequences of most bacteria, possibly reflecting their role in translational pausing and erroneous initiation. The degree of

this under-representation is highly variable across bacterial species, which is suggestive of possible mechanistic differences in the translation machinery[97,153,172,173].

Several other translation initiation mechanisms exist in bacteria, but are less well-studied. Leaderless genes are so-named because the transcript sequence begins at, or very close to, the recognized start codon[147,174–176]. Additionally, translational scanning, and internal ribosome entry sites (IRES) are common eukaryotic initiation mechanisms that may also be functional in prokaryotes under certain circumstances[177,178]. Finally, investigation into almost any genome uncovers a number of genes with seemingly normal length 5′ UTRs, but no discernible SD sequence to signify the beginning of the predicted coding sequence[141]. In these cases, it is thought that weak mRNA structure plays a large role in facilitating start codon recognition, but AT-nucleotide rich sequences upstream may also be involved in this process by binding to the RPS1 protein on the 30S ribosomal subunit[118,120,179–182].

Despite an abundance of research showing that the SD sequence enhances translation initiation and start codon recognition of downstream genes, it is still not known why bacteria use such diverse mechanisms, especially given the high conservation of the aSD sequence in the 16S rRNA[71,143,171]. For example, why is it that roughly 90% of *Bacillus subtilis* genes are preceded by a SD sequence while for *Caulobacter crescentus* the comparable number is closer to 50%[144,146,149]?

Cross-species variation in translation initiation mechanisms may impact genetic isolation and transfer of genetic material, and quantifying the source and extent of

this variation may prove useful in identifying important genes in a genome or ecological community[93,167]. Further, both translation-system engineering and biotechnology applications involving less well-studied microbial species are increasingly popular targets in the synthetic biology community[162,183–187]. These efforts are likely to benefit from a better understanding of factors shaping translation initiation mechanisms.

Here, we use phylogenetic comparative methods in order to isolate independent evolutionary events and show that the proportion of SD-led genes within a genome is strongly related to the growth demands faced by a species. We develop a metric that uses sequence entropy to summarize the presence of over-represented motifs in the UTRs from a given genome, and show that it is predictive of minimum doubling times for 187 bacteria. In order to extend our analysis to species without known minimum doubling times, we use a database of 618 phylogenetically diverse species and show that genome-wide variation in SD sequence utilization is largely predictable at the individual organism level given knowledge of phylogenetic relatedness and a small number of genomic features relating to the strength of selection on translation efficiency.

4.3. Results

4.3.1. Sequence entropy and its relation to SD sequence utilization

Several techniques have been previously used to quantify the magnitude of an individual organisms utilization of the aSD::SD mechanism. In motif based methods,

researchers pre-define several sub-sequences closely related to the canonical SD sequence and search through the protein coding genes within a given genome to determine the fraction of genes that contain one of these motifs within some range upstream of the start codon[141,167]. Similarly, in aSD sequence complementarity based methods, researchers pre-define a range upstream of the start codon to consider, a putative aSD sequence, and a hybridization energy threshold for determining whether a gene is considered to be SD-led[142,144–146,166].

Both of these methods rely on critical assumptions that may be hard to justify when extrapolating across a large-set of diverse organisms. First, both methods rest on a dichotomy between SD- and non-SD-led genes. While this simplification is useful for *discussing* the phenomenon, an abundance of research has shown that there are not two distinct categories but rather a spectrum of sequence complementarity that affects translation initiation in a continuous manner[71,158]. Second, bacterial genomes span a range of GC contents, and previous research has shown that it is critical to compare the proportion of SD-led genes in a genome to appropriate null model expectation[144]. Third, both of these methods carry an assumption that a SD sequence, regardless of its location relative to the start codon, has the same impact on translation initiation. However, experimental approaches have shown that spacing between the SD sequence and start codon can have dramatic effects on translation initiation rates[71,158,168,169].

We sought a complementary approach grounded in information theory that would allow us to investigate hundreds of diverse genomes without having to *a priori* define

aSD sequence or SD sequence motifs. For each genome we extract the $5'$ upstream sequences from all annotated protein coding sequences (see Materials and Methods). From this set, we then sum the information contained in the sequences within the region where SD-type motifs are expected to occur (-20 to -4 relative to the start codon):

$$I_{observed} = \sum_{i=-20}^{-4} \left(\log_2 4 + \sum_{k \in \{A,T,G,C\}} p_{i,k} \log_2 p_{i,k} \right) \tag{4.1}$$

where $p_{i,k}$ is the probability of finding base k at position i. To control for non-uniform GC content in translation initiation regions, we construct a null model where we randomly shuffle each upstream region and calculate I_{random} for one instance of this shuffled set (see Materials and Methods). We repeat this process for 500 randomized sets and compare the sequence information from the real genome to the average of the randomized sets:

$$\Delta I = I_{observed} - \overline{I}_{random} \tag{4.2}$$

By definition, ΔI is a measure of non-randomness in the translation initiation region for a particular genome. Except for the requirement of a pre-defined range upstream of start codons to include in the analysis, it does not require any other assumptions about the the aSD::SD interaction. ΔI is agnostic to *which* sequence motifs are over-represented—thus alleviating the need to pre-define either a putative aSD or SD sequence.

Figure 4.1A displays sequence logos of 5′ UTRs for representative species and illustrates our methodology graphically. Figure 4.1B shows that while SD motif and aSD sequence complementarity based methods (both calculated relative to a randomized null model control, see Materials and Methods) generally correlate well in a large dataset of diverse species, there is a clear departure from a linear relationship for the Firmicute phylum. Further, ΔI correlates strongly with aSD sequence complementarity for most species (Fig. 4.1B). However, we note that ΔI shows marked differences for members of the Bacteroidetes phylum. Prior research identified major alterations in the aSD sequence region of the 16S within this phylum, and our results strongly support the hypothesis of altered sequence preferences by showing that the 5′ upstream region of genes from these genomes exhibit significant variation in ΔI without any significant difference in either SD motif- or aSD sequence complementarity-based metrics[165]. The ΔI metric thus allows us to incorporate Bacteroidetes into future analysis (Fig. 4.1B, red data points). For simplicity, we will refer below to ΔI and SD sequence utilization interchangeably, and make clear when we use alternative measurements of SD sequence utilization.

4.3.2. Translation initiation and organismal growth demands

Given that SD sequences are known to enhance translation initiation and efficiency for individual genes, we speculate that genome-wide variation in the usage of this mechanism may be related to the differential growth demands of species. Vieira-Silva *et al.* (2010) curated a list of minimum doubling times from the literature

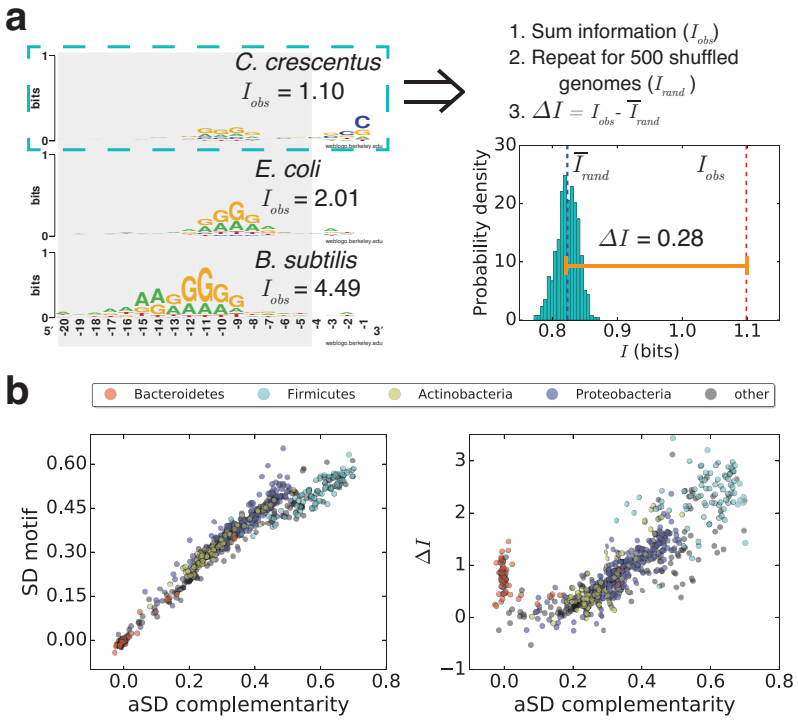

Figure 4.1. **A sequence entropy based metric for genome-wide SD sequence utilization** *A*, Representative sequence logo examples and an illustration of our pipeline for calculating ΔI *B*, Comparison between different ways of summarizing SD sequence utilization. On the left, we show the correlation between SD motif and aSD binding strength methods. On the right, ΔI and aSD binding strength. The four largest phyla are color-coded according to the legend, a scheme which we use throughout.

and showed that a small set of genomic traits can be combined to predict this value from an organism's genome[91,188]. We replicated several of the findings of Vieira-Silva *et al.* (2010) using phylogenetically generalized least squares (PGLS) regression to account for the lack of independence of the data (see Materials and Methods)[189]. We find that rRNA copy number, tRNA copy number, and $\Delta ENC\prime$ (a measurement of relative codon usage bias) all show a significant relationship with minimum doubling times (F-test, $p < 0.005$ for all cases), while overall coding sequence number has a weak but still significant effect ($p = 0.007$)[91].

Next, we turn to translation initiation related metrics and find that ΔI significantly correlates with minimum doubling times in this set of species($p < 10^{-5}$), showing the 2nd strongest correlation of all individual traits that we considered (Table 4.1, Fig. 4.2).

We test several other translation initiation associated metrics and find that, in contrast to SD sequence utilization, the proportion of protein coding genes containing an ATG start, and the average difference in GC content between gene starts and internal regions (Δ GC initiation) (chosen as a proxy for selection on decreased initiation region structure, see Materials and Methods) shows weaker associations with known minimum doubling times in this dataset ($p = 0.056$ and $p = 0.005$ respectively). We constructed a multi-variable PGLS model that combines all of the possible predictors, and observe that only ΔI and $\Delta ENC'$ had statistically significant coefficients ($p < 0.001$, both cases). Overall, a model containing all predictors

Trait	Individual model (R^2)	Complete model (ΔR^2)
$\Delta ENC'$	0.124	**-0.095**
16S gene number	0.056	-0.011
tRNA copy number	0.061	-0.013
CDS number	0.039	-0.002
ΔI	0.108	**-0.077**
ATG start %	0.02	-0.001
Δ GC initiation	0.042	-0.025

Table 4.1. **Parameter contributions for predicting minimum doubling times.** The middle column illustrates the overall goodness-of-fit for individual predictors (left) of minimum doubling time ($p < 0.01$ for all values, except ATG start %). The right column illustrates the change in goodness-of-fit from a model that includes all predictors versus one that excludes only the variable of interest. Bold numbers in this column illustrate the variables with significant coefficients in the complete model ($p < 0.001$). All analyses were performed using Phylogenetic Generalized Least Squares regression to remove the effect of shared ancestry, with independent maximum likelihood fits of Pagel's λ branch length transformation.

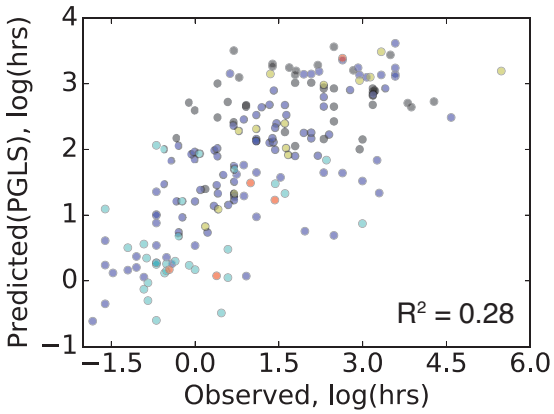

Figure 4.2. **Phylogenetically independent association between genome-scale translation metrics and minimum doubling time** We show the observed and predicted values from a PGLS regression model using all predictors in Table 4.1. Data are colored according to phyla as in Fig. 4.1B.

resulted in an R^2 of 0.28 ($p < 10^{-9}$, $\lambda = 0.94$), while a more parsimonious model containing only the two predictors with statistically significant coefficients resulted in an R^2 of 0.23 ($p < 10^{-10}$, $\lambda = 0.95$), demonstrating the strong relationship between growth demands and SD sequence utilization when controlling for shared ancestry. For reference to prior research, a phylogenetically *agnostic* linear regression model using all predictors yields an R^2 of 0.54 ($p < 10^{-15}$)—though we strongly caution against ignoring the effects of shared ancestry in comparative analyses. SD usage summary statistics calculated via aSD sequence complementarity showed similar results overall.

4.3.3. Relationship between SD sequence utilization and translation efficiency associated traits

We next assembled a much larger and phylogenetically diverse data set consisting of 618 bacterial species—unique at the genus level—for whom we have complete, annotated genome-sequences as well as a high-quality phylogenetic tree describing their relatedness[92]. We confirmed previous results, showing that SD sequence utilization between these species varies considerably both within and between phyla (Fig. 4.3A)[144].

In order to determine whether the different traits that are important for translation efficiency co-vary with one another (independent of phylogeny) we test whether any of the previously analyzed traits could predict SD sequence utilization at the genome-scale. We again applied PGLS regression independently for each variable and

find that ΔI is most significantly correlated with 16S rRNA gene copy numbers(F-test, $p < 10^{-13}$), the fraction of genes containing and ATG start codon($p < 10^{-13}$), and Δ GC initiation ($p < 10^{-8}$) (Table 4.2).

In order to test the overall predictability of SD sequence utilization across this diverse bacterial dataset, we consider a multi-variable PGLS model consisting of all of the traits, resulting in an R^2 of 0.226 ($p < 10^{-15}$, $\lambda = 1.0$)(Fig. 4.3B, Table 4.2). A parsimonious model that included only those traits with significant coefficients in the initial multi-variable regression—16S rRNA gene copy numbers($p < 10^{-8}$), the fraction of genes containing an ATG start($p < 10^{-10}$), and Δ GC initiation($p < 10^{-5}$)—yielded an R^2 of 0.209 ($p < 10^{-15}$, $\lambda = 0.999$). As with the predictions of minimum doubling times, we observe similar findings when using aSD sequence complementarity based summary statistics of SD sequence utilization as the dependent variable.

4.3.4. Relationship between translation initiation mechanisms and optimal growth temperature

Having established that genome-scale SD sequence utilization is part of a suite of traits related to differential organismal growth strategies, we next assess whether other ecological factors relating to an organisms growth conditions may play a role in further constraining the evolution of the SD sequence mechanism. Nakagawa *et al.* 2010 had previously shown no association between SD sequence utilization and optimum growth temperatures[144]. By contrast, our phylogenetically informed

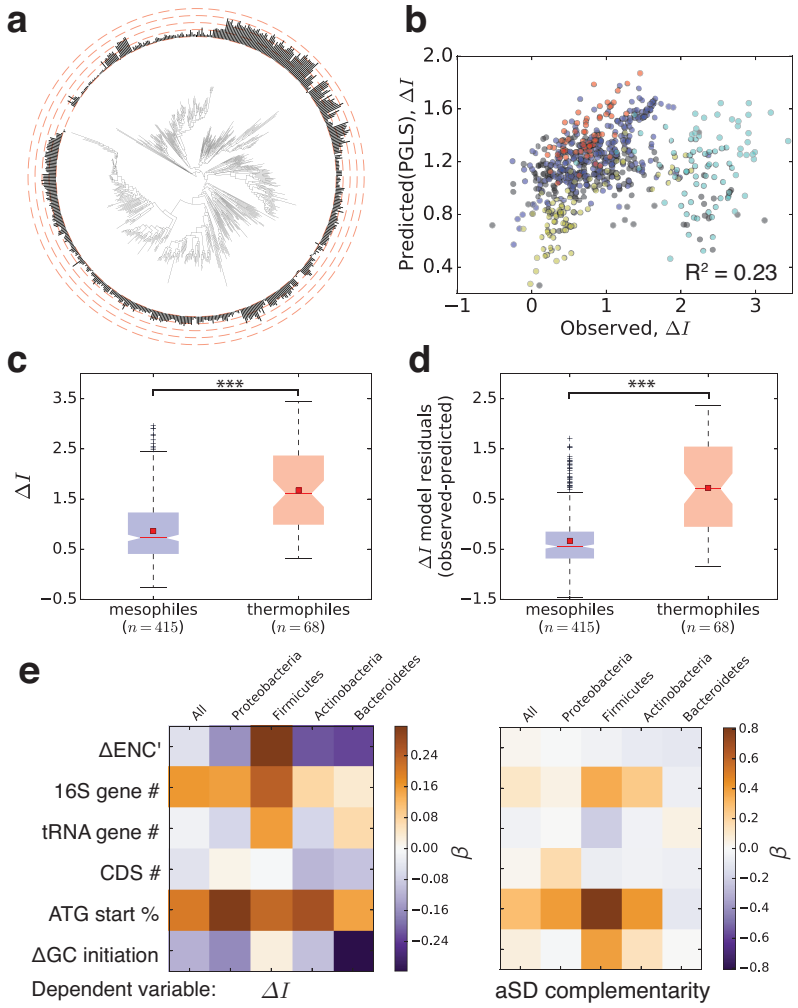

Figure 4.3. **Genomic and environmental predictors of genome-scale SD sequence utilization** *A*, Phylogenetic tree illustrating variation in ΔI *B*, Scatter plot illustrating the observed and predicted values from a PGLS regression model using all predictors in Table 4.2. Data are colored according to phyla as in Fig. 4.1B. *C*, For the subset of species for whom we have optimal growth temperature estimates, we show values of ΔI according to predicted mesophiles and thermophiles ($p = 0.0026$). *D*, As in panel *C* we instead show phylogenetic residuals calculated from the model shown in *B* ($p < 10^{-7}$). *E*, We show the standardized regression coefficients (β) from models using all independent variables to predict SD sequence utilization in the full data set, and individual phyla to illustrate the robustness of our findings. ATG start % and 16S gene copy number show a consistently positive relationship with SD sequence utilization regardless of the clade analyzed or the specific measurement of SD sequence utilization tested.

Trait	Individual model (R^2)	Complete model (ΔR^2)
$\Delta ENC'$	0.033	-0.006
16S gene number	0.091	**-0.049**
tRNA copy number	0.048	-0.001
CDS number	0.001	-0.001
ATG start %	0.115	**-0.058**
Δ GC initiation	0.055	**-0.029**

Table 4.2. **Parameter contributions for predicting ΔI.** The middle column illustrates the overall goodness-of-fit for individual predictors (left) of ΔI ($p < 10^{-6}$ for all cases except CDS number, $p = 0.368$). The right column illustrates the change in goodness-of-fit from a model that includes all predictors to one that excludes only the variable of interest. Bold numbers in this column illustrate the variables with significant coefficients in the complete model($p < 10^{-5}$). All analyses were performed using Phylogenetic Generalized Least Squares regression to remove the effect of shared ancestry, with independent maximum likelihood fits of Pagel's λ branch length transformation.

approach applied to this larger dataset (483 of the 618 species in our dataset had high-confidence growth temperature annotations) finds that temperature constrains genome-wide SD sequence utilization such that the genomes of thermophilic species contain a significantly larger proportion of SD-led genes than mesophilic species (Fig. 4.3C, F-test $p = 0.0026$). This association is even more pronounced when we analyze the phylogenetic residuals of our best fitting model from Fig. 4.3B (Fig. 4.3D, $p < 10^{-7}$) by including temperature as a categorical dummy variable to represent mesophile (0) or thermophile (1). When predicting ΔI, the R^2 of the best fitting PGLS model containing all variables increases from 0.203 to 0.252 for this subset of 483 species.

4.3.5. Clade specific relationships

Thus far, all of our analyses have been performed on the entire bacterial dataset, and we have largely observed similar results with regard to the predictability of minimum doubling times, and the predictability of SD sequence utilization regardless of the different summary statistics of SD sequence utilization that we applied. However, as we noted in Fig. 4.1B, ΔI and aSD sequence complementarity based methods show varying degrees of correlation according to phylum. Here, we test the robustness of our conclusions by independently analyzing relationships in the four largest phyla in our dataset.

Despite the fact that our multi-variable model in Fig. 4.3 shows a fairly strong ability to predict overall SD sequence utilization, there is considerable heterogeneity

in the predictions when performed *independently* for different phyla. This finding may reflect differences in the sample sizes, quality of genome annotations, or true biological variation in mechanisms between independent phylogenetic lineages. Particularly, a PGLS regression model containing the same variables as in Fig. 4.3 is able to account for a strikingly large amount of the variance when only considering Proteobacteria, regardless of the method used to measure SD sequence utilization (Table 4.3, Fig. 4.3E)($p < 10^{-15}$, both cases). We observe similar results for Actinobacteria ($p < 10^{-10}$ for ΔI, $p < 10^{-4}$ for aSD sequence complementarity). However, fitting with our hypothesis that traditional methods are missing important information with regard to the Bacteroidetes phylum, we show that models constructed with ΔI as the dependent variable within this lineage have substantially more explanatory power than comparable models constructed using the aSD sequence complementarity based summary statistic as a dependent variable ($R^2 =$ 0.487 vs 0.05, $p < 10^{-6}$ and $p = 0.81$ respectively) (Table 4.3, Fig. 4.3E). By contrast, we observe the opposite trend for the Firmicutes lineage where models constructed to explain ΔI as the dependent variable perform substantially worse than those constructed to predict the aSD sequence complementarity based summary statistic (R^2 $= 0.181$ and 0.514, $p = 0.018$ and $p < 10^{-9}$ respectively).

These phyla specific results illustrate several critical points. First, the sign on the relationships between the most predictive individual features is extremely robust, regardless of the phylum or SD sequence utilization summary statistic under consideration. Increasing 16S copy number and increasing ATG start codon usage are

		R^2 using all predictors	
	n	ΔI	aSD binding
All	618	0.226	0.24
Proteobacteria	267	0.467	0.477
Firmicutes	82	0.181	0.514
Actinobacteria	78	0.553	0.303
Bacteroidetes	63	0.487	0.05

Table 4.3. Correlation coefficients for different features in single and multiple regression.

consistently associated with increased SD sequence utilization. Second, ΔI is measuring an aspect of translation in the Bacteroidetes phylum that is not accounted for in previous models, which likely reflects novel sequence preferences in this lineage. Third, the formulation of ΔI is *missing* an important aspect of translation initiation that previous methods are able to capture for the Firmicutes phylum. Finally, in the case of Proteobacteria and Actinobacteria, a substantial portion of the variation in SD sequence utilization between genomes can be explained by a small number of independent genome-level predictors regardless of the summary statistic used to identify SD sequence utilization.

4.4. Discussion

To our knowledge, ours is the first study to show that variation in bacterial translation initiation mechanisms is a result of the life-history strategies of individual species. We observe a strong relationship between minimum observed doubling times and SD sequence utilization at the genome-scale, and further show that SD sequence utilization co-varies with several genomic indicators of rapid growth, including the copy number of rRNA genes which has long been established as being necessary for rapid growth. Critically, our analysis is performed in a manner that corrects for phylogenetic non-independence of data-points allowing us to show the robustness of these relationships in independent clades across the bacterial kingdom, while at the same time highlighting clade-specific features that may point to mechanistic differences.

A potential limitation of our findings is the fact that our study strongly relies on genome annotations. As many as 10% of the genes from even relatively well-studied species may have mis-annotated start codons, and our analysis fundamentally relies on extracting the upstream region of annotated coding sequences[149]. However, it seems unlikely that bias in annotation quality would artificially enhance our findings with the more plausible scenario being that poor annotation quality likely dilutes the true significance of the findings that we observe here. Additionally, although we show that the sign of the relationship between SD sequence utilization and genomic traits is largely robust for 4 different phyla, we note that these conclusions are partially limited by the constraint that we are only able to independently analyze phyla for which sufficient numbers of sequenced genomes exist. Future analyses of larger datasets based off the methods presented here will allow researchers to investigate a larger number of phyla and taxonomic groupings at a finer resolution to potentially uncover novel biological variability of the type that we report for Bacteroidetes and Firmicutes.

Our conclusions have several important consequences for researchers moving forward. The observation that a particular species seemingly does not utilize the SD sequence mechanism to a large degree is not evidence that the aSD::SD sequence interaction is either dysfunctional or that it provides little translational advantage for genes within this species. Our results suggest that the selective pressure on highly efficient translation mechanisms may simply be diminished in particular lineages based on their life-history strategies which may include slow growth, growth at

low temperatures, or other unconsidered ecological/population-genetic factors. Thus, researchers working on biotechnological applications in less commonly studied organisms should particularly consider this point when designing vectors for recombinant gene expression[183–187].

Our results also add to the body of knowledge showing that a small number of genomic traits—that includes usage of the SD sequence—can be used to predict variation in minimum doubling times with surprisingly high accuracy. We show that measurements of SD sequence utilization outperform more commonly known associations such as the number of rRNA copies and speculate that this is, in part, a consequence of the evolutionary malleability of different features[188]. Genome-wide usage of the SD sequence mechanism, like codon usage bias, requires hundreds of mutations to substantially alter and thus this trait will evolve much more slowly across a phylogeny when compared to more labile traits that rely on copy number variation such as rRNA and tRNA number. An interesting possibility is that the *difference* in evolutionary rates between different traits could be further be exploited by advanced methods to predict species adapting to novel growth strategies on shorter time scales. Like codon usage biases and in contrast to rRNA and tRNA numbers, we note that summary statistics based on SD sequence utilization do not require complete genome sequences and therefore may be *estimated* with partial genome fragments. The results and methods that we present here may have important applications in

our understanding of novel, uncultivated genomes, environmental meta-genomic sequencing efforts, and the relationship between higher-order genome traits and growth strategies[190].

4.5. Materials and Methods

4.5.1. Data assembly

We first assembled a database of prokaryotic genomes from NCBI using the GBProks software (https://github.com/hyattpd/gbproks), including only 'complete' genomes in our download and subsequent analysis (accessed on: March 10, 2016). From the annotated GenBank files, we excluded pseudogenes and plasmid based sequences from all subsequent analyses and proceeded to compile several traits for each genome. In addition to SD sequence utilization summary statistics, we applied RNammer to each genome in order to compile a list of ribosomal-RNA genes, and tRNAscan-SE to assemble a large set of higher-order genome statistics[191,192]. We wrote custom scripts to calculate codon usage bias $\Delta ENC'$ (having first parsed the gene annotations to find ribosomal protein coding genes), coding sequence number, and the fraction of annotated coding sequences that begin with 'ATG'.

To calculate ΔGC initiation, we restricted our analysis to eliminate coding sequences shorter than 150 nucleotides and/or with ambiguous nucleotides. For each coding sequence that passed this filter, we then calculated the GC content between position -30 to +29 relative to the first base of the start codon (0) and subtracted from this value the GC content from bases +30 to +90. For a genome, we then took

the average value of this difference. We include this metric as a rough proxy for the difference in mRNA structure surrounding the start codon vs the internal region of genes.

For data on minimum doubling time, we downloaded the data table from Vieira-Silva *et al.* (2010), and paired each bacterial species with a complete genome from our database resulting in 187 pairs. To control for shared ancestry in subsequent analyses, we constructed a phylogenetic tree based off the rRNA sequences for this set of species. We first used RNAmmer to extract the 16S and 23S rRNA sequences, followed by MUSCLE (v3.8.31) on each individual rRNA to produce a multiple-sequence alignments[193]. These were concatenated together we conducted a partitioned analysis using RAxML to construct a final tree. We performed 100 rapid Bootstrap searches, 20 ML searches and selected the best ML tree for subsequent analysis[194].

For the larger data-set, we instead relied on a previously computed high-quality dataset published by Hug *et al.* (2016)[92]. We used custom scripts to match entries in this tree with genomes from our complete-genome database, and pruned away all species without a high-quality match resulting in 618 species in our final dataset for subsequent analyses.

For temperature annotations, we matched our existing set of 618 species to the ProTraits database using custom scripts, and restricted our analysis to species with temperature annotations exceeding a precision of 0.9 (equivalent to a FDR < 0.1) for subsequent analyses[195].

4.5.2. Calculating summary statistics of SD sequence utilization

The calculation of ΔI is illustrated mathematically in the main text. Here, we only add that calculation of the randomized sequences is performed by randomizing the upstream region of each gene between -30 to 0. Analysis is then performed as discussed for the -20 to -4 region (with index '0' being the first base of the start codon). Next, we highlight our calculation of the other two methods for calculating SD sequence utilization. For each genome, we extract the -20 to -4 region upstream of the start codon for each gene. In motif based methods, we consider a gene as being SD-led if, in this defined region, any of the following motifs appear: 'GGAA', 'GGAG', 'GAGG', 'AGGA', or 'AAGG'. We repeat this same process for 500 randomized 'genomes' where a randomized genome is defined as the collection of all upstream elements, with the critical difference being that these upstream elements are first randomized between the -30 to 0 region (on a per-gene basis) prior to motif search. We then take the *difference* between the observed number of SD-led genes for a given genome and the average of the randomized cases to determine how many more or less SD sequences appear in a given genome relative to the null model expectation. This number is then divided by the overall CDS number to determine the final metric.

For aSD sequence complementarity, we perform a nearly identical procedure to above with one major difference. Instead of searching the upstream region for motifs, we evaluate the binding energy between each 8 nucleotide segment in the -20 to -4 region and the putative aSD sequence defined as 5′-ACCUCCUU-3′. If any sequence

binds at a threshold of -4.5 kcal/mol or stronger, we consider this gene to be SD-led. Compilation of the final aSD sequence complementarity metric proceeds as above.

4.5.3. Phylogenetically generalized least squares

We utilize phylogenetically generalized least squares (PGLS) regression in order to mitigate the effects of shared ancestry in statistical analyses. Our PGLS analysis relies on the most common null model, which assumes a brownian motion model of trait evolution. For all statistical analyses presented in the paper, we use the R package 'caper' and perform a simultaneous maximum-likelihood estimate of Pagel's λ, a branch length transformation, alongside the coefficients for independent variables of interest in order to control for false-positive, and false-negative rates.

CHAPTER 5

Depletion of Shine-Dalgarno sequences within bacterial coding regions is expression dependent

This work was published with Chuyue Yang (co-first author), Michael C Jewett, and Luís AN Amaral in G3:Genes|Genomes|Genetics, 2016.

5.1. Abstract

Efficient and accurate protein synthesis is crucial for organismal survival in competitive environments. Translation efficiency — the number of proteins translated from a single mRNA in a given time period — is the combined result of differential translation initiation, elongation, and termination rates. Previous research identified the Shine-Dalgarno (SD) sequence as a modulator of translation initiation in bacterial genes, while codon usage biases are frequently implicated as a primary determinant of elongation rate variation. Recent studies have suggested that SD sequences within coding sequences may negatively affect translation elongation speed, but this claim remains controversial. Here, we present a metric to quantify the prevalence of SD sequences in coding regions. We analyze hundreds of bacterial genomes and find that the coding sequences of highly expressed genes systematically contain fewer SD sequences than expected, yielding a robust correlation between the normalized occurrence of SD sites and protein abundances across a range of bacterial taxa. We further

show that depletion of SD sequences within ribosomal protein genes is correlated with organismal growth rates, supporting the hypothesis of strong selection against the presence of these sequences in coding regions and suggesting their association with translation efficiency in bacteria.

5.2. Introduction

Translation of mRNA to protein consumes a vast amount of cellular resources, particularly in rapidly growing unicellular organisms[1,2,196]. Many researchers have hypothesized that efficient — fast and accurate — translation is highly advantageous and should therefore leave a recognizable signature on the genome[35,48,66,80,91,197].

For decades, researchers have focused on understanding the link between tRNA concentration and translation rates of cognate codons, under the assumption that ribosomal dwell-time on a particular codon is partially determined by diffusion limited tRNA binding and competition between near-cognates[23,50,51]. Indeed, multiple lines of evidence strongly support this hypothesis in a multitude of different organisms[52].

Recently, ribosome profiling — a technique that maps transcriptome-wide ribosome occupancy — has been applied to study whether different codons show variation in translation rates, but researchers have come to conflicting conclusions, even when using the same dataset[81–84,97]. One of the most startling findings to emerge from ribosome profiling experiments is the striking degree of heterogeneity in ribosome occupancy across mRNAs, which is punctuated by large peaks suggestive of 'pausing' or 'stalling'[97,148,149]. These pauses — in contrast to known stalling sequences — are orders of magnitude larger than what is expected from basal translation rate

variations due to tRNA concentrations and may instead result from nascent peptide interactions within the ribosomal exit tunnel (such as poly-proline sequences), ribosomal queuing, or trans-interactions between mRNA and ribosomes[70,84,97,100,198].

Using ribosome profiling, Li *et al.* (2012)[97] showed that, in bacteria, translational pauses were significantly associated with sequence binding between the anti-Shine-Dalgarno (aSD) sequence of the 16S ribosomal-RNA and the translating message. This binding interaction is important during the process of translation initiation, where the ribosome binds to the 5′ untranslated region (5′-UTR) to facilitate start codon recognition (Fig. 5.1A). However, the occurrence of these 'Shine-Dalgarno' (SD) sequences within coding sequences had not been previously associated with translational pausing[71,72]. SD sequence mediated pauses have now been documented for several bacterial species and independent ribosomal profiling datasets[97,149,199]. Studies have built on these results by showing SD-associated pauses *in vitro*, negative effects of SD sequences on protein production in engineered sequences, enhanced solubility of recombinant proteins via rational insertion of SD sequences at protein domain boundaries, and enrichment of SD sequences following trans-membrane domains of natural sequences[53,68,200–202].

By contrast, recent results have questioned whether the observed SD-associated pauses are actually an experimental artifact resulting from the ribosome profiling protocol — specifically the differential sizes of sequencing fragments[153,203]. Indeed,

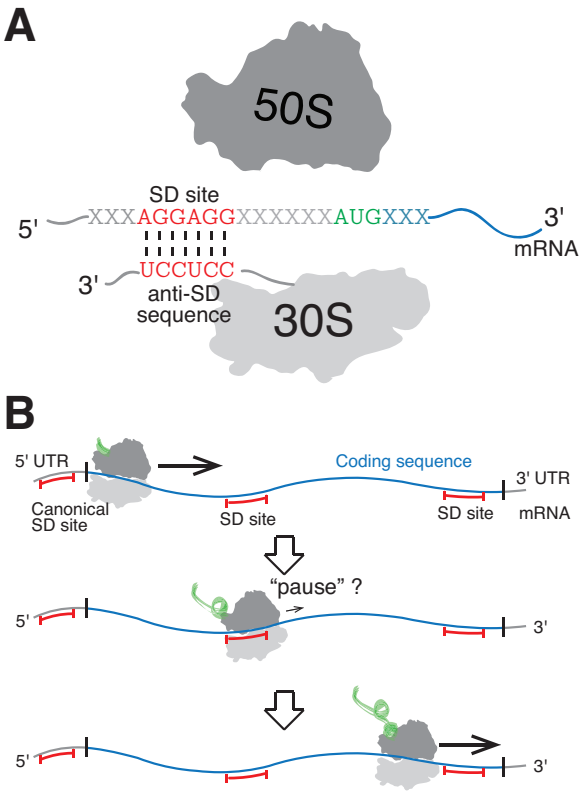

Figure 5.1. **The possible dual impacts of Shine-Dalgarno(SD) sequences on protein synthesis.** *A*, SD sequences in the 5′ untranslated region (UTR) of mRNA are known to facilitate translation initiation in bacteria via binding to the anti-SD sequence on the 3′ tail of the 16S ribosomal RNA. *B*, Recent research suggests that SD sequences within coding sequences may regulate the rate of translation elongation.

the existence of SD mediated pauses has not been confirmed using several other experimental methods[153,204,205]. It thus remains unclear what role, if any, SD sequences within protein coding genes have in modulating translation speed (Fig. 5.1B).

Even though the usage and diversity of SD sequences within the 5′-UTR has been analyzed extensively at the genome-scale[142,144,146], the occurrence pattern of these important sequence motifs within the coding sequences of diverse species has been largely neglected (though see Itzkovitz *et al.* (2010)[45] for an exception). Open questions thus remain as to whether SD sequences are indeed systematically depleted within coding sequences from diverse species, and if so, whether the depletion follows any particular pattern that may provide clues to the evolutionary significance of these sequences.

In order to answer these open questions, we sought to characterize the general occurrence of SD sequences within protein coding genes across a range of bacterial species of known phylogeny. We first present a metric to characterize single mRNA sequences according to their estimated sequence binding propensity with the ribosomal aSD sequence. Using this metric, we show that depletion of SD sequences in coding regions is a hallmark of bacterial genes and that, within a given species, the degree of this depletion is inversely correlated with measured gene expression levels. Finally, we show that variation in SD sequence depletion between different genomes is related to the minimal known doubling time of individual species, suggesting that depletion of SD sequences is driven by evolutionary pressure for greater translation efficiency.

5.3. Results

5.3.1. Quantifying the occurrence of SD sequences within coding sequences

We first counted the number of occurrences of the canonical SD motif (5′-AGGAGG-3′) within the coding sequences of the 187 bacterial species compiled by Viera-Silva *et al.* (2010)[91]. For each genome, we compared the number of SD sequences found within coding sequences to the number expected by chance using a codon-shuffled null model to control for codon usage bias within each gene (see Methods). We found that 175 out of 187 genomes contained fewer canonical SD sequences in their coding sequences than expected by chance (154 were significant at $p < 0.0001$, Monte Carlo hypothesis testing, Fig. 5.3A).

However, single or multiple base mismatches to the canonical SD sequence are frequently assumed to be functional in translation initiation and the strength of aSD sequence binding to different hexamer sequences spans a range of values. To quantify the occurrence of SD sequences on a per-gene basis in a manner that encapsulates the full breadth of this heterogeneity, we estimated the free energy of binding between the aSD sequence and each hexamer within the coding region of each mRNA (Fig. 5.2A, see Methods for details). Since the free energy of binding (ΔG) at a particular site is proportional to the logarithm of the ratio of the association and dissociation rate constants of binding, we define the affinity A of a hexamer $\{n_1...n_6\}$ to the aSD sequence as:

$$A_{\{n_1...n_6\}} \equiv \exp(|\Delta G_{\{n_1,...,n_6\}}|) \tag{5.1}$$

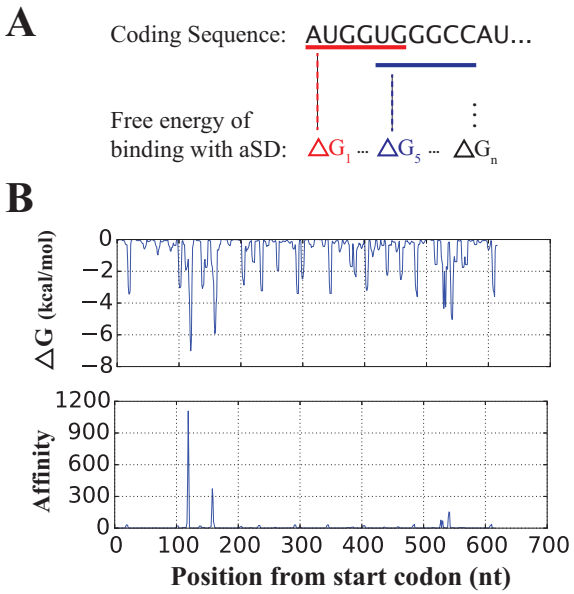

Figure 5.2. **Quantifying aSD sequence binding within coding regions.** *A*, We estimate the free energy of binding for each hexamer within a gene to the core anti-SD sequence (5′-CCUCCU-3′). *B*, Free energy (top) and affinity (bottom) profiles for a typical *E. coli* gene (b3055). The affinity profile amplifies the contribution from strongly binding regions within the gene.

We define the aSD binding score S of a gene as:

$$S_{gene} \equiv \log \bar{A}, \tag{5.2}$$

where $\bar{A}$ is the average affinity over a gene's coding sequence (Fig. 5.2B). The transformations involved in the definition of S aim to lessen the contribution of weak-binding interactions while amplifying the contributions from the strongest aSD binding sequences.

We calculated S_{gene} for each of the coding sequences of 187 bacterial species, and define genome aSD binding score $S_{genome} = \bar{S}_{gene}$. We again compared this empirical value to the expected value for a given genome based off a codon-shuffled null model and found that, similar to the previous analysis, 172 out of 187 genomes had average aSD binding scores lower than expected by chance (167 were significant at $p < 0.0001$, Monte Carlo hypothesis test, Fig. 5.3B). These results demonstrate that genomes contain significantly fewer SD sequences than would be expected from gene-specific codon usage biases and amino acid sequences.

5.3.2. The occurrence of SD sequences in coding regions correlates negatively with *E. coli* gene expression data

S_{gene} allows us to test whether variation in aSD sequence binding between different genes correlates with gene-level features such as expression level. We obtained five genome-scale expression datasets for *Escherichia coli* to ensure the robustness of our results (Supplementary Table C.2) and correlated the gene expression measurements

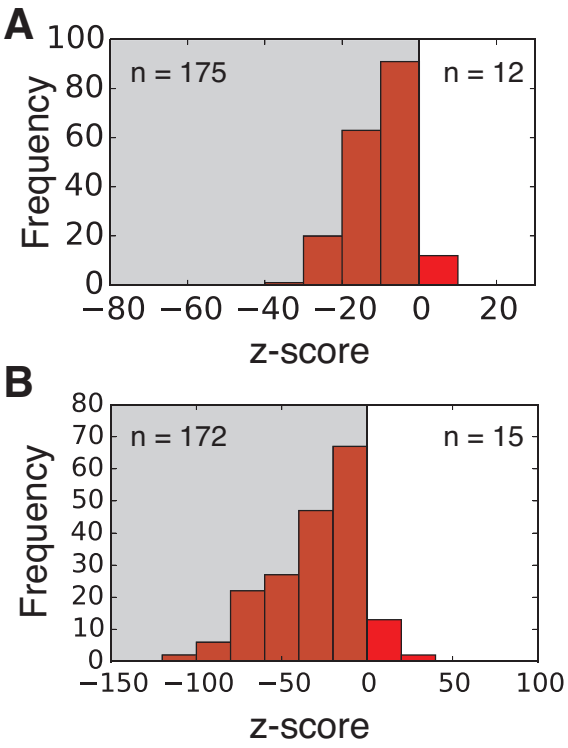

Figure 5.3. **Depletion of SD occurrence in genomes compared to expectation from 1000 randomly generated genomes using our codon-shuffled null model.** *A*, the canonical SD sequence AGGAGG is depleted within coding sequences in most genomes (175 of 187) and *B*, The genome aSD binding score S_{genome} is lower for most organisms (172 of 187). Both distributions are centered significantly to the left of zero showing that the majority of organisms avoid SD sequences according to both metrics.

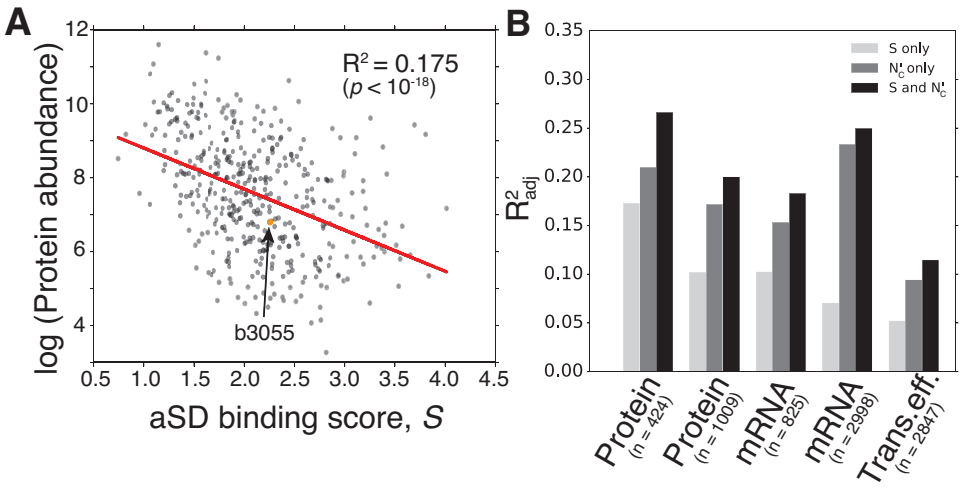

Figure 5.4. **aSD binding scores negatively correlate with gene expression in *E. coli*.** *A*, An example data set showing negative correlation between protein abundance and aSD binding scores for individual *E. coli* genes ($R^2_{adj} = 0.175$, $p < 10^{-18}$). Specifically, coding sequences containing fewer SD sequence motifs have higher protein abundances. *B*, Multivariate regression shows that expression changes cannot be fully explained by codon usage bias, and that additional predictive power is offered by S_{gene}. We chose 5 datasets that provide independent measurements of mRNA, protein, and translation efficiency levels in order to test the robustness of our findings[10,13,127,133].

against the calculated aSD binding score for each gene (Fig. 5.4A)[10,13,127,133]. We observed a highly significant negative relationship in all datasets indicating that the coding sequences of highly expressed genes contain fewer SD sequences ($p < 10^{-18}$, for all cases) (Fig. 5.4A,B).

A number of different factors are known to influence protein abundances, including start codon choice, mRNA structural accessibility, and SD sequence usage at translation initiation sites[132]. Here we wish to focus on the elongation phase of translational control to determine what, if any, additional predictive power is conferred by the effect of aSD sequence binding within coding sequences. Prior studies have established that the codon usage bias of individual genes is highly correlated with protein levels[52]. In order to investigate whether the observed correlation between S_{gene} and gene expression is driven solely by codon usage bias, we conducted multivariable linear regression using both S and an established method for quantifying codon usage bias to predict expression levels (N_c')[206]. If S were solely a consequence of codon usage bias, the adjusted-R^2 (R^2_{adj}) should *decrease* when S is included as an independent variable along with N_c'. On the contrary, we observe that the best model for all datasets includes both N_c' and S as predictors of expression (Fig. 5.4B, Supplementary Table C.2). While the enhancement in predictive power is not additive, this is not uncommon when evaluating models with multiple co-varying predictors.

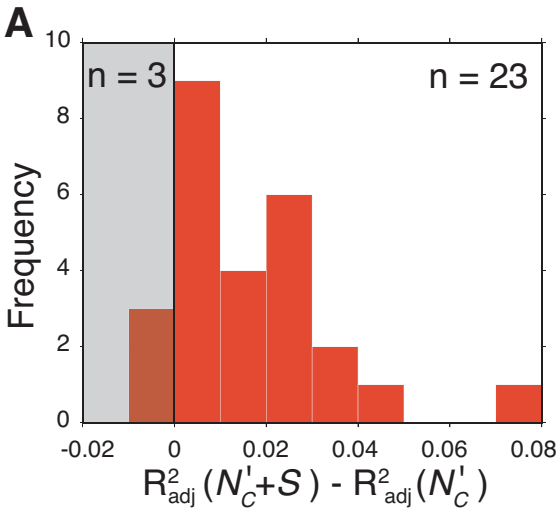

Figure 5.5. **Shine-Dalgarno sequence depletion is correlated with protein abundances in a diverse set of bacterial taxa.** Distribution of differences between the R^2_{adj} for models which do and do not contain the S score. For 23 of the 26 organisms, inclusion of aSD binding score as an independent variable enhances predictive power. The full data table including organism names and values is available in Supplementary Table C.3

5.3.3. The occurrence of SD sequences within coding regions correlates negatively with protein abundances in diverse bacterial taxa

To determine the generality of the previous finding, we expanded our analysis to 26 diverse bacteria for whom protein expression data was previously collected by Wang *et al.* (2015)[207] (see Methods). For 19 out of 26 datasets, we observed that S was significantly negatively correlated ($p < 0.01$) with protein abundances (Fig. 5.5, Supplementary Table C.3). As in the previous subsection, we also implemented a multivariate model to determine whether the observed correlation was solely a consequence of codon usage bias. For 23 out of 26 datasets we saw an improved R^2_{adj} value when S_{gene} is added as a predictor along with estimates of codon usage bias (Fig. 5.5).

We further confirmed the observation that the more complex multivariate model resulted in a better fit to the data by using AIC and BIC to evaluate model fits. For 22 and 18 organisms, respectively, the multivariate model provided a better fit to the data than a linear model based on codon usage bias alone (Supplementary Fig. C.1).

5.3.4. Ribosomal protein coding sequences contain fewer SD sequences than other genes

To overcome the limited availability of bacterial protein expression datasets, we next investigated whether ribosomal protein coding sequences contain fewer SD sequences than other genes within a genome. Ribosomal proteins are essential for all organisms

and they are generally expressed at high levels making them some of the most likely genes to show selection for accurate and efficient translation.

In *E. coli*, we observed that aSD binding scores for the 58 ribosomal protein genes are significantly lower than that of all other genes (Fig. 5.6A). To quantify the magnitude of this difference we define the normalized SD bias within a genome, B_{SD}, as:

$$B_{SD} = \frac{\bar{S}_{ribosome\,genes} - \bar{S}_{genome}}{\bar{S}_{genome}} * 100\% \tag{5.3}$$

where $\bar{S}_{ribosome\,genes}$ is the averaged S_{gene} for ribosomal protein coding genes, and $\bar{S}_{genome}$ is the averaged S_{gene} for all genes within a genome. When $B_{SD} < 0$, ribosomal protein genes contain fewer SD sequences than would be expected based on the genome-wide average. We opt for this approach for two primary reasons. First, the S values of ribosomal protein coding genes themselves would be heavily influenced by the underlying genomic GC content. Normalizing to the genome-wide average should help to mitigate this effect. Second, research has shown that at higher growth rates, ribosomal protein genes make up an increasingly larger fraction of bacterial proteomes[208]. Thus, relative differences in S between ribosomal protein coding genes and the genome as a whole should reflect the selective pressure for increased ribosomal protein production during periods of rapid growth.

Of the other 187 diverse bacteria spanning different genomic GC contents, growth environments and growth rates, 173 have $B_{SD} < 0$, suggesting that the vast majority

of bacteria have a larger depletion of SD sequences in their ribosomal protein coding genes relative to the genome as a whole (Fig. 5.6B). The systematic depletion of SD sequences in ribosomal protein coding sequences further suggests that these motifs negatively impact gene expression and/or cellular fitness in a wide-diversity of bacteria.

Previous studies have shown the relative codon usage bias of ribosomal genes compared to the rest of the genome is correlated with the minimum observed doubling time for particular species[91]. This finding is mechanistically assumed to be a consequence of the fact that, at rapid growth rates, ribosomal proteins constitute an increasingly large fraction of the proteome; selection for translational accuracy or efficiency within these genes relative to the genome thus likely reflects the evolutionary history driven by growth rate demands. We therefore hypothesized that B_{SD} scores may also be related to the growth rate demands of individual species. Indeed, we found that B_{SD} is positively correlated with the minimum known doubling times of this set of 187 bacteria — fewer SD sequences within the ribosomal protein coding sequences relative to the genome is associated with faster maximal growth rates (Spearman-rank: $\rho = 0.530$, $p < 10^{-14}$)(Fig. 5.6C). We further confirmed the robustness of this finding via phylogenetic generalized least squares regression (see Methods)($\lambda = 0.978$: $R^2_{adj} = 0.07$, $p = 0.0002$). This finding strongly suggests that SD motifs within coding sequences are detrimental to growth and reproduction likely via negatively impacting translation.

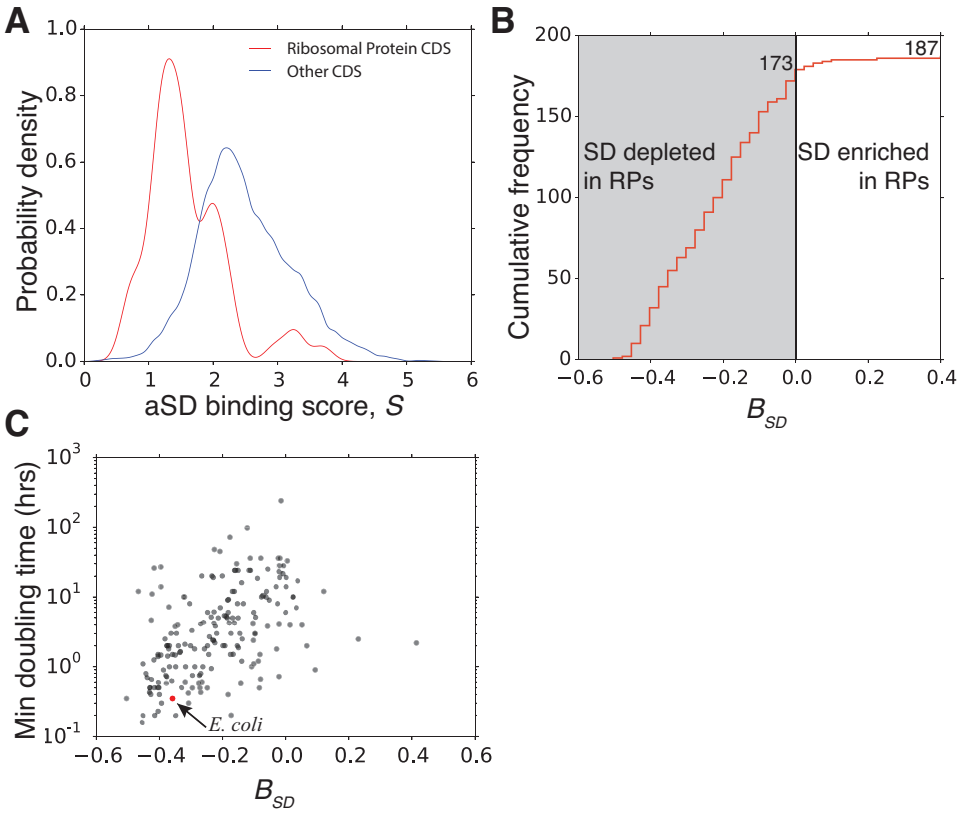

Figure 5.6. **Depletion of SD sequences within ribosomal protein coding genes is widespread throughout the bacterial kingdom and associated with organismal growth.** *A*, Distribution of aSD binding scores of ribosomal protein coding sequences in *E. coli*, compared to that of all other protein coding sequences. We characterize SD sequence usage bias in a genome with equation (3). *B*, Distribution of genome SD bias index for 187 bacteria genomes. Ribosomal proteins have significantly lower aSD binding scores, as compared to the rest of the genome, in the majority of bacterial species. *C*, SD bias is correlated with minimum generation time in 187 organisms (Spearman-rank: $\rho = 0.530$, $p < 10^{-14}$). Depletion of internal-SD sequences in ribosomal protein genes is associated with faster growth.

5.4. Discussion

Prior research into translation elongation has focused on codon usage as the primary means of modulating elongation speed, but recently, researchers have proposed that anti-Shine-Dalgarno mediated sequence interactions are a dominant source of translational pausing in bacteria[97,128]. If true, this finding has important consequences for our understanding of the basic mechanisms of translation as well as practical implications for coding sequence design for synthetic biology and biotechnological purposes. By quantifying the usage of SD sequences within coding sequences in a diverse set of bacterial taxa, we have shown a consistent trend whereby SD sequences within coding regions are systematically depleted. Specifically, this effect is strongest in the most highly expressed genes across a variety of genomes. We further show that the level of biased depletion of SD sequences is strongest in organisms capable of very rapid growth where selection for translation efficiency has previously been shown to produce a variety of genome-scale hallmarks[91].

Recently, Diwan and Agashe (2016)[172] published an elegant analysis of 'internal-SD-like' sequence usage in prokaryotes. Our results largely confirm the major finding of this study that showed internal-SD-like sequences are depleted in >80% of the species analyzed. While their results found a number of species that were exceptions to this rule, we note that many of these exceptions are Archaea, whose translation initiation mechanisms remain elusive and are therefore excluded from our analysis. Further, our results build on these findings in important ways. By developing a metric of S, which is defined at the single-gene level, our analysis provides insight into

within-genome variation and the selective pressures governing the usage of internal-SD sequences as it relates to gene expression costs. This within-genome analysis allows us to show that avoidance of SD sequences is highly related to the maximal growth rates of organisms using a method that controls for GC content variation, which Diwan and Agashe (2016)[172] found to impose an important constraint on the appearance of internal-SD sequences. Our analysis does not focus on temperature or variation in internal-SD usage with regard to position within genes, but the thorough results of Diwan and Agashe (2016)[172] likely hold within our data set.

There are several possible limitations to our methodology that readers should be aware of when interpreting our findings. First, our study relies on an assumed anti-Shine-Dalgarno sequence of 5′-CCUCCU-3′ to calculate aSD binding strength scores for individual genes. It is possible, and evidence strongly suggests, that in particular lineages the aSD sequence may be slightly altered or extended compared to this canonical sequence[165]. We may therefore be mis-characterizing the aSD sequence for several species in our dataset, or not encompassing the full breadth of possible sequence interactions. Future work can refine our findings to account for this aSD heterogeneity as more aSD sequences will be empirically determined, but we opt here for a conservative approach likely to be applicable for the majority of organisms in our dataset. Second, while our study relies on the precise definition of coding sequences bounds in existing genome annotations, prior research has shown that these annotations are likely spurious for up to ~10% of annotated genes[149,209].

However, reliable N-terminal mapping is currently available for only a small fraction of bacterial genomes; until better computational models are developed to refine translational start site predictions, this will remain a limitation that adds noise to any computational genome-scale analysis, such as the one we perform here.

SD sequences may be avoided within coding sequences for several different — and non-mutually exclusive — reasons. These sequences may: (i) result in erroneous internal translation initiation leading to the production of truncated protein products (ii) temporarily sequester ribosomes, thus limiting the number available for proper translation initiation (iii) encourage translational frameshifting or (iv) substantially slow down translation elongation[97,168,210,211]. In all of these cases, we would expect SD sequences within coding sequences to be largely detrimental and thus avoided. In particular, given that the consequences of any of the above explanations is amplified by high mRNA copy numbers, avoidance of these SD sequences would also be expected to manifest particularly in the most highly expressed genes.

Although our results indicate that SD sequences are by and large detrimental, we also wish to clarify that some proportion of the SD sites within coding sequences may serve important functions. Owing to the compact nature of bacterial genomes, the translation initiation site of many genes within operons will occur within the 3' terminus of the preceding coding sequence. Further, the presence of multiple translation initiation sites may serve a regulatory role for certain proteins, allowing for the production of distinct isoforms depending on the N-terminal sequence or controlling protein folding rates[68,149,202,212].

One benefit of our large-scale analysis is that exceptions to the rules can point to interesting cases for further study. In Fig. 5.5 we found three species where S did not appear to enhance predictions of protein abundance: *Mycoplasma pneumoniae*, *Shigella flexneri*, and *Leptospira interrogans*. Although none of these species are known to use non-canonical aSD sequences[165], all are pathogenic species, suggesting that a possible relationship may exist between ecological strategies, effective population size, and the selection against SD sequences. However, owing to the large number of pathogenic species in this dataset, this finding will require further detailed investigation. Additionally, several species analyzed in Fig. 5.6 showed an enhancement of SD sequence usage within ribosomal proteins relative to the genome. Nearly all of these cases come from 3 distinct orders (phyla), pointing to likely mechanistic changes in the aSD interaction in particular clades: Rickettsia (Alphaproteobacteria), Mollicutes (Tenericutes) and Spirochaetes (Spirochaete) (both *M. pneumoniae* and *L. interrogans*, mentioned above, fall within one of these orders). Future ribosome profiling experiments on species from within these clades may provide clues on the evolution of the aSD sequence interaction.

The patterns that we observe provide significant insight into the debate surrounding the usage of SD sequences within protein coding genes. Moreover, our results are fully orthogonal to ribosome profiling based conclusions. It is clear from this bioinformatic analysis that SD sequences are largely avoided across the bacterial kingdom, and that this avoidance is likely due to deleterious effects on translation. We thus conclude that even if SD mediated elongation pausing is an artifact of the

ribosomal profiling protocol as suggested by Mohammad *et al.* (2016)[153], care should be taken to avoid SD sequences when designing coding sequences for recombinant protein production applications.

5.5. Materials and Methods

5.5.1. Codon-shuffled null model

We randomly generated null model genomes that preserve codon usage and primary amino acid sequence at the gene level. For each gene, we constructed a list of all codons used in the original sequence. Given the primary amino acid sequence of the gene, we then randomly selected a codon from the pool of available synonymous codons for that particular amino acid without replacement. The start and stop codons are not affected by this process and thus remain fixed during the shuffling process. We repeated this procedure for every gene within a given genome in order to create one instance of a randomized genome for null model comparison. For statistical comparison using Monte Carlo hypothesis testing, we created 1000 randomized genomes in this manner. Using our metric, we calculated the mean and standard deviation in these randomized genomes for each organism, and then calculated a z-score for the real genome along with the resulting p-value, which we report in the main text.

5.5.2. aSD hybridization

We predicted thermodynamic interactions between the anti-SD (aSD) sequence and each six-nucleotide long sequence using the RNA co-fold method of the ViennaRNA Package 2.0 with default parameters[163]. For this study, we have chosen to use the canonical core aSD sequence of 5′-CCUCCU-3′ for all species, owing to the fact that this core sequence is nearly universally conserved. Further, the 3′-tail of 16s rRNAs is slightly variable and poorly annotated[144,165] making it difficult to empirically determine the precise aSD sequence for each individual species.

5.5.3. Pax-Db data collection

We collected the complete bacterial dataset from the Protein Abundance Across Organisms Database (Pax-Db) in August 2015[207]. This resource contains protein abundance measurements for 26 different bacteria. When multiple datasets were available for a particular organism, we chose the 'Integrated' dataset, which is the result of Pax-Db curators integrating the various protein abundance data sources based on coverage and quality. The full set of data that we analyzed for each species is available upon request.

5.5.4. Growth-rate dataset and phylogenetic relatedness

We obtained growth rate measurements (minimum doubling time, measured in hours) from Viera-Silva *et al.* (2010)[91]. For each species in their data table, we matched the name of the species provided in the original data source to the species name

in a local copy of the NCBI genbank complete genome sequences. This resulted in 187 matches for bacteria (Archaeal species, which were provided in the original dataset, were ignored for the purposes of this study). Within each of these bacterial genomes, we relied on annotations in the genbank files to find ribosomal proteins by searching the 'product' field for 'ribosomal subunit', or perturbations thereof. Full data including genbank files for all relevant organisms, and ribosomal protein 'locus_tags' used in this study is available upon request.

To construct a phylogenetic tree from these species, we extracted the 23S and 16S gene sequences using RNAmmer-1.2[192]. When multiple sequences were available for a given genome we randomly chose one of each for alignment. We then individually aligned 23S and 16S sequences using MUSCLE[193]. Finally, we concatenated the 16S and 23S alignments for each organism and constructed a maximum likelihood (ML) tree using RAxML with a partitioned analysis that separately fit rate models for the 16S and 23S sequences. We used a GTRGAMMA evolutionary model with 100 rapid bootstrap searches and 20 ML searches and selected the best fitting ML tree[194].

5.5.5. Regression analyses

With one exception noted below, all statistical analyses were performed using the SciPy (version: 0.16.0) and StatsModels (version: 0.6.1) packages in Python.

To control for phylogenetic effects in our growth rates regression analysis, we used the PGLS function from the 'caper' package in R, choosing the optimal lambda value to transform our input tree via maximum likelihood search.

CHAPTER 6

Concluding Remarks

In the Introduction, I discussed the fact that a relatively small 100 amino acid protein can be coded for by $\approx 10^{48}$ unique synonymous gene sequences. Assuming the same rate of transcription, each of these *possible* constructs would nevertheless produce variable amounts of active protein over time owing largely to organism-specific rates of translation efficiency, translation accuracy, and transcript degradation. Over the course of my research, I have focused on bacterial translation efficiency and have advanced our understanding of its link with mRNA sequence in several fundamental ways. When my work began, my intention was to investigate the process of translation elongation, specifically: how the choice of synonymous codons could affect elongation rates. However, in my initial studies described in Chapter 2, I showed that the demands for efficient translation initiation—weak mRNA structure surrounding the start codon—constrain synonymous codon usage. Previous models of codon usage bias, which are applied frequently throughout the biological literature, have ignored this effect, but my work showed how position-dependent codon usage bias can be incorporated existing models to increase statistical power and accuracy.

Based on this initial work, I switched my focus away from codon usage bias to instead look at the sequence control of translation initiation. In Chapters 3 & 4, I

investigated the causes and consequences of variation in the prevalence of a prominent translation initiation motif—known as the Shine-Dalgarno sequence—within and between bacterial genomes. I utilized existing ribosome profiling measurements to validate and extend long-standing results relating to the control of translation efficiency by the Shine-Dalgarno sequence. I showed that genes with intermediate levels of complementarity to the anti-Shine-Dalgarno sequence show the highest levels of translation efficiency, and that the extents of the anti-Shine-Dalgarno sequence binding interaction can be defined in a data-driven manner. I next built on this work by analyzing a large data-set of whole-genome sequences, and applied phylogenetic comparative methods to show that the overall levels of Shine-Dalgarno sequence occurrences within a genome is largely predictable based on the maximal growth-rates of organisms.

Finally, in Chapter 5 I returned to look more closely at coding sequences. Having analyzed the role of the Shine-Dalgarno sequence in governing initiation rates and the evolutionary pressures that shape variation in genome-wide SD sequence utilization, I asked: to what degree do coding sequences themselves avoid these important motifs and what are the likely consequences for elongation when internal-Shine-Dalgarno sites occur? Mentoring an undergraduate researcher throughout this project, we showed that internal-Shine-Dalgarno sites constrain coding sequence evolution at the individual gene and genome levels. *Genes* with more internal Shine-Dalgarno sequences tend to be expressed at lower levels, and *genomes* with more internal-Shine-Dalgarno sequences tend to be from organisms that grow slowly.

6.1. Common threads

Taken together, several common threads run through and unite this research program which are worthy of discussion here. First, it is essential to note that none of this research would have been possible a decade ago. Throughout all of these chapters, much of my work relied on the availability of high-quality experimental datasets, whether they be measurements of mRNA and protein abundances or estimates of the minimum doubling time for individual species. During the course of my PhD, the technique of ribosome profiling was first applied to bacterial species providing an even richer set of data and novel opportunities for researchers interested in understanding translational regulation. On top of this, nearly all of my work relied on the availability of high-quality genome-sequences from evolutionarily diverse genomes.

Next, it is deserving of note that the abundance of data sources currently available for analysis presents its own set of unique challenges. *Integrating* fragments of data from diverse sources in the search for truth is an essential part of the scientific endeavor. Knowing which sources to trust, and deciphering which may provide novel insight into a particular biological problem is essential to the work of data-driven computational biology. Further, as experimental protocols (such as ribosome profiling) grow increasingly complicated, advanced statistical techniques are essential to account for bias, sources of error, and to limit the prevalence of false positive and false negative results. A fundamental advance of every chapter in this thesis was the development of a new way to quantify and summarize a complex biological phenomenon. This work is far from done, and in many respects I suspect we are just barely

scratching the surface. Summarizing a complex biological phenomenon like translation initiation or elongation rates into a single (or even a small set of) meaningful number/s in the hope that we will uncover a meaningful biological relationship is a monumental challenge.

Finally, my work has been united by the effort to quantify *how much we know* about particular phenomena. This point is perhaps best illustrated with an example: using classical experimental techniques, researchers have shown for decades that the Shine-Dalgarno sequence is important for translation initiation. Researchers can (and have) taken genes such as GFP or lacZ, put them on a plasmid, and holding all else as constant as reasonably possible, shown what happens to the protein levels in two parallel experiments when they manipulate some facet of the SD sequence interaction[71]. Moving the Shine-Dalgarno sequence too close or too far to the start codon results in decreased protein expression, and disrupting the Shine-Dalgarno sequence through point mutations can progressively abolish the ability of mRNAs to be translated[134,136,137]. Despite all of these findings, however, open questions remained as to how repeatable these findings were if different genes were used, if the genes were on plasmids or chromosomally integrated, if they were placed behind a strong or weak promoter, if the cells were grown in rich or minimal medium, etc. In short, we've known for a long time that the Shine-Dalgarno sequence *matters* for translation and that is a critical first step. But *how much* it matters in the context of all the other variation that affects gene expression was an open—and in my mind just as critical—question. As I describe in Chapter 3, the conclusions here are rough and

subject to the limitation that ribosome profiling based measurements of translation efficiency are very noisy. But researchers including myself were nevertheless surprised to find that at best, knowledge of the Shine-Dalgarno sequence interaction explains roughly 10% of the variation in genome-wide translation efficiency. Armed with that knowledge, researchers can now ask a host of other important questions relating to the conditions under which utilization of this sequence might evolve for particular genes, and what other factors may be largely responsible for the 'missing' 90% of variation.

6.2. The contribution of systems and synthetic biology

Systems biology is a sub-discipline of biology that likely has as many definitions as practitioners. But counting myself among the latter, I view the challenge of systems biology as 'putting the pieces back together'. Reductionist approaches have cataloged a host of important factors relating to the growth and maintenance of organisms. The task that many in the biological community are attempting under the guise of systems biology is to see how these anecdotes and fragments can fit together to quantitatively describe the behavior of individual pathways, cells, and organisms. This task relies strongly on researchers having laid the groundwork to identify key players involved in a system. But at a certain point, to advance our understanding of whole-cells and systems, someone has to zoom back-out and ask *'how well are we doing?'* in order to see the gaps and guide future studies. My on-going research philosophy is that for biologists to interrogate and ultimately understand increasingly complex systems, we must move towards codifying the decades of important anecdotes and research

findings into a predictive, quantitative framework. The work presented in this thesis represents my attempts at doing just that for the field of bacterial translational control.

Synthetic biology, often mentioned in the same breath of systems biology, has grown into its own formidable sub-discipline and is allowing researchers to test a host of novel hypotheses at unprecedented scale. While many practitioners of synthetic biology are driven by end goals of producing particular molecules for biotechnological purposes, this practical goal can lead to important basic science findings. In the context of translational regulation, it is researchers in the synthetic biology community who have driven knowledge integration by measuring libraries of tens of thousands of different constructs and asking how much variation in one factor— such as presence or absence of the Shine-Dalgarno sequence—can explain the results. Reductionist approaches are essential for finding molecules and pathways of interest that are necessary and sufficient for certain phenomenon. Synthetic approaches, by contrast, allow researchers to *test the numbers*. By transferring genes or pathways to different organisms, researchers can test the extent to which we observe the predicted behavior and how much variation in the behavior of interest can be described by the purported mechanism. In essence, large-scale synthetic biology studies continue to provide the most stringent test in terms of determining the state of our knowledge and whether we truly understand a system. The results don't always look pretty. Despite decades of research that has been cited throughout this thesis relating to the importance of codon usage bias, synthetic approaches are beginning to show that it is *relatively*

easy to remove entire codons from an organism's genome with little apparent consequence for fitness[213–215]. Such findings should be cause for serious alarm that we're missing something important in our existing understanding of codon usage, and it's difficult to see how this knowledge would have been gained from anything other than large-scale synthetic attempts to recode an entire genome.

6.3. A modest proposal

While the previous section focused on some larger issues relating to science, the state of biology, and the common themes that unite my research program, here I wish to turn more directly back to protein translation and ask what the future may hold. Dramatic reductions in the cost associated with sequencing technologies have resulted in an explosive growth in the number of genome-wide applications that leverage sequencing data in order to better understand cellular processes. Additionally, near-constant breakthroughs in microscopy and mass-spectrometry continue to make it easier for researchers to measure, with increasing accuracy, the abundance of individual molecules through time and space. It is clear that technological breakthroughs will continue, and speculation past a few short years is likely a fools-errand. But in the short term, several advances seem relatively predictable and provide reason for excitement about our future understanding of the sequence based determinants of translation efficiency.

As a thought experiment, I like to imagine what experimental datasets I wish that I had access to in an ideal world—without time or budgetary constraints. As a starting point, I would want to know the abundances of all the mRNAs and protein

molecules for a particular species in a particular growth condition. Since the link between these two molecular species can be accomplished in a number of ways, I would also like to know the instantaneous rates of transcription, translation, and degradation of both molecular species, which can be accomplished through a variety of existing methods. I would also want to have structural models for both the mRNA and protein species for all genes. This hypothetical dataset could provide a population level snapshot into nearly every aspect of gene expression, and now I could think about extending it from the population level to acquiring each of these measurements at the level of single-cells. I'd still be left with a snapshot, so I'd really need a time-course of all this information to understand dynamic changes. Still, I'd be left with an abundance of data about a single organism in a particular set of growth conditions, so I could also envision repeating the very same process for the same organism in different growth conditions (that would hopefully mimic native conditions as much as possible). Oh, and repeating the entire process for a few different closely related strains/species (10s? 100s? 1000s?) would be pretty nice in order to link genetic variation to observed phenotypic differences.

This is a very 'systems biology' view. The parallel 'synthetic biology' dream experiments would look a little bit different and are equally interesting to note. Rather than studying natural evolution, the same measurements noted above could be performed on recombinant constructs. So I'd start with GFP, for instance, and look at perhaps thousands or tens of thousands of synonymous variants. For each variant, I would want single-cell distributions of both mRNA and protein abundances,

of course. Ideally, I'd have those measurements alongside instantaneous transcription, translation and degradation rates as above. Essential to these studies would be measurements of growth rates and fitness for the different constructs to determine how they individually impact the fitness of cells. Rather than choosing synonymous constructs at random, a perhaps richer source of data that would more closely mimic natural evolutionary processes would be to look at perhaps 100 different synonymous constructs, and for each construct develop deep mutational scanning libraries to explore the *local* fitness landscapes consisting of single or double synonymous mutations. Of course, the reality is that I would like generalizable results so the same process would have to be repeated for the same protein under several promoter::ribosome binding site libraries and genomic integration sites, and in several different growth media, temperatures, etc. Finally, as lovely as all this data would be, a single recombinant protein can only be generalized so far. I'd want to repeat this for at least several different proteins, particularly from different structural classes and preferably ones that provide a functional output for the organism.

The reality is that nearly every experiment that I mentioned in both of these dream scenarios is highly possible, and has been done to some degree by different labs. The challenge is simply combining these techniques in a single-lab experiment under common and repeatable growth conditions—although, in reality I'd want each experiment performed by a few different labs for reasons of error-estimation and robustness. Well, the other challenge is, of course, what on earth one would *do* with all this data. How could we analyze it to get meaningful knowledge? I would argue

that answering this question *before* we actually have the problem of dealing with all of this data is essential to limiting researcher degrees of freedom and producing robust, reproducible results. It seems unlikely that someone with enough time on their hands to actually perform all of these experiments would have also had the time to dedicate a decade of their life to learning the proper statistical analyses required to deal with data having this level of complexity, and so computational biologists can and should play a large role in anticipating these future experiments and developing robust analytical pipelines.

6.4. Limitations of existing approaches

Perhaps the biggest issue surrounding many current and past studies related to sequence based control of gene expression is the sheer number of possible synonymous constructs to evaluate. Quite frankly, it is not sufficient to take a given gene, make two or three synonymous versions using 'optimal' codons and two or three using 'sub-optimal' codons, and expect to answer an interesting functional question. To be sure, you may find differences that *appear* interesting, but under-powered studies of this sort are highly susceptible to false positives. A researcher could choose any combination from the $> 10^{25}$ possible synonymous coding sequences that use 'preferred' codons and test their hypothesis against any combination of the other $> 10^{25}$ sequences that do not. How a researcher chooses their constructs, or how they even *should* in an ideal case is an interesting question. But if the claim is to make a mechanistic link related to something as diverse as codon usage biases, it's absolutely essential to show these results for tens, hundreds, or thousands of constructs; there

are simply to many moving parts that can't be controlled for in smaller studies. Constructs that vary in their codon usage will inevitably vary in their RNA secondary structure, presence of particular binding motifs, etc. Without large numbers to minimize these other factors, it is impossible to determine the ultimate cause with any degree of reliability. This is precisely why I think the era of systems and synthetic biology is uniquely suited to studying how individual properties of coding sequences influence translation efficiency. Even a high-throughput approach that studies 10^8 constructs wouldn't begin to scratch the surface of the possible synonymous diversity for an amino acid sequence, but it is nevertheless far more difficult to get spurious false positive results with datasets of this size.

Another important area that biologists of all sub-disciplines must improve upon is benchmarking different methods to set specific methodological standards. If a researcher is comparing the mean values of some trait between two populations, there are but a small number of accepted ways to quantify the statistical difference between them—first and foremost being the t-test. For more complex problems such as measuring codon usage bias, differential expression from RNA-sequencing data, codon occupancy from ribosome profiling data, etc. there are absolutely no accepted standards. Researchers instead must seemingly choose at random from a tremendous array of different methods. Unless they are highly skilled in understanding the statistical assumptions and limitations behind each method, users of these methods seemingly must hope that their analysis and chosen method is appropriate. To be sure, there have been efforts to benchmark a large variety of techniques including

differential expression analysis, but poorly performing methods continue to be used. Different methods applied to the same data can support different conclusions, making it critical that researchers understand the methods they use and—when possible—that these methods are standardized.

6.5. Parting words

Despite the notes of pessimism embedded in the preceding discussion, I remain supremely optimistic that the near future will provide answers to many of the biggest problems surrounding translation efficiency. Experimental and computational methods are continually improving and providing insight into fundamental biological processes in ways that researchers could scarcely have imagined even a decade ago. This entire thesis, is simply a reminder that no one can do it by themselves. This is a computational biology thesis, and the work that I've described would not have been possible without the efforts of countless experimental researchers. As much as computational biologists may rely on the data of experimentalists, so to must experimentalists increasingly rely on computational biologists to design better experiments, and to analyze those experiments in a manner that is free from bias and utilizes as much of the available information as possible in a statistically rigorous manner. *Interpreting* results and asking how they fit into, and/or disrupt our existing knowledge of molecules, pathways, cells, organisms, populations, and ecosystems is what makes a researcher a biologist, no matter the methods.

References

1. Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–92, jul 2005.

2. Irit Shachrai, Alon Zaslaver, Uri Alon, and Erez Dekel. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Molecular Cell*, 38(5):758–67, jun 2010.

3. Matt Eames and Tanja Kortemme. Cost-Benefit Tradeoffs in Engineered lac operons. *Science*, 339(August 2011):911–915, 2012.

4. Moshe Kafri, Eyal Metzl-Raz, Ghil Jona, and Naama Barkai. The Cost of Protein Production. *Cell Reports*, 14(1):22–31, 2016.

5. Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

6. Ali Mortazavi, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):1–8, 2008.

7. Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):1344–1349, 2008.

8. Ryan Lister, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, 133(3):523–536, 2008.

9. Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mRNA and protein in complex biological samples. *FEBS Letters*, 583(24):3966–3973, 2009.

10. Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, jul 2010.

11. Christine Vogel, Raquel de Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce A Shapiro, Suzanne C Burns, Devraj Sandhu, Daniel R Boutz, Edward M Marcotte, and Luiz O Penalva. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6(400):1–9, aug 2010.

12. Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, mar 2012.

13. Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology*, 25(1):117–24, jan 2007.

14. Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, may 2011.

15. Jingyi Jessica Li, Peter J Bickel, and Mark D Biggin. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2:e270, 2014.

16. Gábor Csárdi, Alexander Franks, David S. Choi, Edoardo M. Airoldi, and D. Allan Drummond. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLOS Genetics*, 11(5):e1005206, 2015.

17. Zhe Cheng, Guoshou Teo, Sabrina Krueger, Tara M Rock, Hiromi Wl Koh, Hyungwon Choi, and Christine Vogel. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Molecular systems biology*, 12(1):855, 2016.

18. Fredrik Edfors, Frida Danielsson, Björn M Hallström, Lukas Käll, Emma Lundberg, Fredrik Pontén, Björn Forsström, and Mathias Uhlén. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology*, 12(883):1–10, 2016.

19. Douglas W Selinger, Rini Mukherjee Saxena, Kevin J Cheung, George M Church, and Carsten Rosenow. Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation. *Genome Research*, pages 216–223, 2003.

20. Jonathan A Bernstein, Pei-hsun Lin, Stanley N Cohen, and Sue Lin-chao. Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *PNAS*, 101(9):2758–2763, 2004.

21. Simen M Kristoffersen, Chad Haase, M Ryan Weil, Karla D Passalacqua, Faheem Niazi, Stephen K Hutchison, Brian Desany, Anne-Brit Kolstø, Nicolas J Tourasse, Timothy D Read, and Ole Andreas Økstad. Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome biology*, 13(4):R30, 2012.

22. Tobias Maier, Alexander Schmidt, Marc Güell, Sebastian Kühner, Anne-Claude Gavin, Ruedi Aebersold, and Luis Serrano. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7(511):1–12, jul 2011.

23. T Ikemura. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, 151(3):389–409, sep 1981.

24. Stefan Klumpp, Matthew Scott, Steen Pedersen, and Terence Hwa. Molecular crowding limits translation and cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42):16754–9, 2013.

25. Idan Frumkin, Dvir Schirman, Aviv Rotman, Fangfei Li, Zahavi Liron, Ernest Mordret, Omer Asraf, Song Wu, Sasha F Levy, and Yitzhak Pilpel. Gene Architectures that Minimize Cost of Gene Expression. *Molecular cell*, 65:142–153, 2017.

26. Xiaoshu Chen and Jianzhi Zhang. No gene-specific optimization of mutation rate in Escherichia coli. *Molecular Biology and Evolution*, 30(7):1559–1562, 2013.

27. Xiaoshu Chen, Jian Rong Yang, and Jianzhi Zhang. Nascent RNA folding mitigates transcriptionassociated mutagenesis. *Genome Research*, 26(1):50–59, 2016.

28. Ben Lehner. Conflict between noise and plasticity in yeast. *PLoS Genetics*, 6(11), 2010.

29. Gajinder Pal Singh. Coupling between noise and plasticity in E. coli. *G3 (Bethesda, Md.)*, 3(12):2115–20, 2013.

30. Yuheng Huang and Aneil F. Agrawal. Experimental Evolution of Gene Expression and Plasticity in Alternative Selective Regimes. *PLoS Genetics*, 12(9):1–23, 2016.

31. Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin K O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, 2006.

32. Daniel L Jones, Robert C Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science (New York, N.Y.)*, 346(6216):1533–1536, 2014.

33. Noreen Walker, Philippe Nghe, and Sander J Tans. Generation and filtering of gene expression noise by the bacterial cell cycle. *BMC biology*, 14(1):11, 2016.

34. Michael Lynch and John S Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, 2003.

35. Nina Stoletzki and Adam Eyre-Walker. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular Biology and Evolution*, 24(2):374–81, feb 2007.

36. Premal Shah and Michael a Gilchrist. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10231–6, jun 2011.

37. Brian Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(March):195–205, 2009.

38. Paul M Sharp, Laura R Emery, and Kai Zeng. Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B*, 365:1203–1212, 2010.

39. Way Sung, Matthew S Ackerman, Samuel F Miller, Thomas G Doak, and Michael Lynch. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(45):18488–92, 2012.

40. Craig A Fogle, James L Nagle, and Michael M Desai. Clonal Interference, Multiple Mutations and Adaptation in Large Asexual Populations. *Genetics*, 180(December):2163–2173, 2008.

41. Gregory I Lang, Daniel P Rice, Mark J Hickman, Erica Sodergren, George M Weinstock, David Botstein, and Michael M Desai. Pervasive genetic hitchiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.

42. Stephen J. Freeland and Laurence D. Hurst. The genetic code is one in a million. *Journal of Molecular Evolution*, 47(3):238–248, 1998.

43. S J Freeland, R D Knight, L F Landweber, and L D Hurst. Early fixation of an optimal genetic code. *Molecular biology and evolution*, 17(4):511–8, 2000.

44. Shalev Itzkovitz and Uri Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, 17(4):405–412, 2007.

45. Shalev Itzkovitz, Eran Hodis, and Eran Segal. Overlapping codes within protein-coding sequences. *Genome Research*, 20(11):1582–9, nov 2010.

46. W. Fiers, R. Contreras, F. Duerinck, G. Haegmean, J. Merregaert, W. Min Jou, A. Raeymakers, G. Volckaert, M. Ysebaert, J. Van de Kerckhove, F. Nolf, and M. Van Montagu. A-Protein gene of bacteriophage MS2. *Nature*, 256:273–278, 1975.

47. Paul M Sharp and Wen-hsiung Li. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic*

Acids Research, 15(3):1281–1295, 1987.

48. Paul M Sharp, Elizabeth Bailes, Russell J Grocock, John F Peden, and R Elizabeth Sockett. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33(4):1141–53, jan 2005.

49. Grzegorz Kudla, Andrew W Murray, David Tollervey, and Joshua B Plotkin. Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, 324(April):255–258, 2009.

50. Eduardo P C Rocha. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 14(11):2279–86, nov 2004.

51. Mario dos Reis, Renos Savva, and Lorenz Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32(17):5036–44, jan 2004.

52. Tamir Tuller, Yedael Y Waldman, Martin Kupiec, and Eytan Ruppin. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences*, 107(8):3645–50, feb 2010.

53. Fabienne F V Chevance, Soazig L Guyon, and Kelly T Hughes. The effects of codon context on in vivo translation speed. *PLoS Genetics*, 10(6):e1004392, 2014.

54. Caitlin E Gamble, Christina E Brule, Kimberly M Dean, Stanley Fields, and Elizabeth J Grayhack. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell*, 166(3):679–690, 2016.

55. Chris A Brackley, M Carmen Romano, and Marco Thiel. The dynamics of supply and demand in mRNA translation. *PLoS Computational Biology*, 7(10):e1002203, oct 2011.

56. Sebastian Pechmann and Judith Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology*, 20(2):237–243, dec 2012.

57. Wenfeng Qian, Jian-Rong Yang, Nathaniel M. Pearson, Calum Maclean, and Jianzhi Zhang. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. *PLoS Genetics*, 8(3):e1002603, mar 2012.

58. Cristian Del Campo, Alexander Bartholomäus, Ivan Fedyunin, and Zoya Ignatova. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genetics*, pages 1–23, 2015.

59. David H Burkhardt, Silvi Rouskin, Yan Zhang, Gene-wei Li, Jonathan S Weissman, and Carol A Gross. Operon mRNAs are organized into ORF- centric structures that predict translation efficiency. *eLife*, 6(e22037):1–23, 2017.

60. Ariel Hecht, Jeff Glasgow, Paul R Jaschke, Lukmaan A Bawazer, Matthew S Munson, Jennifer R Cochran, Drew Endy, and Marc Salit. Measurements of translation initiation from all 64 codons in E . coli. *Nucleic Acids Research*, pages 1–12, 2017.

61. Vladimir Presnyak, Najwa Alhusaini, Brenton R Graveley, Jeff Coller, Vladimir Presnyak, Najwa Alhusaini, Ying-hsin Chen, Sophie Martin, Nathan Morris, Nicholas Kline, Sara Olson, David Weinberg, Kristian E Baker, Brenton R Graveley, and Jeff Coller. Codon Optimality Is a Major Determinant of mRNA Stability. *Cell*, 160(6):1111–1124, 2015.

62. Grégory Boël, Reka Letso, Helen Neely, Nicholson Price, Kam-ho Wong, Min Su, Jon D Luff, Mayank Valecha, John F Hunt, John K Everett, Thomas B Acton, Rong Xiao, Gaetano T Montelione, Daniel P Aalberts, and John F Hunt. Codon influence on protein expression in E . coli correlates with mRNA levels. *Nature*, 529(7586):358–363, 2016.

63. Premal Shah and Michael a Gilchrist. Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS genetics*, 6(9):e1001128, sep 2010.

64. Emily B Kramer and Philip J Farabaugh. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*, 13:87–96, 2007.

65. Beatrice C Ortego, Jeremiah J Whittenton, Hui Li, Shiao-chun Tu, and Richard C Willson. In Vivo Translational Inaccuracy in Escherichia coli: Missense Reporting Using Extremely Low Activity Mutants of Vibrio harveyiLuciferase. *Biochemistry*, 46:13864–13873, 2007.

66. D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52, jul 2008.

67. Gong Zhang, Magdalena Hubalewska, and Zoya Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology*, 16(3):274–80, mar 2009.

68. Nir Fluman, Sivan Navon, Eitan Bibi, and Yitzhak Pilpel. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *eLife*, 3:1–19, jan 2014.

69. Chien-hung Yu, Yunkun Dang, Zhipeng Zhou, Cheng Wu, Fangzhou Zhao, Matthew S Sachs, and Yi Liu. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell*, 59(5):744–754, 2015.

70. Premal Shah, Yang Ding, Malwina Niemczyk, Grzegorz Kudla, and Joshua B Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–601, 2013.

71. Howard M Salis, Ethan A Mirsky, and Christopher A Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–50, oct 2009.

72. J Shine and L Dalgarno. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences*, 71(4):1342–1346, 1974.

73. Mark Welch, Sridhar Govindarajan, Jon E Ness, Alan Villalobos, Austin Gurney, Jeremy Minshull, and Claes Gustafsson. Design parameters to control synthetic gene expression in Escherichia coli. *PloS One*, 4(9):e7002, jan 2009.

74. Fran Supek and Tomislav Šmuc. On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. *Genetics*, 185(3):1129–34, jul 2010.

75. Sriram Kosuri, Daniel B Goodman, Guillaume Cambray, Vivek K Mutalik, Yuan Gao, Adam P Arkin, Drew Endy, and George M Church. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences*, 110(34):14024–9, aug 2013.

76. Vivek K Mutalik, Joao C Guimaraes, Guillaume Cambray, Quynh-Anh Mai, Marc Juul Christoffersen, Lance Martin, Ayumi Yu, Colin Lam, Cesar Rodriguez, Gaymon Bennett, Jay D Keasling, Drew Endy, and Adam P Arkin. Quantitative estimation of activity and quality for collections of functional genetic elements. *Nature methods*, 10(4):347–53, apr 2013.

77. J Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, and Steffen Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–7, jun 2008.

78. Iris Bahir, Menachem Fromer, Yosef Prat, and Michal Linial. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Molecular Systems Biology*, 5(311):311, jan 2009.

79. Tamir Tuller, Yana Girshovich, Yael Sella, Avi Kreimer, Shiri Freilich, Martin Kupiec, Uri Gophna, and Eytan Ruppin. Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Research*, 39(11):4743–55, jun 2011.

80. Maya Botzman and Hanah Margalit. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology*, 12(10):R109, jan 2011.

81. Alexandra Dana and Tamir Tuller. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research*, 42(14):9171–9181, jul 2014.

82. Justin Gardin, Rukhsana Yeasmin, Alisa Yurovsky, Ying Cai, Steve Skiena, and Bruce Futcher. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3:1–20, jan 2014.

83. Jeffrey A. Hussmann, Stephanie Patchett, Arlen Johnson, Sara Sawyer, and William H. Press. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genetics*, 11(12):1–25, 2015.

84. David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B. Plotkin, and David P. Bartel. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports*, 14(7):1787–1799, 2016.

85. Paige S Spencer, Efraín Siller, John F Anderson, and José M Barral. Silent Substitutions Predictably Alter Translation Elongation Rates and Protein Folding Efficiencies. *Journal of Molecular Biology*, 422(3):328–335, 2012.

86. Eric D Kelsic, Hattie Chung, Niv Cohen, Jimin Park, Harris H Wang, and Roy Kishony. RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq Article RNA. *Cell Systems*, 3(6):563–571, 2016.

87. Gerrit Brandis and Diarmaid Hughes. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genetics*, 12(3):e1005926, 2016.

88. Daniel B Goodman, George M Church, and Sriram Kosuri. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(October):475–479, 2013.

89. Elad Firnberg, Jason W Labonte, Jeffrey J Gray, and Marc Ostermeier. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular biology and evolution*, 31(6):1581–1592, 2014.

90. Konrad U Foerstner, Christian von Mering, Sean D Hooper, and Peer Bork. Environments shape the nucleotide composition of genomes. *EMBO reports*, 6(12):1208–13, 2005.

91. Sara Vieira-Silva and Eduardo P C Rocha. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genetics*, 6(1), 2010.

92. Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nature Microbiology*, 1(5):16048, 2016.

93. Anita Krisko, Tea Copic, Toni Gabaldón, Ben Lehner, and Fran Supek. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biology*, 15(3):R44, 2014.

94. Swaine L Chen, William Lee, Alison K Hottes, Lucy Shapiro, and Harley H McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3480–5, mar 2004.

95. Peter a Lind and Dan I Andersson. Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(46):17878–83, nov 2008.

96. Ruth Hershberg and Dmitri A Petrov. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics*, 6(9), sep 2010.

97. Gene-Wei Li, Eugene Oh, and Jonathan S. Weissman. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484(7395):538–541, mar 2012.

98. M A Sørensen and S Pedersen. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *Journal of Molecular Biology*, 222(2):265–80, nov 1991.

99. Shuntaro Takahashi, Kentaro Tsuji, Takuya Ueda, and Yoshio Okahata. Traveling Time of a Translating Ribosome along Messenger RNA Monitored Directly on a Quartz Crystal Microbalance. *Journal of the American Chemical Society*, 134:6793–6800, 2012.

100. Catherine A. Charneski and Laurence D. Hurst. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biology*, 11(3):e1001508, mar 2013.

101. Manqing Li, Elaine Kao, Xia Gao, Hilary Sandig, Kirsten Limmer, Mariana Pavon-Eternod, Thomas E Jones, Sebastien Landry, Tao Pan, Matthew D Weitzman, and Michael David. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*, 11:4–9, sep 2012.

102. Johan Elf, Daniel Nilsson, Tanel Tenson, and Mans Ehrenberg. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, 300(5626):1718–22, jun 2003.

103. Arvind R Subramaniam, Tao Pan, and Philippe Cluzel. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6):2419–2424, dec 2012.

104. Milana Frenkel-Morgenstern, Tamar Danon, Thomas Christian, Takao Igarashi, Lydia Cohen, Ya-Ming Hou, and Lars Juhl Jensen. Genes adopt non-optimal

codon usage to generate cell cycle-dependent oscillations in protein levels. *Molecular Systems Biology*, 8(572):1–10, feb 2012.

105. Mian Zhou, Jinhu Guo, Joonseok Cha, Michael Chae, She Chen, Jose M Barral, Matthew S Sachs, and Yi Liu. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439):111–115, feb 2013.

106. Yao Xu, Peijun Ma, Premal Shah, Antonis Rokas, Yi Liu, and Carl Hirschie Johnson. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature*, 495(7439):116–120, feb 2013.

107. Hans Liljenstrom and Gunnar von Heijne. Translation Rate Modification By Preferential Codon Usage: Intragenic Position Effects. *Journal of Theoretical Biology*, 124:43–55, 1987.

108. Michael Bulmer. Codon Usage and Intragenic Position. *Journal of Theoretical Biology*, 133:67–71, 1988.

109. G F Chen and M Inouye. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Research*, 18(6):1465–73, mar 1990.

110. A Eyre-Walker and M Bulmer. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Research*, 21(19):4599–603, sep 1993.

111. A Eyre-Walker and M Bulmer. Synonymous substitution rates in enterobacteria. *Genetics*, 140(4):1407–12, aug 1995.

112. Hong Qin, Wei Biao Wu, Josep M Comeron, Martin Kreitman, and Wen-Hsiung Li. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168(4):2245–60, dec 2004.

113. Gina Cannarozzi, Nicol N Schraudolph, Mahamadou Faty, Peter von Rohr, Markus T Friberg, Alexander C Roth, Pedro Gonnet, Gaston Gonnet, and Yves Barral. A role for codon order in translation dynamics. *Cell*, 141(2):355–67, apr 2010.

114. Kajetan Bentele, Paul Saffert, Robert Rauscher, Zoya Ignatova, and Nils Blüthgen. Efficient translation initiation dictates codon usage at gene start. *Molecular systems biology*, 9(675):675, jan 2013.

115. H Ohno, H Sakai, T Washio, and M Tomita. Preferential usage of some minor codons in bacteria. *Gene*, 276(1-2):107–15, oct 2001.

116. Tamir Tuller, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, and Yitzhak Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54, apr 2010.

117. Jesper Vind, Michael A. Sorensen, Michael D. Rasmussen, and Steen Pedersen. Synthesis of Proteins in Escherichia coli is Limited by the Concentration of Free Ribosomes. *Journal of Molecular Biology*, 231:678–688, 1993.

118. Wanjun Gu, Tong Zhou, and Claus O Wilke. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology*, 6(2):e1000664, feb 2010.

119. Tong Zhou and Claus O Wilke. Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. *BMC Evolutionary Biology*, 11(1):59, jan 2011.

120. Thomas E Keller, S David Mis, Kevin E Jia, and Claus O Wilke. Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. *Genome Biology and Evolution*, 4(2):80–8, jan 2012.

121. Tamir Tuller, Isana Veksler-Lublinsky, Nir Gazit, Martin Kupiec, Eytan Ruppin, and Michal Ziv-Ukelson. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biology*, 12(11):R110, 2011.

122. Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transations on Automatic Control*, Ac-19(6):716–723, 1974.

123. Gal Lenz, Adi Doron-faigenboim, Eliora Z Ron, Tamir Tuller, and Uri Gophna. Sequence Features of E . coli mRNAs Affect Their Degradation. *PloS one*, 6(12):1–6, 2011.

124. Chungoo Park, Xiaoshu Chen, Jian-Rong Yang, and Jianzhi Zhang. 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):E678–86, feb 2013.

125. A. Carbone, A. Zinovyev, and F. Kepes. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16):2005–2015, oct 2003.

126. Douglas W Raiford, Dan E Krane, Travis E Doom, and Michael L Raymer. Automated isolation of translational efficiency bias that resists the confounding effect of GC(AT)-content. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2):238–50, 2010.

127. Katsuyuki Shiroguchi, Tony Z Jia, Peter a Sims, and X Sunney Xie. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–52, jan 2012.

128. Hila Gingold and Yitzhak Pilpel. Determinants of translation efficiency and accuracy. *Molecular Systems Biology*, 7(481):1–13, apr 2011.

129. Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, jan 2011.

130. K. P. Burnham and R.P. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

131. I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, jul 2003.

132. Joao C. Guimaraes, Miguel Rocha, and Adam P. Arkin. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic acids research*, 42(8):4791–9, apr 2014.

133. Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–35, apr 2014.

134. D Barrick, K Villanueba, J Childs, R Kalil, T D Schneider, C E Lawrence, L Gold, and G D Stormo. Quantitative analysis of ribosome binding sites in E.coli. *Nucleic Acids Research*, 22(7):1287–95, apr 1994.

135. Hongyun Chen, Matthew Bjerknes, Ravindra Kumar, and Ernest Jay. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Research*, 22(23):4953–4957, 1994.

136. Maarten H de Smit and Jan van Duin. Translation initiation on structured messengers: another role for the Shine-Dalgarno interaction. *Journal of Molecular Biology*, 235:173–184, 1994.

137. Maarten H. de Smit and Jan van Duin. Secondary structure of the ribosome binding site determines translational efficiency : A quantitative analysis. *Proceedings of the National Academy of Sciences*, 87(October):7668–7672, 1990.

138. Jutta Rinke-Appel, Nicole Junke, Richard Brimacombe, Inna Lavrik, Svetlana Dokudovskaya, Olga Dontsova, and Alexei Bogdanov. Contacts between 16S ribosomal RNA and mRNA , within the spacer region separating the AUG initiator codon cross-linking study. *Nucleic acids research*, 22(15):3018–3025, 1994.

139. Amin Espah Borujeni, Anirudh S Channarasappa, and Howard M Salis. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research*, 42(4):2646–59, feb 2014.

140. Adam J Hockenberry, M Irmak Sirer, Luís A Nunes Amaral, and Michael C Jewett. Quantifying position-dependent codon usage bias. *Molecular Biology and Evolution*, 31(7):1880–93, jul 2014.

141. Bill Chang, Saman Halgamuge, and Sen Lin Tang. Analysis of SD sequences in completed microbial genomes: Non-SD-led genes are as common as SD-led genes. *Gene*, 373(1-2):90–99, 2006.

142. Jiong Ma, Allan Campbell, and Samuel Karlin. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *Journal of Bacteriology*, 184(20):5733–5745, 2002.

143. Dokyun Na, Sunjae Lee, and Doheon Lee. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Systems Biology*, 4:71, jan 2010.

144. So Nakagawa, Yoshihito Niimura, Kin-Ichiro Miura, and Takashi Gojobori. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proceedings of the National Academy of Sciences*, 107(14):6382–6387, 2010.

145. H Sakai, C Imamura, Y Osada, R Saito, T Washio, and M Tomita. Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. *Journal of Molecular Evolution*, 52(2):164–170, 2001.

146. J Starmer, A Stomp, M Vouk, and D Bitzer. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Computational Biology*, 2(5):454–466, 2006.

147. Xiaobin Zheng, Gang-Qing Hu, Zhen-Su She, and Huaiqiu Zhu. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics*, 12(1):361, 2011.

148. Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23, apr 2009.

149. Jared M Schrader, Bo Zhou, Gene-Wei Li, Keren Lasker, W Seth Childers, Brandon Williams, Tao Long, Sean Crosson, Harley H McAdams, Jonathan S Weissman, and Lucy Shapiro. The coding and noncoding architecture of the Caulobacter crescentus genome. *PLoS Genetics*, 10(7):e1004463, jul 2014.

150. Gene-Wei Li. How do bacteria tune translation efficiency? *Current opinion in microbiology*, 24:66–71, 2015.

151. Nicholas F Lahens, Ibrahim Halil Kavakli, Ray Zhang, Katharina Hayer, Michael B Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S Thomas, Gregory R Grant, and John B Hogenesch. IVT-seq reveals extreme bias in RNA-sequencing. *Genome biology*, 15(6):R86, jun 2014.

152. T. P. Miettinen and M. Bjorklund. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids Research*, 43(2):1019–1034, dec 2014.

153. Fuad Mohammad, Christopher J. Woolstenhulme, Rachel Green, and Allen R. Buskirk. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Reports*, 14:686–694, 2016.

154. Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, Stylianos E Antonarakis, Jonas Behr, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12):1177–84, dec 2013.

155. Anze Zupanic, Catherine Meplan, Sushma N Grellscheid, John C Mathers, Tom B L Kirkwood, John E Hesketh, and Daryl P Shanley. Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA (New York, N.Y.)*, 20(10):1507–18, oct 2014.

156. Arvind R Subramaniam, Aaron DeLoughery, Niels Bradshaw, Yun Chen, Erin O'Shea, Richard Losick, and Yunrong Chai. A serine sensor for multicellularity in a bacterium. *eLife*, 2:1–17, 2013.

157. Anastassia V Komarova, Ludmila S Tchufistova, Elena V Supina, and Irina V Boni. Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *RNA*, 8(9):1137–1147, 2002.

158. Vladimir Vimberg, Age Tats, Maido Remm, and Tanel Tenson. Translation initiation region sequence preferences in Escherichia coli. *BMC Molecular Biology*, 8:100, 2007.

159. S Ringquist, T Jones, E E Snyder, T Gibson, I Boni, and L Gold. High-affinity RNA ligands to Escherichia coli ribosomes and ribosomal protein S1: comparison of natural and unnatural binding sites. *Biochemistry*, 34(11):3640–3648, 1995.

160. William H. Mather, Jeff Hasty, Lev S. Tsimring, and Ruth J. Williams. Translational Cross Talk in Gene Networks. *Biophysical Journal*, 104(11):2564–2572, 2013.

161. Wenlin An and Jason W Chin. Synthesis of orthogonal transcription-translation networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8477–82, may 2009.

162. Cédric Orelle, Erik D. Carlson, Teresa Szal, Tanja Florin, Michael C. Jewett, and Alexander S. Mankin. Protein synthesis by ribosomes with tethered subunits. *Nature*, 524:119–124, 2015.

163. Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, and Ivo L Hofacker. The Vienna RNA websuite. *Nucleic acids research*, 36(Web

Server issue):W70–4, jul 2008.

164. Xizeng Mao, Qin Ma, Chuan Zhou, Xin Chen, Hanyuan Zhang, Jincai Yang, Fenglou Mao, Wei Lai, and Ying Xu. DOOR 2.0: Presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research*, 42(D1):654–659, 2014.

165. K. Lim, Y. Furuta, and I. Kobayashi. Large variations in bacterial ribosomal RNA genes. *Molecular Biology and Evolution*, 29(10):2937–2948, 2012.

166. Y Osada, R Saito, and M Tomita. Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics*, 15(1996):578–581, 1999.

167. Damilola Omotajo, Travis Tate, Hyuk Cho, and Madhusudan Choudhary. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics*, 16(1):604, 2015.

168. Aishwarya Devaraj and Kurt Fredrick. Short spacing between the Shine-Dalgarno sequence and P codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed frameshifting. *Molecular Microbiology*, 78(6):1500–1509, 2010.

169. Adam J Hockenberry, Adam R Pah, Michael C Jewett, and Luís A. N. Amaral. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. *Open Biology*, 7(160239), 2017.

170. Amin Espah Borujeni and Howard M. Salis. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *Journal of the American Chemical Society*, 138(22):7016–7023, 2016.

171. Mads T Bonde, Margit Pedersen, Michael S Klausen, Sheila I Jensen, Tune Wulff, Scott Harrison, Alex T Nielsen, Markus J Herrgård, and Morten O A Sommer. Predictable tuning of protein expression in bacteria. *Nature Methods*, 13(3):2230–226, 2016.

172. Gaurav D Diwan and Deepa Agashe. The Frequency of Internal Shine-Dalgarno – Like Motifs in Prokaryotes. *Genome Biology and Evolution*, 8(6):1722–1733, 2016.

173. Chuyue Yang, Adam J Hockenberry, Michael C Jewett, and Luis A N Amaral. Depletion of Shine-Dalgarno Sequences within Bacterial Coding Regions Is Expression Dependent. *G3*, 6(November):3467–74, 2016.

174. Teresa Cortes, Olga T Schubert, Graham Rose, Kristine B Arnvig, Iñaki Comas, Ruedi Aebersold, and Douglas B Young. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Reports*, 5(4):1121–31, nov 2013.

175. Scarlet S. Shell, Jing Wang, Pascal Lapierre, Mushtaq Mir, Michael R. Chase, Margaret M. Pyle, Richa Gawande, Rushdy Ahmad, David A. Sarracino, Thomas R. Ioerger, Sarah M. Fortune, Keith M. Derbyshire, Joseph T. Wade, and Todd A. Gray. Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genetics*, 11(11):1–31, 2015.

176. Piet Kramer, Katrin Gäbel, Friedhelm Pfeiffer, and Jörg Soppa. Haloferax volcanii, a prokaryotic species that does not use the Shine Dalgarno mechanism for translation initiation at 5'-UTRs. *PloS one*, 9(4):e94979, jan 2014.

177. Hiroshi Yamamoto, Daniela Wittek, Romi Gupta, Bo Qin, Takuya Ueda, Roland Krause, Kaori Yamamoto, Renate Albrecht, Markus Pech, and Knud H Nierhaus. 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 113(9):E1180–9, 2016.

178. Timothy M Colussi, David a Costantino, Jianyu Zhu, John Paul Donohue, Andrei a Korostelev, Zane a Jaafar, Terra-dawn M Plank, Harry F Noller, and Jeffrey S Kieft. Initiation of translation in bacteria by a structured eukaryotic IRES RNA. *Nature*, 519(7541):110–113, 2015.

179. Lars B Scharff, Liam Childs, Dirk Walther, and Ralph Bock. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genetics*, 7(6):e1002155, jun 2011.

180. Mélodie Duval, Alexey Korepanov, Olivier Fuchsbauer, Pierre Fechter, Andrea Haller, Attilio Fabbretti, Laurence Choulier, Ronald Micura, Bruno P Klaholz, Pascale Romby, Mathias Springer, and Stefano Marzi. Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biology*, 11(12):e1001731, dec 2013.

181. Pamela A. Barendt, Najaf A. Shah, Gregory A Barendt, Parth A. Kothari, and Casim A. Sarkar. Evidence for context-dependent complementarity of non-shine-dalgarno ribosome binding sites to Escherichia coli rRNA. *ACS Chemical Biology*, 8(5):958–966, 2013.

182. Anastassia V. Komarova, Ludmila S. Tchufistova, Marc Dreyfus, and Irina V. Boni. AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in Escherichia coli. *Journal of Bacteriology*, 187(4):1344–1349, 2005.

183. Sarah Guiziou, Vincent Sauveplane, Hung-Ju Chang, Caroline Cler E, Nathalie Declerck, Matthieu Jules, and Jerome Bonnet. A part toolbox to tune genetic expression in Bacillus subtilis. *Nucleic Acids Research*, 44(10):7495–7508, 2016.

184. Andrew L. Markley, Matthew B. Begemann, Ryan E. Clarke, Gina C. Gordon, and Brian F. Pfleger. Synthetic Biology Toolbox for Controlling Gene Expression in the Cyanobacterium Synechococcus sp. strain PCC 7002. *ACS Synthetic Biology*, 4(5):595–603, 2015.

185. Christopher Tauer, Stefan Heinl, Esther Egger, Silvia Heiss, and Reingard Grabherr. Tuning constitutive recombinant gene expression in Lactobacillus plantarum. *Microbial Cell Factories*, 13(1):150, 2014.

186. Matthew T Weinstock, Eric D Hesek, Christopher M Wilson, and Daniel G Gibson. Vibrio natriegens as a fast-growing host for molecular biology. *Nature Methods*, 13(10):1–39, 2016.

187. Jeong Sang Yi, Min Woo Kim, Minsuk Kim, Yujin Jeong, Eun-Jung Kim, Byung-Kwan Cho, and Byung-Gee Kim. A Novel Approach for Gene Expression Optimization through Native Promoter and 5' UTR Combinations Based on RNA-seq, Ribo-seq, and TSS-seq of Streptomyces coelicolor. *ACS Synthetic Biology*, page acssynbio.6b00263, 2016.

188. Benjamin R. K. Roller, Steven F. Stoddard, and Thomas M. Schmidt. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nature Microbiology*, 1(September):16160, 2016.

189. L J Revell. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1(4):319–329, 2010.

190. Christopher T Brown, Matthew R Olm, Brian C Thomas, and Jillian F Banfield. Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology*, 34(12):057992, 2016.

191. Todd M Lowe and Patricia P Chan. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, 44(W1):W54–7, 2016.

192. Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans Henrik Stærfeldt, Torbjørn Rognes, and David W. Ussery. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108, 2007.

193. Robert C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

194. Alexandros Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

195. Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Research*, 44(21):10074–10090, 2016.

196. Andreas Wagner. Energy Constraints on the Evolution of Gene Expression. *Molecular Biology and Evolution*, 22(6):1365–1370, 2005.

197. Fran Supek, Nives Skunca, Jelena Repar, Kristian Vlahovicek, and Tomislav Smuc. Translational selection is ubiquitous in prokaryotes. *PLoS Genetics*, 6(6):e1001004, jun 2010.

198. Christopher J. Woolstenhulme, Nicholas R. Guydosh, Rachel Green, and Allen R. Buskirk. High-Precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Reports*, 11(1):13–21, 2015.

199. Xiaoqiu Liu, Huifeng Jiang, Zhenglong Gu, and Jeffrey W Roberts. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proceedings of the National Academy of Sciences*, 110(29):11928–33, 2013.

200. Deepa Agashe, N. Cecilia Martinez-Gomez, D. Allan Drummond, and Christopher J. Marx. Good codons, bad transcript: Large reductions in gene expression

and fitness arising from synonymous mutations in a key enzyme. *Molecular Biology and Evolution*, 30(3):549–560, 2013.

201. Jin Chen, Alexey Petrov, Magnus Johansson, Albert Tsai, Seán E O'Leary, and Joseph D Puglisi. Dynamic pathways of -1 translational frameshifting. *Nature*, 512(7514):328–32, 2014.

202. Kevin A. Vasquez, Taylor A. Hatridge, Nicholas C. Curtis, and Lydia M. Contreras. Slowing translation between protein domains by increasing affinity between mRNAs and the ribosomal anti-Shine-Dalgarno sequence improves solubility. *ACS Synthetic Biology*, 5:133–145, 2015.

203. Patrick B F O'Connor, Gene-Wei Li, Jonathan S Weissman, John F Atkins, and Pavel V Baranov. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. *Bioinformatics*, 29(12):1488–91, 2013.

204. Anneli Borg and Måns Ehrenberg. Determinants of the rate of mRNA translocation in bacterial protein synthesis. *Journal of Molecular Biology*, 427(9):1835–1847, 2015.

205. Yuhei Chadani, Tatsuya Niwa, Shinobu Chiba, Hideki Taguchi, and Koreaki Ito. Integrated in vivo and in vitro nascent chain profiling reveals widespread translational pausing. *Proceedings of the National Academy of Sciences*, 113(7):E829–E838, 2016.

206. John A Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution*, 19(8):1390–1394, 2002.

207. Mingcong Wang, Christina J. Herrmann, Milan Simonovic, Damian Szklarczyk, and Christian von Mering. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–3168, 2015.

208. Olivier Borkowski, Anne Goelzer, Marc Schaffer, Magali Calabre, U. Ma der, Stéphane Aymerich, Matthieu Jules, and Vincent Fromion. Translation elicits a growth rate-dependent, genome-wide, differential protein production in Bacillus subtilis. *Molecular Systems Biology*, 12(870), 2016.

209. Kenji Nakahigashi, Yuki Takai, Michiko Kimura, Nozomi Abe, Toru Nakayashiki, Yuh Shiwa, Hirofumi Yoshikawa, Barry L. Wanner, Yasushi Ishihama, and Hirotada Mori. Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. *DNA Research*, 23(March):193–201, 2016.

210. Dominique Chu, David J. Barnes, and Tobias Von Der Haar. The role of tRNA and ribosome competition in coupling the expression of different mRNAs in Saccharomyces cerevisiae. *Nucleic acids research*, 39(15):6705–14, aug 2011.

211. Weston R Whitaker, Hanson Lee, Adam P Arkin, and John E Dueber. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synthetic Biology*, 4:249–257, 2014.

212. Amanda J Ozin, Teresa Costa, Adriano O Henriques, and Charles P Moran Jr. Alternative Translation Initiation Produces a Short Form of a Spore Coat Protein in Bacillus subtilis Alternative Translation Initiation Produces a Short Form of a Spore Coat Protein in Bacillus subtilis. *Journal of Bacteriology*, 183(6):2032–2040, 2001.

213. M J Lajoie, S Kosuri, J A Mosberg, C J Gregg, D Zhang, and G M Church. Probing the Limits of Genetic Recoding in Essential Genes. *Science*, 342(October):361–364, 2013.

214. Nili Ostrov, Matthieu Landon, Marc Guell, Gleb Kuznetsov, Jun Teramoto, Natalie Cervantes, Minerva Zhou, Kerry Singh, Michael G Napolitano, Mark Moosburner, Ellen Shrock, Benjamin W Pruitt, Nicholas Conway, Daniel B Goodman, Cameron L Gardner, Gary Tyree, Alexandra Gonzales, Barry L Wanner, Julie E Norville, Marc J Lajoie, and George M Church. Design, synthesis, and testing toward a 57-codon genome. *Science*, 353(6301):819–822, 2016.

215. Michael G Napolitano, Matthieu Landon, Christopher J Gregg, Marc J Lajoie, Lakshmi Govindarajan, Joshua A Mosberg, Gleb Kuznetsov, Daniel B Goodman, Oscar Vargas-Rodriguez, Farren J Isaacs, Dieter Soll, and George M Church. Emergent rules for codon choice elucidated by editing rare arginine codons in Escherichia coli. *Proc Natl Acad Sci USA*, pages E5588–E5597, 2016.

APPENDIX A

Supporting information to Chapter 2

codon	ChiSq Value	ChiSq Pvalue	ChiSq DOF	median	median Zscore	auc	auc Zscore	dvalue	dvalue Pvalue	Sig. in 3 of 4	Sig in 4 of 4
AAA	127.9641	0.0311	100	172	0.1087	0.905	2.7689	0.0031	0		
AAC	178.1378	0	100	195	5.9422	0.894	-6.2962	0.0135	0	x	x
AAG	127.9641	0.0311	100	166	0.4676	0.906	-2.7689	0.0103	0		
AAT	178.1378	0	100	166	-6.8943	0.906	6.2962	0.0166	0	x	x
ACA	434.0803	0	100	149	-5.7725	0.909	2.785	0.0455	0	x	
ACC	264.8102	0	100	177	-0.1293	0.903	2.8729	0.0111	0		
ACG	195.9198	0	100	178	4.0861	0.9	-6.3574	0.0144	0	x	x
ACT	177.5995	0	100	163	-1.8434	0.906	1.4265	0.0119	0.0201		
AGA	444.905	0	100	137	-2.2043	0.917	-1.6772	0.0832	0		
AGC	175.1767	0	100	177	4.6877	0.9	-6.1675	0.0127	0	x	x
AGG	161.8724	0.0001	100	147	1.1216	0.912	-1.4514	0.0441	0		
AGT	114.0159	0.1599	100	166	0.3874	0.905	-0.2131	0.0151	0		
ATA	426.8494	0	100	136	-6.2105	0.919	3.649	0.0559	0	x	
ATC	199.7121	0	100	177	2.7448	0.904	-1.8019	0.0092	0		
ATT	123.2836	0.0571	100	167	0.3095	0.91	0.1793	0.0021	0.7866		
CAA	358.2179	0	100	164	-5.7557	0.907	5.2793	0.0205	0	x	x
CAC	145.3595	0.0021	100	179	1.7609	0.902	0.6488	0.0038	0.764		
CAG	358.2179	0	100	181	5.9895	0.899	-5.2793	0.0109	0	x	x
CAT	145.3595	0.0021	100	168	-0.7991	0.908	-0.6488	0.0029	0.764		
CCA	119.4929	0.0894	100	173	1.6362	0.905	-1.5159	0.0088	0.1718		
CCC	409.6223	0	100	137	-14.895	0.918	14.9674	0.0731	0	x	x
CCG	580.0727	0	100	193	12.646	0.895	-15.073	0.0294	0	x	x
CCT	233.154	0	100	151	-7.7896	0.913	8.8962	0.0411	0	x	x
CGA	151.2388	0.0007	100	155	-2.6476	0.913	3.1366	0.032	0		
CGC	205.5766	0	100	178	-3.0107	0.903	5.3818	0.0107	0	x	
CGG	176.6578	0	100	174	2.1245	0.902	-3.0312	0.0138	0.0259		
CGT	122.1438	0.0656	100	185	2.5295	0.898	-4.096	0.0073	0.0216		
CTA	146.646	0.0017	100	159	-2.5401	0.911	2.0695	0.0249	0		
CTC	158.0108	0.0002	100	160	-7.2841	0.91	10.1871	0.0278	0	x	
CTG	962.8651	0	100	183	11.623	0.901	-15.234	0.022	0	x	x
CTT	367.5968	0	100	145	-11.375	0.916	10.4749	0.0439	0	x	x
GAA	155.174	0.0003	100	179	-0.7899	0.902	6.0903	0.0047	0		
GAC	96.5478	0.5791	100	180	-0.8384	0.899	1.818	0.0039	0.4281		
GAG	155.174	0.0003	100	182	2.4628	0.898	-6.0903	0.0104	0		
GAT	96.5478	0.5791	100	177	0.926	0.901	-1.818	0.0023	0.4281		
GCA	208.6315	0	100	170	1.9355	0.905	-3.228	0.0102	0		
GCC	197.6553	0	100	169	-6.5429	0.907	9.1629	0.0177	0	x	x
GCG	309.3065	0	100	185	7.2372	0.9	-7.5829	0.0185	0	x	x
GCT	172.2946	0	100	163	-3.7639	0.907	1.8237	0.0186	0		
GGA	214.8093	0	100	163	-0.338	0.909	0.2316	0.0229	0		
GGC	91.9616	0.7043	100	181	-1.8576	0.901	3.0845	0.005	0.0277		
GGG	119.2033	0.0924	100	175	-0.7468	0.903	0.1922	0.0085	0.0936		
GGT	98.1357	0.534	100	182	2.581	0.899	-3.6456	0.007	0.0216		
GTA	222.3969	0	100	172	0.9514	0.905	-0.4236	0.0159	0		
GTC	180.7344	0	100	163	-8.1232	0.907	7.6787	0.0234	0	x	x
GTG	501.6292	0	100	187	9.7697	0.899	-12.641	0.025	0	x	x
GTT	181.3435	0	100	164	-5.5019	0.907	6.8606	0.0159	0	x	x
TAC	103.5041	0.3852	100	190	3.1483	0.897	-2.4395	0.0095	0		
TAT	103.5041	0.3852	100	174	-3.337	0.903	2.4395	0.0072	0		
TCA	186.5467	0	100	157	-1.6692	0.91	1.0257	0.0186	0		
TCC	124.7843	0.0473	100	167	-4.9891	0.906	5.6959	0.0177	0	x	
TCG	156.7292	0.0003	100	182	1.9039	0.9	-2.5485	0.0175	0		
TCT	119.7204	0.0872	100	165	-1.8393	0.908	4.1201	0.0112	0.0465		
TGC	107.6422	0.283	100	161	1.8409	0.912	-1.0002	0.0093	0.0211		
TGT	107.6422	0.283	100	151	-1.5305	0.915	1.0002	0.0117	0.0211		
TTA	647.4117	0	100	149	-7.439	0.916	5.6637	0.0373	0	x	x
TTC	318.2339	0	100	188	9.7178	0.899	-12.989	0.0281	0	x	x
TTG	108.4854	0.2641	100	170	2.0954	0.907	-4.3668	0.01	0.0424		
TTT	318.2339	0	100	155	-9.6702	0.914	12.9885	0.0209	0	x	x

Figure A.1. **Significance of statistical tests for uniformity in the** *E. coli* **genome.** Shown are the test-statistic, where applicable, and the significance values for 4 separate statistical tests of non-uniformity for the 59 redundant codons. Values colored in red are significant and (x) in the final column marks whether three out of four, or four out of four (respectively), tests are significant for a given codon. Column headings: auc refers to the area under the curve that is observed after constructing the CDF function, and aucZscore refers to the z-score of this auc value compared to 200 randomized genomes. The dvalue column refers to the maximum distance (y-axis displacement) between the observed cumulative distribution and the average of 200 cumulative distributions and the dvaluePvalue column shows how this result compares to the individual 200 randomizations (since D is always positive, the p-value is the number of empirically observed values $>$ D divided by the number of randomizations).

	Log-likelihood (parameter #)			
	Uniform(1)	Linear (2)	Step-wise (3)	Exponential (3)
GAC	-2794.47	-2789.48	-2789.68	-2789.32
AGC	-2775.43	-2752.32	-2745.27	-2740.28
ATC	-2842.97	-2779.41	-2788.63	-2762.78
CAC	-2171.51	-2156.10	-2158.19	-2154.73
CCC	-2160.52	-2062.76	-2106.42	-1937.82
ATT	-2747.34	-2727.15	-2730.35	-2722.97
GTC	-2730.50	-2713.43	-2728.87	-2692.17
ACC	-2930.96	-2928.90	-2841.52	-2928.42
CGA	-1801.75	-1770.94	-1770.44	-1749.32
CCA	-2279.92	-2277.18	-2255.97	-2250.77
GTA	-2563.16	-2561.96	-2518.62	-2501.20
TTA	-2932.73	-2817.41	-2720.85	-2649.27
GGA	-2448.38	-2415.51	-2417.87	-2354.15
AGT	-2447.40	-2446.15	-2432.08	-2426.60
GAG	-2888.66	-2877.53	-2855.17	-2877.58
TCC	-2338.55	-2335.93	-2335.59	-2335.89
AAT	-2735.21	-2631.52	-2674.72	-2572.07
ACG	-2766.73	-2749.01	-2736.70	-2736.04
TAC	-2342.80	-2324.19	-2329.56	-2317.34
CGT	-2840.95	-2799.26	-2823.09	-2790.93
TAT	-2342.80	-2324.19	-2326.04	-2317.34
TCT	-2309.37	-2298.37	-2300.23	-2292.24
CAA	-2748.81	-2703.93	-2623.16	-2611.38
TGT	-1628.75	-1625.76	-1623.83	-1612.71
GCC	-3106.17	-3083.55	-3093.74	-3083.22
CGG	-2180.64	-2180.58	-2173.83	-2180.28
ATA	-2031.83	-2001.75	-1922.25	-1839.26
AGA	-1672.13	-1647.89	-1511.63	-1470.99
TCG	-2465.60	-2448.03	-2438.27	-2410.75
GAA	-2888.66	-2877.53	-2856.03	-2877.27
GCG	-3408.48	-3323.48	-3303.03	-3174.56
GTT	-2837.32	-2810.69	-2800.89	-2772.86
AAC	-2735.21	-2631.52	-2621.25	-2572.07
CCT	-2273.83	-2203.16	-2241.84	-2132.78
AAA	-2411.35	-2411.10	-2408.03	-2411.09
CAG	-2748.81	-2703.93	-2613.32	-2611.38
GCT	-2847.28	-2829.07	-2804.45	-2758.80
AAG	-2411.35	-2411.10	-2408.83	-2404.27
ACT	-2431.21	-2423.97	-2406.10	-2400.95
AGG	-1232.83	-1226.68	-1202.38	-1191.80
TTC	-2719.94	-2515.59	-2594.28	-2448.72
CTG	-4027.73	-3726.43	-3598.45	-3418.53
ACA	-2470.09	-2453.33	-2296.07	-2256.96
GGT	-2985.37	-2973.14	-2981.98	-2971.71
CTA	-1913.40	-1907.63	-1890.16	-1881.54
GGC	-3051.30	-3048.90	-3033.90	-3027.68
CAT	-2171.51	-2156.10	-2158.69	-2154.73
TTG	-2662.91	-2662.26	-2662.09	-2662.25
TTT	-2719.94	-2515.59	-2539.36	-2448.72
GAT	-2794.47	-2789.48	-2786.39	-2789.31
TGC	-1628.75	-1625.76	-1622.39	-1612.71
GTG	-3267.58	-3180.87	-3101.29	-2989.82
TCA	-2234.97	-2216.67	-2195.23	-2188.20
GCA	-3021.36	-3019.15	-2968.09	-2939.11
CTT	-2656.59	-2572.62	-2588.59	-2461.84
CGC	-2830.84	-2830.48	-2794.77	-2830.45
CCG	-3060.24	-2834.63	-2887.45	-2612.59
CTC	-2517.82	-2501.71	-2511.31	-2492.55
GGG	-2541.45	-2539.34	-2539.55	-2536.80

Figure A.2. **Log likelihood values for individual codons in the *E. coli* genome.** Larger numbers, i.e. less negative, indicate the model with a better fit to the data. For nearly all codons, the exponential model has the highest log likelihood. This is to be partially expected due to increased parameter number in the exponential decay model compared to the uniform model, which we take into account during our model selection phase. The most likely model is highlighted in red text.

	AIC (parameter #)			
	Uniform(1)	Linear (2)	Step-wise (3)	Exponential (3)
GAC	5590.93	5582.96	5585.36	5584.64
AGC	5552.87	5508.65	5496.54	5486.56
ATC	5687.94	5562.82	5583.26	5531.57
CAC	4345.02	4316.19	4322.37	4315.46
CCC	4323.05	4129.52	4218.85	3881.64
ATT	5496.68	5458.31	5466.70	5451.95
GTC	5462.99	5430.86	5463.74	5390.34
ACC	5863.93	5861.80	5689.04	5862.84
CGA	3605.50	3545.88	3546.88	3504.64
CCA	4561.83	4558.37	4517.94	4507.54
GTA	5128.32	5127.92	5043.24	5008.39
TTA	5867.47	5638.83	5447.69	5304.54
GGA	4898.75	4835.02	4841.73	4714.29
AGT	4896.80	4896.30	4870.16	4859.19
GAG	5779.32	5759.06	5716.33	5761.17
TCC	4679.11	4675.86	4677.18	4677.78
AAT	5472.43	5267.05	5355.44	5150.14
ACG	5535.46	5502.02	5479.39	5478.08
TAC	4687.59	4652.39	4665.12	4640.67
CGT	5683.90	5602.53	5652.17	5587.86
TAT	4687.59	4652.39	4658.07	4640.67
TCT	4620.75	4600.75	4606.45	4590.48
CAA	5499.61	5411.86	5252.32	5228.77
TGT	3259.49	3255.53	3253.65	3231.43
GCC	6214.34	6171.09	6193.49	6172.44
CGG	4363.27	4365.15	4353.66	4366.57
ATA	4065.65	4007.50	3850.51	3684.52
AGA	3346.26	3299.78	3029.25	2947.99
TCG	4933.19	4900.06	4882.53	4827.49
GAA	5779.32	5759.06	5718.06	5760.53
GCG	6818.97	6650.96	6612.05	6355.12
GTT	5676.65	5625.38	5607.79	5551.72
AAC	5472.43	5267.05	5248.50	5150.14
CCT	4549.67	4410.31	4489.68	4271.56
AAA	4824.71	4826.20	4822.07	4828.17
CAG	5499.61	5411.86	5232.64	5228.77
GCT	5696.56	5662.14	5614.90	5523.61
AAG	4824.71	4826.20	4823.66	4814.54
ACT	4864.43	4851.94	4818.20	4807.89
AGG	2467.65	2457.36	2410.77	2389.61
TTC	5441.89	5035.18	5194.57	4903.44
CTG	8057.46	7456.86	7202.90	6843.06
ACA	4942.18	4910.65	4598.14	4519.91
GGT	5972.73	5950.28	5969.97	5949.43
CTA	3828.79	3819.27	3786.33	3769.08
GGC	6104.61	6101.80	6073.81	6061.37
CAT	4345.02	4316.19	4323.37	4315.46
TTG	5327.82	5328.51	5330.18	5330.50
TTT	5441.89	5035.18	5084.72	4903.44
GAT	5590.93	5582.96	5578.77	5584.63
TGC	3259.49	3255.53	3250.79	3231.43
GTG	6537.15	6365.74	6208.59	5985.63
TCA	4471.95	4437.35	4396.47	4382.40
GCA	6044.73	6042.29	5942.18	5884.22
CTT	5315.18	5149.24	5183.17	4929.67
CGC	5663.68	5664.96	5595.55	5666.91
CCG	6122.48	5673.25	5780.90	5231.17
CTC	5037.64	5007.42	5028.61	4991.09
GGG	5084.91	5082.68	5085.11	5079.59

Figure A.3. **AIC values for each model for individual codons in the *E. coli* genome.** Small numbers indicate that the given model is a better fit to the empirical data. Red text indicates the best fitting model for a given codon.

species	AIC_mean	AIC_linear	AIC_step-wise	AIC_exp
B. subtilis	374355.552	373589.971	371743.169	369328.286
C. jejuni	179164.559	178987.147	178820.388	178771.106
C. crescentus	255774.937	250097.602	244808.780	234878.952
C. botulinum	225176.261	224575.347	224609.334	224138.508
E. coli	312674.758	308250.155	305774.703	302254.963
H. influenzae	224756.294	223877.934	223763.599	223123.465
H. pylori	220469.890	220297.955	219487.046	219202.906
M. janaschii	184551.548	184100.167	183786.520	183479.849
L. johnsonii	260477.432	257690.455	257966.250	257158.966
L. lacti	248164.842	247762.575	246376.179	246005.036
L. interrogans	309803.699	308606.406	308625.767	307625.167
L. monocytogenes	259221.680	257328.985	256603.597	254971.737
P. aeruginosa	357453.098	355801.308	337398.014	330336.789
S. cerevisiae	653864.750	648938.769	651779.163	646124.387
S. enterica	314332.338	310524.853	307182.460	301771.057
S. aureus	261564.265	260817.230	260636.098	259244.467
S. pneumoniae	266061.017	263857.885	262093.112	260548.064
V. cholerae	310292.930	308448.254	306891.006	303137.144
Y. pestis	399517.998	397444.534	396135.214	393541.304

Figure A.4. **AIC values for each organism tested.** Small numbers indicate that the given model is a better fit to the empirical data. Red text indicates the best fitting model for a given genome.

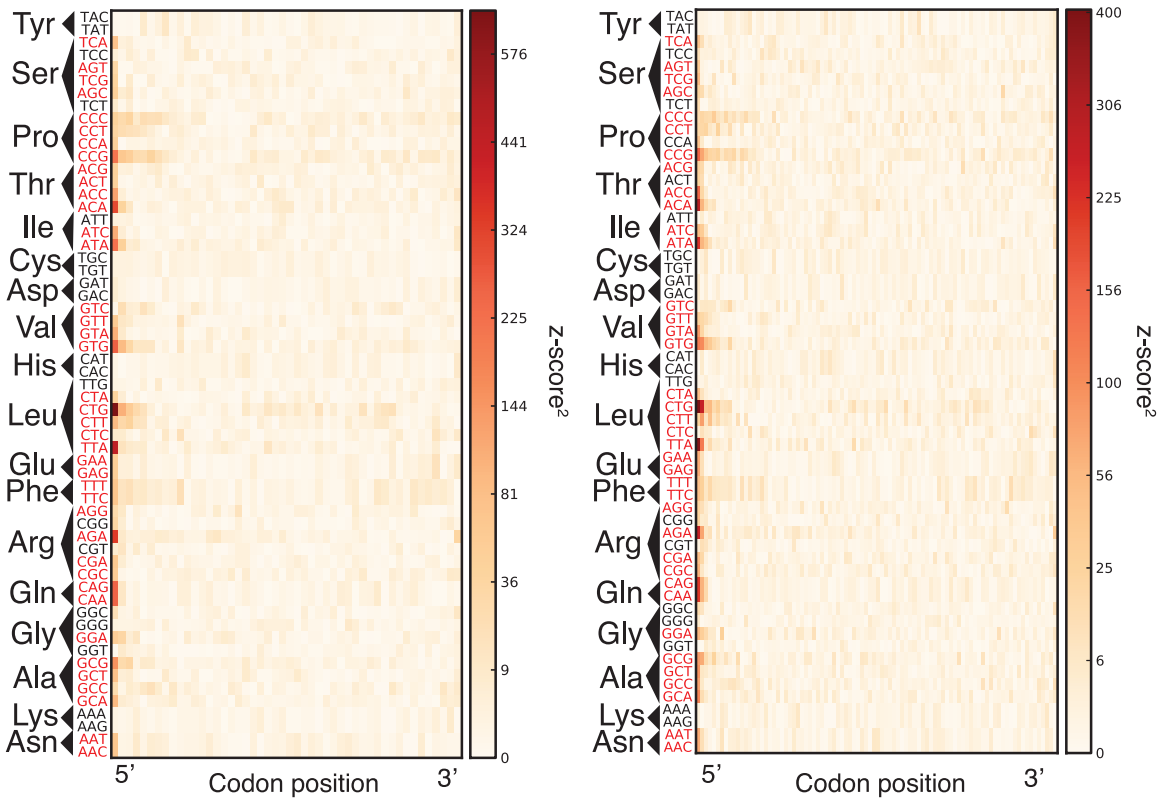

Figure A.5. χ^2 **test for two different bin schemes.** As opposed to Fig. 2.1 of the main manuscript, here we depict significance of codons when dividing the genome into 50 (left) and 100 bins (right) to demonstrate robustness to bin size.

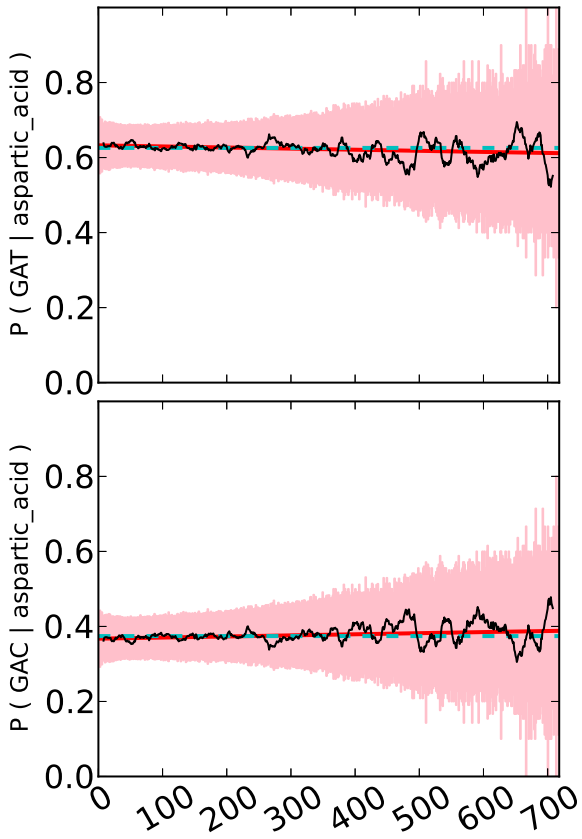

Figure A.6. **Aspartic acid codon usage in the** *E. coli* **genome.**

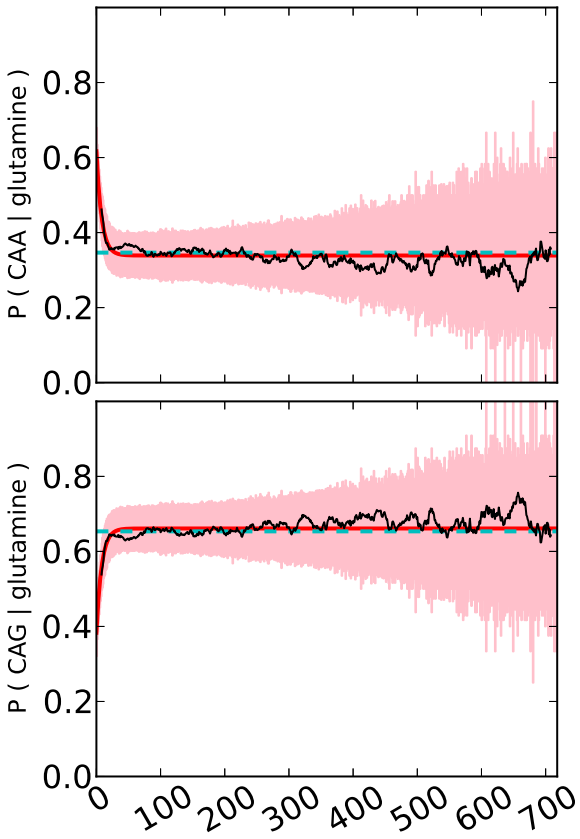

Figure A.7. **Glutamine codon usage in the** *E. coli* **genome.**

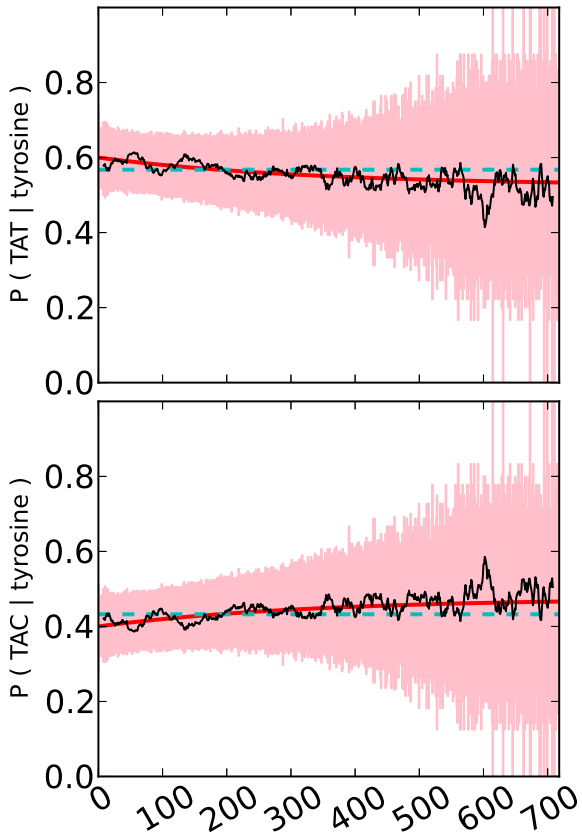

Figure A.8. **Tyrosine codon usage in the** *E. coli* **genome.**

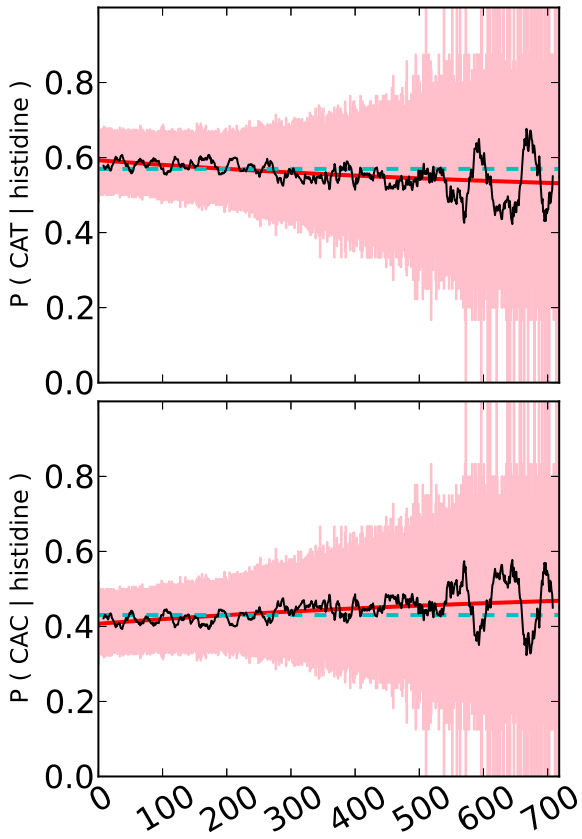

Figure A.9. **Histidine codon usage in the** *E. coli* **genome.**

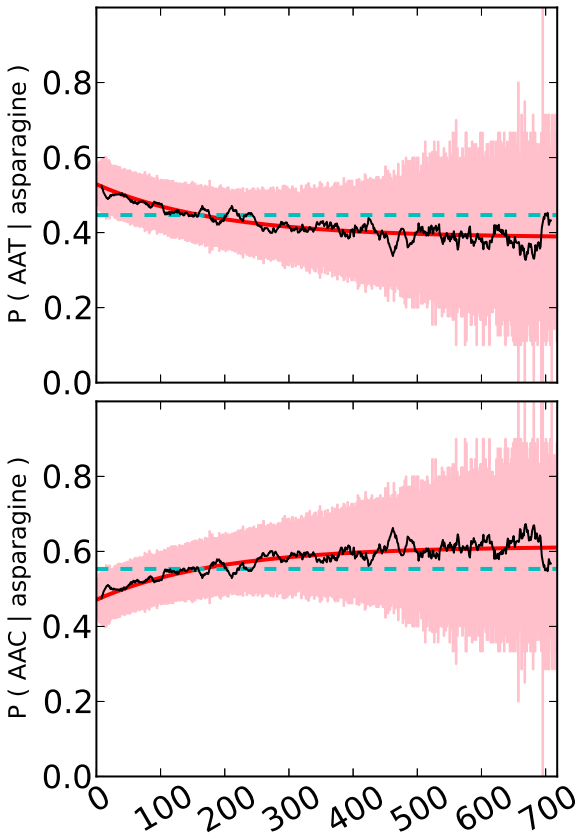

Figure A.10. **Asparagine codon usage in the** *E. coli* **genome.**

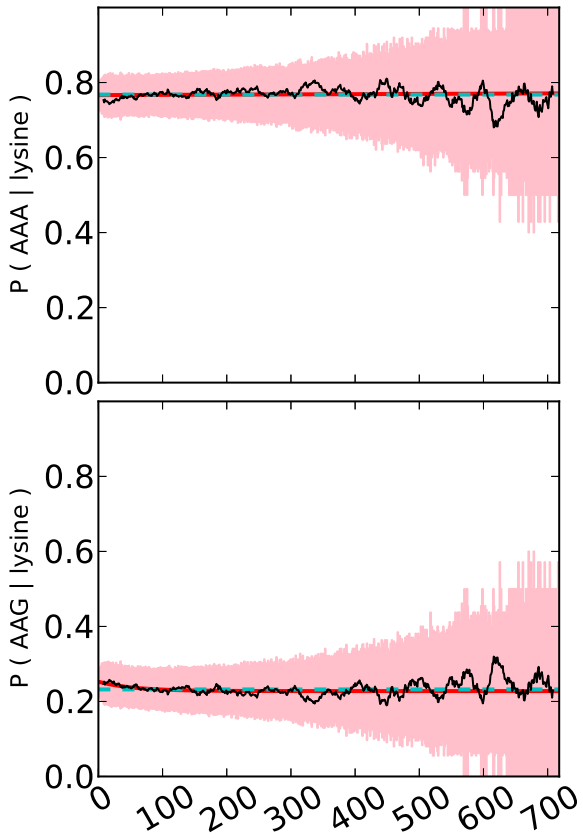

Figure A.11. **Lysine codon usage in the** *E. coli* **genome.**

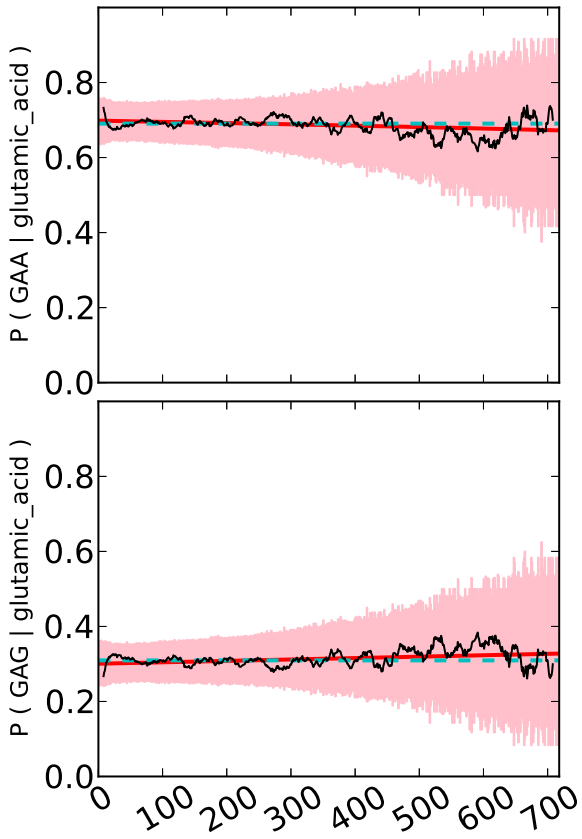

Figure A.12. **Glutamic acid codon usage in the** *E. coli* **genome.**

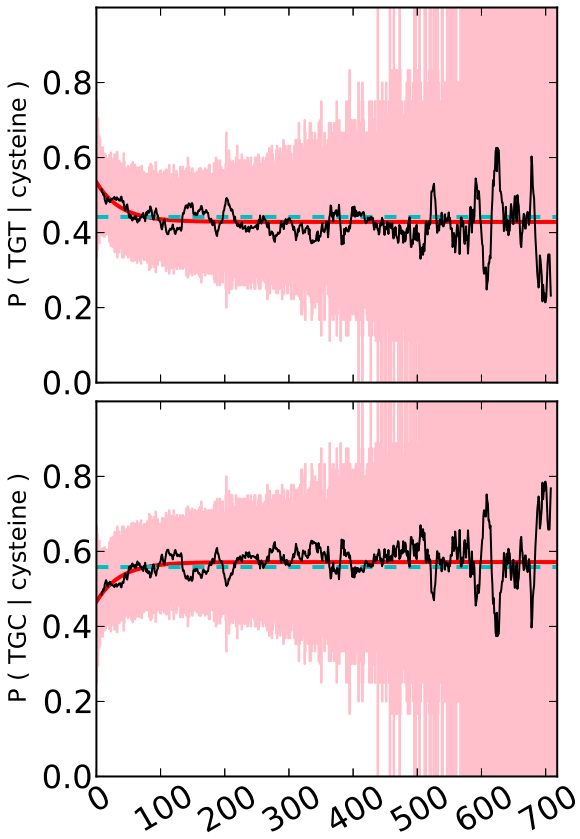

Figure A.13. **Cysteine codon usage in the** *E. coli* **genome.**

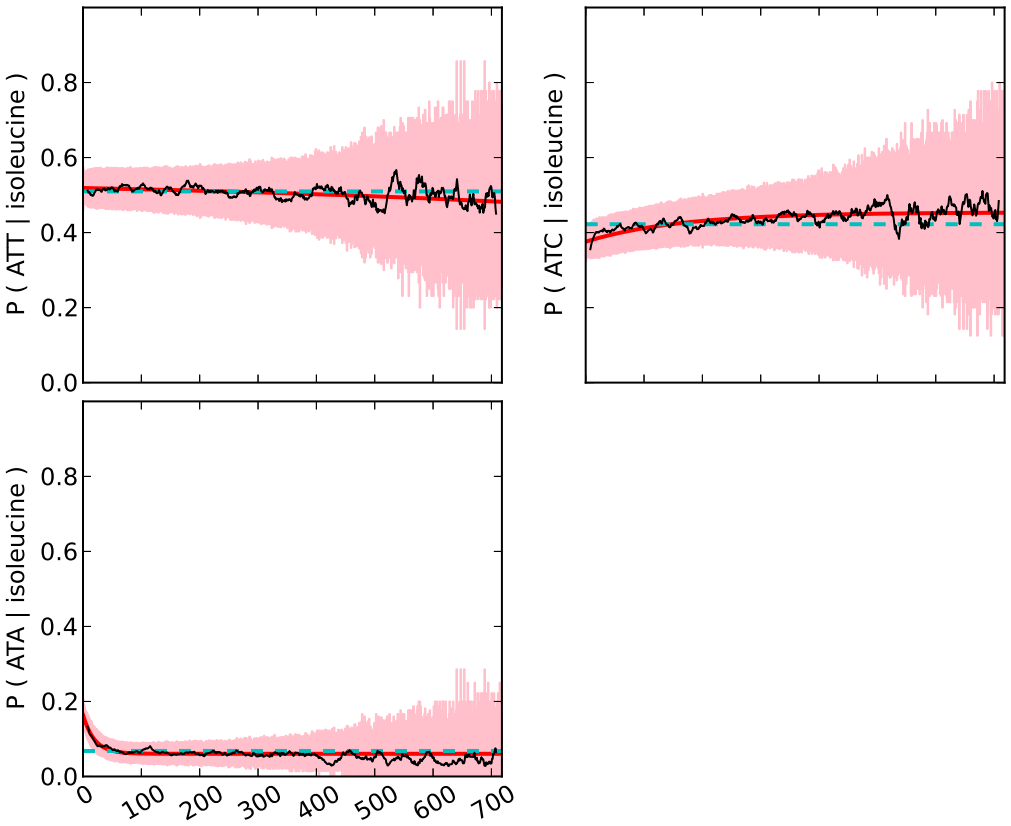

Figure A.14. **Isoleucine codon usage in the** *E. coli* **genome.**

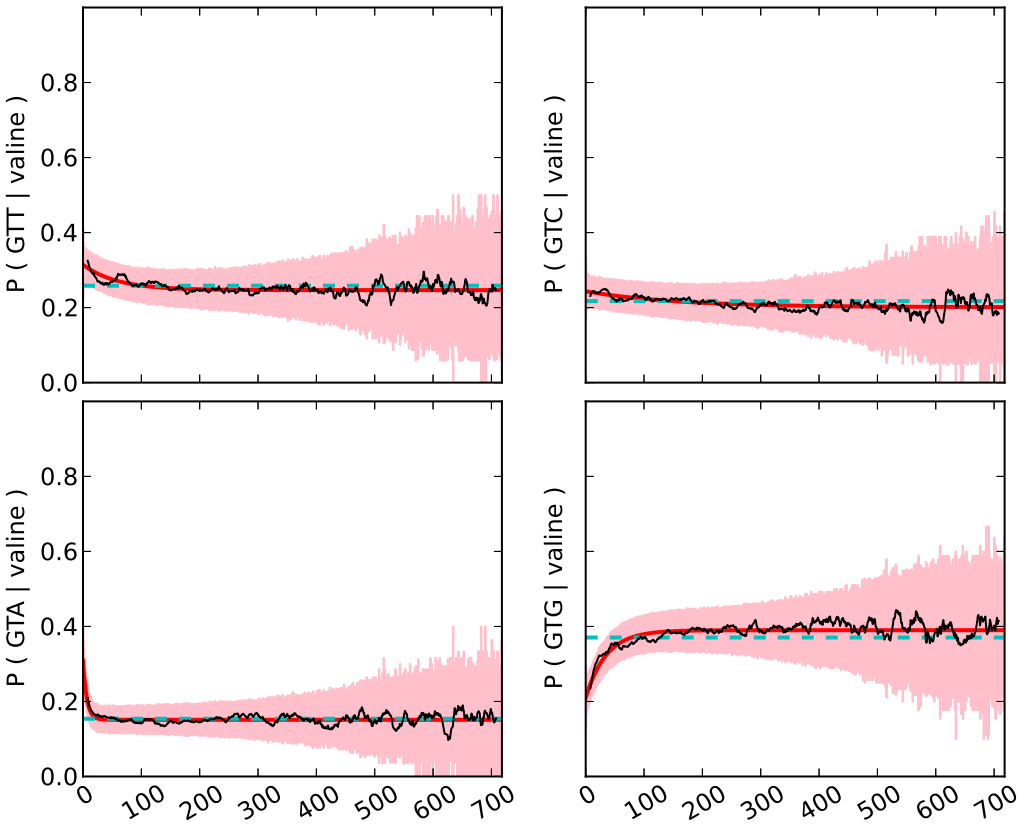

Figure A.15. **Valine codon usage in the** *E. coli* **genome.**

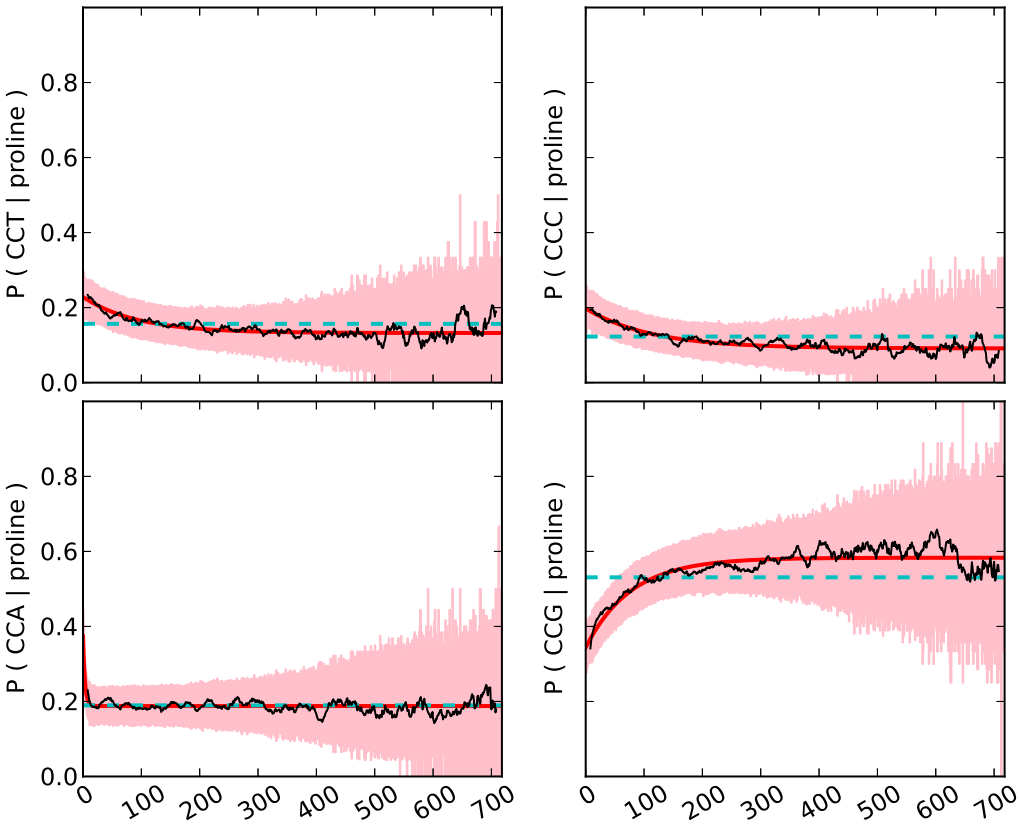

Figure A.16. **Proline codon usage in the** *E. coli* **genome.**

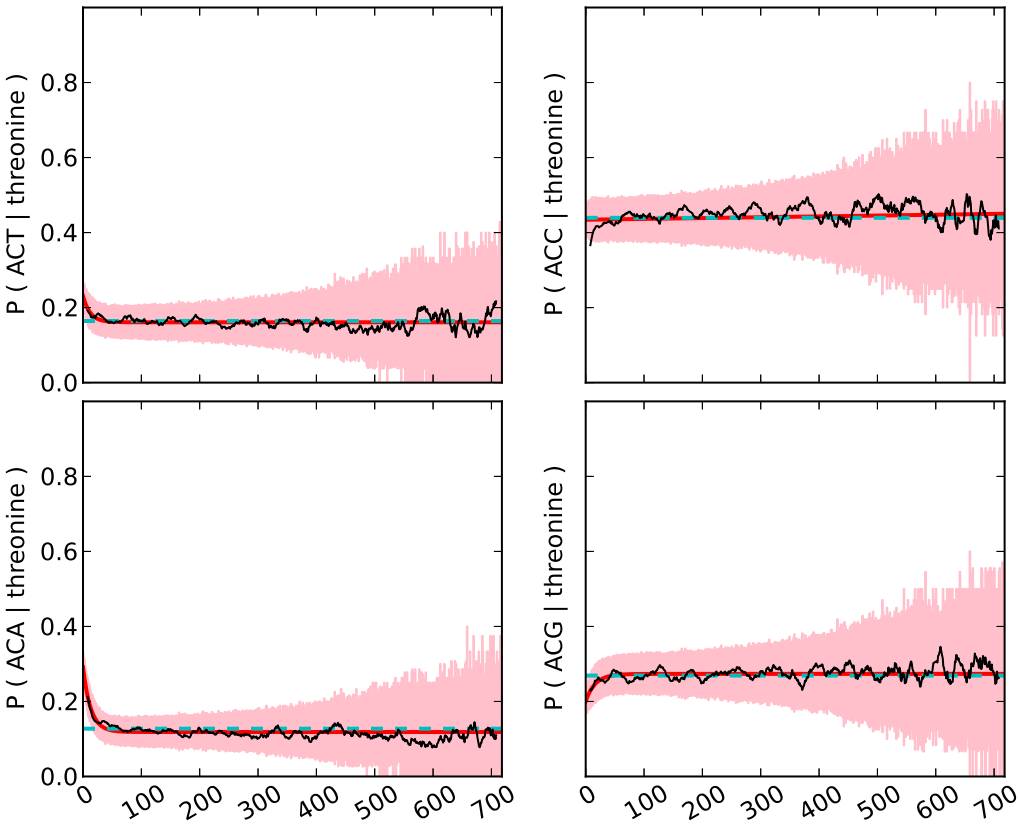

Figure A.17. **Threonine codon usage in the** *E. coli* **genome.**

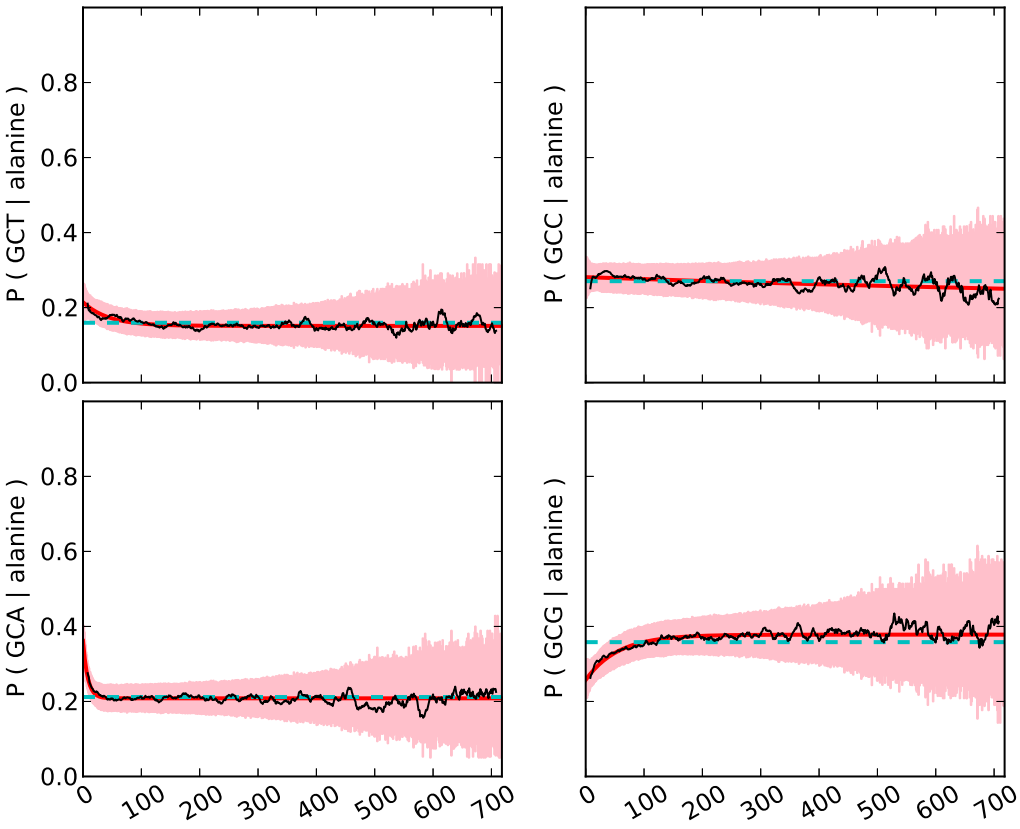

Figure A.18. **Alanine codon usage in the** *E. coli* **genome.**

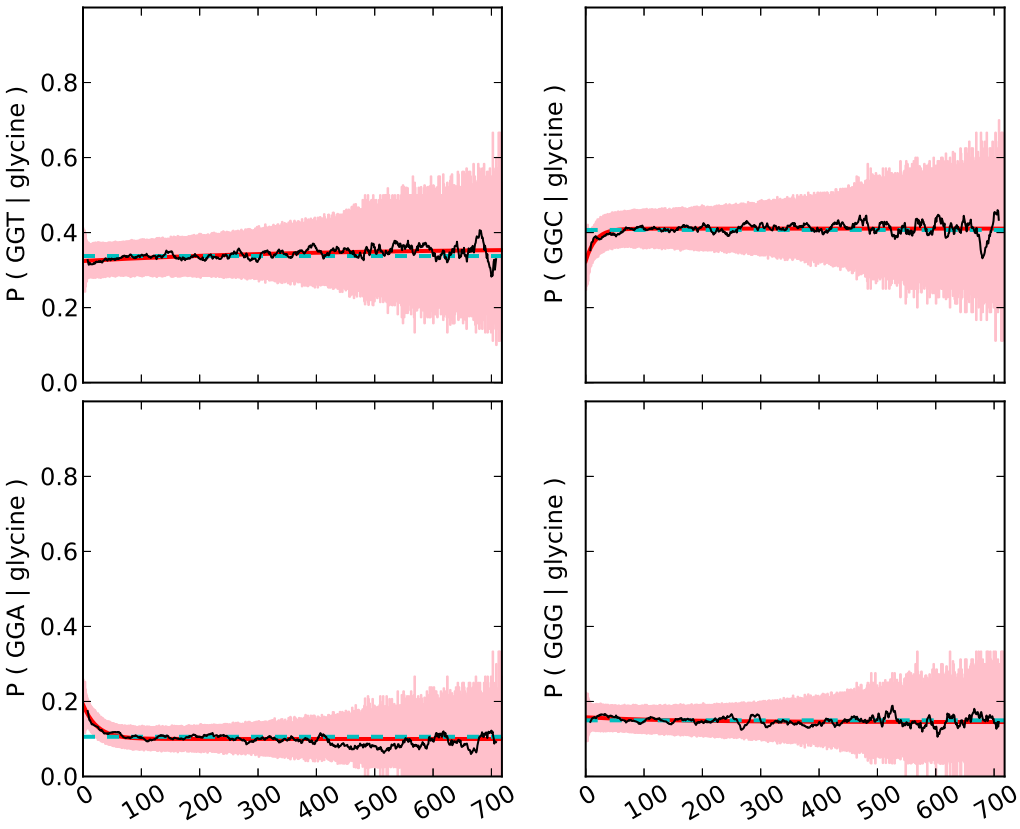

Figure A.19. **Glycine codon usage in the** *E. coli* **genome.**

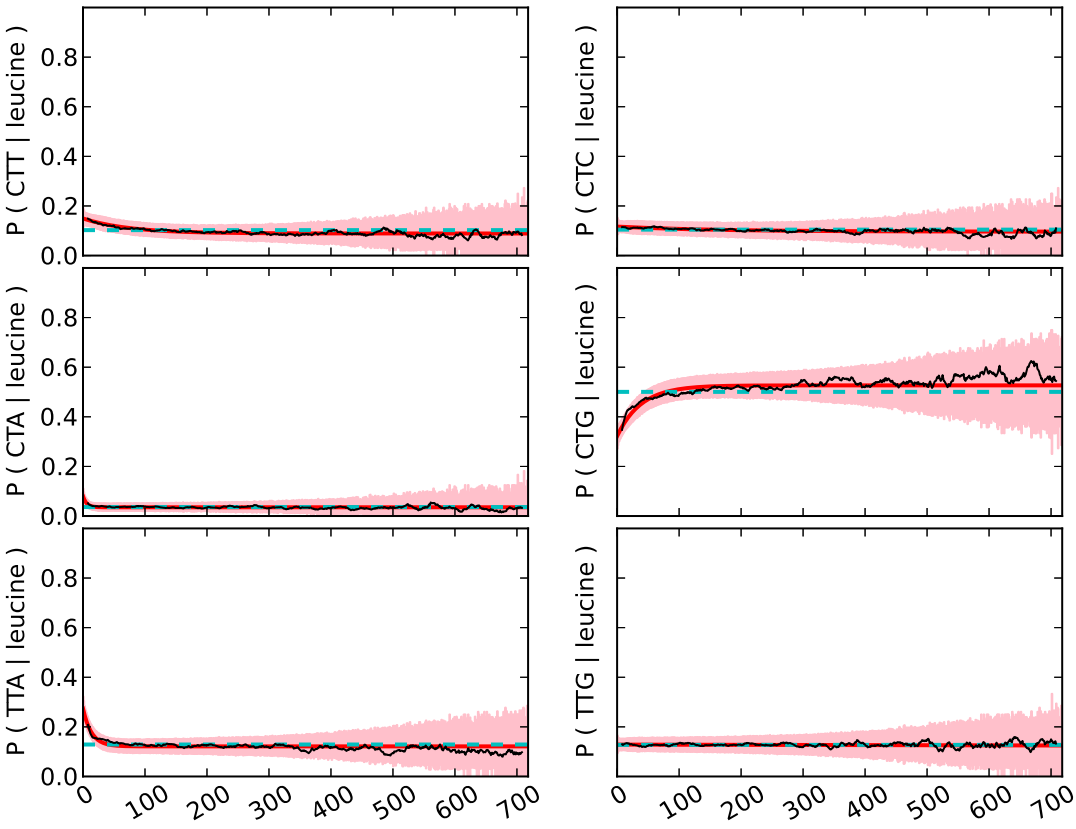

Figure A.20. **Leucine codon usage in the** *E. coli* **genome.**

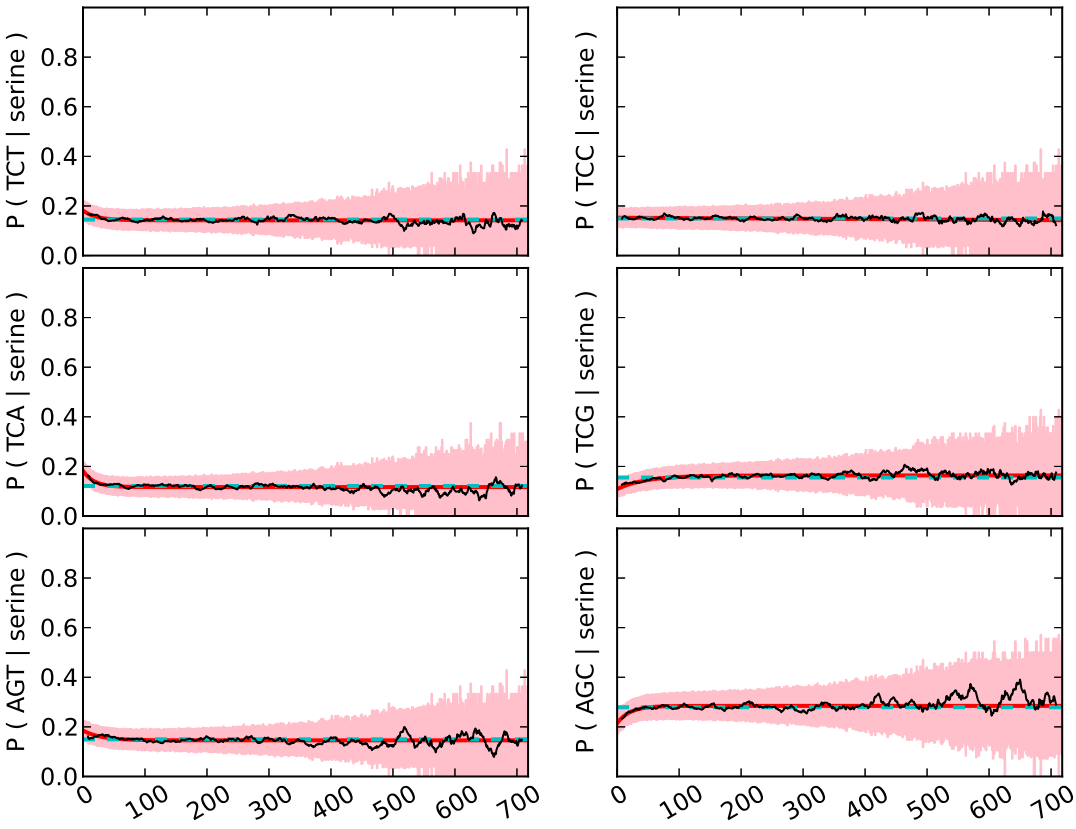

Figure A.21. **Serine codon usage in the** *E. coli* **genome.**

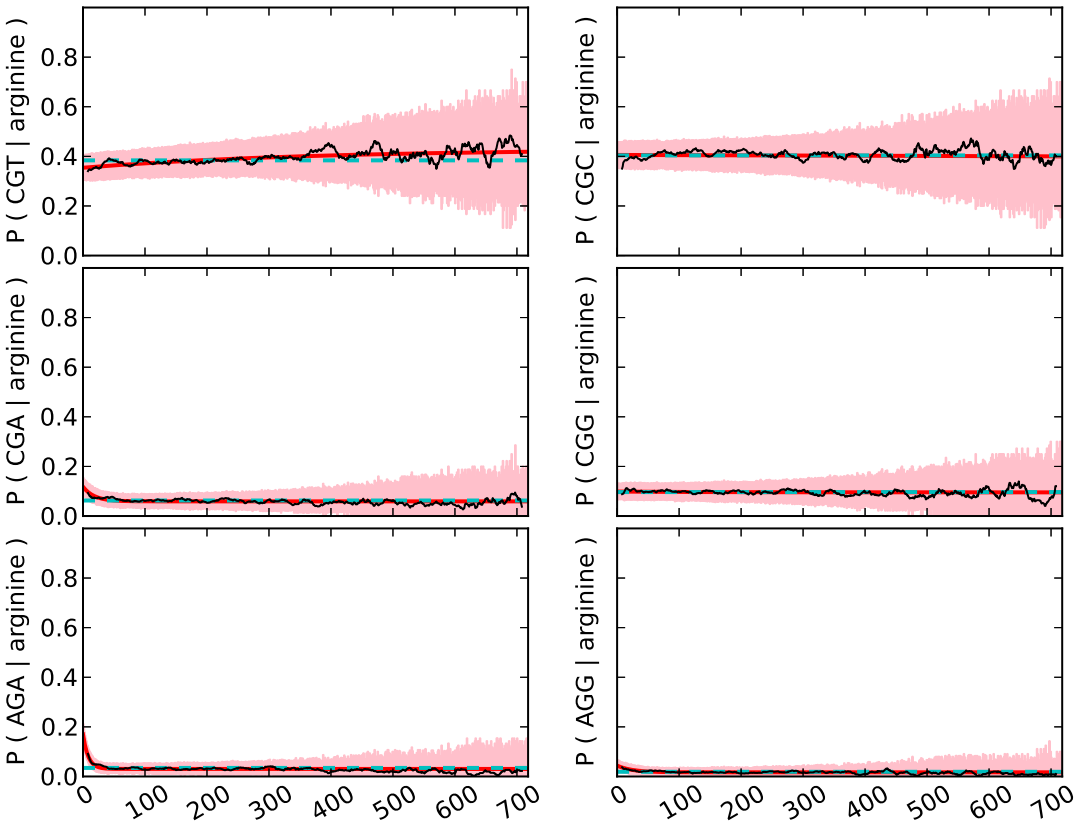

Figure A.22. **Arginine codon usage in the** *E. coli* **genome.**

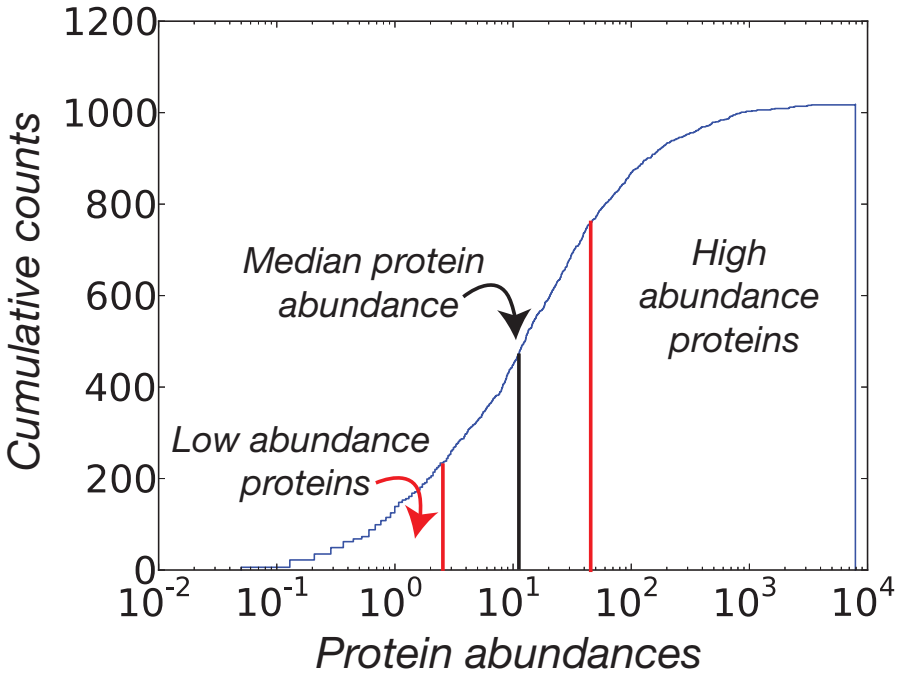

Figure A.23. **Summary of gene expression bins.** For the protein expression dataset used in Fig. 2.3 of the main text, we show the cumulative distribution of expression highlighting the quartiles used to classify low and high abundance proteins, as well as the median that was used to perform the same analysis in Supplementary Fig. A.25.

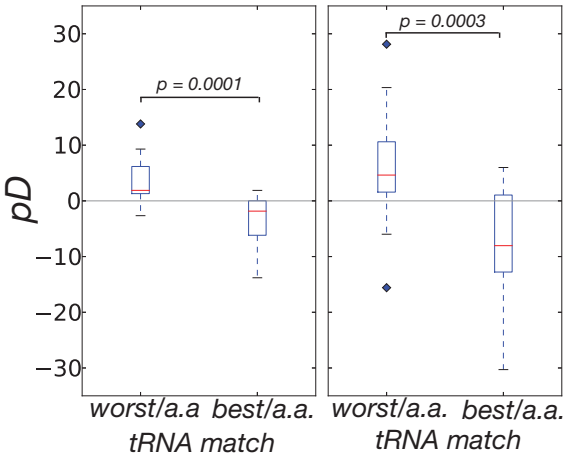

Figure A.24. **Classification of codons based off of tRNA Adaptation Index.** Instead of separating codons based on frequency of occurrence in a highly expressed reference set, we separate codons based on the codon value in the tRNA Adaptation Index(tAI). For each of the 18 redundantly coded amino acids, we select the best and worst codon and separate these into 'low' and 'high' tAI categories (n =18 and 18, respectively) and compare the pD values between the two sets.

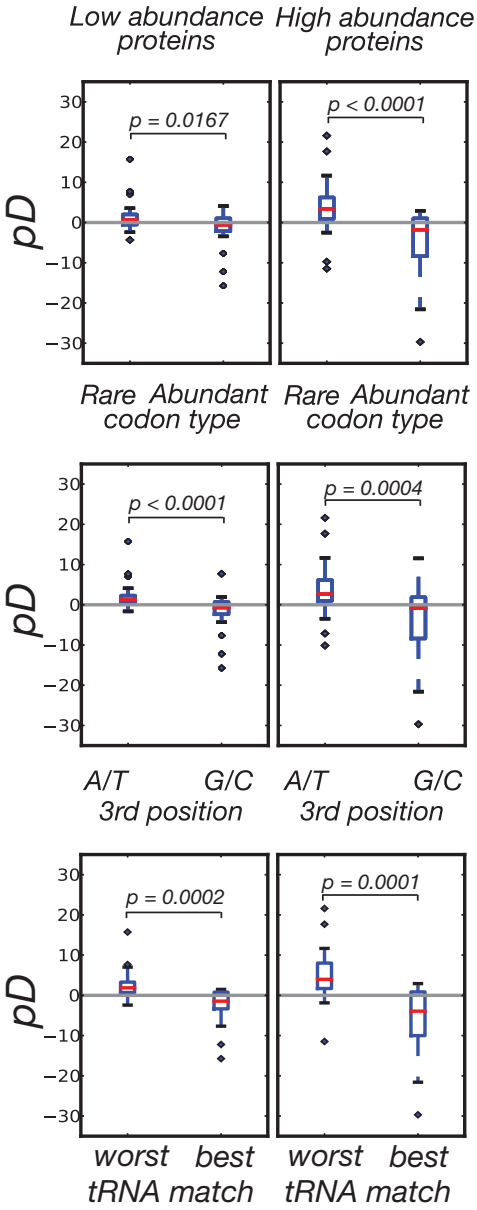

Figure A.25. **As in Fig. 2.3 of main text, using median gene expression to delineate low and high abundance proteins.** The results of Fig. 2.3 are significant when defining lowly and highly abundant proteins using the median of the expression set rather than the top and bottom quartiles.

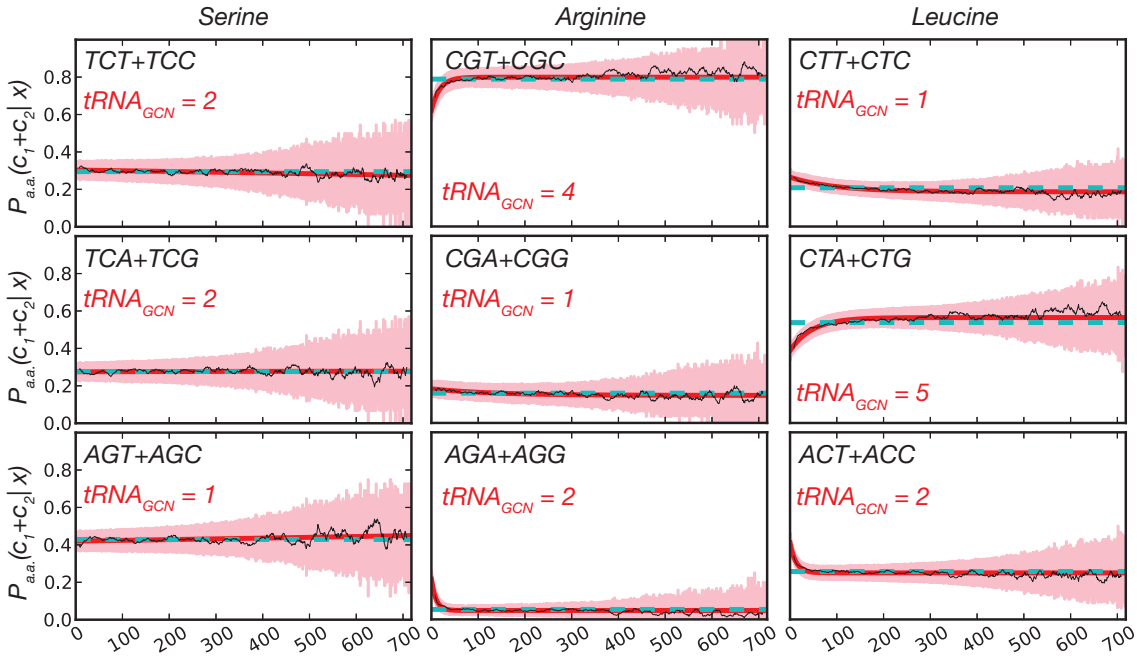

Figure A.26. **As in Fig. 2.5 of main text.** Shown are codon groupings for the 6-fold redundant amino acids.

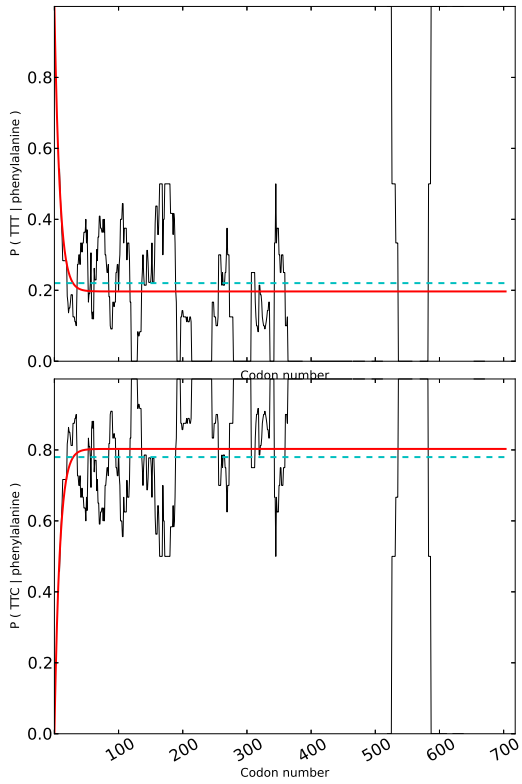

Figure A.27. **Spatial usage of Phenylalanine in the CAI reference set.** In contrast to Fig. 2.2B in the main text, the codons coding for phenylalanine show an extreme skewness that highlights the fact that high expressing genes (as evidenced by the reference set) also use the 'disfavored' codons in the $5'$ region.

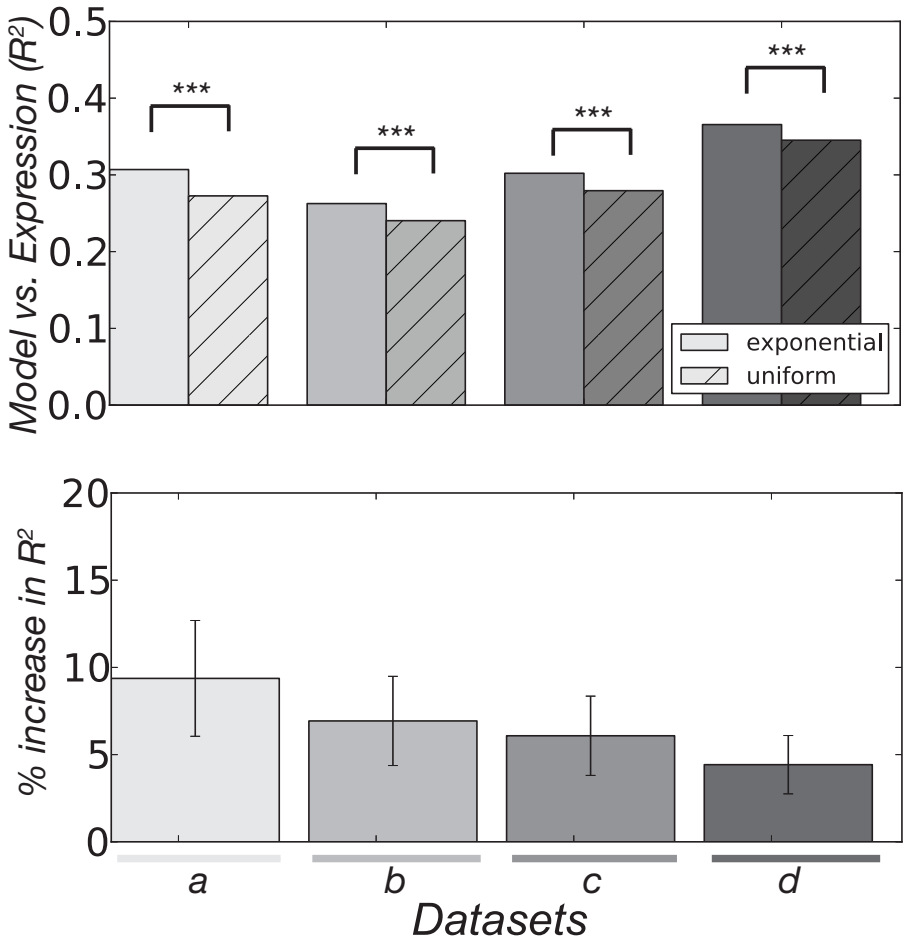

Figure A.28. **CAI calculations with different reference set.** In the manuscript, we show the traditionally used 27 gene reference set. To demonstrate robustness, here we show the same figures and calculations using a distinct reference set of highly expressed genes. In general, this reference set performs more poorly at predicting transcript/protein abundances and percent increases between our exponential fits and the traditional uniform calculation are slightly lower, albeit still positive and highly significant.

APPENDIX B

Supporting information to Chapter 3

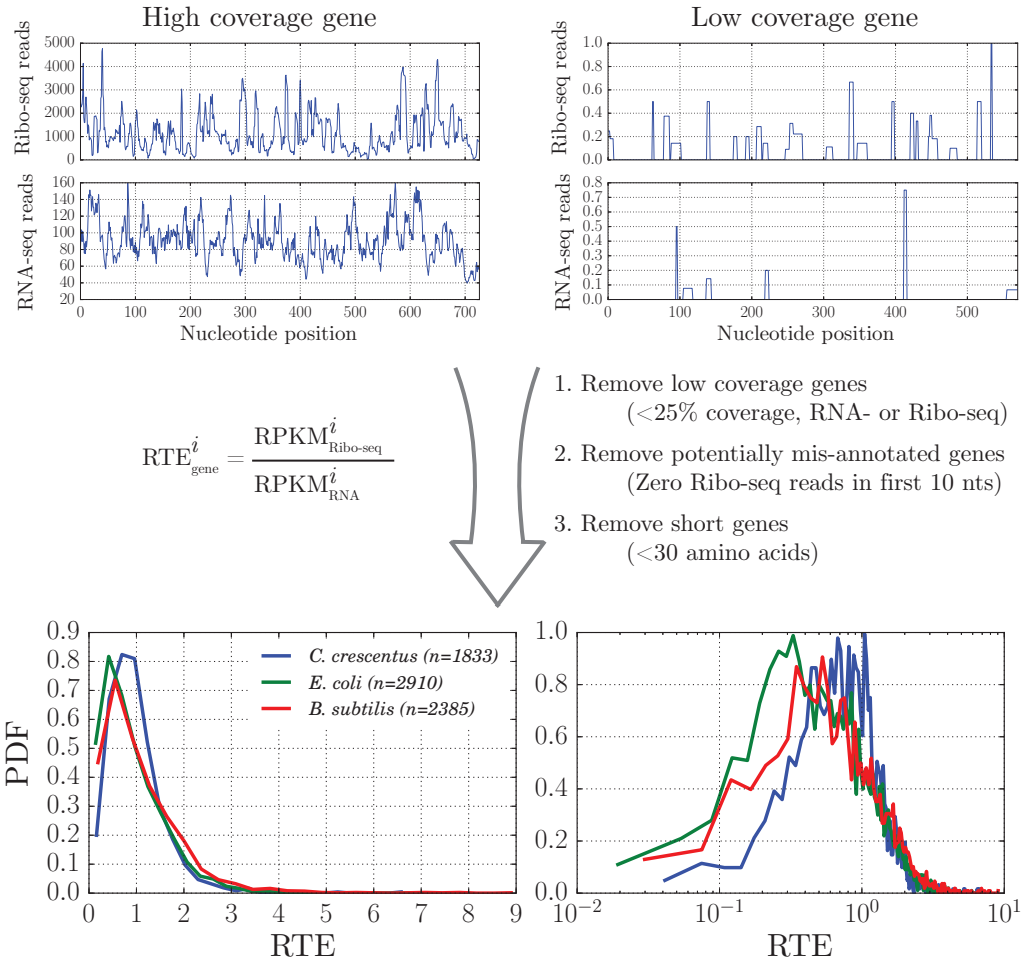

Figure B.1. **Example gene profiles showing mapped RNA- and Ribo-seq reads that are used as input to calculate** RTE**.** Our pipeline first removes a subset of the total genes based off of coverage, annotation, and length requirements resulting in RTE measurements for 2910, 1833, and 2385 genes in *E. coli*, *C. crescentus* and *B. subtilis*. Distributions of the RTE values on normal and log-scale show that RTE is approximately log-normally distributed and comparable between the three datasets studied.

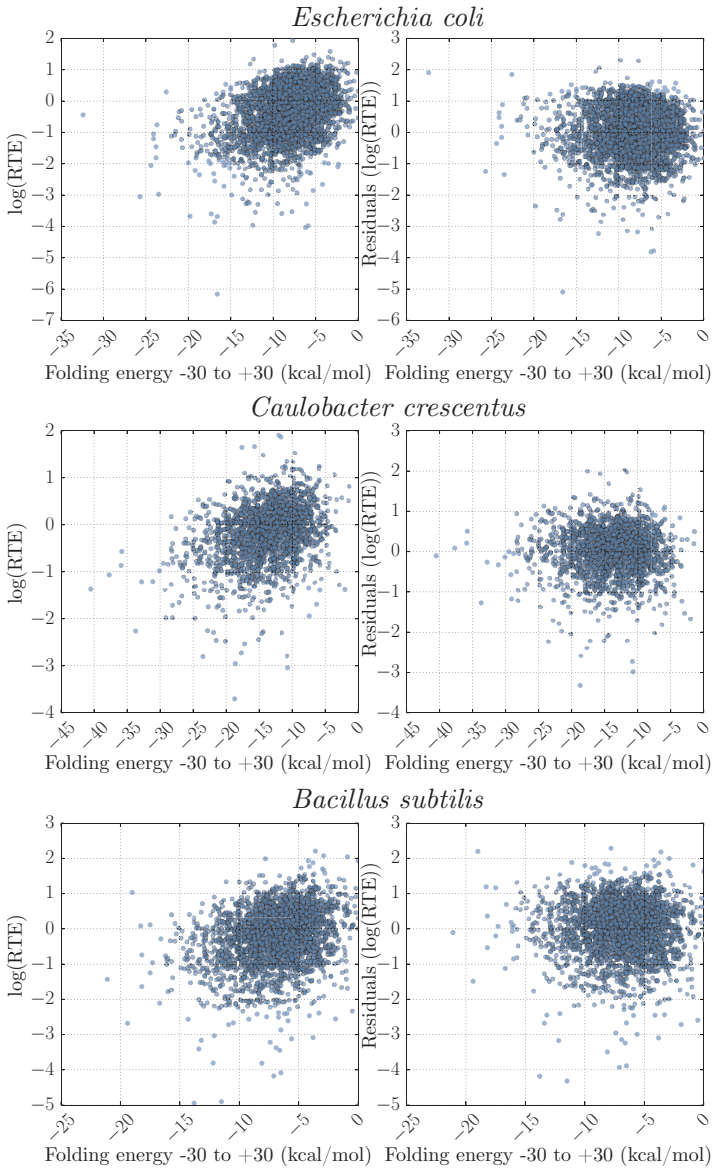

Figure B.2. **Correlation between the free energy of RNA folding around the start codon (-30 to +30) and $log(RTE)$ for three different organisms studied.** Left, $R^2 = 0.13, 0.10, 0.08$ for *E. coli*, *C. crescentus*, and *B. subtilis* respectively; for all cases $p < 10^{-43}$). For RTE in the main text we utilize the residuals from the best fitting linear model based off this regression for each organism, effectively removing the influence of mRNA structure on RTE (right).

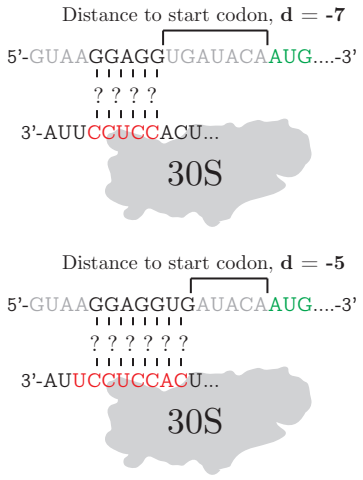

Figure B.3. **Extended illustration of our numbering scheme for distance.** This example highlights that distance is not absolute and, rather, is calculated relative to the aSD sequence being considered. For the same mRNA sequence (top and bottom), the distance decreases by two due to the fact that the example in the bottom extends the hypothetical aSD sequence 2 bases in the 5′ direction. Numbering is always relative to the 5′ end of the aSD sequence so varying this sequence. Other numbering systems to compare between different putative aSDs will all suffer from this problem unless an absolute point is used as an anchor (such as the middle U of 5′-CCUCC-3′). We opted for our scheme because we feel that, at the conclusion of the process, our scheme is simpler to interpret for a given aSD sequence.

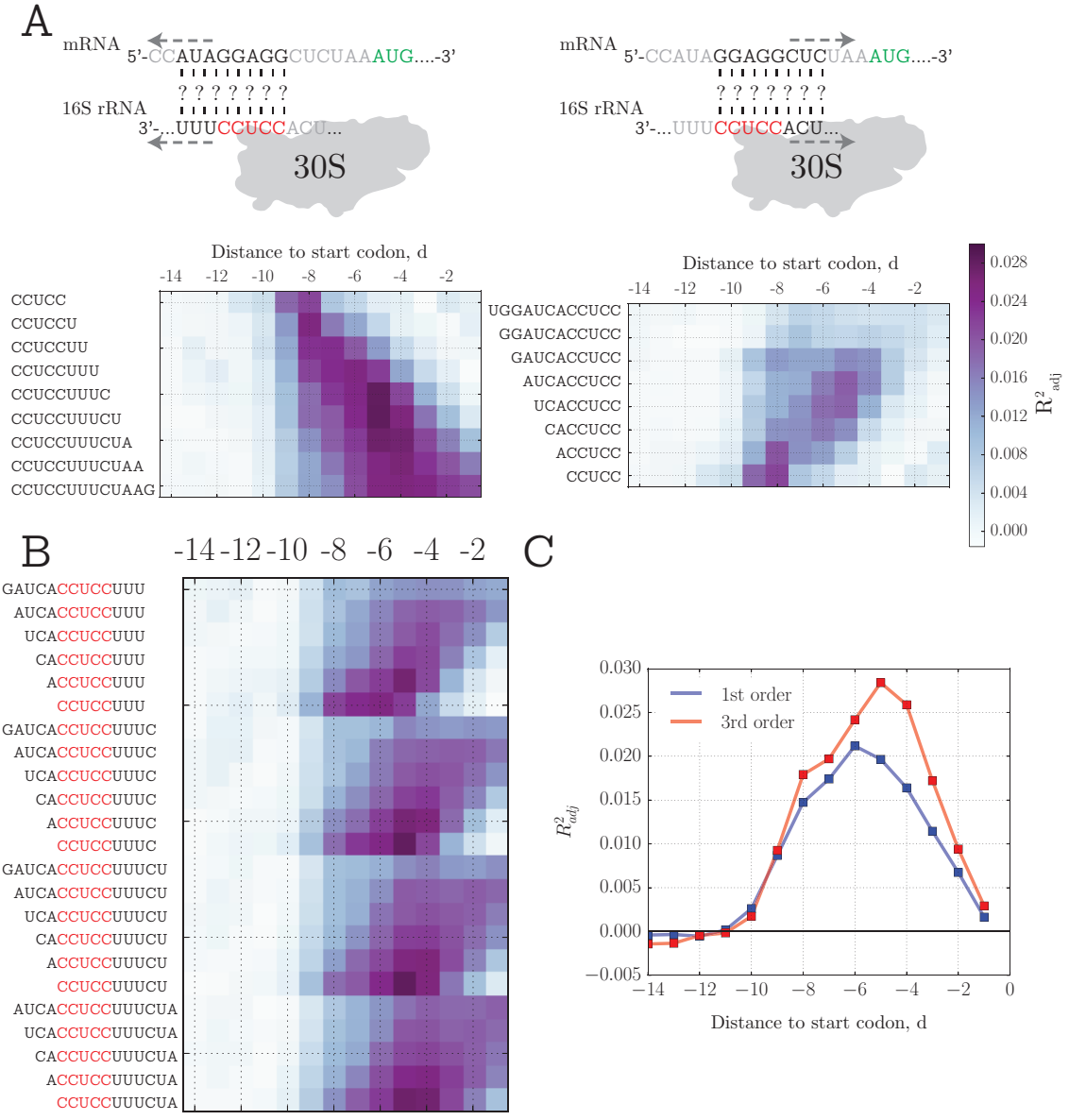

Figure B.4. **As in Fig. 3.3 of main text.** *A*, R^2_{adj} from the 3rd order model at different distances to the start codon and various 3′ and 5′ extensions to the core aSD for *C. crescentus*. *B*, Combination of best fitting putative aSDs from (A) to determine the optimal aSD sequence. *B*, Comparison of R^2_{adj} between the 1st and 3rd order polynomial models from best performing aSD sequence.

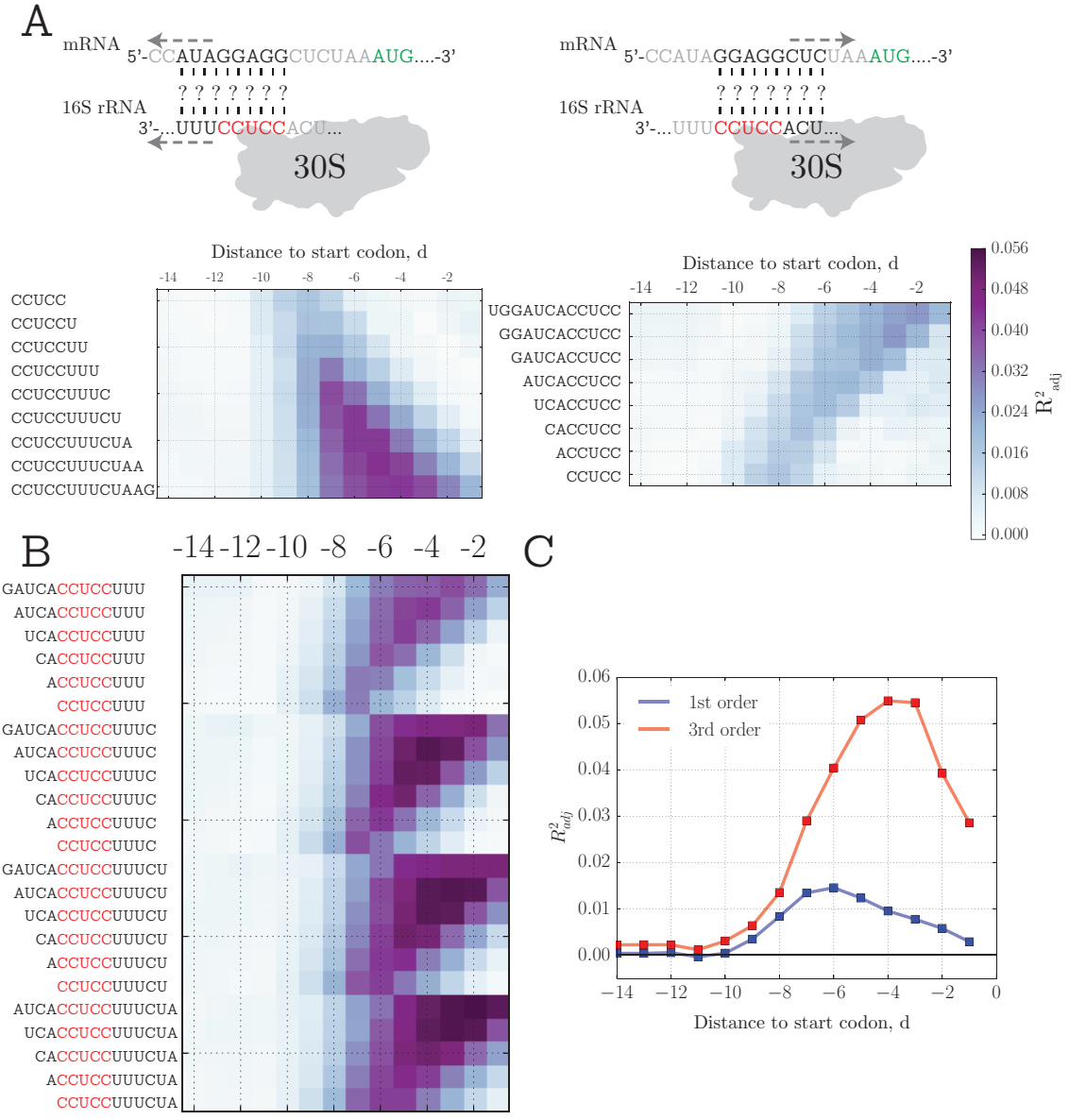

Figure B.5. **As in Fig. 3.3 of main text.** *A*, R^2_{adj} from the 3rd order model at different distances to the start codon and various 3′ and 5′ extensions to the core aSD for *B. subtilis*. *B*, Combination of best fitting putative aSDs from (A) to determine the optimal aSD sequence. *B*, Comparison of R^2_{adj} between the 1st and 3rd order polynomial models from best performing aSD sequence.

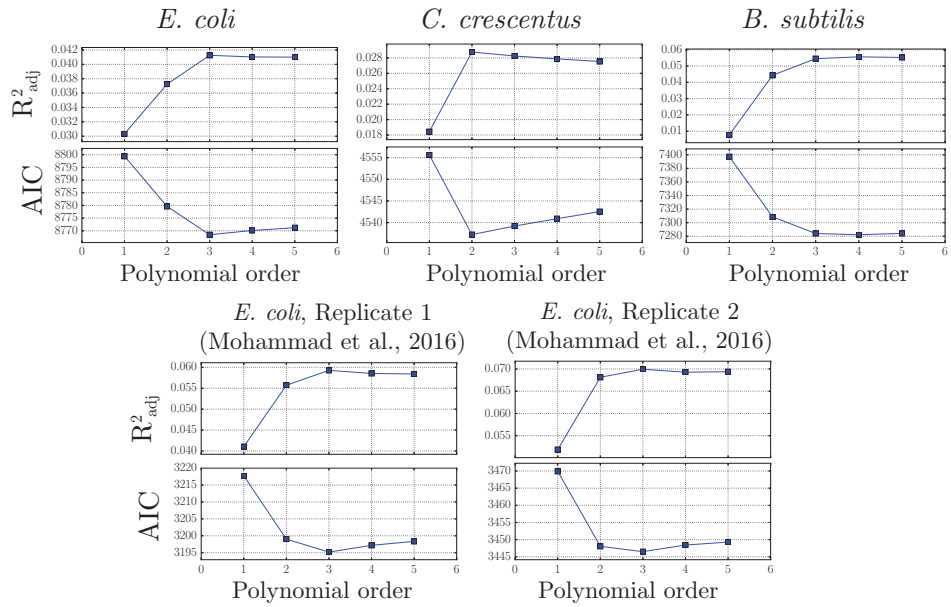

Figure B.6. **Support to Figs. 3.4,3.5 of main text.** For *E. coli* data (left), given the optimal distance and aSD parameters, we show the effect of increasingly complex polynomial fits on the R^2_{adj} (top) and Akaike Information Criterion (AIC)(bottom), two statistical methods commonly used for model selection. Both metrics penalize models with increasing parameter number through different statistical means in order to prevent over-fitting; the best model, according to the R^2_{adj}, should be the one that maximizes this metric while for the AIC the best model should minimize this value. The data are also shown for *C. crescentus* (center) and *B. subtilis* (right) data, in all cases this data was calculated using the optimal aSD and spacing values indicated in Fig. 3.4 of the main text for each organism. Bottom row shows this same data for *E. coli* based on the data used in Fig. 3.5 of the main text, from Mohammad et al. (2016).

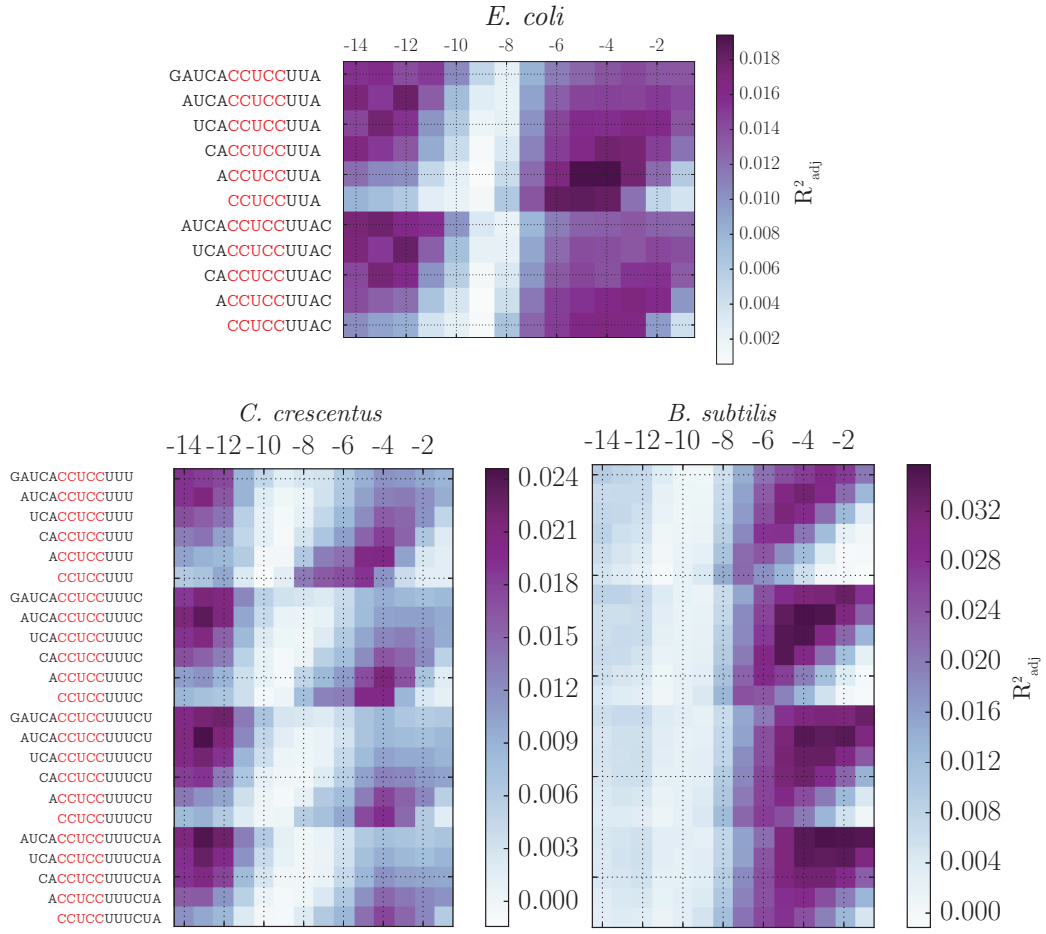

Figure B.7. **As in main text Fig. 3.3 and Supporting Figs. B.4& B.5.** Whereas the previous results display correlations between aSD sequence complementarity and residual *RTE* values (calculated by removing the effect of predicted mRNA structure), here we repeated our algorithm to choose the best fitting aSD sequence and distance parameters given the raw log-transformed *RTE* values. For each of the three different organism datasets displayed, the qualitative conclusions are similar with decreases in the overall magnitude and significance of the observed effect but clear peaks for particular aSD and distance combinations which closely align with the conclusions in the main text. We attribute the increasing significance on the left side of the *E. coli* and *C. crescentus* data to the fact that the aSD sequence complementarity is likely measuring GC content in this region and thus mRNA structure by proxy.

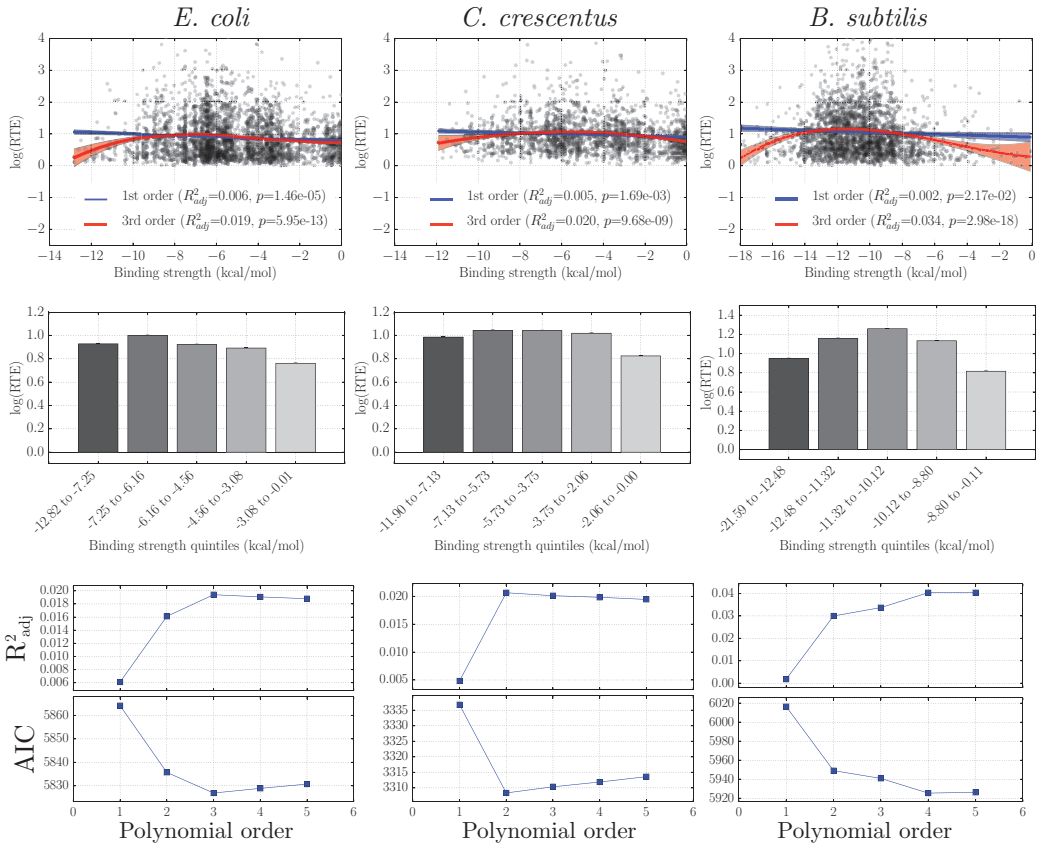

Figure B.8. **As in main text Fig. 3.4.** We repeated our analysis of the relationship between aSD sequence complementarity and translation efficiency by looking at the log-transformed RTE values (rather than the residual RTE values calculated by removing the effect of predicted mRNA structure) for the best fitting aSD sequence and distance parameters discovered in the main text. In all cases, a 3rd order model is highly statistically significant, and fits the data better than a 1st order model. Further, quintile bins again show that the strongest binding quintile of genes for all datasets exhibits reduced RTE values.

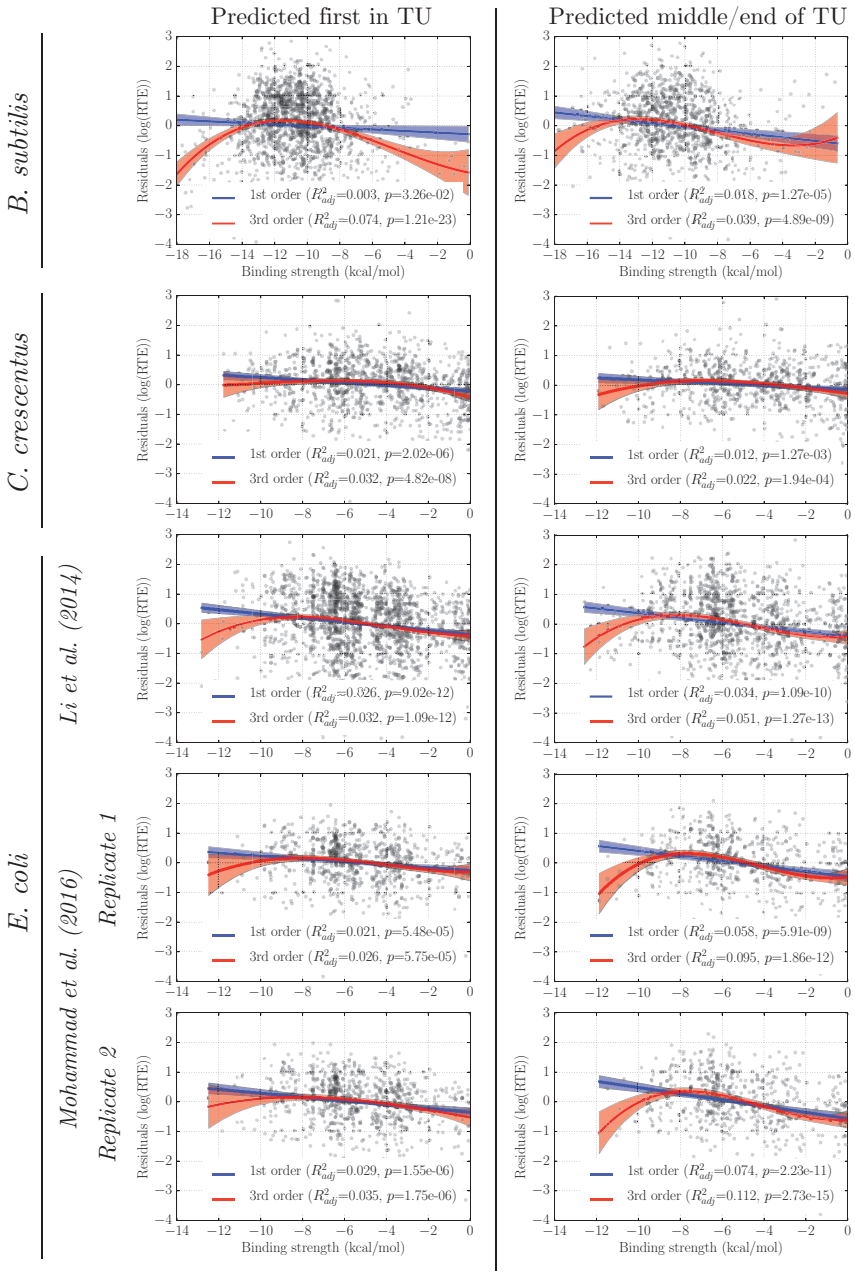

Figure B.9. **Robustness of the results with respect to gene position within operons for the indicated datasets.** aSD sequences and distances for each dataset are as in Figs. 3.4 & 3.5 of the main text.

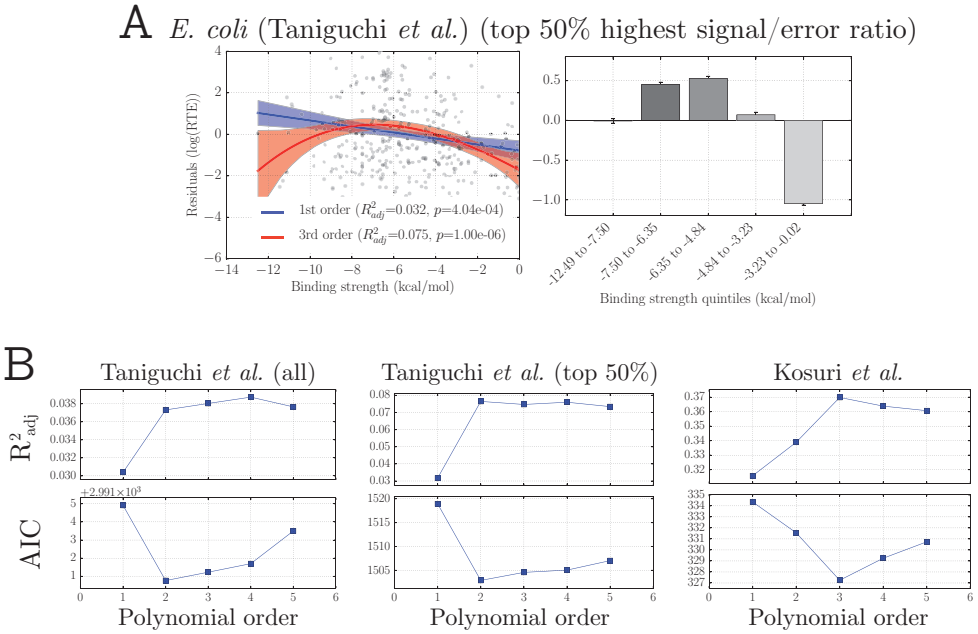

Figure B.10. **As in main text Fig. 3.6.** We analyzed the Taniguchi *et al.* dataset but here restrict the analysis to the top 50% genes in this dataset with the highest signal/error ratio and observe similar trends to Fig. 3.6A but with stronger predictive power of the underlying polynomial model as assessed by the R^2_{adj} value. B) The relevant AIC and and R^2_{adj} values for the indicated datasets from Fig. 3.6.

APPENDIX C

Supporting information to Chapter 5

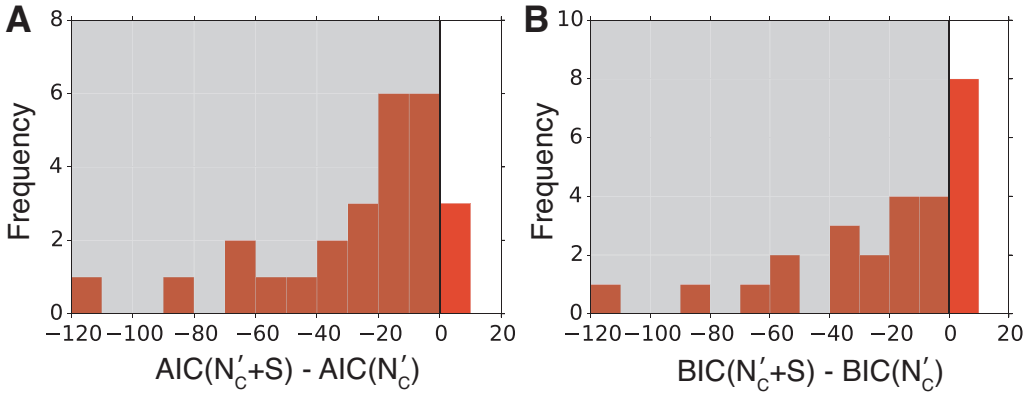

Figure C.1. Inclusion of the aSD binding strength score, S, to multivariate regression between protein abundance and codon usage bias (N'_C) enhances total predictive power of mulitple regression models as evidenced by lower (A) AIC and (B) BIC scores.

Dataset	n	R^2_{SDonly}	p	slope	$R^2_{N'_Conly}$	p	slope	R^2_{both}	p	$slope_{SD}$	$slope_{N'_C}$
Protein abundance (Lu *et al.* 2007).	424	0.173	2.179e-19	-1.111	0.209	1.441e-23	-4.975	0.266	1.971e-29	-0.713	-3.702
Mean protein abundance (Taniguchi *et al.* 2011)	1009	0.102	1.675e-25	-1.031	0.172	2.431e-43	-6.335	0.200	5.695e-50	-0.599	-5.223
mean mRNA level (Taniguchi *et al.* 2011)	825	0.103	2.308e-21	-0.827	0.153	8.828e-32	-4.663	0.183	3.321e-37	-0.495	-3.716
mRNA digital counts (Shiroguchi *et al.* 2012)	3002	0.071	7.657e-50	-0.595	0.233	2.097e-175	-5.615	0.250	1.805e-188	-0.305	-5.151
Translation efficiency (Li *et al.* 2014)	2877	0.052	1.427e-35	-0.280	0.094	4.521e-64	-1.908	0.115	3.339e-77	-0.185	-1.629

Figure C.2. Single and multivariable regression outputs for *E.coli* gene expression data, aSD binding score (S) and codon usage bias (N'_C).

Organism	n	R^2_{SD}	p	slope	$R^2_{N'_C}$	p	slope	R^2_{both}	p	$slope_{SD}$	$slope_{N'_C}$
Escherichia coli str. K-12 substr. MG1655	3736	0.0590	1.64E-051	-1.15	0.202	3.85E-185	-11.1	0.215	1.13E-197	-0.580	-10.2
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	2392	0.00176	0.0224	0.130	0.0788	8.90E-045	-8.01	0.0784	1.57E-043	0.0128	-7.99
Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819	769	0.00706	0.0112	-0.203	0.0718	2.48E-014	-5.05	0.0801	4.73E-015	-0.217	-5.09
Bartonella henselae Houston-1	1252	-0.000633	0.649	-0.0264	0.0943	5.63E-029	-6.18	0.0968	8.83E-029	-0.117	-6.34
Pseudomonas aeruginosa PAO1	2417	0.0565	1.30E-032	-0.564	0.0994	3.89E-057	-5.24	0.121	1.51E-068	-0.366	-4.44
Bacillus anthracis Sterne	1386	0.0672	5.99E-023	-0.720	0.208	1.72E-072	-6.63	0.225	1.27E-077	-0.377	-6.04
Staphylococcus aureus subsp. aureus Mu50	1683	0.0152	2.39E-007	-0.353	0.248	2.38E-106	-7.80	0.253	1.29E-107	-0.212	-7.68
Mycoplasma pneumoniae FH	381	-0.000267	0.344	0.102	0.00324	0.136	-1.29	0.00282	0.216	0.0985	-1.27
Streptococcus pyogenes M1 GAS	1284	0.0613	1.33E-019	-0.793	0.332	9.36E-115	-10.4	0.343	7.32E-118	-0.345	-9.89
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	743	0.0366	9.02E-008	-0.442	0.0990	9.47E-019	-7.72	0.124	2.22E-022	-0.368	-7.29
Microcystis aeruginosa NIES-843	4090	0.000138	0.211	-0.0504	0.0902	2.92E-086	-6.74	0.0911	6.45E-086	-0.0848	-6.77
Shewanella oneidensis MR-1	1316	0.101	1.36E-032	-0.788	0.182	1.07E-059	-6.49	0.207	3.36E-067	-0.427	-5.42
Helicobacter pylori 26695	1410	0.00260	0.0309	0.166	0.00805	0.000433	-2.31	0.0101	0.000289	0.152	-2.24
Bacteroides thetaiotaomicron VPI-5482	768	0.112	9.15E-022	-1.00	0.115	2.77E-022	-6.68	0.162	1.82E-030	-0.713	-4.82
Desulfovibrio vulgaris str. Hildenborough	956	0.0755	2.98E-018	-0.645	0.160	2.82E-038	-6.53	0.174	1.00E-040	-0.311	-5.63
Lactococcus lactis subsp. lactis Il1403	1293	0.0898	1.83E-028	-0.812	0.246	1.38E-081	-7.23	0.286	1.39E-095	-0.554	-6.60
Neisseria meningitidis MC58	411	0.0109	0.0194	-0.198	0.00755	0.0430	-1.20	0.0112	0.0370	-0.151	-0.708
Deinococcus deserti VCD115	1300	0.0717	5.21E-023	-0.672	0.139	2.24E-044	-6.55	0.164	1.80E-051	-0.418	-5.62
Mycobacterium tuberculosis H37Rv	3320	0.00546	1.19E-005	0.275	0.0295	1.27E-023	-5.31	0.0338	6.09E-026	0.246	-5.21
Shigella flexneri 2a str. 301	1648	0.0156	2.09E-007	-0.470	0.391	1.85E-179	-11.1	0.390	7.10E-178	0.0253	-11.1
Bacillus subtilis subsp. subtilis str. 168	3914	0.0106	5.90E-011	-0.259	0.138	6.91E-129	-7.89	0.143	2.44E-132	-0.176	-7.76
Acidithiobacillus ferrooxidans ATCC 23270	1302	0.00655	0.00201	-0.158	0.0915	3.49E-029	-4.59	0.0942	4.54E-029	-0.107	-4.51
Listeria monocytogenes EGD-e	469	0.0545	1.88E-007	-0.703	0.172	3.57E-021	-6.19	0.193	6.76E-023	-0.457	-5.69
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	2561	0.0936	8.04E-057	-0.812	0.143	3.65E-088	-6.11	0.183	3.50E-113	-0.556	-5.07
Yersinia pestis CO92	1197	0.0498	3.46E-015	-0.640	0.0450	7.45E-014	-4.39	0.0692	9.21E-020	-0.483	-3.13
Synechocystis sp. PCC 6803	1183	0.0770	1.46E-022	-0.622	0.0294	1.70E-009	-3.29	0.101	1.73E-028	-0.602	-2.99

Figure C.3. Single and multivariable regression outputs between protein abundance, aSD binding score (S) and codon usage bias (N'_C) in 26 bacteria.