The Effects of Differential Selection for Translation Efficiency: A Comparative Analysis of Bacterial Translation Machineries

Heba Farookhi

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the degree
Master of Science in Biology (Specialization in Bioinformatics)

Department of Biology
Faculty of Science
University of Ottawa

Abstract

Different bacterial species have dramatically different generation times, from 20–30 min in *Escherichia coli* to about two weeks in *Mycobacterium leprae*. The translation machinery in a cell needs to synthesize all proteins for a new cell in each generation. The three subprocesses of translation, i.e., initiation, elongation, and termination, are expected to be under stronger selection pressure to optimize in short-generation bacteria (SGB) such as *Vibrio natriegens* than in the long-generation *Mycobacterium leprae*. The initiation efficiency depends on the start codon decoded by the initiation tRNA, the optimal Shine–Dalgarno (SD) decoded by the anti-SD (aSD) sequence on small subunit rRNA, and the secondary structure that may embed the initiation signals and prevent them from being decoded. The elongation efficiency depends on the tRNA pool and codon usage. The termination efficiency in bacteria depends mainly on the nature of the stop codon and the nucleotide immediately downstream of the stop codon. By contrasting SGB with long-generation bacteria (LGB), we predict (1) SGB to have more ribosome RNA operons to produce ribosomes, and more tRNA genes for carrying amino acids to ribosomes, (2) SGB to have a higher percentage of genes using AUG as the start codon and UAA as the stop codon than LGB, (3) SGB to exhibit better codon and anticodon adaptation than LGB, and (4) SGB to have a weaker secondary structure near the translation initiation signals than LGB. These differences between SGB and LGB should be more pronounced in highly expressed genes than the rest of the genes. We present empirical evidence in support of these predictions. An introduction of the bacterial translation machinery is presented in Chapter 1, a modified version of the article titled “Differential Selection for Translation Efficiency Shapes Translation Machineries in Bacterial Species” is presented in Chapter 2, and Chapter 3 consists of a brief discussion.

Résumé[1]

Les temps de génération varient considérablement d'une espèce bactérienne à l'autre, allant de 20 à 30 minutes chez *Escherichia coli* à environ deux semaines chez *Mycobacterium leprae*. La machinerie de traduction d'une cellule doit synthétiser toutes les protéines d'une nouvelle cellule à chaque génération. Les trois sous-processus de la traduction, à savoir l'initiation, l'élongation et la terminaison, devraient être soumis à une pression de sélection plus forte pour les optimiser chez les bactéries à génération courte (SGB) telles que *Vibrio natriegens* que chez *Mycobacterium leprae* à génération longue. L'efficacité de l'initiation dépend du codon de départ décodé par l'ARNt d'initiation, de la séquence Shine-Dalgarno (SD) optimale décodée par la séquence anti-SD (aSD) sur la petite sous-unité d'ARNr, et de la structure secondaire qui peut intégrer les signaux d'initiation et les empêcher d'être décodés. L'efficacité de l'élongation dépend du pool d'ARNt et de l'utilisation des codons. L'efficacité de la terminaison chez les bactéries dépend principalement de la nature du codon stop et du nucléotide situé immédiatement en aval du codon stop. En comparant les SGB aux bactéries de longue génération (LGB), nous prévoyons (1) que les SGB ont plus d'opérons d'ARN ribosomique pour produire des ribosomes et plus de gènes d'ARNt pour transporter les acides aminés vers les ribosomes, (2) que les SGB ont un pourcentage plus élevé de gènes utilisant AUG comme codon de départ et UAA comme codon d'arrêt que les LGB, (3) que les SGB présentent une meilleure adaptation des codons et des anticodons que les LGB, et (4) que les SGB ont une structure secondaire plus faible que les LGB à proximité des signaux d'initiation de la traduction. Ces différences entre SGB et LGB devraient être plus prononcées dans les gènes fortement exprimés que dans le reste des gènes. Nous présentons des preuves empiriques à l'appui de ces prédictions. Une introduction à la

[1] Abstract translated from English to French using DeepL Translator

machinerie de traduction bactérienne est présentée au chapitre 1, une version modifiée de l'article intitulé "Differential Selection for Translation Efficiency Shapes Translation Machineries in Bacterial Species" est présentée au chapitre 2, et le chapitre 3 consiste en une brève discussion.

Table of Contents

Acknowledgements

I would like to thank my supervisor, Dr. Xuhua Xia for his belief in me and for his generosity with his time, my Thesis Advisory Committee members, Dr. Stephane Aris-Brosou and Dr. John Lewis, for their help and advice, all members of the Xia Lab, especially Parisa Aris, Alibek Kruglikov, Yulong Wei, and Mahbubeh Askarirad for their encouragement, and my family, especially my brother Dayyan, who I could not have completed this thesis without.

List of Abbreviations

A	Adenosine
A site	Aminoacyl site
aaRS	Aminoacyl-tRNA synthetase
AC	Anticodon
aSD	Anti-Shine-Dalgarno sequence
C	Cytosine
CAI	Codon Adaptation Index
DAMBE	Data Analysis in Molecular Biology and Evolution
DtoStart	Distance to Start codon
E site	Exit site
EF	Elongation Factor
ENC	Effective Number of Codons
fMet	Formyl-methionine
G	Guanine
GT	Generation Time
GTP	Guanosine triphosphate
HEG	Highly Expressed Gene
HGT	Horizontal Gene Transfer
I	Inosine
IC	Initiation Complex
IF	Initiation Factor
IRES	Internal Ribosomal Entry Site

ITE	Index of Translation Elongation
Leaderless mRNA	lmRNA
LEG	Lowly Expressed Gene
LGB	Long-Generation Bacteria
MFE	Minimum Folding Energy
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
nt	Nucleotide
OGT	Optimal Growth Temperature
ORF	Open Reading Frame
oriC	Origin of Chromosomal Replication
P site	Peptidyl site
RankGT	Ranked Generation Time
RBS	Ribosome Binding Site
RF	Release Factor
rRNA	Ribosomal RNA
S	Svedberg unit
SD	Shine-Dalgarno sequence
SGB	Short-Generation Bacteria
ssu	Small Subunit
T	Thymine
tRNA	Transfer RNA

U	Uracil
UTR	Untranslated Region

List of Figures and Tables

CHAPTER 1:

Introduction

Translation is the process of protein synthesis (Rodnina, 2018). In bacteria, canonical translation occurs via the subprocesses of initiation, elongation, and termination. The efficiency of subprocesses in canonical translation is characterized by key interactions between "coder" elements which carry a signal and "decoder elements" which recognize the signal carried by the coder element, in combination with ribosome dynamics. In initiation, the keystone coder:decoder pairs are the start codons (i.e., AUG, GUG, and UUG) and the initiator fMet-tRNA as well as the SD sequence within the mRNA and aSD (Xia, 2018b). In elongation, the keystone pairs are sense codons and the tRNA species bearing their cognate amino acids (Xia, 2018a). Tolerance for degeneracy at the third codon site allows for extended pairings via wobble interactions between sense codons and tRNA anticodons. In termination, the keystone pairs are stop codons and either RF1 or RF2 (Xia, 2018c).

1.1. Bacterial Genome Organization

In general, the bacterial genome is composed of a conserved core set of essential, highly expressed housekeeping genes supplemented by a highly diverse accessory complement. Of the former group, the ribosomal protein-coding genes are notable for constitutively high expression (Karlin et al., 2001), and of the latter group, conditionally essential genes that play a role in stress response (Zhang et al., 2012) may be preferentially translated via noncanonical mechanisms (Babitzke & O'Connor, 2017). Gene order and content are conserved within clusters of functionally-linked genes among related bacteria, although gene order is not wholly conserved during evolution (Casjens, 2003; Rocha, 2004; Soler-Bistué et al., 2020; Tillier & Collins, 2000).

Beyond the chromosomal genome, bacteria harbor a distinct mobilome consisting of plasmids, prophages, transposons and other mobile genetic elements, many of which code for adaptive traits such as virulence or antibiotic resistance (Morgado & Vicente, 2021). Genome-wide base composition appears to be determined based on the interplay between selective and mutational forces (Lind & Andersson, 2008; Raghavan et al., 2012). To clarify, AT- mutation bias has been described as a universal feature amongst bacteria (Hershberg & Petrov, 2010; Van Leuven & McCutcheon, 2012). As a trend, conflicting selective forces appear to shift equilibrium base composition towards increasing GC content (Lind & Andersson, 2008; Raghavan et al., 2012), although there is ongoing debate regarding the relative importance of selection for increased GC content versus recombination-driven GC-biased gene conversion (Lassalle et al., 2015; Weissman et al., 2019). Genomic GC content varies from about 14% to about 75% GC among bacteria (McCutcheon & Moran, 2010).

Across the domain Bacteria, genome sizes vary about 10-fold (Mira et al., 2001). Due to relative scarcity of intergenic sequences and duplicate genes, variation in genome size translates "almost directly" into "biochemical, physiological and organismal complexity" (Mira et al., 2001). A major driver of genome expansion is horizontal gene transfer, which has provoked a counteractive deletion bias that reduces the bacterial genome as a whole (Kuo & Ochman, 2009; Mira et al., 2001) in order to maintain genome size and preserve resources.

Lifestyle and environmental factors mediate genome-wide selection pressures (Kirchberger et al., 2020). For example, genome-streamlining of some marine bacteria in stable,

nutrient-poor environments appears to be the result of strong negative selection (Kirchberger et al., 2020). Conversely, genome erosion of host-dependent bacteria, particularly those in stable, nutrient-rich environments, occurs under the dominant influence of drift (Kirchberger et al., 2020). To illustrate a third alternative, in soil-dwelling bacteria that endure environmentally variable conditions, genome expansion appears to be the result of strong positive selection for a diverse repertoire of genes involved in adaptive response (Kirchberger et al., 2020).

In our experimental framework, we classify selected bacterial species as SGB or LGB on the basis of generation time, but we will briefly describe the conceptually overlapping concept of r-strategist/K-strategist classification, which is used widely in the field of microbial ecology, in order to provide additional context.

The r-strategist/K-strategist classification system provides a useful theoretical guideline for interpreting the influence of selection on bacterial traits. Bacteria designated as r-strategists tend to optimize for growth and fast replication in uncrowded, nutrient-abundant environments. In contrast, the slow-growing K-strategists are typically adapted for survival in nutrient-limited environments as well as competition within crowded yet stable niches (Andrews & Harris, 1986). Whole phyla may be usefully discussed as r-strategists (e.g. Proteobacteria, including *E. coli*, *V. natriegens*, and *V. natriegens*) or K-strategists (e.g. *Actinobacteria*, including *M. smegmatis*, *M. abscessus*, *M. tuberculosis*, and *M. leprae*) (Brzeszcz et al., 2016). In practice, the r-strategist and K-strategist designations represent a position along a spectrum of aggregated life history and phenotypic traits rather than a dichotomy, are expected to shift in response to fluctuating r-selective and K-selective environmental regimes, and are often superseded by

niche-specific adaptations that redefine selective strategy. It is also important to note disruptive exceptions to the classification system such as obligate intracellular pathogens, which exhibit slow growth and yet have compact genomes (Beare et al., 2011; Merhej et al., 2009). Unlike free-living bacteria, which may evolve streamlined genomes as the result of selection for metabolic efficiency (Kirchberger et al., 2020; Marais et al., 2008), convergent evolution of obligate intracellular pathogens has resulted in a conserved array of specialized genome features, including reduced genomes, minimal complement of rRNA genes, (Merhej et al., 2009), and pseudogene prevalence during periods of ongoing gene loss (Beare et al., 2011). As an aside, relaxed selection pressure in intracellular genomes often results in elevated genetic drift (Beare et al., 2011), which, in turn, elevates proteome-wide nonsynonymous mutation rates (Kirchberger et al., 2020).

1.2. Characterization of Short-Generation Bacteria and Long-Generation Bacteria

Differences in translation machinery between fast-replicating and slow-replicating bacteria appear to be shaped by distinct selective priorities. Given that generation time and growth speed are tightly coupled, we will refer to fast-replicating bacteria, fast-growing bacteria, and fast-generating bacteria interchangeably with SGB, and will refer to slow-replicating bacteria, slow-growing bacteria, and slow-generating bacteria interchangeably with LGB. For fast-replicating bacteria, translational machinery appears to be characterized by optimization for maximal translation speed, “with selection for accuracy playing a small supplementary role” (Ran & Higgs, 2012). The translation machinery of SGB in comparison to LGB can(Ran & Higgs, 2012; Rocha, 2004a; Ruusala & Kurland, 1984; Saxena et al., 2021)be characterized by elevated expression of ribosome components and tRNAs (Ran & Higgs, 2012), stronger codon usage bias (Ran & Higgs, 2012; Sharp et al., 2005), decreased prioritization of translational

accuracy, and less versatile, less robust responsiveness to stressors and adverse environmental conditions (Saxena et al., 2021; Starosta et al., 2014; Wu et al., 2022; Zhu et al., 2019).

In comparison to LGB genomes, SGB genomes feature increased copy numbers of rRNA and tRNA genes, increased operonic organization of rRNA and ribosomal proteins (Aiyar et al., 2002; Wagner, 1994), and localization of genes involved in transcription and translation near the origin of replication, where they benefit from dosage-dependent increases in gene expression stemming from multifork replication (Soler-Bistué et al., 2020). Further, SGBs tend to encode isoacceptor tRNAs that feature U at the first anticodon position and are able to promiscuously decode a greater number of codon types within each synonymous codon family via wobble interactions, a finding that is in alignment with the wobble versatility hypothesis (Carullo & Xia, 2008). Thus, the SGB tRNA gene complement may be described as a highly expressed yet modestly diverse set of generalist decoders bearing a limited set of unique anticodons. Overall, translational capacity tends to be elevated in SGB relative to LGB.

Next, in comparison to LGB genomes, SGB genomes feature stronger codon usage bias, which reflects better co-adaptation between abundance of tRNA isoacceptors and frequency of usage of cognate codons within each synonymous codon family (Sharp & Li, 1987; Vieira-Silva & Rocha, 2010), as well as better co-adaptation between tRNA gene copy number and amino acid usage (Du et al., 2017), both of which result in apparent selection for speed. Overall, in SGB, selection for translational speed appears to be a “dominant effect in driving codon usage bias”, while “selection for accuracy plays a small supplementary role” (Ran & Higgs, 2012). It

should also be noted that the effect of tRNA-mediated selection for codon usage is not always greater in SGB than in LGB. (Wei et al., 2019)

In contrast to SGB genomes, LGB genomes may tend to experience greater selective pressure favouring translational accuracy (Saxena et al., 2021), although the evidence is ultimately mixed. A potential relationship between increased translational accuracy and slower growth, albeit with the causal arrow pointing in the opposite direction, is supported by experimental evidence showing that hyperaccurate ribosomes inhibit growth (Rocha, 2004b; Ruusala & Kurland, 1984). A more indirect connection, by way of the assertion that strategic pauses in ribosome processivity are useful in regulating translation accuracy, continues to build the assertion; compared to LGB, SGB are enriched in translation-stalling proline repeat motifs (Brewer & Wagner, 2021).

In addition, slow-growing bacteria are notable for greater robustness and increased diversity in stress-tolerance mechanisms in comparison to fast-growing bacteria. Specifically, these LGB-specific regulatory features include increased drug resistance, increased prevalence of toxin/antitoxin systems, and increased persistence during dormancy, (Saxena et al., 2021).

1.3. Overview of Core Bacterial Translation Machinery

1.3.1. Translation Initiation Machinery

Of the translation stages, "translation initiation is often the limiting step" (Bulmer, 1990; Kudla et al., 2009; Liljenström & von Heijne, 1987; Xia, 2018b). The core translation initiation signal suite consists of characteristic mRNA unfolding localized to the gene region surrounding the start codon, including the start-proximal region of the 5' UTR, the start codon itself, and a

region comprising several nucleotides downstream of the start codon (Nakamoto, 2006; Scharff et al., 2011). An additional signal in the core suite is the SD sequence, which initially recruits the ribosome and aids in its proper positioning relative to the start codon (Hockenberry et al., 2017; Li, 2015; Osterman et al., 2013; Shine & Dalgarno, 1974; Wen et al., 2021; Xia, 2018b). The last core signal is the start codon itself, which can be specified by coder variants of differential signal strength (i.e. AUG, UUG, and GUG) and which is bound by the initiator tRNA-fMet (Asano, 2014; Xia, 2018b).

The minimal segment of the initiation site that must be "essentially free" of secondary structure has been determined to be 30 nucleotides. This 30nt+ segment engages in tight complementary base-pairing interactions with the ribosome (Nakamoto, 2006).

The overall strength of the SD signal can be broken down into three subsignals: (1) distance between the 3' end of the 16S small subunit rRNA and the nucleotide just before the start codon, measured in the number of nucleotide bases (denoted DtoStart), (2) distance between the SD and start codon on the mRNA (denoted "leash length"), and (3) strength of SD/aSD base-pairing (Abolbaghaei et al., 2017; Prabhakaran et al., 2015; Wei et al., 2017; Xia, 2018b).

The role of the Shine-Dalgarno sequence in translation is ultimately "ancillary" (Nakamoto, 2006). While the SD sequence is an important signal for translation initiation, widely conserved as a feature (Hockenberry et al., 2018), highly influential in modulating the rate of translation initiation of the genes it is present in (Hockenberry et al., 2017); with the length of the spacer sequence between the SD and the start codon alone dictating a 100-fold

range in modulation of protein yield (Komarova et al., 2020), and also has a significant association with predicted highly expressed genes (Ma et al., 2002), it is “not obligatory” for translation initiation (Nakamoto, 2006).

With relation to the SD sequence, there are two classes of genes which are thought to undergo canonical translation initiation, both of which share the 5’ UTR as a common feature. Genes featuring an SD sequence within their 5’UTR are classified as SD-led, while genes in which an SD sequence is absent are classified as SD-independent (Hockenberry et al., 2018; Scharff et al., 2011; Zheng et al., 2011). It has been proposed that SD-independent genes rely more on structural features than regions of complementarity between segments of the 5’UTR and the ribosome in order to recruit the ribosome to the initiation site (Scharff et al., 2011). The third major class of genes are the leaderless genes, which harbor a truncated, nonfunctional 5’UTR or lack a 5’UTR altogether (Beck & Moll, 2018; Leiva & Katz, 2022; Zheng et al., 2011).

In general, the AUG start codon is the primary start codon and the most efficient start codon for bacteria. The initiation tRNA carries a 5’-CAU-3’ anticodon, such that AUG forms a highly stable complementary base-pairing interaction, which corresponds to its high initiation efficiency (Antoun et al., 2006; Milón & Rodnina, 2012), whereas the alternative start codons, of which GUG and UUG are predominant (Asano, 2014), pair less stably with the initiation tRNA, leading to lower translation initiation efficiency (Hecht et al., 2017; Komarova et al., 2020). Typically, GUG is associated with higher initiation efficiency than UUG is (Hartz et al., 1991; Hecht et al., 2017).

1.3.2. Translation Elongation Machinery

The 61 sense codons, the particular amino acids each codon specifies, and the pairings between a limited number of tRNA species, some of which are capable of isoacceptor activity, and their cognate codons, are the features defining the translation elongation polypeptide-building step (Rodnina, 2018).

In the bacterial genetic code (i.e. default translation table 11), Methionine and Tryptophan are encoded by single-codon families, Isoleucine is encoded by a 3-fold degenerate codon family, and all other codons are encoded by 2-fold, 4-fold, or 6-fold degenerate codon families. 6-fold generate codon families can be subdivided into 2-fold and 4-fold codon subfamilies.

Codon usage bias is a term which refers to the differential usage of synonymous codons within a synonymous codon family. Codon usage bias is chiefly determined by two factors: mutation bias and tRNA-mediated selection(Bulmer, 1987; Carullo & Xia, 2008; Ikemura, 1981, 1982; Muto & Osawa, 1987; Xia, 2018a). It should be noted that amino acid usage, which is intrinsically tied to codon usage bias, is also a key feature determining translation elongation efficiency (Dufton, 1997; Xia, 2018a). Genomic GC bias results in differential availability of nucleotide building blocks for different species, such that a species with a GC-rich genome is expected to favour usage of GC-rich codons over AT-rich codons, and vice versa for a species with an AT-rich genome. Similarly, there is differential availability of tRNA species available to act as decoders within the cell (Ikemura, 1981; Novoa & Ribas de Pouplana, 2012; Sharp et al., 2005; Xia, 1998). Codon-anticodon adaptation refers to the coevolution between mRNA codons and tRNA anticodons, such that preferential codon usage within synonymous codon families reflects the availability of the most abundant cognate tRNAs (Bulmer, 1991; Xia, 1998, 2018b),

and this preference is particularly prominent in HEGs compared to LEGs. (Carullo & Xia, 2008; Coghlan & Wolfe, 2000; Comeron & Aguade, 1998; Duret & Mouchiroud, 1999; Gouy & Gautier, 1982; Ikemura, 1981, 1982; Xia, 1998).

Codon usage preferences can be measured using codon usage indices, including the Effective Number of Codons (ENC) (Sun et al., 2013; Wright, 1990), the Codon Adaptation Index (CAI) (Sharp & Li, 1987), and the Index of Translation Elongation (ITE) (Xia, 2015). ENC and CAI are widely-used gene-specific codon usage indices, while ITE is a relatively recently developed gene-specific codon usage index (Xia, 2015). ENC measures deviation from equal codon usage for each gene (Sun et al., 2013; Wright, 1990), while CAI and ITE measure the intensity of tRNA-mediated selection on codon usage bias using a reference set of HEGs (Sharp & Li, 1987; Xia, 2015), and where both CAI and ITE may take on values between 0 and 1, such that a higher value indicates higher intensity of tRNA-mediated selection (Xia, 2018a). ITE differs from CAI and other codon usage indices in that translationally optimal codons are defined as the codon within a codon family that is preferred in HEGs even after taking into account background mutation bias (Xia, 2018a). To illustrate, in cases where mutation favours a particular codon and drives up its frequency in HEG and nonHEG gene sets relative to other codons, it is possible to assess which of the codons within the R-ending or Y-ending synonymous codon subfamily is favoured by tRNA-mediated selection by determining which codon is relatively more frequent in HEGs relative to non-HEGs (Xia, 2018a). In cases where there is no background mutation bias, ITE and CAI will be identical (Xia, 2018a).

1.3.3. Translation Termination Machinery

The core translation termination signal suite consists of the stop codon itself, which can be specified by coder variants (i.e. UAA, UGA, and UAG) (Xia, 2018c), the RF1 and RF2 decoders (which each have different stop codon specificity, such that RF1 decodes UAA and UAG, while RF2 decodes UAA and UGA) (Craigen & Caskey, 1986; Dahlgren & Rydén-Aulin, 2000; Xia, 2018c), and a reduction in mRNA secondary structure strength in the region preceding the stop codon, owing to the finding that increased mRNA secondary structure can render the stop codon inaccessible (Tate & Brown, 1992; Tate & Mannering, 1996; Xia, 2018c). RF1 and RF2 achieve their discriminatory capacity through use of a PXT tripeptide anticodon motif in the former case and an SPF anticodon motif in the latter one (Korkmaz & Sanyal, 2017; Xia, 2018c). Among the stop codons, UAA is associated with the lowest readthrough error rate (Korkmaz & Sanyal, 2017; Poole et al., 1997; Ryden & Isaksson, 1984).

Notably, the +4 site downstream of the stop codon as well as additional nucleotides flanking the stop codon participate in modulation of termination efficiency for all stop codons, and may influence the rate of readthrough error (Betney et al., 2010; Wei & Xia, 2017; Xia, 2018c). To illustrate the importance of the +4 site, we will describe the case of RF2 autoregulation (Betney et al., 2010; Xia, 2018c). RF2 autoregulation is a RF2 abundance is managed via an autoregulation cycle involving functional RF2, the gene encoding RF2, which is *prfB*, and two competitor tRNAs. The *prfB* gene contains an in-frame UGA stop codon at CS26 (Betney et al., 2010; Smith, 2011; Xia, 2018c). When RF2 is low, tRNALeu/GAG or $tRNA^{Asp/GUC}$ tend to interact unconventionally with 4-nucleotide sequences overlapping the UGA stop codon, such that a +1 frameshift is induced (Betney et al., 2010; Craigen & Caskey, 1986; Smith, 2011; Xia, 2018c). Frameshifting rate at this site may vary from ~30% to 100%

(Márquez et al., 2004). To elaborate, the GAG anticodon of $tRNA^{Leu/GAG}$, in addition to an adjacent G nucleotide on the stem loop of the tRNA, forms a 3' GAGG 5' sequence (Betney et al., 2010; Smith, 2011; Xia, 2018c). This 3' GAGG 5' sequence is able to wobble-pair with the 5' CUUU 3' sequence that lies upstream of the UGA at CS26 within *prfB* (Betney et al., 2010; Smith, 2011; Xia, 2018c). Specifically, the 3' U of the 5' CUUU 3' sequence comprises the first nucleotide of UGA. Alternatively, the GUC anticodon of tRNAAsp/GUC, in addition to an adjacent A nucleotide on the stem loop of the tRNA, forms a 3' GUCA 5' sequence which is able to wobble-pair with 5' UGAC 3' (Betney et al., 2010; Smith, 2011; Xia, 2018c). The 5' UGAC 3' sequence includes the UGA and a downstream C nucleotide. In both cases, the +1 frameshift allows the UGA at CS26 within *prfB* to be skipped, and for full-length, functional RF2 to be produced using the altered reading frame (Betney et al., 2010; Craigen & Caskey, 1986; Smith, 2011; Xia, 2018c). Once a high concentration of RF2 has accumulated within the cell, successful decoding of UGA by RF2 becomes likely (Betney et al., 2010; Smith, 2011; Xia, 2018c). When RF2 decodes the UGA at CS26, *prfB* is translated as a truncated, nonfunctional peptide. In this way, the RF2 concentration begins to decrease, and the autoregulation cycle continues. (Betney et al., 2010; Craigen et al., 1985; Craigen & Caskey, 1986; Xia, 2018c). Outside of this case, use of +4C downstream of UGA is generally avoided (Brown et al., 1990; Tate et al., 1995; Wei & Xia, 2017). Further, +4U downstream of UAA appears to enhance termination efficiency, and is overrepresented in HEGs compared to LEGS (Tate et al., 1996; Tate & Brown, 1992; Wei & Xia, 2017; Xia, 2018c), but it is unclear whether this apparent pattern may also be attributable to genomic enrichment of overlapping genes with a "UAAUG" configuration (Eyre-Walker, 1996; Xia, 2018c), such that what appears to be a +4U is actually the first nucleotide of an AUG start codon which does not directly participate in translation termination.

It is also worth discussing the way selection and mutation govern stop codon preferences. When genomic GC bias is high, mutation favours usage of UAG and UGA over UAA, and RF2 abundance tends to be high, which favours usage of UGA over UAG. When genomic GC content is intermediate, UAA usage predominates, as UAA is favoured by selection and is not disfavoured by mutation bias (Wei et al., 2016; Xia, 2018c). When genomic AT bias is high, mutation favours usage of UAA over UGA and UAG, and because high AT genomes are associated with very low RF2 abundances, UGA usage is selected against while UAG usage is not (Wei et al., 2016; Xia, 2018c). As a result, across all three cases, UAG usage tends to be maintained at a low, consistent level. (Wei et al., 2016; Xia, 2018c).

1.4. Thesis Objectives and Significance

In this thesis, we aim to investigate how differences in bacterial generation time influence the evolution of translation machinery. Specifically, we investigate the influence of differential intensities of selection for efficiency of translation on the translation machinery of 9 bacterial study species of differing generation times, focusing on core translation signals relating to initiation, elongation, and termination. We expect that bacteria with shorter generation times will be subject to more intense selection pressure for translation efficiency than bacteria with longer generation times. As a secondary goal, we aim to investigate how the level of gene expression influences translation efficiency, as reflected in translation machinery composition. Specifically, we investigate the differences between HEG gene sets (consisting of ribosomal protein genes and select genes previously characterized as highly expressed in several bacterial species (Karlin et

al., 2001) and REST gene (consisting of all the remaining protein-coding genes in the genome). We expect that HEGs will be subject to more intense selection pressure for translation efficiency than REST genes.

The 9 study species chosen for our analyses were selected due to the decisive and pronounced demarcations between their reported generation times. Based on these generation times, we have chosen the terms "Short-Generation Bacteria" (SGB) to refer to *Vibrio natriegens, Vibrio cholerae, Escherichia coli, Bacillus subtilis, Haemophilus influenzae,* and *Mycolicibacterium smegmatis* and "Long-Generation Bacteria" to refer to *Mycobacterioides abscessus, Mycobacterium tuberculosis,* and *Mycobacterium leprae*.

1.5. References

Abolbaghaei, A., Silke, J. R., & Xia, X. (2017). How Changes in Anti-SD Sequences Would Affect SD Sequences in *Escherichia coli* and *Bacillus subtilis*. *G3 (Bethesda, Md.)*, *7*(5), 1607–1615. https://doi.org/10.1534/G3.117.039305

Aiyar, S. E., Gaal, T., & Gourse, R. L. (2002). rRNA promoter activity in the fast-growing bacterium *Vibrio natriegens*. *Journal of Bacteriology*, *184*(5), 1349–1358. https://doi.org/10.1128/JB.184.5.1349-1358.2002/ASSET/F2E295AF-A382-4FAE-9FB9-E978F8EACA53/ASSETS/GRAPHIC/JB0521265006.JPEG

Andrews, J. H., & Harris, R. F. (1986). r- and K-Selection and Microbial Ecology. *Advances in Microbial Ecology*, 99–147. https://doi.org/10.1007/978-1-4757-0611-6_3

Antoun, A., Pavlov, M. Y., Lovmar, M., & Ehrenberg, M. (2006). How Initiation Factors Maximize the Accuracy of tRNA Selection in Initiation of Bacterial Protein Synthesis. *Molecular Cell*, *23*(2). https://doi.org/10.1016/j.molcel.2006.05.030

Asano, K. (2014). Review Article: Why is start codon selection so precise in eukaryotes? *Translation*, *2*(1), e28387. https://doi.org/10.4161/TRLA.28387

Babitzke, P., & O'Connor, M. (2017). Noncanonical Translation Initiation Comes of Age. *Journal of Bacteriology*, *199*(14), 295–312. https://doi.org/10.1128/JB.00295-17

Beare, P. A., Sandoz, K. M., Omsland, A., Rockey, D. D., & Heinzen, R. A. (2011). Advances in Genetic Manipulation of Obligate Intracellular Bacterial Pathogens. *Frontiers in Microbiology*, *2*(MAY). https://doi.org/10.3389/FMICB.2011.00097

Beck, H. J., & Moll, I. (2018). Leaderless mRNAs in the Spotlight: Ancient but Not Outdated! *Regulating with RNA in Bacteria and Archaea*, 155–170. https://doi.org/10.1128/9781683670247.CH10

Betney, R., De Silva, E., Krishnan, J., & Stansfield, I. (2010). Autoregulatory systems controlling translation factor expression: thermostat-like control of translational accuracy. *RNA (New York, N.Y.)*, *16*(4), 655–663. https://doi.org/10.1261/RNA.1796210

Brewer, T. E., & Wagner, A. (2021). Translation stalling proline motifs are enriched in slow-growing, thermophilic, and multicellular bacteria. *The ISME Journal 2021 16:4*, *16*(4), 1065–1073. https://doi.org/10.1038/s41396-021-01154-y

Brown, C. M., Stockwell, P. A., Trotman, C. N. A., & Tate, W. P. (1990). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Research*, *18*(21), 6339–6345. https://doi.org/10.1093/NAR/18.21.6339

Brzeszcz, J., Steliga, T., Kapusta, P., Turkiewicz, A., & Kaszycki, P. (2016). r-strategist versus K-strategist for the application in bioremediation of hydrocarbon-contaminated soils. *International Biodeterioration & Biodegradation*, *106*, 41–52. https://doi.org/10.1016/J.IBIOD.2015.10.001

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature 1987 325:6106*, *325*(6106), 728–730. https://doi.org/10.1038/325728a0

Bulmer, M. (1990). The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res*, *18*, 2869–2873.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, *129*(3). https://doi.org/10.1093/genetics/129.3.897

Carullo, M., & Xia, X. (2008). An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. *Journal of Molecular Evolution*, *66*(5). https://doi.org/10.1007/s00239-008-9102-8

Casjens, S. (2003). THE DIVERSE AND DYNAMIC STRUCTURE OF BACTERIAL GENOMES. *Https://Doi.Org/10.1146/Annurev.Genet.32.1.339*, *32*, 339–377. https://doi.org/10.1146/ANNUREV.GENET.32.1.339

Coghlan, A., & Wolfe, K. H. (2000). Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast*, *16*, 1131–1145.

Comeron, J. M., & Aguade, M. (1998). An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol*, *47*, 268–274.

Craigen, W. J., & Caskey, C. T. (1986). Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, *322*(6076). https://doi.org/10.1038/322273a0

Craigen, W. J., Cook, R. G., Tate, W. P., & Caskey, C. T. (1985). Bacterial peptide chain release factors: Conserved primary structure and possible frameshift regulation of release factor 2. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(11). https://doi.org/10.1073/pnas.82.11.3616

Dahlgren, A., & Rydén-Aulin, M. (2000). A novel mutation in ribosomal protein S4 that affects the function of a mutated RF1. *Biochimie*, *82*(8). https://doi.org/10.1016/S0300-9084(00)01160-3

Du, M. Z., Wei, W., Qin, L., Liu, S., Zhang, A. Y., Zhang, Y., Zhou, H., & Guo, F. B. (2017). Co-adaption of tRNA gene copy number and amino acid usage influences translation rates in three life domains. *DNA Research*, *24*(6). https://doi.org/10.1093/dnares/dsx030

Dufton, M. J. (1997). Genetic Code Synonym Quotas and Amino Acid Complexity: Cutting the Cost of Proteins? *Journal of Theoretical Biology*, *187*(2), 165–173. https://doi.org/10.1006/JTBI.1997.0443

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. U S A*, *96*, 4482–4487.

Eyre-Walker, A. (1996). The close proximity of *Escherichia coli* genes: Consequences for stop codon and synonymous codon use. *Journal of Molecular Evolution*, *42*(2). https://doi.org/10.1007/BF02198830

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, *10*(22). https://doi.org/10.1093/nar/10.22.7055

Hartz, D., McPheeters, D. S., & Gold, L. (1991). Influence of mRNA determinants on translation initiation in *Escherichia coli*. *Journal of Molecular Biology*, *218*(1), 83–97. https://doi.org/10.1016/0022-2836(91)90875-7

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D., & Salit, M. (2017). Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Research*, *45*(7), 3615. https://doi.org/10.1093/NAR/GKX070

Hershberg, R., & Petrov, D. A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics*, *6*(9), e1001115. https://doi.org/10.1371/JOURNAL.PGEN.1001115

Hockenberry, A. J., Pah, A. R., Jewett, M. C., & Amaral, L. A. N. (2017). Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. *Open Biology*, *7*(1). https://doi.org/10.1098/RSOB.160239

Hockenberry, A. J., Stern, A. J., Amaral, L. A. N., & Jewett, M. C. (2018). Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. *Molecular Biology and Evolution*, *35*(3). https://doi.org/10.1093/molbev/msx310

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol*, *146*, 1–21.

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

Karlin, S., Mrázek, J., Campbell, A., & Kaiser, D. (2001). Characterizations of highly expressed genes of four fast-growing bacteria. *Journal of Bacteriology*, *183*(17), 5025–5040. https://doi.org/10.1128/JB.183.17.5025-5040.2001/ASSET/F63B6988-E3D0-4158-A250-3480C806FCD5/ASSETS/GRAPHIC/JB1710100001.JPEG

Kirchberger, P. C., Schmidt, M. L., & Ochman, H. (2020). The Ingenuity of Bacterial Genomes. In *Annual Review of Microbiology* (Vol. 74). https://doi.org/10.1146/annurev-micro-020518-115822

Komarova, E. S., Chervontseva, Z. S., Osterman, I. A., Evfratov, S. A., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Bogdanov, A. A., Gelfand, M. S., Dontsova, O. A., & Sergiev, P. V. (2020). Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in *Escherichia coli*. *Microbial Biotechnology*, *13*(4), 1254–1261. https://doi.org/10.1111/1751-7915.13561

Korkmaz, G., & Sanyal, S. (2017). R213I mutation in release factor 2 (RF2) is one step forward for engineering an omnipotent release factor in bacteria *Escherichia coli*. *Journal of Biological Chemistry*, *292*(36). https://doi.org/10.1074/jbc.M117.785238

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science (New York, N.Y.)*, *324*(5924), 255–258. https://doi.org/10.1126/SCIENCE.1170160

Kuo, C.-H., & Ochman, H. (2009). Deletional bias across the three domains of life. *Genome Biology and Evolution*, *1*, 145–152. https://doi.org/10.1093/GBE/EVP016

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., & Daubin, V. (2015). GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLOS Genetics*, *11*(2), e1004941. https://doi.org/10.1371/JOURNAL.PGEN.1004941

Leiva, L. E., & Katz, A. (2022). Regulation of Leaderless mRNA Translation in Bacteria. *Microorganisms 2022, Vol. 10, Page 723*, *10*(4), 723. https://doi.org/10.3390/MICROORGANISMS10040723

Li, G. W. (2015). How do bacteria tune translation efficiency? In *Current Opinion in Microbiology* (Vol. 24). https://doi.org/10.1016/j.mib.2015.01.001

Liljenström, H., & von Heijne, G. (1987). Translation rate modification by preferential codon usage: Intragenic position effects. *Journal of Theoretical Biology*, *124*(1). https://doi.org/10.1016/S0022-5193(87)80251-5

Lind, P. A., & Andersson, D. I. (2008). Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 17878–17883. https://doi.org/10.1073/PNAS.0804445105/SUPPL_FILE/0804445105SI.PDF

Ma, J., Campbell, A., & Karlin, S. (2002). Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures. *Journal of Bacteriology*, *184*(20), 5733. https://doi.org/10.1128/JB.184.20.5733-5745.2002

Marais, G. A. B., Calteau, A., & Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, *134*(2), 205–210. https://doi.org/10.1007/S10709-007-9226-6/METRICS

Márquez, V., Wilson, D. N., Tate, W. P., Triana-Alonso, F., & Nierhaus, K. H. (2004). Maintaining the Ribosomal Reading Frame. *Cell*, *118*(1). https://doi.org/10.1016/j.cell.2004.06.012

McCutcheon, J. P., & Moran, N. A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. *Genome Biology and Evolution*, *2*(1). https://doi.org/10.1093/gbe/evq055

Merhej, V., Royer-Carenzi, M., Pontarotti, P., & Raoult, D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct*, *4*(1), 1–25. https://doi.org/10.1186/1745-6150-4-13/FIGURES/6

Milón, P., & Rodnina, M. V. (2012). Kinetic control of translation initiation in bacteria. *Critical Reviews in Biochemistry and Molecular Biology*, *47*(4), 334–348. https://doi.org/10.3109/10409238.2012.678284

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *TRENDS in Genetics*, *17*(10). http://tig.trends.com0168

Morgado, S. M., & Vicente, A. C. P. (2021). Comprehensive in silico survey of the Mycolicibacterium mobilome reveals an as yet underexplored diversity. *Microbial Genomics*, *7*(3). https://doi.org/10.1099/mgen.0.000533

Muto, A., & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA*, *84*, 166–169.

Nakamoto, T. (2006). A unified view of the initiation of protein synthesis. *Biochemical and Biophysical Research Communications*, *341*(3), 675–678. https://doi.org/10.1016/J.BBRC.2006.01.019

Novoa, E. M., & Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends in Genetics*, *28*(11), 574–581. https://doi.org/10.1016/J.TIG.2012.07.006

Osterman, I. A., Evfratov, S. A., Sergiev, P. V., & Dontsova, O. A. (2013). Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Research*, *41*(1), 474–486. https://doi.org/10.1093/NAR/GKS989

Poole, E. S., Brimacombe, R., & Tate, W. P. (1997). Decoding the translational termination signal: The polypeptide chain release factor in *Escherichia coli* crosslinks to the base following the stop codon. *RNA*, *3*(9).

Prabhakaran, R., Chithambaram, S., & Xia, X. (2015). *Escherichia coli* and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. *The Journal of General Virology*, *96*(Pt 5), 1169–1179. https://doi.org/10.1099/VIR.0.000050

Raghavan, R., Kelkar, Y. D., & Ochman, H. (2012). A selective force favoring increased G+C content in bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14504–14507. https://doi.org/10.1073/PNAS.1205683109/-/DCSUPPLEMENTAL/PNAS.201205683SI.PDF

Ran, W., & Higgs, P. G. (2012). Contributions of Speed and Accuracy to Translational Selection in Bacteria. *PLOS ONE*, *7*(12), e51652. https://doi.org/10.1371/JOURNAL.PONE.0051652

Rocha, E. P. C. (2004a). Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, *14*(11), 2279–2286. https://doi.org/10.1101/GR.2896904

Rocha, E. P. C. (2004b). The replication-related organization of bacterial genomes. *Microbiology*, *150*(6), 1609–1627. https://doi.org/10.1099/MIC.0.26974-0

Rodnina, M. V. (2018). Translation in Prokaryotes. *Cold Spring Harbor Perspectives in Biology*, *10*(9), a032664. https://doi.org/10.1101/CSHPERSPECT.A032664

Ruusala, T., & Kurland, C. G. (1984). Streptomycin preferentially perturbs ribosomal proofreading. *MGG Molecular & General Genetics*, *198*(1). https://doi.org/10.1007/BF00328707

Ryden, S. M., & Isaksson, L. A. (1984). A temperature-sensitive mutant of *Escherichia coli* that shows enhanced misreading of UAG/A and increased efficiency for some tRNA nonsense suppressors. *Mol Gen Genet*, *193*, 38–45.

Saxena, S., Spaink, H. P., & Forn-Cuní, G. (2021). Drug resistance in nontuberculous mycobacteria: Mechanisms and models. In *Biology* (Vol. 10, Issue 2). https://doi.org/10.3390/biology10020096

Scharff, L. B., Childs, L., Walther, D., & Bock, R. (2011). Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genetics*, *7*(6). https://doi.org/10.1371/JOURNAL.PGEN.1002155

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, *33*(4). https://doi.org/10.1093/nar/gki242

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. https://doi.org/10.1093/NAR/15.3.1281

Shine, J., & Dalgarno, L. (1974). 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 of the United States of America*, *71*(4), 1342–1346. https://doi.org/10.1073/PNAS.71.4.1342

Smith, E. (2011). *Is release factor 2 uniquely regulated because it is a key transcriptional regulator of other genes*. https://api.semanticscholar.org/CorpusID:86184518

Soler-Bistué, A., Aguilar-Pierlé, S., Garcia-Garcerá, M., Val, M. E., Sismeiro, O., Varet, H., Sieira, R., Krin, E., Skovgaard, O., Comerci, D. J., Rocha, E. P. C., & Mazel, D. (2020). Macromolecular crowding links ribosomal protein gene dosage to growth rate in *Vibrio cholerae*. *BMC Biology*, *18*(1), 1–18. https://doi.org/10.1186/S12915-020-00777-5/FIGURES/6

Starosta, A. L., Lassak, J., Jung, K., & Wilson, D. N. (2014). The bacterial translation stress response. In *FEMS Microbiology Reviews* (Vol. 38, Issue 6). https://doi.org/10.1111/1574-6976.12083

Sun, X. Y., Yang, Q., & Xia, X. (2013). An Improved Implementation of Effective Number of Codons (Nc). *Molecular Biology and Evolution*, *30*, 191–196.

Tate, W. P., & Brown, C. M. (1992). Translational Termination: "Stop" for Protein Synthesis or "Pause" for Regulation of Gene Expression. *Biochemistry*, *31*(9). https://doi.org/10.1021/bi00124a001

Tate, W. P., & Mannering, S. A. (1996). Three, four or more: The translational stop signal at length. In *Molecular Microbiology* (Vol. 21, Issue 2). https://doi.org/10.1046/j.1365-2958.1996.6391352.x

Tate, W. P., Poole, E. S., Dalphin, M. E., Major, L. L., Crawford, D. J. G., & Mannering, S. A. (1996). The translational stop signal: Codon with a context, or extended factor recognition element? *Biochimie*, *78*(11–12). https://doi.org/10.1016/S0300-9084(97)86716-8

Tate, W. P., Poole, E. S., Horsfield, J. A., Mannering, S. A., Brown, C. M., Moffat, J. G., Dalphin, M. E., McCaughan, K. K., Major, L. L., & Wilson, D. N. (1995). Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. *Https://Doi.Org/10.1139/O95-118*, *73*(11–12), 1095–1103. https://doi.org/10.1139/O95-118

Tillier, E. R. M., & Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nature Genetics 2000 26:2*, *26*(2), 195–197. https://doi.org/10.1038/79918

Van Leuven, J. T., & McCutcheon, J. P. (2012). An AT mutational bias in the tiny GC-rich endosymbiont genome of Hodgkinia. *Genome Biology and Evolution*, *4*(1), 24–27. https://doi.org/10.1093/GBE/EVR125

Vieira-Silva, S., & Rocha, E. P. C. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genetics*, *6*(1). https://doi.org/10.1371/journal.pgen.1000808

Wagner, R. (1994). The regulation of ribosomal RNA synthesis and bacterial cell growth. In *Archives of Microbiology* (Vol. 161, Issue 2). https://doi.org/10.1007/BF00276469

Wei, Y., Silke, J. R., & Xia, X. (2017). Elucidating the 16S rRNA 3′ boundaries and defining optimal SD/aSD pairing in *Escherichia coli* and *Bacillus subtilis* using RNA-Seq data. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-17918-6

Wei, Y., Silke, J. R., & Xia, X. (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-39369-X

Wei, Y., Wang, J., & Xia, X. (2016). Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. *Molecular Biology and Evolution*, *33*(9), 2357–2367. https://doi.org/10.1093/molbev/msw107

Wei, Y., & Xia, X. (2017). The role of +4U as an extended translation termination signal in bacteria. *Genetics*, *205*(2), 539–549. https://doi.org/10.1534/GENETICS.116.193961/-/DC1

Weissman, J. L., Fagan, W. F., & Johnson, P. L. F. (2019). Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLOS Genetics*, *15*(11), e1008493. https://doi.org/10.1371/JOURNAL.PGEN.1008493

Wen, J. Der, Kuo, S. T., & Chou, H. H. D. (2021). The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation. In *RNA Biology* (Vol. 18, Issue 11). https://doi.org/10.1080/15476286.2020.1861406

Wright, F. (1990). The ‘effective number of codons’ used in a gene. *Gene*, *87*(1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Wu, C., Balakrishnan, R., Braniff, N., Mori, M., Manzanarez, G., Zhang, Z., & Hwa, T. (2022). Cellular perception of growth rate and the mechanistic origin of bacterial growth law. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(20), e2201585119. https://doi.org/10.1073/PNAS.2201585119/SUPPL_FILE/PNAS.2201585119.SAPP.PDF

Xia, X. (1998). How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and Saccharomyces cerevisiae? *Genetics*, *149*(1). https://doi.org/10.1093/genetics/149.1.37

Xia, X. (2015). A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics*, *199*(2), 573–579. https://doi.org/10.1534/GENETICS.114.172106

Xia, X. (2018a). Bioinformatics and Translation Elongation. In *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics* (pp. 197–238). Springer. https://doi.org/10.1007/978-3-319-90684-3_9

Xia, X. (2018b). Bioinformatics and Translation Initiation. In *Bioinformatics and the Cell: modern computational approaches in genomics, proteomics and transcriptomics,* (pp. 173–195). Springer. https://doi.org/10.1007/978-3-319-90684-3_8

Xia, X. (2018c). Bioinformatics and Translation Termination in Bacteria. In *Bioinformatics and the Cell: modern computational approaches in genomics, proteomics and transcriptomics* (pp. 239–254). Springer.

Zhang, Y. J., Ioerger, T. R., Huttenhower, C., Long, J. E., Sassetti, C. M., Sacchettini, J. C., & Rubin, E. J. (2012). Global Assessment of Genomic Regions Required for Growth in *Mycobacterium tuberculosis*. *PLoS Pathogens*, *8*(9). https://doi.org/10.1371/journal.ppat.1002946

Zheng, X., Hu, G. Q., She, Z. S., & Zhu, H. (2011). Leaderless genes in bacteria: Clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics*, *12*. https://doi.org/10.1186/1471-2164-12-361

Zhu, M., Pan, Y., & Dai, X. (2019). (p)ppGpp: the magic governor of bacterial growth economy. *Current Genetics*, *65*(5), 1121–1125. https://doi.org/10.1007/S00294-019-00973-Z

CHAPTER 2

Differential Selection for Translation Efficiency Shapes Translation Machineries in Bacterial Species

A version of this manuscript is published in the journal MDPI *Microorganisms*, with modifications to the formatting:

Farookhi, H., & Xia, X. (2024). Differential Selection for Translation Efficiency Shapes Translation Machineries in Bacterial Species. *Microorganisms*, *12*(4), 768. https://doi.org/10.3390/MICROORGANISMS12040768

2.1. Author Contributions:

Heba Farookhi [1] and Xuhua Xia [1,2,*]

1. Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486.

2. Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada K1H 8M5.

Conceptualization, H.F. and X.X.; methodology, H.F. and X.X.; software, H.F. and X.X.; validation, X.X.; formal analysis, H.F. and X.X.; investigation, H.F. and X.X.; resources, X.X.; data curation, H.F. and X.X.; writing—original draft preparation, H.F. and X.X.; writing—review and editing, H.F. and X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, X.X.

2.2. Abstract

Different bacterial species have dramatically different generation times, from 20–30 min in *Escherichia coli* to about two weeks in *Mycobacterium leprae*. The translation machinery in a cell needs to synthesize all proteins for a new cell in each generation. The three subprocesses of translation, i.e., initiation, elongation, and termination, are expected to be under stronger selection pressure to optimize in short-generation bacteria (SGB) such as *Vibrio natriegens* than in the long-generation *Mycobacterium leprae*. The initiation efficiency depends on the start codon decoded by the initiation tRNA, the optimal Shine–Dalgarno (SD) decoded by the anti-SD (aSD) sequence on small subunit rRNA, and the secondary structure that may embed the initiation signals and prevent them from being decoded. The elongation efficiency depends on the tRNA pool and codon usage. The termination efficiency in bacteria depends mainly on the nature of the stop codon and the nucleotide immediately downstream of the stop codon. By contrasting SGB with long-generation bacteria (LGB), we predict (1) SGB to have more ribosome RNA operons to produce ribosomes, and more tRNA genes for carrying amino acids to ribosomes, (2) SGB to have a higher percentage of genes using AUG as the start codon and UAA as the stop codon than LGB, (3) SGB to exhibit better codon and anticodon adaptation than LGB, and (4) SGB to have a weaker secondary structure near the translation initiation signals than LGB. These differences between SGB and LGB should be more pronounced in highly expressed genes than the rest of the genes. We present empirical evidence in support of these predictions.

2.3. Introduction

"The dream of a bacterium is to become two bacteria" (Jacob, 2013), but the realization of this dream is often delayed by the rate of biosynthesis, especially the rate of translation (Bulmer, 1991; Kudla et al., 2009; Liljenström & von Heijne, 1987; Prabhakaran et al., 2015; Xia, 1998, 2015, 2021, 2023b), partly because most of the cellular dry weight is contributed by proteins. In well-studied enterobacteria, proteins account for more than half of the cell dry mass (Bremer & Dennis, 1996; Valgepea et al., 2013). In *Escherichia coli*, the ratio of dry weight to wet weight is 0.2294 (Glazyrina et al., 2010), and the ratio of protein weight to wet weight is about 0.2 (Milo, 2013). Thus, proteins contribute about 87% (=0.2/0.2294) of the cellular dry biomass. It is not surprising that the growth rate of *E. coli* increases with the rate of protein production (Valgepea et al., 2013). One, therefore, would expect a strong selection to optimize the translation machinery and mRNA features to increase translation efficiency.

2.3.1. Empirical Studies on Natural Selection for Optimizing Translation Efficiency

Efficient translation requires efficient initiation, elongation, and termination, as well as mRNA stability, and limits the rate of biosynthesis in both bacteria (Kudla et al., 2009; Tuller et al., 2010; Xia, 1998, 2015; Zhu & Dai, 2018) and phages (Chithambaram et al., 2014a, 2014b; Prabhakaran et al., 2015). Translation initiation is often the limiting step (D. I. Andersson & Kurland, 1983; Bulmer, 1990, 1991; Gualerzi & Pon, 2015; Liljenström & von Heijne, 1987). Efficient initiation in bacteria generally requires (1) AUG as a start codon, (2) a well-positioned base-pairing between the Shine–Dalgarno (SD) sequence and the anti-SD (aSD) of the free 3′ end of the small ribosomal rRNA (Hui & De Boer, 1987; Shine & Dalgarno, 1974b, 1974a, 1975; Steitz & Jakes, 1975; Taniguchi & Weissmann, 1978), and (3) no strong secondary

structure that would embed the start codon or SD and consequently obscure it from being decoded by initiation tRNA or aSD, respectively (Calogero et al., 1988; Nakamoto, 2006; Studer & Joseph, 2006).

With efficient translation initiation, translation elongation becomes rate-limiting (Tuller et al., 2010; Xia, 1998). Codon–anticodon adaptation is invariably observed in rapidly replicating organisms (Bulmer, 1991; Ikemura, 1981; Xia, 1998, 2018a), especially in highly expressed genes (Coghlan & Wolfe, 2000; Comeron & Aguade, 1998; Duret & Mouchiroud, 1999; Gouy & Gautier, 1982; Ikemura, 1981, 1982; Xia, 1998, 2015, 2018a). The same pattern was also found in phages (Chithambaram et al., 2014a, 2014b; Prabhakaran et al., 2014, 2015; Xia, 2023b). Experimental replacement of minor codons by major codons or vice versa typically leads to an increased or decreased translation rate in bacteria (D. I. Andersson et al., 1982; S. G. Andersson et al., 1984; S. G. Andersson & Kurland, 1990; Robinson et al., 1984; Sorensen et al., 1989; Xia, 2015) and viruses (Haas et al., 1996; Ngumbela et al., 2008). Eukaryotic viruses such as HIV-1 tend to have the codon usage of their early and late genes adapted to their respective tRNA pools (Van Weringh et al., 2011). These results have led to the explicit formulation of codon–anticodon coevolution and adaptation theory, e.g., (Akashi, 1994; Moriyama & Powell, 1997; Ran & Higgs, 2012; Xia, 1998, 2008), and codon adaptation indices (Sharp & Li, 1987; Sun et al., 2013; Xia, 2007, 2015). Although the relationship between codon usage optimization and protein production was briefly challenged (Kudla et al., 2009), the relationship was fully reestablished by more detailed data analysis (Tuller et al., 2010; Xia, 2015).

Translation termination in bacterial species is mediated by one or two release factors RF1 (decoding UAA and UAG) and RF2 (decoding UAA and UGA). All three stop codons can be misread by tRNAs in bacterial species (Davies et al., 1966; Ryden & Isaksson, 1984). The readthrough frequency is at least 10^{-3} to 10^{-2} for UGA in *Salmonella typhimurium* (Roth, 1970) and *E. coli* (Sambrook et al., 1967) and 1.1×10^{-4} to 7×10^{-3} for UAG, depending on the nature of the downstream nucleotides (Bossi, 1983; Bossi & Roth, 1980; Ryden & Isaksson, 1984; Strigini & Brickman, 1973). The readthrough rate is the lowest for UAA, at frequencies from 9×10^{-4} to less than 1×10^{-5} (Ryden & Isaksson, 1984). The readthrough error rate in the order of UGA > UAG > UAA has been consistently observed in multiple studies (César Sánchez et al., 1998; Geller & Rich, 1980; Jorgensen et al., 1993; Meng et al., 1995; Miller & Albertini, 1983; Parker, 1989; W. P. Tate et al., 1999). Empirical evidence supports the hypothesis that highly expressed genes prefer UAA stop codons (Wei et al., 2016).

Termination efficiency also depends on the nucleotide immediately downstream of the stop codon (W. P. Tate & Brown, 1992; W. P. Tate & Mannering, 1996; Wei & Xia, 2017), leading to the proposal of the tetranucleotide stop signal including the +4 site (Brown et al., 1990; Konecki et al., 1977; McCaughan et al., 1995; W. P. Tate et al., 1995, 1996; W. P. Tate & Brown, 1992; Wei & Xia, 2017; Xia, 2018c). Crosslinking was detected between RF2 and the +4 site in *E. coli* (Poole et al., 1997, 1998; W. Tate et al., 1990)and translation termination efficiency changes when different nucleotides were placed at the +4 site in *E. coli* (Poole et al., 1995). In particular, the nucleotide usage bias at the +4 site is stronger in highly expressed genes than in low-expressed genes (Cridge et al., 2006; Wei & Xia, 2017). The best-documented case involves the translation of *prfB* mRNA (encoding RF2) in *E. coli* in which an inframe UGA stop codon is

followed by nucleotide C (Craigen et al., 1985; Craigen & Caskey, 1986; Curran & Yarus, 1988). When RF2 is abundant, the inframe UGA is decoded correctly to terminate translation, generating a short non-functional peptide. When RF2 is rare, the inframe UGA is not decoded. A +1 frameshift leads to the translation of GAC at a different coding frame, generating a functional RF2.

Almost all studies reviewed above focus on the optimization of mRNA but not on the translation machinery, which consists mainly of ribosomes, translation initiation factors, tRNAs, tRNA-charging enzymes, and release factors, as well as the energy that drives the translation machinery (Weaver, 2012; Xia, 2023b). Given that ribosomes represent parallel translation machines, there are two lines of empirical evidence suggesting that short-generation bacteria (SGB) should produce more ribosomal proteins, rRNA (Klappenbach et al., 2000; Sharp et al., 2005), and tRNA (Rocha, 2004) than LGB. First, the bacterial growth rate increases with the number of ribosomes in the cell (Kjeldgaard & Kurland, 1963; Moller et al., 1995; Poulsen et al., 1993; Schaechter et al., 1995). Second, the number of ribosomes increases with rRNA abundance (Yang et al., 2008). More ribosomes imply a greater need for tRNAs (Ran et al., 2014). There should be more rRNA operons and tRNA genes in SGB than in LGB.

2.3.2. Two Approaches in Studying the Effect of Selection on Translation Optimization

The first approach to identifying the effect of selection on translation optimization is by within-species comparisons, especially by contrasting highly expressed and low-expressed genes (HEGs and LEGs) in rapidly replicating bacteria (Gouy & Gautier, 1982; Prabhakaran et al., 2015; Wei et al., 2016; Xia, 2017, 2021, 2023b). Because AUG and UAA represent stronger start

and stop signals, respectively, than their alternatives, one would predict that HEGs should exhibit a stronger preference for AUG and UAA as start and stop codons than LEGs (Prabhakaran et al., 2015; Wei et al., 2016; Xia, 2021, 2023b). Because a strong secondary structure embeds translation signals such as a start codon, SD sequence, and stop codon, one would predict that HEGs should have a weaker secondary structure in sequence flanking these translation signals than LEGs (Xia, 2017). Because early genes and late genes are translated with dramatically different tRNA pools, the two groups of genes should exhibit different codon optimization degrees (Van Weringh et al., 2011). The studies reviewed above constitute empirical tests of such predictions involving different groups of genes within individual species.

This paper focuses on differential selection for translation optimization among species. Relatively few studies are available, partly because of the difficulty in identifying differential selection. Occasionally, codon usage itself was taken as a proxy for selection (Sharp et al., 2005), which is not a satisfactory approach. However, in at least four cases, such differential selection was identified with little controversy. The first case involves mitochondrial tRNA and protein-coding genes. Mitochondrial genomes of most multicellular eukaryotes encode only a single $tRNA^{Met/CAU}$ (where Met is the amino acid methionine and CAU is the tRNA anticodon for a Watson–Crick base-pairing with the Met codon AUG) to decode both Met codons AUG and AUA (through wobble-pairing). In contrast, mitochondrial genomes in some bivalve species and tunicate species encode an additional $tRNA^{Met/UAU}$ gene, which can decode AUA codons without wobble-pairing. One can predict that AUA codons should be used more frequently in mitochondrial protein-coding genes in those bivalve and tunicate species, which is true (Xia, 2012, 2018b; Xia et al., 2007)

The second case of identifiable differential selection involves stop codons and release factors (RF1 decoding UAA and UAG, and RF2 decoding UAA and UGA) in bacteria (Wei et al., 2016). If bacterial species X has more RF2 but less RF1 than species Y, then we would predict more UGA usage in species X than in species Y. This predicted pattern has also been documented in multiple bacterial species (Wei et al., 2016).

The third case involves phages differing in the presence of a lysogenic phase. Phage genomes in a lysogenic phase generally are not under selection for translation efficiency because the phage genome replicates by hitchhiking on the host genome. In contrast, phages without a lysogenic phase will be more commonly selected for translation optimization. This leads to the prediction that the former should exhibit stronger translation optimization than the latter, which is consistent with the empirical evidence (Chithambaram et al., 2014a; Prabhakaran et al., 2015). The fourth case involves a key component of the translation machinery, i.e., a small subunit rRNA whose functions depend on stable secondary structures. Thermophiles are expected to have longer and more GC-rich stems than mesophiles, and this prediction is empirically supported (Wang et al., 2006). The difference in growth temperature between thermophiles and mesophiles also affects the proportion of SD-led genes (Hockenberry et al., 2018).

2.3.3. Hypothesized Impact of Differential Selection on Translation among Bacterial Species

Bacterial species differ dramatically in generation time. Under optimal culture conditions, the generation time is ~10 min in *Vibrio natriegens* (Eagon, 1962; Yin et al., 2020), 16–20 min in

Vibrio cholerae (Dryselius et al., 2008), 20–30 min in *Escherichia coli* (Sezonov et al., 2007), 30–70 min *Bacillus subtilis* (Sharpe et al., 1998), 103–107 min in *Haemophilus influenzae* (Artman et al., 1983), about 2 h in *Mycobacterium smegmatis* (Cook et al., 2009), 4–5 h in *M. abscessus* (Cortes et al., 2010), 20–30 h in *M. tuberculosis* (Cole, 2002; Gengenbacher & Kaufmann, 2012; Zhu & Dai, 2018), and more than 7 days in *Mycobacterium leprae* (Change et al., 1967; Zhu & Dai, 2018). Such dramatic differences in generation time have been dichotomized into r- and K-selection (Andrews & Harris, 1986; Brzeszcz et al., 2016). However, the evolutionary consequence of such differential selection has rarely been explored at the molecular level.

Translation machinery that needs to complete the task of protein synthesis in 10 min, as in the case of *V. natriegens*, should be under stronger selection for optimizing translation than that completing the task in 7 days, as in the case of *M. leprae*. The fitness of some parasitic bacteria is more related to survival against host attacks than to efficient translation. For example, *M. tuberculosis* cells build their private niches essentially isolated from the surrounding environment. This protective environment, while beneficial for the survival of the pathogen, also makes it difficult for the pathogen to gain nutrients and oxygen. If protein production is not the limiting factor in growth and reproduction, then one would not expect strong selection optimizing the translation machinery. Thus, we predict that selection for optimizing translation should leave much stronger signatures in rapidly replicating species such as *V. natriegens* than in slowly replicating species such as *M. tuberculosis* and *M. leprae.*

2.4. Materials and Methods

We include nine bacterial species with completely sequenced genomes and a well-documented generation time from the shortest to the longest among mesophiles under optimal growth conditions (Table 1). Generation time has typically been quantified only in model organisms (e.g., *E. coli* as a model species for Gram-negative bacteria and *Bacillus subtilis* as a model species for Gram-positive bacteria) and very harmful pathogens, so the sample of nine species is not representative of all bacteria. Also, the optimal growth condition from an experimenter's perspective may not be the true optimal, so the generation time in Table 1 might be overestimated. However, the rank of the relative length of generation time (Table 1) should be correct, and the subsequent analysis will mainly be based on this rank. Our general prediction is that SGB should have more efficient protein-production systems than long-generation bacteria (LGB).

Table 1. Nine species with complete genomes and well-documented generation times (GTs) under optimal growth conditions.

Species	Accession [1]	OGT [2]	GT [3]	Rank [4]	Ref. [5]
Vibrio natriegens	NZ_CP009977, NZ_CP009978.1	37 °C	10 min.	1	(Eagon, 1962; Yin et al., 2020)
Vibrio cholerae	NZ_CP043554, NZ_CP043556.1	37 °C	16–20 min.	2	(Dryselius et al., 2008)
Escherichia coli	NC_000913.3	37 °C	20–30 min.	3	(Sezonov et al., 2007)
Bacillus subtilis	NC_000964.3	37 °C	30–70 min.	4	(Sharpe et al., 1998)
Haemophilus influenzae	NZ_CP007470.1	37 °C	103–107 min.	5	(Artman et al., 1983)
Mycolicibacterium smegmatis	NZ_CP054795.1	37 °C	~2 h	6	(Cook et al., 2009)
Mycobacterioides abscessus	NZ_CP034181.1	36 °C	4–5 h	7	(Cortes et al., 2010)
Mycobacterium tuberculosis	NC_000962.3	37 °C	20–30 h	8	(Cole, 2002; Gengenbacher & Kaufmann, 2012; Zhu & Dai, 2018)
Mycobacterium leprae	NZ_CP029543.1	30 °C	7 days	9	(Change et al., 1967; Zhu & Dai, 2018)

[1] GenBank accession number. [2] Optimal growth temperature. [3] Generation time under optimal growth conditions. [4] Ranking of generation time from smallest to largest. [5] References pertaining to the generation time.

The optimal growth temperature for the first eight species in Table 1 is about 37°C (or slightly lower). This is true not only for human pathogens and commensals but also for the free-living *B. subtilis* and *M. smegmatis*. The optimal growth temperature is about 30°C for *M. leprae* when cultured with mouse foot pads (Shepard, 1965) (as it has never been grown in vitro). *M. leprae* exhibits reduced growth above 33°C presumably because of the lack of a heat shock response (Cook et al., 2009).

The RefSeq genomic sequences for the nine species were downloaded from GenBank by using the accession numbers in Table 1. The software DAMBE 7.3.0 (Xia, 2018e) was used to extract coding sequences (CDSs), rRNA, and tRNA genes from the GenBank files. We classified the bacterial host genes into two expression groups: (1) known highly expressed protein-coding genes (HEGs) including small and large ribosomal protein genes, RNA polymerases, and some other genes known to be highly expressed (Supplemental Table S1), and (2) the rest of protein-coding genes not in the HEG group (REST). Pseudogenes were excluded from all analyses.

The RNA secondary structure could embed important translation signals such as the start codon, SD sequence, and stop codon and prevent them from being decoded by the translation machinery. We measured secondary structural stability based on the minimum folding energy (MFE) implemented in DAMBE (Xia, 2018e), which uses the Vienna RNA fold library (Hofacker, 2003) for secondary structural characterization. A sliding window of 40 nt along the CDSs was used to characterize the change in MFE along the sequence. Codon adaptation was measured by the index of translation efficiency (I_{TE}), which generalizes the conventional codon

adaptation index (CAI) (Sharp & Li, 1987; Xia, 2007) to accommodate the background mutation rate (Xia, 2015).

The base-pairing between the SD sequence on mRNA and the aSD sequence on small-subunit (ssu) rRNA was hypothesized to position the start codon at the P-site to pair with the initiation tRNA (Prabhakaran et al., 2015; Xia, 2018b, 2023b). The distance between the end of the ssu rRNA and the start codon ($D_{toStart}$) is highly constrained in bacterial genes, especially highly expressed ones. A narrow distribution of $D_{toStart}$ values suggests a stronger selection in terms of the SD/aSD pairing than a wide distribution. DAMBE (Xia, 2017) implements the calculation of $D_{toStart}$.

2.5. Results

2.5.1. Differential Investment in Translation Machinery

Ribosomes represent parallel protein-production factories and their abundance in *E. coli* increases linearly with the growth rate (Gausing, 1977; Verma et al., 1999). Short-generation species may have not only more efficient factories but also more factories than LGB. Each ribosome features a set of 16S, 23S, and 5S rRNAs that are transcribed from the same operon and processed into individual rRNAs (Deutscher, 2015). *E. coli* has seven rRNA operons (*rrnA* to *rrnE*, *rrnG*, *rrnH*) with promotors that are almost identical to the −10 and −35 consensus (Keener et al., 1996; Murray et al., 2003), suggesting a high demand for rRNA molecules met by both efficient and parallel transcription of multiple rRNA operons. The production of ribosomes in *E. coli* is limited by rRNA production (Nomura et al., 1984), which explains why *E. coli* maintains multiple *rrn* operons in its genome for parallel transcription. A generalization of this

would lead to the prediction that SGB should have more *rrn* operons than LGB. This prediction should also apply to tRNA genes because, with more ribosomes, more tRNA molecules are needed (Du et al., 2017).

The two predictions are both supported by empirical evidence (Table 2), with the number of rRNA and tRNA genes decreasing highly significantly with an increasing generation time ($p < 0.0001$ for any rank-based nonparametric tests). The observation that SGB maintain more rRNA and tRNA genes in their genomes than LGB is consistent with the interpretation of stronger selection for translation efficiency in SGB than in LGB.

Table 2. Short generation times are associated with increased genomic investment in ribosomal RNA and tRNA.

Species	GT [(1)]	L_{Genome} [(2)]	N_{CDS} [(3)]	N_{rrn} [(4)]	N_{tRNA} [(5)]
Vibrio natriegens	1	5,175,153	4496	11	129
Vibrio cholerae	2	4,089,299	3628	10	102
Escherichia coli	3	4,641,652	4298	7	86
Bacillus subtilis	4	4,215,606	4237	10	86
Haemophilus influenzae	5	1,846,259	1714	6	59
Mycolicibacterium smegmatis	6	6,993,871	6540	3	47
Mycobacterioides abscessus	7	5,067,231	4938	1	47
Mycobacterium tuberculosis	8	4,411,532	3905	1	45
Mycobacterium leprae	9	3,187,112	2328	1	45

[(1)] Ranking of generation times from the shorted to the longest. [(2)] Genome length. The two Vibrio species each have two chromosomes. The genome length is the sum of the two chromosomes. [(3)] Number of protein-coding genes. [(4)] Number of *rrn* operons in the genome. [(5)] Number of tRNA genes in the genome.

Note that an organism needs at least one *rrn* operon for translation. Ribosomal RNAs form the core of ribosomes with all important sites such as A, P, and E sites, with ribosomal proteins padding the surface of a ribosome (Noller, 2012). Thus, the number of *rrn* operons

cannot be less than 1. Also, there should be a minimum set of tRNA genes to decode all 61 sense codons. One might use the following two equations to model the numbers of rrn operons and tRNA genes, respectively:

$$N_{rrn} = 1 + ae^{-bx} \quad (1)$$

$$N_{tRNA} = c + ae^{-bx} \quad (2)$$

where x is RankGT in Table 2. Equation (1) ensures a minimum N_{rrn} of 1. Equation (2) ensures a minimum N_{tRNA} of c, which is estimated by the least-squares approach to be 30 (Figure 1B). The minimum number of tRNA genes required for decoding all sense codons in a natural translation system is observed in a vertebrate mitochondrial genome that encodes 22 tRNA genes.

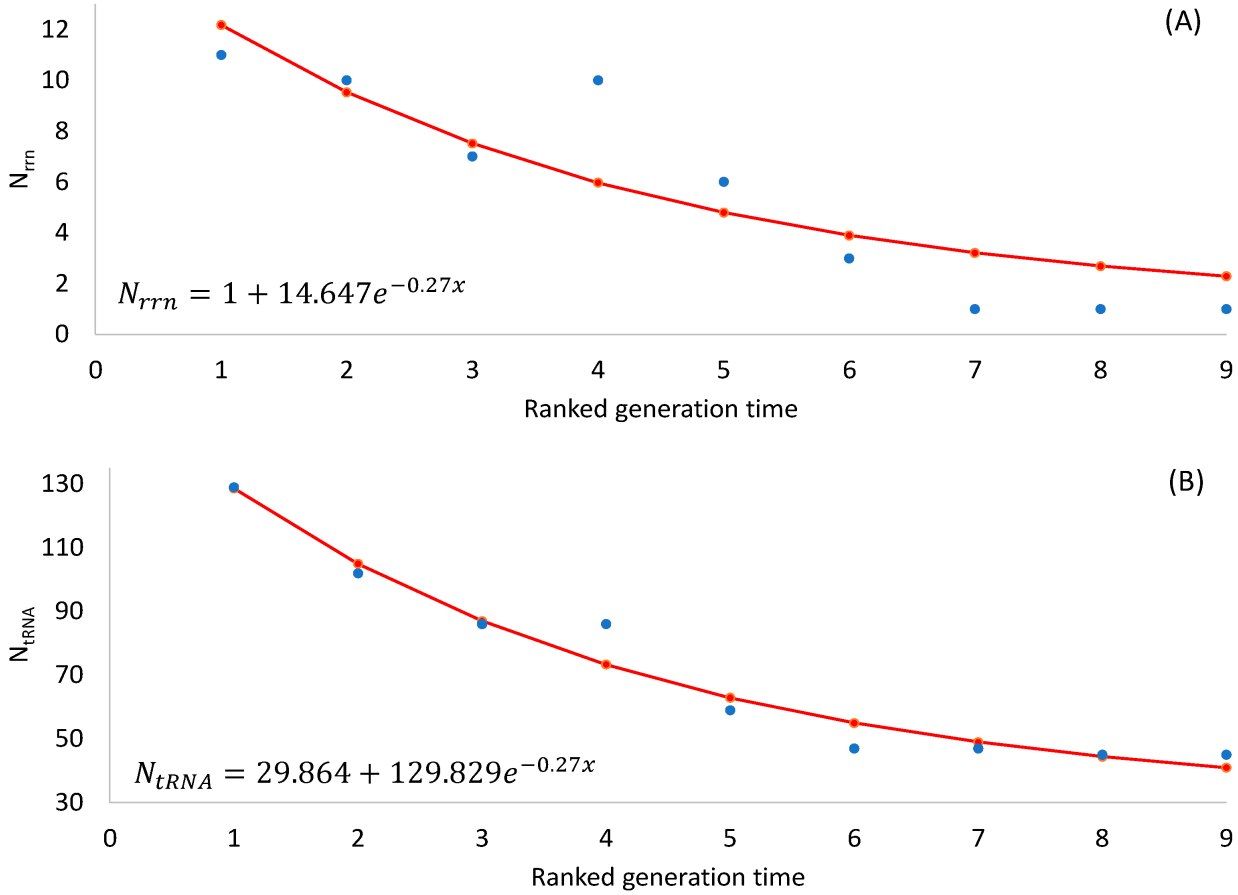

Figure 1. Fitted nonlinear equations to the observed data (blue dots). (**A**) Relationship between N_{rrn} and RankGT in Table 2. (**B**) Relationship between N_{tRNA} and RankGT in Table 2.

2.5.2. Differential Preference for Start Codon AUG

Among the three canonical start codons (AUG, GUG, and UUG), the translation initiation efficiency is consistently in the order of AUG > GUG > UUG (Hartz et al., 1991; Hecht et al., 2017). HEGs in bacteria and bacteriophages tend to prefer AUG as a start codon (Belinky et al., 2017; Gualerzi & Pon, 2015; Hartz et al., 1991; Hecht et al., 2017; Prabhakaran et al., 2015; Xia, 2023b). This is true for all nine bacterial species in which the percentage of AUG (AUG%, Table 3) is consistently higher in HEGs than those in the REST (Table 3), which includes all genes not in the HEG group and consequently contains both highly expressed and low-expressed genes. This suggests that AUG is the most efficient start codon, as is consistent with previous empirical studies based on within-species comparisons (Prabhakaran et al., 2015; Xia, 2021, 2023b).

Table 3. Start codon usage and percentage of AUG (AUG%) in highly expressed protein-coding genes (HEGs) and the rest of the protein-coding genes (REST).

Species	HEGs					REST					
	AUG	**GUG**	**YUG** [(1)]	**N_{HEG}** [(2)]	**AUG%**	**AUG**	**GUG**	**YUG** [(1)]	**AUH** [(1)]	**N_{REST}** [(2)]	**AUG%**
V. natriegens	71	3	2	76	93.42	3976	297	124	15	4412	90.12
V. cholerae	70	3	3	76	92.11	3166	253	131	18	3568	88.73
E. coli	69	3	1	73	94.52	3805	335	81	4	4225	90.06
B. subtilis	58	5	3	66	87.88	3225	382	555	9	4171	77.32
H. influenzae	68	1	1	70	97.14	1554	51	28	10	1643	94.58
M. smegmatis	54	18	0	72	75.00	4255	1998	187	26	6466	65.81
M. abscessus	53	14	1	68	77.94	3255	1458	145	14	4872	66.81
M. tuberculosis	49	13	0	62	79.03	2357	1306	177	4	3844	61.32
M. leprae	48	16	3	67	71.64	1162	782	276	43	2263	51.35

[(1)] YUG includes CUG and UUG; AUH includes AUA, AUC, and AUU. [(2)] N_{HEG}: the number of highly expressed genes; H_{REST}: the number of all other genes not in the HEG group.

Given that AUG is the most efficient start codon, one would predict that SGB should use more AUG codons as start codons than LGB. The AUG% in Table 3 is indeed strongly associated with the ranked generation time in Table 1, as predicted. However, the AUG% is also

affected by the genomic GC% because GC-rich genomes tend to have more GC-rich codons (Xia, 2018a). This is best illustrated by focusing on start codons AUG and GUG. AT-rich genomes tend to have AT-biased mutations, which favor AUG over GUG. Among the nine species, the *H. influenzae* genome is the most AT-rich and is expected to have a high AUG% because both mutation and selection favor AUG over GUG. In contrast, GC-rich genomes have GC-biased mutations, which will favor GUG over AUG. Because the four LGB all have a higher GC% than the five species with relatively short generations, they may use fewer AUG and more GUG codons as start codons simply because of their genomic GC-richness. For this reason, it is necessary to include genomic GC as a control variable. Also, HEGs and REST genes may differ in the relationship between AUG% and the generation time. A conceptually more comprehensive and explicit model is, therefore, needed.

Here, the dependent variable is AUG% and independent variables include ranked generation time (RankGT), genomic GC content (GC%, which is taken as a proxy for a genomic mutation shared by both HEGs and REST genes), and gene expression (GE with two categories, HEGs and REST genes, encoded as 0 and 1, respectively). The input data (Table 4) are used to fit the model.

Table 4. The percentage of AUG codons (AUG%) acting as start codons depends on the generation time (RankGT), genomic GC content (GC%), and gene expression (GE).

Species	AUG%	RankGT	GC%	GE
V. natriegens	93.4211	1	45.0	HEG
V. cholerae	92.1053	2	47.3	HEG
E. coli	94.5205	3	50.8	HEG
B. subtilis	87.8788	4	43.5	HEG
H. influenzae	97.1429	5	38.2	HEG
M. smegmatis	75.0000	6	67.4	HEG
M. abscessus	77.9412	7	64.1	HEG
M. tuberculosis	79.0323	8	65.6	HEG
M. leprae	71.6418	9	57.8	HEG
V. natriegens	90.1179	1	45.0	REST
V. cholerae	88.7332	2	47.3	REST
E. coli	90.0592	3	50.8	REST
B. subtilis	77.3196	4	43.5	REST
H. influenzae	94.5831	5	38.2	REST
M. smegmatis	65.8058	6	67.4	REST
M. abscessus	66.8103	7	64.1	REST
M. tuberculosis	61.3163	8	65.6	REST
M. leprae	51.3478	9	57.8	REST

A regression analysis of the input data in Table 4 showed that all three independent variables were statistically significant, but their interaction terms were not (Table 5). The regression model was then fitted without interaction terms. This reduced model accounted for 82.0% of the variation in the dependent variable AUG%. The two regression equations derived from the regression coefficients in Table 5, one for HEGs and the other for the REST, are

$$\text{For HEGs: AUG\%} = 125.86 - 2.54\text{RankGT} - 0.52\text{GC\%} \qquad (3)$$

$$\text{For REST: AUG\%} = 116.68 - 2.54\text{RankGT} - 0.52\text{GC\%} \qquad (4)$$

Table 5. Regression analysis of data in Table 4, with AUG% as the dependent variable, and ranked generation time (RankGT), genomic GC content (GC%), and gene expression (with HEG and REST encoded as 0 and 1, respectively) as the independent variables.

	Coefficient	Standard Error	t Stat	*p*-Value
Intercept	125.85528	8.60272	14.62970	0.00000
RankGT	−2.53863	0.76412	−3.32229	0.00503
GC%	−0.52069	0.19491	−2.67147	0.01825
GE	−9.17674	2.94435	−3.11673	0.00758

Equations (3) and (4) and Table 5 show that (1) the AUG% decreases highly significantly with the generation time, and (2) the AUG% is significantly higher in HEGs than in REST genes (Table 5). However, the GC% also has a significant effect on the AUG% ($p = 0.01825$, Table 5), with the AUG% decreasing with an increasing GC%. This is consistent with the interpretation of stronger selection operating on SGB than on LGB. That is, a non-AUG start codon mutating to AUG is more strongly favored by natural selection in SGB than in LGB.

We should emphasize that Equations (3) and (4) are descriptive models. They do not explicitly prevent AUG% from taking values smaller than 0 or larger than 1. A sigmoidal function would have been more appropriate if there were enough data for parameter estimation.

2.5.3. Differential Preference for Stop Codon UAA

As reviewed previously, the stop codon UAA exhibits the smallest readthrough error rate among the three nonsense codons (César Sánchez et al., 1998; Geller & Rich, 1980; Jorgensen et al., 1993; Meng et al., 1995; Parker, 1989; W. P. Tate et al., 1999). Consequently, HEGs favor stop codon UAA over other stop codons in multiple bacterial species (Wei et al., 2016; Xia, 2018b, 2023b). This is also true for all nine bacterial species. UAA was preferred by HEGs in *E.*

coli, *B. subtilis*, *M. tuberculosis* (Wei et al., 2016), and *M. abscessus* (Xia, 2023b). Of the five remaining species, the proportions of UAA in REST genes and HEGs were 0.653 and 0.8816, respectively, in *V. natriegens*; 0.6365 and 0.9474 in *V. cholerae*; 0.7602 and 0.9429 in *H. influenzae*; 0.0585 and 0.2222 in *M. smegmatis*; and 0.2364 and 0.2836 in *M. leprae*. Thus, UAA is consistently preferred in HEGs relative to REST genes.

Given that UAA is the best stop codon (Adamski et al., 1994; Crawford et al., 1999), we predicted that SGB should use more UAA codons as stop codons than LGB. Similar to our analysis of the start codon AUG usage, the dependent variable now was UAA% and the independent variables included the ranked generation time (RankGT), genomic GC content (GC%), and gene expression (GE with two categories, HEGs and REST genes, encoded as 0 and 1, respectively). The genomic GC% is particularly relevant in studying UAA usage because protein-coding genes in a GC-rich genome tend to use more UGA and UAG stop codons than those in an AT-rich genome (Wei et al., 2016).

A regression analysis showed that all three independent variables were statistically significant (Table 6). The model accounted for 94.7% of the total variation in UAA%. The two regression equations derived from the regression coefficients in Table 6, one for HEGs and the other for the REST, are

$$\text{For HEGs: UAA\%} = 197.95 - 3.78\text{RankGT} - 2.22\text{GC\%} \qquad (5)$$

$$\text{For REST: UAA\%} = 180.60 - 3.78\text{RankGT} - 2.22\text{GC\%} \qquad (6)$$

Table 6. Regression analysis of the impact of UAA% on three independent variables: ranked generation time (RankGT), genomic GC content (GC%), and gene expression (with HEG and REST encoded as 0 and 1, respectively).

	Coefficient	**Standard Error**	**t Stat**	***p*-Value**
Intercept	197.94875	11.48080	17.24172	0.00000
RankGT	−3.77935	1.01976	−3.70611	0.00235
GC%	−2.21876	0.26012	−8.52992	0.00000
GE	−17.35326	3.92940	−4.41627	0.00059

Equations (5) and (6) and Table 6 show that the UAA% decreases highly significantly with an increasing generation time, and that HEGs use UAA significantly more frequently than REST genes (Table 6, a 17.35326% difference between the two). This is consistent with the interpretation of stronger selection operating on SGB than on LGB. A non-UAA stop codon mutating to UAA is more strongly favored by natural selection in SGB than in LGB.

UAA usage also decreases significantly with an increasing genomic GC% (Table 6), which is understandable. As the genomic GC% increases, GC-biased mutations will favor UAG and UGA codons over UAA codons (Wei et al., 2016). It is for this reason that the genomic GC content needs to be taken into consideration when assessing codon usage bias.

2.5.4. Differential Selection on Sense Codons

It is difficult to evaluate the impact of tRNA-mediated selection on codon usage across species because the codon adaptation index (CAI) (Sharp & Li, 1987; Xia, 2007) and the index of translation efficiency (I_{TE}) (Xia, 2015) are both for comparing genes within species. The effective number of codons (ENC) (Sun et al., 2013; Wright, 1990) could potentially be used for among-species comparisons, but all these indices are strongly affected by genomic mutation bias (Palidwor et al., 2010; Xia, 2018d). For example, protein-coding genes in strongly GC-biased

genomes will have mostly G-ending and C-ending codons, leading to a reduced ENC that has little to do with selection.

It is reasonable to assume that mutation bias affects both HEGs and REST genes. Thus, the difference in mean I_{TE} between HEGs and REST genes, i.e.,

$$D_{I_{TE}} = (\bar{I}_{TE.HEG} - \bar{I}_{TE.REST}), \qquad (7)$$

should be relatively independent of mutation bias. $D_{I_{TE}}$ should increase with selection intensity for translation elongation efficiency. If there is no selection for translation elongation efficiency, then $D_{I_{TE}}$ is expected to be 0. With an increasingly strong selection for translation elongation efficiency, $D_{I_{TE}}$ should also increase because the selection is expected to be stronger for HEGs than for REST genes. In other words, strong selection for translation efficiency in SGB should drive HEGs towards better codon adaptation than the REST genes, thereby increasing $D_{I_{TE}}$. The $D_{I_{TE}}$ values and its ranks (Rank $D_{I_{TE}}$, Table 7) depend strongly on generation time (RankGT, Table 7). The relationship is best illustrated with two ranked variables, i.e., ranked generation time (RandGT) and ranked $D_{I_{TE}}$ (Figure 2). The fitted regression line accounts for 87.84% of the total variation in ranked $D_{I_{TE}}$ (Figure 2).

Table 7. Differences in mean I_{TE} values between HEGs and REST genes ($D_{I_{TE}}$) for the nine bacterial species at different generation times (RankGT).

Species	RankGT	$D_{I_{TE}}$	Rank $D_{I_{TE}}$
V. natriegens	1	0.2516	9
V. cholerae	2	0.2380	6.5
E. coli	3	0.2465	8
B. subtilis	4	0.2380	6.5
H. influenzae	5	0.2092	5
M. smegmatis	6	0.1266	3
M. abscessus	7	0.1806	4
M. tuberculosis	8	0.0267	1
M. leprae	9	0.0689	2

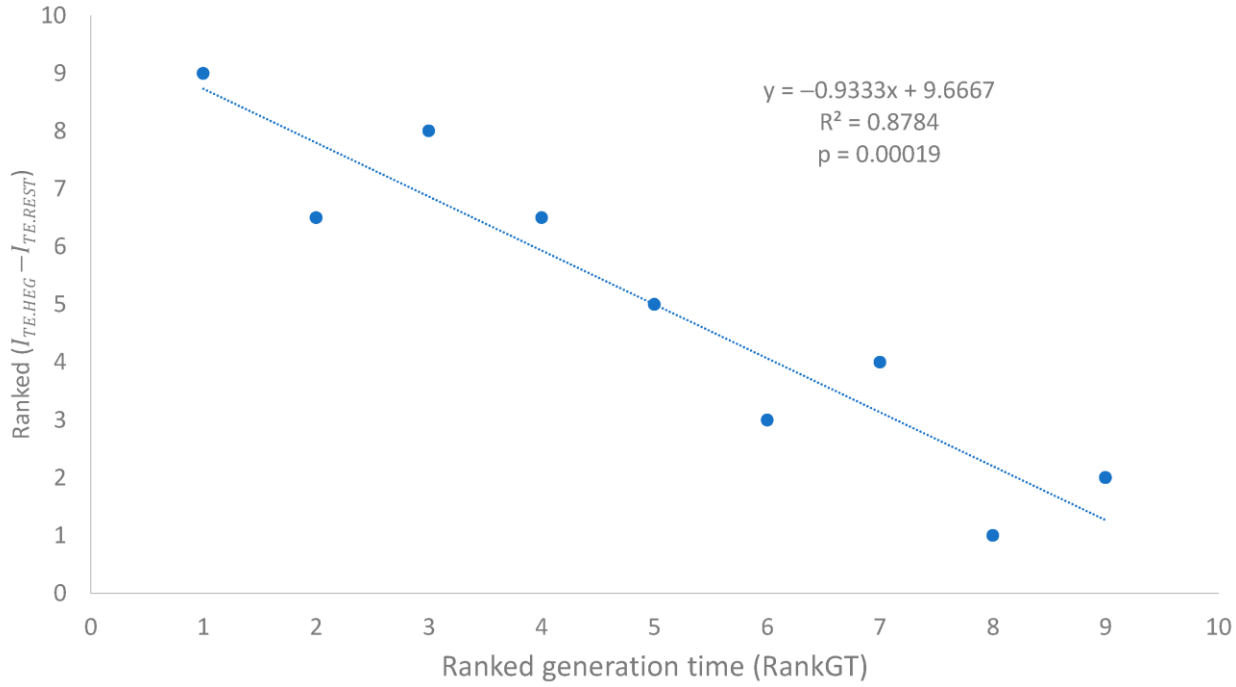

Figure 2. Selection for codon optimization, measured by ranked $D_{I_{TE}}$ ($= \bar{I}_{TE.HEG} - \bar{I}_{TE.REST}$) in Equation (7), decreases with increasing generation time in nine bacterial species.

2.5.5. Differential Selection Drives tRNA Adaptation

Codons and tRNAs are expected to coevolve and adapt to each other (Bulmer, 1991; Ikemura, 1981; Xia, 1998, 2018a), especially in highly expressed genes (Carullo & Xia, 2008; Coghlan & Wolfe, 2000; Comeron & Aguade, 1998; Duret & Mouchiroud, 1999; Gouy & Gautier, 1982; Ikemura, 1981, 1982; Xia, 1998, 2015, 2018a). Given the better codon

optimization in SGB than in LGB (Figure 2 and Table 7), one would predict more tRNA genes for highly used codons than rarely used codons. If we focus on the anticodons of tRNA genes, then the prediction above implies a smaller effective number of anticodons (N_{AC}), equivalent to the concept of the effective number of codons (Sun et al., 2013; Wright, 1990), in SGB than in LGB. Specifically, N_{AC} should increase with RankGT (ranked generation time).

We calculated N_{AC} in the same way the effective number of codons is calculated (Sun et al., 2013), except that codons in coding sequences were replaced by anticodons in tRNA genes. The ranked N_{AC} increased highly significantly with ranked RankGT (Table 8 and Figure 3A), consistent with our prediction that SGB should have a smaller N_{AC} than LGB. One should note the difference between a codon replacement and an anticodon replacement. A codon replacement may have only a minor effect on the translation of a single gene, but an anticodon replacement will affect the translation of numerous codons. For this reason, anticodons are strongly constrained and much less affected by genomic GC%.

Table 8. The effective number of anticodons (N_{AC} and its ranked N_{AC}) increases with ranked generation time (RankGT).

Species	RankGT	N_{AC}	Rank N_{AC}
V. natriegens	1	40.2129	1
V. cholerae	2	44.2802	2
E. coli	3	49.0127	4
B. subtilis	4	45.8531	3
H. influenzae	5	50.4305	5
M. smegmatis	6	59.0667	7
M. abscessus	7	58.8153	6
M. tuberculosis	8	59.3744	8.5
M. leprae	9	59.3744	8.5

We have previously used Rank $D_{I_{TE}}$ (Table 7) as a species-level measure of codon adaptation. This $D_{I_{TE}}$ is expected to be negatively associated with the ranked N_{AC} for the following reason. Better codon adaptation (high Rank $D_{I_{TE}}$) implies higher usage of major

codons, which requires more tRNAs with the corresponding decoding anticodon to translate these overused major codons, leading to a decreased N_{AC} in species with a high degree of codon adaptation. This expected relationship between N_{AC} and $D_{I_{TE}}$ was substantiated empirically (Figure 3B, p = 0.00015).

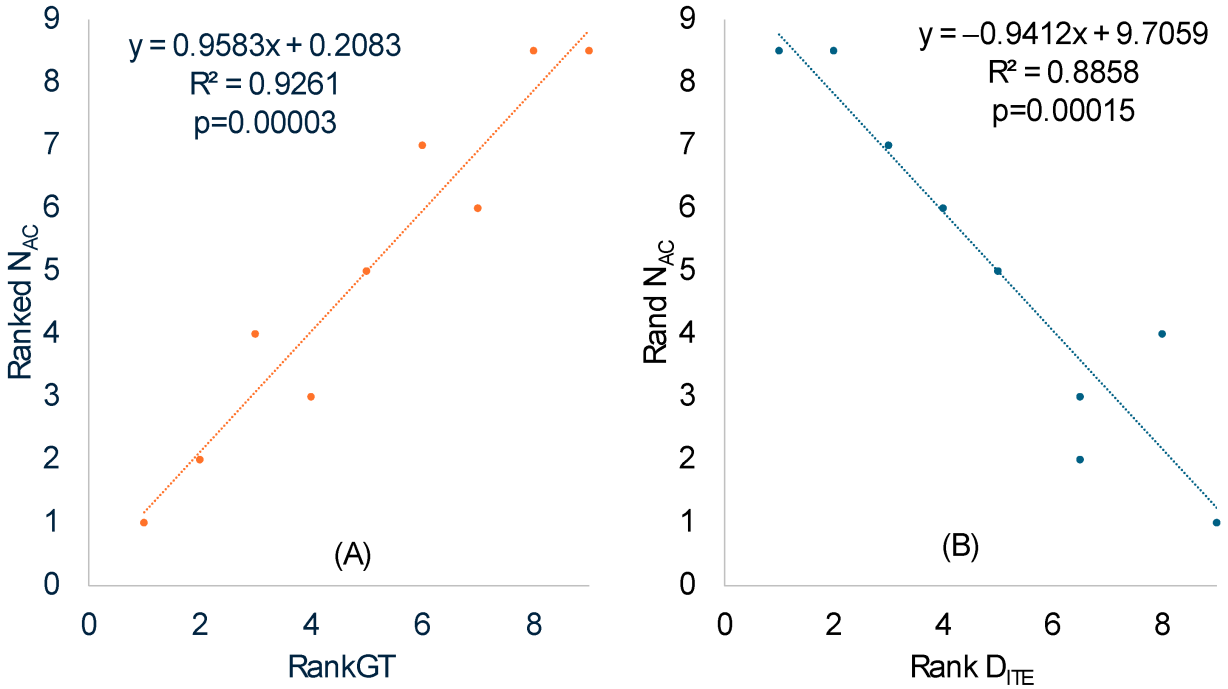

Figure 3. The ranked effective number of anticodons (Ranked N_{AC}) increases with ranked generation time (RankGT) in (**A**), and decreases with increasing codon adaptation (Rank $D_{I_{TE}}$) in (**B**).

The analysis above assumes that tRNA gene copy numbers in bacterial genomes are proportional to the abundances of tRNA molecules in the cell. With the availability of transcriptomic data, it has been found that the assumption is generally true, i.e., the copy number of a tRNA is highly correlated with the transcriptomic representation of the tRNA (Wei et al., 2019).

2.5.6. Secondary Structural Stability near the Start and Stop Codons

Because the 30S ribosomal subunit requires a single-stranded mRNA region for binding (Calogero et al., 1988; Studer & Joseph, 2006), sequences immediately flanking translation signals in bacteria (e.g., Shine–Dalgarno sequence and start and stop codons) are expected to have reduced secondary structures, to avoid embedding translation signals in a stable secondary structure (De Smit & Duin, 1990; Gualerzi & Pon, 2015; Hartz et al., 1991; Nakamoto, 2006; Scharff et al., 2011; Xia, 2018b). This pattern has also been observed in bacteriophage genes (Prabhakaran et al., 2015; Xia, 2023b). The weakening of the secondary structure near, or immediately upstream of, the start codon has also been observed in highly expressed eukaryotic mRNAs (Xia, 2021; Xia et al., 2011), especially in mRNAs requiring internal ribosome entry for translation(Xia & Holcik, 2009).

Secondary structural stability in RNA is typically measured by the minimum folding energy (MFE). An MFE equal to 0 means no secondary structure, and a stronger secondary structure corresponds to a more negative MFE value. Experiments involving engineered *E. coli* genes have shown that the translation initiation efficiency depends heavily on the MFE of the sequence upstream and including the start codon (Kudla et al., 2009; Tuller et al., 2010; Xia, 2015). We followed the convention of previous studies (Prabhakaran et al., 2015; Xia, 2023b) and measured the MFE with a sliding window of 40 nt along mRNA sequences, to quantify the change in secondary structural stability.

Secondary structural stability, as measured by MFE, decreased near the start codon, but the weakest secondary structure was observed slightly upstream of the start codon (Figure 4),

corresponding to the SD sequence. This pattern has been observed before in bacteriophage genes and their host genes (Prabhakaran et al., 2015; Xia, 2023b) and is consistent with the interpretation that a strong secondary structure embedding the SD sequence or the start codon is selected against because it prevents the translation initiation signal (SD and start codon) from being decoded by the aSD sequence and the initiation tRNA, respectively.

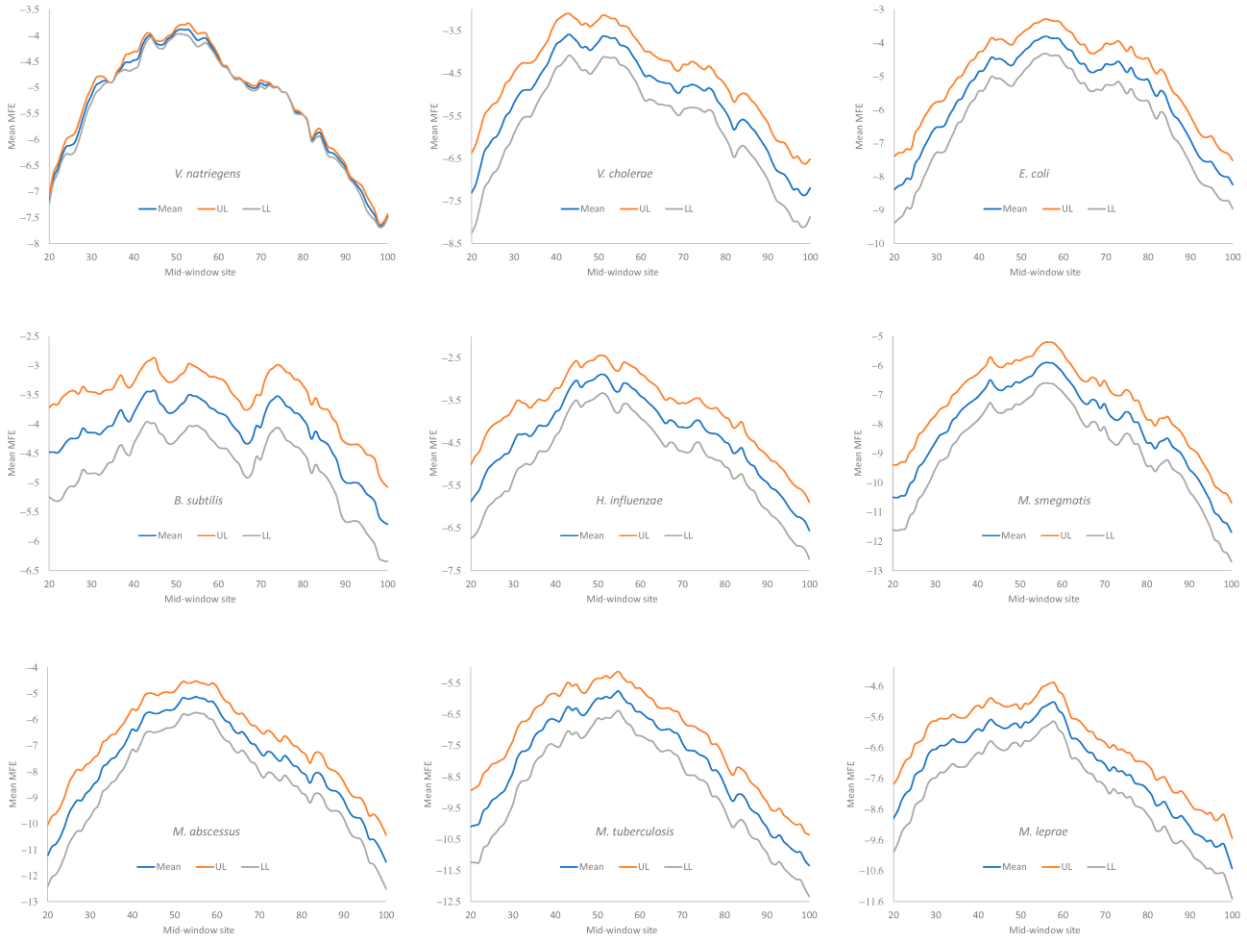

Figure 4. Change in MFE (minimum folding energy) over a sliding window of 40 nt in HEGs of the nine bacterial species (one sub-figure for each species). The start codon occupies sites 61–63. The mid-window site (horizontal axis) indicates the middle of the sliding window of 40 nt. The middle blue curve is the mean MFE of all HEGs (e.g., each point in the mean curve for *V. natriegens* is the average of 76 HEGs). The two curves above and below the mean curve are the 95% upper and lower limits (UL and LL).

One might argue against the interpretation that the reduced secondary structure near the translation initiation serves to avoid embedding crucial translation initiation signals such as SD sequences and start codons in a stable secondary structure. The SD sequences are purine-rich and cannot form a secondary structure within them. Thus, the reduced secondary structure near the SD sequence (Figure 4) could be a direct consequence of the purine-richness in the SD sequences, with nothing to do with the hypothesized avoidance of the secondary structure embedding important translation initiation signals. We thus have two hypotheses. Hypothesis 1 invokes selection against secondary structural stability near the translation initiation signals. Hypothesis 2 states that the weakening of the secondary structure near the translation initiation signal is a simple consequence of the purine-rich SD, with no selection specifically against the secondary structure.

One way to differentiate these two hypotheses is to consider the observation that SD sequences are mainly G-rich. Because G will base-pair with C, Hypothesis 1 (invoking selection against secondary structural stability) will predict a stronger avoidance of nucleotide C relative to nucleotide U. In contrast, Hypothesis 2 gives no reason to expect an avoidance of nucleotide C relative to U near the translation initiation signals. Figure 5 plots the position weight matrix (PWM) scores for the 60 nucleotides immediately upstream of the start codon for two species, *Bacillus subtilis* and *Vibrio cholarae.* PWM scores measure nucleotide usage bias relative to the background nucleotide frequencies. A value of 0 means unbiased usage, a positive value means overuse, and a negative value means avoidance. Figure 5A,B contrast the HEGs and the REST genes in *B. subtilis.* If Hypothesis 1 is correct, then we expect stronger avoidance of C relative to U in HEGs than in the REST genes. The nucleotide usage patterns in Figure 5A,B are consistent

with the prediction from Hypothesis 1. An approximate statistical test can be performed as follows. Within sites 45 to 55, there 36 nucleotide Cs and 129 nucleotide Us in the HEGs. The corresponding numbers are 2952 C and 7069 U. The percent of C is 21.818% in the former and 29.458% in the latter. The two are significantly different (likelihood ratio chi-square = 4.84, DF = 1, $p = 0.0278$). If we narrow the range to sites 48 to 53, then the difference becomes more significant. The nucleotide usage patterns for HEGs and REST genes in *V. cholerae* are similar (Figure 5C,D).

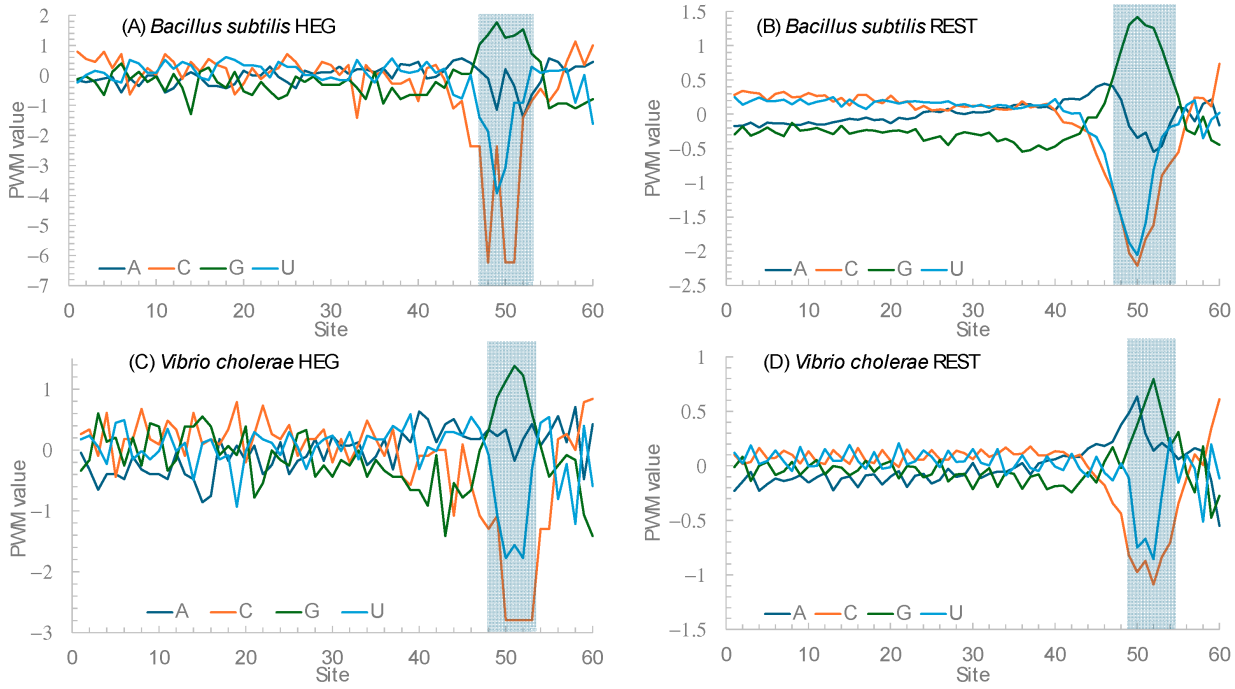

Figure 5. Dramatic reduction in nucleotide C near the Shine–Dalgarno sequence (shaded). Plotted are position weight matrix (PWM) scores in sequences immediately upstream of the start codon (at sites 61–63) in HEGs (**A**) and REST genes (**B**) in *Bacillus subtilis*, and in HEGS (**C**) and REST genes (**D**) in *Vibrio cholerae.*

All nine bacterial species exhibited a decreased secondary structure near the translation initiation signals (Figure 4). However, we did not take into account the effect of the GC%. Increasing the GC% is expected to increase secondary structural stability in mRNA. As we can see from Table 9, *H. influenzae* has the lowest genomic GC% (38.2%), and its mean MFE is the

closest to 0 (which means no secondary structure). In contrast, *M. smegmatis* and *M. tuberculosis* have the highest genomic GC%, and their mRNAs tend to have more negative MFE values (Figure 4). Accordingly, without controlling for GC%, the observation that the LGB have a stronger secondary structure than the SGB cannot be attributed to a reduced selection in these LGB against a stable secondary structure. Note that the MFE and GC% were calculated for each sequence, and their averages are presented in Table 9 as the MeanMFE and GC%, respectively.

Table 9. Secondary structural stability (measured by MeanMFE) is expected to change with generation time (RankGT), GC%, and gene expression (GE).

Species	RankGT	GC%	GE	MeanMFE
V. natriegens	1	42.8557	HEG	−4.2351
V. cholerae	2	43.7813	HEG	−4.0721
E. coli	3	46.5955	HEG	−4.2153
B. subtilis	4	38.7556	HEG	−3.7484
H. influenzae	5	37.9329	HEG	−3.2660
M. smegmatis	6	60.4224	HEG	−6.4462
M. abscessus	7	60.4048	HEG	−5.5966
M. tuberculosis	8	61.5043	HEG	−6.2934
M. leprae	9	56.5344	HEG	−5.7794
V. natriegens	1	40.4234	REST	−3.5238
V. cholerae	2	43.3125	REST	−4.0330
E. coli	3	45.6418	REST	−4.4059
B. subtilis	4	38.6014	REST	−3.7406
H. influenzae	5	33.7986	REST	−2.5466
M. smegmatis	6	63.6954	REST	−8.6708
M. abscessus	7	61.2667	REST	−8.3538
M. tuberculosis	8	63.0856	REST	−9.1141
M. leprae	9	57.1858	REST	−7.9865

Given that a weak secondary structure near the SD sequence and the start codon is favorable (Figure 4), one would predict that SGB should have weaker secondary structures (larger MFE values) than LGB. In order to test this prediction while accommodating the effect of the GC%, we characterized the MFE plots of sliding windows for each species using a single value for each. That is, for each plot in Figure 4, we calculated the mean value for mid-window

sites 46–65 (with the start codon occupying sites 58–60). These sites included both the SD sequence and the start codon. The mean MFE (Table 9) was now the dependent variable. It was expected to (1) decrease (i.e., more stable secondary structure) with an increasing generation time (RankGT) and increasing GC%, and (2) increase with gene expression (GE, i.e., be greater for HEGs than for REST genes). These three independent variables are also included in Table 9.

The best model, which accounts for 98.2% of the total variation in MeanMFE in Table 9, included the three dependent variables and an interaction term (Table 10). The two-tailed p for RankGT was 0.064 (Table 10). However, because we had an explicit one-tailed prediction of a negative slope (i.e., MeanMFE should decrease with increasing RankGT), p was half of 0.064, i.e., 0.032. Other regression terms for GC%, GE, and their interaction were highly significant and consistent with the predictions (Table 10).

Table 10. Regression output based on data in Table 9, with GE encoded as a binary dummy variation (0 for HEG and 1 for REST). "GC%* GE" is the interaction term.

	Coefficients	**Standard Error**	**t Stat**	***p*-Value**
Intercept	0.57408	0.61265	0.93704	0.36582
RankGT	−0.08382	0.04142	−2.02386	0.06404
GC%	−0.10038	0.01414	−7.09756	0.00001
GE	4.46040	0.75315	5.92236	0.00005
GC%*GE	−0.10972	0.01483	−7.39673	0.00001

As before, we give the two regression equations separately for HEGs and the REST from the regression coefficients in Table 10:

$$\text{For HEGs: MeanMFE} = 0.57408 - 0.08382\ \text{RankGT} - 0.10038\ \text{GC\%} \quad (8)$$

$$\text{For REST: MeanMFE} = 5.03448 - 0.08382\ \text{RankGT} - 0.21009\ \text{GC\%} \quad (9)$$

Equations (8) and (9) show that, for both HEGs and REST genes, secondary structural stability increases with the generation time (MFE becomes more negative with an increasing generation time). This is consistent with our prediction that selection against the secondary structure near the translation start signals (SD sequence and start codon) is stronger in SGB than LGB.

MeanMFE decreases more sharply with the GC% in Equation (9) than in Equation (8), i.e., secondary structural stability increases more rapidly with the GC% for REST genes than for HEGs. This is easy to understand if the selection against secondary structural stability is on average stronger in HEGs than in REST genes. The MeanMFE decreases by only 0.10038 (Equation (8)) for a unit increase in the GC% with the strong selection of HEGs, but decreases by 0.21009 (Equation (9)) for the same unit change in the GC% with the relatively weak selection of REST genes.

The selectionist interpretation above does not consider the effect of mutations, which offers an alternative interpretation. In general, spontaneous mutations in AT-rich genomes tend to be AT-biased, based on (1) comparisons between pseudogenes and their functional counterparts (W. H. Li et al., 1981; W.-H. Li, 1983), (2) the mutation patterns of pathogenic bacteria with relaxed selection (Hershberg & Petrov, 2010; Lind & Andersson, 2008), and (3) nucleotide bias at the three codon sites across multiple bacterial species (Muto & Osawa, 1987). *H. influenzae* has an AT-rich genome, suggesting AT-biased mutation, in contrast to *M. smegmatis,* which has a GC-rich genome. However, protein-coding genes in both species have SD sequences that are purine-rich (especially G-rich) (Figure 6). The G-rich SD will form base-pairs with nearby C

nucleotides, so the MFE will not fall to 0. In pseudogenes where selection for maintaining the G-rich SD sequence is absent, or in low-expressed genes where the selection is weak, the AT-rich *H. influenzae* will lose these G nucleotides in the SD sequence, leading to an MFE closer to 0. Indeed, the MeanMFE value for the 19 pseudogenes in *H. influenzae* is −2.2887, closer to 0 than all MeanMFE values in Table 9. Similarly, the MeanMFE is −3.2660 for HEGs and −2.5466 for REST genes (Table 9). This is consistent with the interpretation that the G-rich SD is more likely to be hit by G→A and G→T mutations and lose G/C base-pairs in REST genes than in HEGs. For example, *H. influenzae* has a GC% of 37.9329% for HEGs but only 33.7986% for REST genes, leading to a MeanMFE value closer to 0 in REST genes than in HEGs.

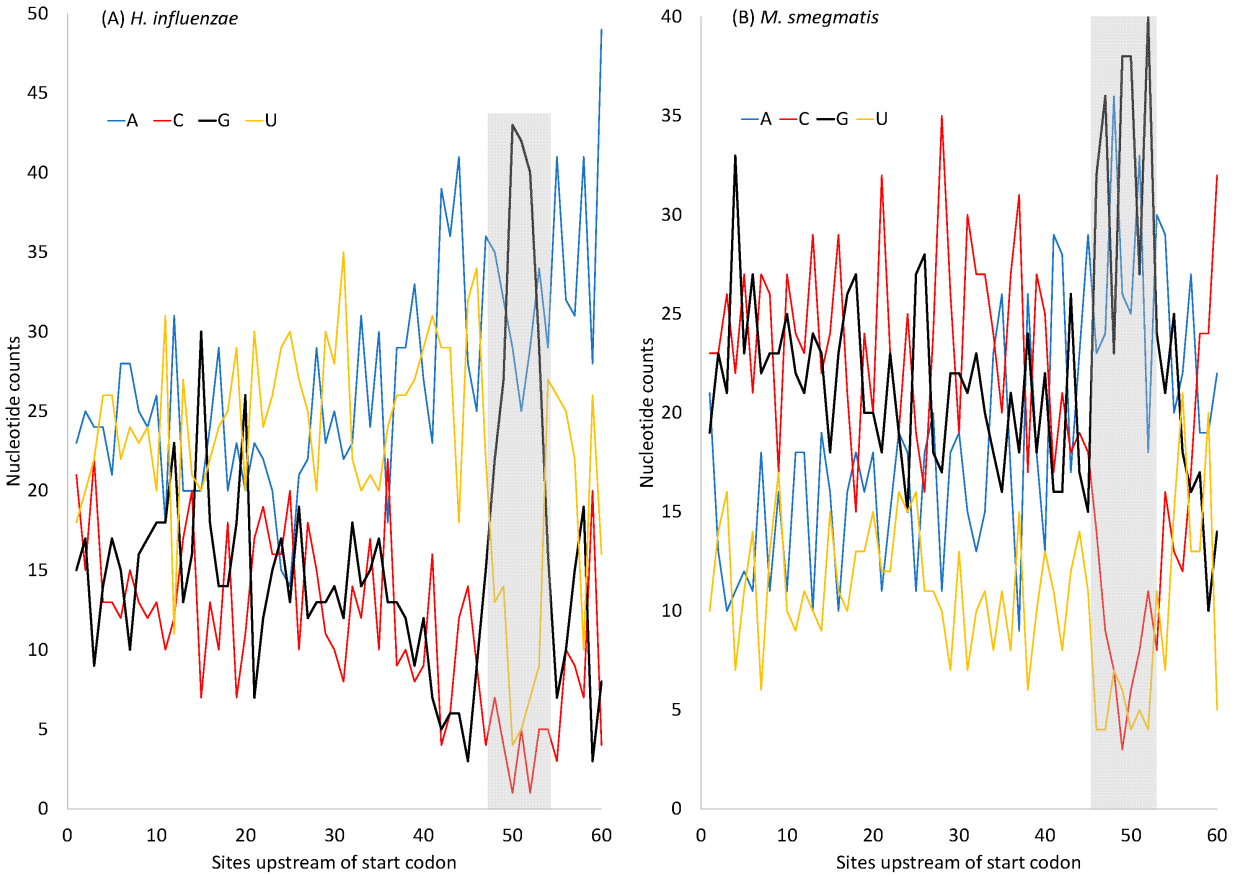

Figure 6. Changes in nucleotide frequencies in sequences immediately upstream of the start codon (at sites 61–63) in highly expressed genes (HEGs). Note the sharp increase in nucleotide

G and concurrent decrease in nucleotides C and U near site 50 corresponding to the Shine–Dalgarno (SD) sequence (shaded). (**A**) AT-rich *H. influenzae.* (**B**) GC-rich *M. smegmatis.*

In short, although the purine-rich SD sequences (Figure 6) can hardly form secondary structures within themselves, the dramatically increased G nucleotides within the SD sequence could base-pair with the neighboring C nucleotides and contribute to secondary structural stability. If there is no selection maintaining the G-richness in the SD sequences, then these G nucleotides may be replaced by A and T, leading to a further decrease in secondary structural stability. Thus, both selection and mutation could contribute to secondary structural stability in sequences near the translation initiation signals (the SD sequence and the start codon). The models in Equations (8) and (9) are, therefore, oversimplified and should be interpreted with caution. The secondary structure in sequences near the stop codon exhibits a similar pattern to those near the start codon (Figure 7).

The decrease in secondary structural stability may not necessarily be related to the avoidance of embedding SD sequences and start codons. Efficiently translated yeast mRNAs (i.e., mRNAs in polysomes with high ribosome densities) often have a short poly(A) tract before the start codon (Xia et al., 2011), with the poly(A) interpreted as long enough to recruit translation initiation factors but short enough to avoid binding by the poly(A)-binding proteins. However, the presence of poly(A) also weakens the secondary structural stability as a secondary consequence.

2.6 Discussion

Our assumption that SGB are under stronger selection for translation efficiency than LGB appears to be valid because multiple predictions based on the assumption are consistent with the empirical evidence. First, the number of ribosome RNA operons, as well as the number of tRNA genes, increases with a decreasing generation time (Table 2 and Figure 1). Second, AUG is known to be the most efficient start codon, and SGB genes exhibit a stronger preference for AUG as a start codon than LGB, especially in highly expressed genes (HEGs) (Tables 3–5). This is also true in the usage of stop codon UAA (Table 6), which is known to be the most efficient termination signal with the smallest readthrough error rate as a stop codon. Third, SGB, especially their HEGs, exhibits better codon and anticodon adaptation than LGB (Table 7 and Figure 2 for codons, and Table 8 and Figure 3). Finally, SGB genes have weaker secondary structures near translation initiation signals than LGB (Figure 4 and Tables 9 and 10). A similar pattern was observed with sequence secondary structures near the stop codon (Figure 7). However, as we discussed previously, the selectionist interpretation is sometimes confounded by biased mutations.

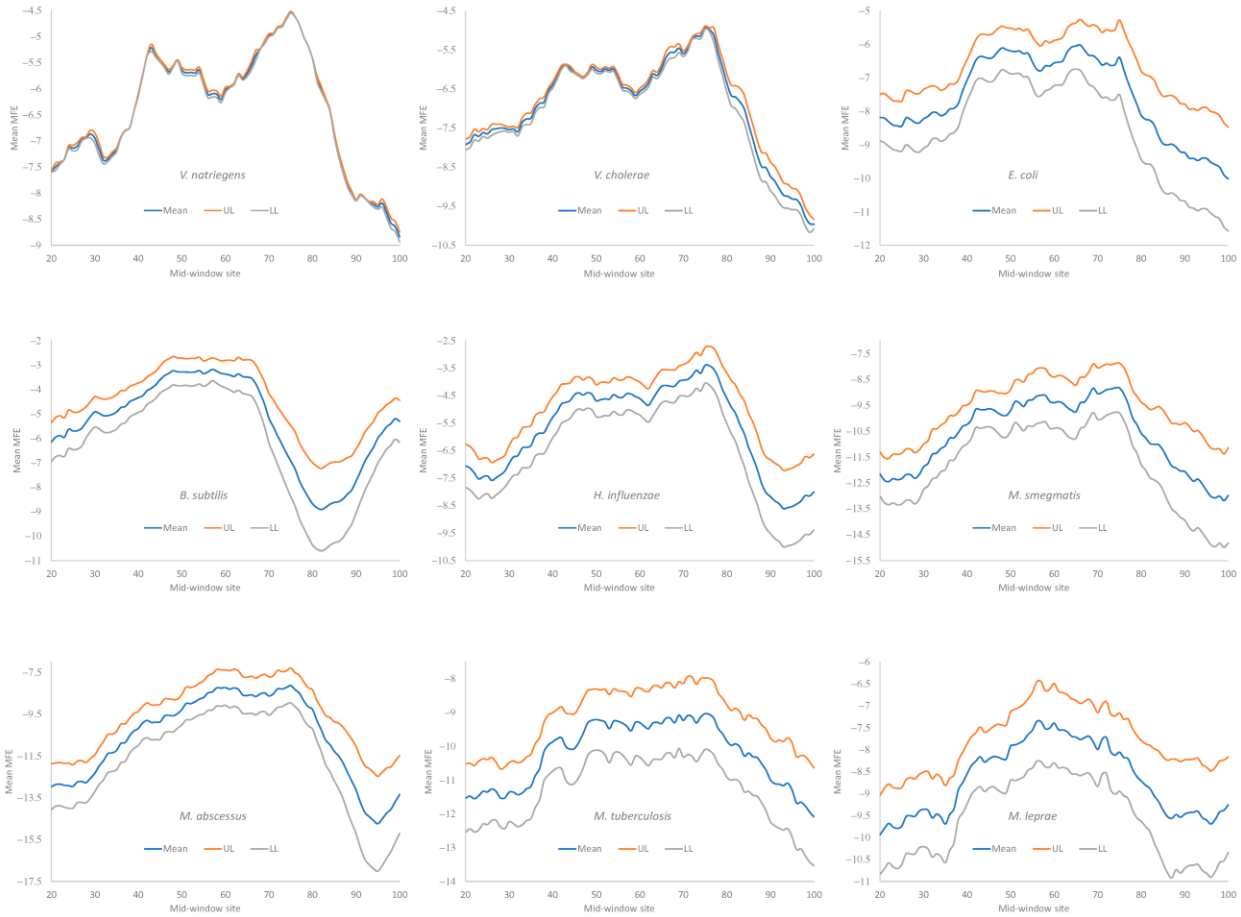

Figure 7. Change in MFE (minimum folding energy) over a sliding window of 40 nt in HEGs of the nine bacterial species. The stop codon occupies sites 58–60. Other annotations are identical to those in Figure 4.

One may ask why there should ever be LGB with weak selection for translation optimization, given natural selection operating to maximize growth and reproduction. There are different environments in which translation efficiency may not be a limiting factor for growth and reproduction. For example, *M. tuberculosis* wraps itself with a thick layer of mycolic acids that serves two functions (Lee & Engels, 2014): (1) to prevent antibiotics from reaching the cell membrane (Nataraj et al., 2015), and (2) to evade the attack of the host immune system (Zhao et al., 2015). However, after being phagocytosed by pulmonary macrophages and confined within the phagosome and the granuloma, *M. tuberculosis* survives in an extremely nutrient-limited

environment and adopts a prolonged stage of dormancy with no chance of rapid growth (Gengenbacher & Kaufmann, 2012). Mutations still occur during this latent stage of dormancy (Colangeli et al., 2014; Shan Chang & Guan, 2021) in both protein-coding genes (Comín et al., 2022) and rRNA genes (Honore et al., 1995; Kawashima et al., 2014). Thus, mutations without the checking of natural selection will lead to a suboptimal translation machinery in *M. tuberculosis.* Such a suboptimal machinery cannot perform efficient translation even when the bacteria are not nutrient-limited. This is in dramatic contrast to *E. coli,* which experiences rapid alternation of feast and famine cycles several times each day, with natural selection eliminating those mutants that cannot translate efficiently during the feast period.

Modern biological research aims to formulate and validate quantitative and mechanistic models. In this context, this study has several shortcomings. The first shortcoming is the inherent inaccuracy of generation time. The experimentally measured generation time sometimes varies widely among strains and among different studies. For example, the generation time is ~21 h in drug-sensitive strains of *M. tuberculosis*, but ~35 h in the multi-drug-resistant strains (Mukherjee et al., 2019). This suggests a cost to the pathogen in developing multi-drug resistance, i.e., the resistance is at the cost of longer generations. However, the observation also highlights the inherent variation in measured generation time. Particularly controversial is what generation time to use. For example, the generation time in *E. coli* is only about 20–30 min under favorable culture conditions but could be substantially longer in the natural habitat of the mammalian intestine, estimated by the rate of mutation accumulation (Gibson et al., 2018). To model the joint effect of mutation and selection, the generation time under natural conditions would seem more appropriate than that under optimal experimental conditions. Unfortunately, the generation

time in bacterial populations in nature cannot be measured accurately. For example, the estimated generation of 15 h for *E. coli* (Gibson et al., 2018) is associated with a 95% confidence interval of 0–30 h. Another uncertainty with generation time is that we do not know if long-generation species such as *M. leprae* really cannot replicate fast or if they could grow fast but microbiologists have not been able to shift them into the fact-growing mode. Because of the uncertainty in generation times among the bacterial species, we ranked the generation times in the hope that, relative to one another, they would be in the correct order. It is for this reason that we chose species with widely different generation times characterized experimentally, so that our ranking of the generation times would not be controversial.

The second shortcoming of this study is the small number of bacterial species, partly because of our conscious effort to avoid species with uncertain generation times. There are various tables of bacterial generation times for more than nine bacterial species, but they often do not include original references or do not have fully sequenced genomes. Some bacterial species in Mollicutes have known generation times, e.g., ~6 h in *Mycoplasma pneumoniae* (Kashyap & Sarkar, 2010; Waites & Talkington, 2004) , as well as fully sequenced genomes. However, their genetic code (i.e., translation table 4) differs from the rest of the bacteria (translation table 11). This complicates comparisons. Consequently, they are not included in this study. The small number of bacterial species included here results in two limitations. First, it does not permit the validation of parameter-rich models. Second, it does not allow for phylogeny-based inference (Felsenstein, 1985; Harvey & Pagel, 1991; Xia, 2013) to alleviate the issue of data dependence. We have previously used such phylogeny-based inference to quantify the relationships between body temperature and genome size (Xia, 1995) and between the optimal growth temperature and

ribosomal RNAs' secondary structural stability (i.e., the stem length and GC% in the stem–loop structure in rRNAs) (Wang et al., 2006). The phylogenetically independent contrasts showed that the stem–loop structure in bacterial thermophiles tended to have longer and more GC-rich stems than that in mesophiles (Wang et al., 2006) and that poikilotherms in warm climates tended to have smaller genomes than those in cold climates (Xia, 1995). In this context, we may highlight two points. First, our results are highly consistent with the prediction that increasing the generation time decreases the intensity of selection on translation efficiency. Second, the bacterial species we use are highly divergent. Even the two most closely related species, *M. tuberculosis* and *M. leprae*, have an evolutionary distance of more than 0.3 even for conserved ribosomal protein genes (Xia, 2023a), so phylogenetic inertia might not affect the quantified relationships, i.e., the data points might be considered to be roughly independent.

The third shortcoming is the lack of a general conceptual framework for the impact of mutation and selection on translation optimization, one that can be used across species. Protein production from an mRNA depends on the ribosomal recruitment rate, the efficiency in forming the 70S initiation complex, the elongation efficiency and accuracy, the termination efficiency and accuracy, the stability of the mRNA, the differential amino acid and tRNA availability, and the energy level of the cell (Liljenström & von Heijne, 1987; Maitra & Dill, 2015; Rodnina, 2018; Xia, 1996). How do these variables interact with each other to affect protein production? For example, when translation is not efficient, codon usage optimization has little effect on protein production. However, protein production increases significantly with codon optimization in mRNAs with efficient initiation (Kudla et al., 2009; Tuller et al., 2010; Xia, 2015). Similarly, there are conflicts between maximizing transcription efficiency and translation efficiency. An

RNA will be transcribed efficiently if it maximizes the usage of the abundant nucleotide A and minimizes the rare nucleotide C (Takata et al., 2017; Xia et al., 2006). However, this will drive up the usage of A-ending codons that may not be the optimal codons for translation. There are also factors affecting codon usage that are not related to translation. For example, mammalian species have zinc-finger antiviral proteins (ZAPs), which work against RNA viruses by targeting CpG dinucleotides in the viral RNA (Ficarelli et al., 2019, 2020; Greenbaum et al., 2008; Meagher et al., 2019; Nchioua et al., 2020; Yap et al., 2003). Many human RNA viruses exhibit much reduced CpG dinucleotides (Atkinson et al., 2014; Ficarelli et al., 2020; Greenbaum et al., 2008, 2009; Xia, 2020), with SARS-CoV-2 being the extreme among coronaviruses (Betney et al., 2010; Wei et al., 2020). Most of the CpG reduction occurs at the di-codon configuration, from NNC GNN to NNT GNN (i.e., a synonymous replacement). This NNC to NNU change has nothing to do with codon optimization. So, what will be the functional relationship among all these variables? The model is even more complicated if we consider not only HEGs but also genes whose optimal protein level is not the maximal, such as the autoregulated level of release factor 2 in *E. coli* (Craigen et al., 1985; Craigen & Caskey, 1986) or many others (Betney et al., 2010). We highlight such questions in the hope that they will motivate researchers to address them.

The last point we wish to discuss is the usage of ranked variables such as the ranked generation time (RankGT). Specifically, the difference in generation time (GT) between the fast-growing *Vibrio natriegens* and *V. cholerae* is only a few minutes, whereas that between the slow-growing *Mycobacterium tuberculosis* and *M. leprae* is about six days. After completing a ranking, the difference in RankGT between the two fast-growing species becomes one (=2–1),

which is the same as the difference in RankGT between the two slow-growing species (=9–8). This implicitly assumes that the difference of a few minutes in GT between the two fast-growing species is roughly equivalent to the differences of about six days in GT between the two slow-growing species.

Our results seem to substantiate the assumption above. For example, *Mycobacterium tuberculosis* and *M. leprae* have one *rrn* operon in their genomes. Any species with a much longer generation time in months or even years will not reduce the number of *rrn* operons below one. Thus, adding many days to the generation time in a slow-growing species will not change the number of *rrn* operons, but shortening the generation time by a few minutes in a fast-growing species may well increase the number of *rrn* operons. This suggests that a difference of a few minutes in GT in fast-growing species may actually result in a greater effect than a difference of a few days in slow-growing species. The same applies to the number of tRNA genes (Figure 1) and might be applicable to other translation-related features.

2.7 Conclusions

Detection of the impact of selection on translation efficiency is mostly performed by contrasting highly expressed genes and low-expressed genes within rapidly replicating species. We generalized such investigations to determine how differential selection for translation efficiency among different species will leave its footprints on the species-specific translation machinery. Our results suggest that selection for translation optimization is stronger in short-generation species than long-generation species, and that this differential selection strongly

shapes the evolutionary trajectories of translation machineries in these species, affecting translation initiation, elongation, and termination.

2.8 References

Adamski, F. M., McCaughan, K. K., Jørgensen, F., Kurland, C. G., & Tate, W. P. (1994). The concentration of polypeptide chain release factors 1 and 2 at different growth rates of *Escherichia coli*. *Journal of Molecular Biology*, *238*(3). https://doi.org/10.1006/jmbi.1994.1293

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

Andersson, D. I., Bohman, K., Isaksson, L. A., & Kurland, C. G. (1982). Translation rates and misreading characteristics of rpsD mutants in *Escherichia coli*. *Mol Gen Genet*, *187*, 467–472.

Andersson, D. I., & Kurland, C. G. (1983). Ram ribosomes are defective proofreaders. *Mol Gen Genet*, *191*, 378–381.

Andersson, S. G., Buckingham, R. H., & Kurland, C. G. (1984). Does codon composition influence ribosome function? *Embo J*, *3*, 91–94.

Andersson, S. G., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev*, *54*, 198–210.

Andrews, J. H., & Harris, R. F. (1986). r- and K-Selection and Microbial Ecology. *Advances in Microbial Ecology*, 99–147. https://doi.org/10.1007/978-1-4757-0611-6_3

Artman, M., Domenech, E., & Weiner, M. (1983). Growth of *Haemophilus influenzae* in simulated blood cultures supplemented with hemin and NAD. *Journal of Clinical Microbiology*, *18*(2), 376–379. https://doi.org/10.1128/JCM.18.2.376-379.1983

Atkinson, N. J., Witteveldt, J., Evans, D. J., & Simmonds, P. (2014). The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Research*, *42*(7). https://doi.org/10.1093/nar/gku075

Belinky, F., Rogozin, I. B., & Koonin, E. V. (2017). Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Scientific Reports 2017 7:1*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-12619-6

Betney, R., De Silva, E., Krishnan, J., & Stansfield, I. (2010). Autoregulatory systems controlling translation factor expression: thermostat-like control of translational accuracy. *RNA (New York, N.Y.)*, *16*(4), 655–663. https://doi.org/10.1261/RNA.1796210

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

Bossi, L., & Roth, J. R. (1980). The influence of codon context on genetic code translation. *Nature*, *286*, 123.

Bremer, H., & Dennis, P. (1996). Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli* and *Salmonella*: cellular and molecular biology. In N. ed. & F.C. (Eds.), *Escherichia coli and Salmonella: Cellular and Molecular Biology 2nd* (pp. 1553–1568). American Society for Microbiology.

Brown, C. M., Stockwell, P. A., Trotman, C. N. A., & Tate, W. P. (1990). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Research*, *18*(21), 6339–6345. https://doi.org/10.1093/NAR/18.21.6339

Brzeszcz, J., Steliga, T., Kapusta, P., Turkiewicz, A., & Kaszycki, P. (2016). r-strategist versus K-strategist for the application in bioremediation of hydrocarbon-contaminated soils. *International Biodeterioration & Biodegradation*, *106*, 41–52. https://doi.org/10.1016/J.IBIOD.2015.10.001

Bulmer, M. (1990). The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res*, *18*, 2869–2873.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, *129*(3). https://doi.org/10.1093/genetics/129.3.897

Calogero, R. A., Pon, C. L., Canonaco, M. A., & Gualerzi, C. O. (1988). Selection of the mRNA translation initiation region by *Escherichia coli* ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(17). https://doi.org/10.1073/pnas.85.17.6427

Carullo, M., & Xia, X. (2008). An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. *Journal of Molecular Evolution*, *66*(5). https://doi.org/10.1007/s00239-008-9102-8

César Sánchez, J., Padrón, G., Santana, H., & Herrera, L. (1998). Elimination of an HuIFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J Biotechnol*, *63*, 179–186.

Change, Y. T., Andersen, R. N., & Vaituzis, Z. (1967). Growth of *Mycobacterium lepraemurium* in cultures of mouse peritoneal macrophages. *J Bacteriol*, *93*, 1119–1131.

Chithambaram, S., Prabhakaran, R., & Xia, X. (2014a). Differential Codon Adaptation between dsDNA and ssDNA Phages in *Escherichia coli*. *Molecular Biology and Evolution*, *31*, 1606–1617.

Chithambaram, S., Prabhakaran, R., & Xia, X. (2014b). The Effect of Mutation and Selection on Codon Adaptation in *Escherichia coli* Bacteriophage. *Genetics*, *197*, 301–315.

Coghlan, A., & Wolfe, K. H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, *16*, 1131–1145.

Colangeli, R., Arcus, V. L., Cursons, R. T., Ruthe, A., Karalus, N., Coley, K., Manning, S. D., Kim, S., Marchiano, E., & Alland, D. (2014). Whole genome sequencing of *Mycobacterium*

tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0091024

Cole, S. T. (2002). Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *Microbiology*, *148*, 2919–2928.

Comeron, J. M., & Aguade, M. (1998). An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol*, *47*, 268–274.

Comín, J., Cebollada, A., Iglesias, M. J., Ibarz, D., Viñuelas, J., Torres, L., Sahagún, J., Lafoz, M. C., Esteban de Juanas, F., Malo, M. C., & Samper, S. (2022). Estimation of the mutation rate of *Mycobacterium tuberculosis* in cases with recurrent tuberculosis using whole genome sequencing. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-21144-0

Cook, G. M., Berney, M., Gebhard, S., Heinemann, M., Cox, R. A., Danilchanka, O., & Niederweis, M. (2009). Physiology of Mycobacteria. *Advances in Microbial Physiology*, *55*, 81–319. https://doi.org/10.1016/S0065-2911(09)05502-7

Cortes, M. A., Nessar, R., & Singh, A. K. (2010). Laboratory maintenance of *Mycobacterium abscessus*. *Current Protocols in Microbiology*, *Chapter 10, Unit 10D.11*.

Craigen, W. J., & Caskey, C. T. (1986). Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, *322*(6076). https://doi.org/10.1038/322273a0

Craigen, W. J., Cook, R. G., Tate, W. P., & Caskey, C. T. (1985). Bacterial peptide chain release factors: Conserved primary structure and possible frameshift regulation of release factor 2. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(11). https://doi.org/10.1073/pnas.82.11.3616

Crawford, D. J. G., Ito, K., Nakamura, Y., & Tate, W. P. (1999). Indirect regulation of translational termination efficiency at highly expressed genes and recoding sites by the factor recycling function of *Escherichia coli* release factor RF3. *EMBO Journal*, *18*(3). https://doi.org/10.1093/emboj/18.3.727

Cridge, A. G., Major, L. L., Mahagaonkar, A. A., Poole, E. S., Isaksson, L. A., & Tate, W. P. (2006). Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. *Nucleic Acids Research*, *34*(7). https://doi.org/10.1093/nar/gkl074

Curran, J. F., & Yarus, M. (1988). Use of tRNA suppressors to probe regulation of *Escherichia coli* release factor 2. *J Mol Biol*, *203*, 75–83.

Davies, J., Jones, D. S., & Khorana, H. G. (1966). A further study of misreading of codons induced by streptomycin and neomycin using ribopolynucleotides containing two nucleotides in alternating sequence as templates. *J Mol Biol*, *18*, 48–57.

De Smit, M. H., & Duin, J. Van. (1990). Control of Prokaryotic Translational Initiation by mRNA Secondary Structure. *Progress in Nucleic Acid Research and Molecular Biology*, *38*(C), 1–35. https://doi.org/10.1016/S0079-6603(08)60707-2

Deutscher, M. P. (2015). Twenty years of bacterial RNases and RNA processing: how we've matured. *RNA*, *21*, 597–600.

Dryselius, R., Izutsu, K., Honda, T., & Iida, T. (2008). Differential replication dynamics for large and small Vibrio chromosomes affect gene dosage, expression and location. *BMC Genomics*, *9*(559). https://doi.org/10.1186/1471-2164-9-559/FIGURES/5

Du, M. Z., Wei, W., Qin, L., Liu, S., Zhang, A. Y., Zhang, Y., Zhou, H., & Guo, F. B. (2017). Co-adaption of tRNA gene copy number and amino acid usage influences translation rates in three life domains. *DNA Research*, *24*(6). https://doi.org/10.1093/dnares/dsx030

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. U S A*, *96*, 4482–4487.

Eagon, R. G. (1962). Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *Journal of Bacteriology*, *83*(4), 736–737. https://doi.org/10.1128/JB.83.4.736-737.1962

Felsenstein, J. (1985). Phylogenies and the comparative method. *Amer. Nat*, *125*, 1–15.

Ficarelli, M., Antzin-Anduetza, I., Hugh-White, R., Firth, A. E., Sertkaya, H., Wilson, H., Neil, S. J. D., Schulz, R., & Swanson, C. M. (2020). CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. *Journal of Virology*, *94*(6). https://doi.org/10.1128/jvi.01337-19

Ficarelli, M., Wilson, H., Galão, R. P., Mazzon, M., Antzin-Anduetza, I., Marsh, M., Neil, S. J. D., & Swanson, C. M. (2019). KHNYN is essential for the zinc finger antiviral protein (ZAP) to restrict HIV-1 containing clustered CpG dinucleotides. *ELife*, *8*. https://doi.org/10.7554/eLife.46767

Gausing, K. (1977). Regulation of ribosome production in *Escherichia coli*: synthesis and stability of ribosomal RNA and of ribosomal protein messenger RNA at different growth rates. *J Mol Biol*, *115*, 335–354.

Geller, A. I., & Rich, A. (1980). A UGA termination suppression tRNATrp active in rabbit reticulocytes. *Nature*, *283*, 41–46.

Gengenbacher, M., & Kaufmann, S. H. (2012). *Mycobacterium tuberculosis*: success through dormancy. *FEMS Microbiol Rev*, *36*, 514–532.

Gibson, B., Wilson, D. J., Feil, E., & Eyre-Walker, A. (2018). The distribution of bacterial doubling times in the wild. *Proc Biol Sci*, 285.

Glazyrina, J., Materne, E. M., Dreher, T., Storm, D., Junne, S., Adams, T., Greller, G., & Neubauer, P. (2010). High cell density cultivation and recombinant protein production with *Escherichia coli* in a rocking-motion-type bioreactor. *Microb Cell Fact*, *9*, 42.

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, *10*(22). https://doi.org/10.1093/nar/10.22.7055

Greenbaum, B. D., Levine, A. J., Bhanot, G., & Rabadan, R. (2008). Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathogens*, *4*(6). https://doi.org/10.1371/journal.ppat.1000079

Greenbaum, B. D., Rabadan, R., & Levine, A. J. (2009). Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. *PLoS ONE*, *4*(6). https://doi.org/10.1371/journal.pone.0005969

Gualerzi, C. O., & Pon, C. L. (2015). Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cellular and Molecular Life Sciences : CMLS*, *72*(22), 4341–4367. https://doi.org/10.1007/S00018-015-2010-3

Haas, J., Park, E.-C., & Seed, B. (1996). Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr. Biol*, *6*, 315–324.

Hartz, D., McPheeters, D. S., & Gold, L. (1991). Influence of mRNA determinants on translation initiation in *Escherichia coli*. *Journal of Molecular Biology*, *218*(1), 83–97. https://doi.org/10.1016/0022-2836(91)90875-7

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford University Press: Oxford.

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D., & Salit, M. (2017). Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Research*, *45*(7), 3615. https://doi.org/10.1093/NAR/GKX070

Hershberg, R., & Petrov, D. A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics*, *6*(9), e1001115. https://doi.org/10.1371/JOURNAL.PGEN.1001115

Hockenberry, A. J., Stern, A. J., Amaral, L. A. N., & Jewett, M. C. (2018). Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. *Molecular Biology and Evolution*, *35*(3). https://doi.org/10.1093/molbev/msx310

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, *31*(13), 3429. https://doi.org/10.1093/NAR/GKG599

Honore, N., Marchal, G., & Cole, S. T. (1995). Novel mutation in 16S rRNA associated with streptomycin dependence in *Mycobacterium tuberculosis*. In *Antimicrobial Agents and Chemotherapy* (Vol. 39, Issue 3). https://doi.org/10.1128/AAC.39.3.769

Hui, A., & De Boer, H. A. (1987). *Specialized ribosome system: Preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli (Shine-Dalgarno sequence/rrnB operon)*. *84*, 4762–4766. https://www.pnas.org

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol*, *146*, 1–21.

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

Jacob, F. (2013). *Génétique cellulaire: Leçon inaugurale prononcée le vendredi 7 mai 1965*.

Jorgensen, F., Adamski, F. M., Tate, W. P., & Kurland, C. G. (1993). Release factor-dependent false stops are infrequent in *Escherichia coli*. *J Mol Biol*, *230*, 41–50.

Kashyap, S., & Sarkar, M. (2010). Mycoplasma pneumonia: Clinical features and management. *Lung India*, *27*, 75–85.

Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M., & Chanfreau, G. F. (2014). Widespread Use of Non-productive Alternative Splice Sites in *Saccharomyces cerevisiae*. *PLoS Genetics*, *10*(4). https://doi.org/10.1371/journal.pgen.1004249

Keener, J., Nomura, M., Biology, N., F.C. III, R. C., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., & Riley. (1996). Regulation of ribosome synthesis. In S. M., U. M., & J.E. (Eds.), *Escherichia coli and Salmonella: cellular and molecular* (Vol. 1, pp. 1417–1428). ASM Press.

Kjeldgaard, N., & Kurland, C. (1963). The distribution of soluble and ribosomal RNA as a function of growth rate. *J. Mol. Biol*, *6*, 341–348.

Klappenbach, J. A., Dunbar, J. M., & Schmidt, T. M. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, *66*, 1328–1333. https://doi.org/10.1128/AEM.66.4.1328-1333.2000

Konecki, D. S., Aune, K. C., Tate, W., & Caskey, C. T. (1977). Characterization of reticulocyte release factor. *J. Biol. Chem*, *252*, 4514–4520.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science (New York, N.Y.)*, *324*(5924), 255–258. https://doi.org/10.1126/SCIENCE.1170160

Lee, W., & Engels, B. (2014). The protonation state of catalytic residues in the resting state of kasa revisited: Detailed mechanism for the activation of KasA by its own substrate. *Biochemistry*, *53*(5). https://doi.org/10.1021/bi401308j

Li, W. H., Gojobori, T., & Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, *292*, 237–239.

Li, W.-H. (1983). *Evolution of duplicate genes and pseudogenes*. Sunderland, MA.

Liljenström, H., & von Heijne, G. (1987). Translation rate modification by preferential codon usage: Intragenic position effects. *Journal of Theoretical Biology*, *124*(1). https://doi.org/10.1016/S0022-5193(87)80251-5

Lind, P. A., & Andersson, D. I. (2008). Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 17878–17883. https://doi.org/10.1073/PNAS.0804445105/SUPPL_FILE/0804445105SI.PDF

Maitra, A., & Dill, K. A. (2015). Bacterial growth laws reflect the evolutionary importance of energy efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(2), 406–411. https://doi.org/10.1073/PNAS.1421138111/SUPPL_FILE/PNAS.201421138SI.PDF

McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J., & Tate, W. P. (1995). Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. U S A*, *92*, 5431–5435.

Meagher, J. L., Takata, M., Gonçalves-Carneiro, D., Keane, S. C., Rebendenne, A., Ong, H., Orr, V. K., MacDonald, M. R., Stuckey, J. A., Bieniasz, P. D., & Smith, J. L. (2019). Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48). https://doi.org/10.1073/pnas.1913232116

Meng, S. Y., Hui, J. O., Haniu, M., & Tsai, L. B. (1995). Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem Biophys Res Commun*, *211*, 40–48.

Miller, J. H., & Albertini, A. M. (1983). Effects of surrounding sequence on the suppression of nonsense codons. *J Mol Biol*, *164*, 59–71.

Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, *35*, 1050–1055.

Moller, S., Kristensen, C. S., Poulsen, L. K., Carstensen, J. M., & Molin, S. (1995). Bacterial growth on surfaces: automated image analysis for quantification of growth rate-related parameters. *Appl Environ Microbiol*, *61*, 741–748.

Moriyama, E. N., & Powell, J. R. (1997). Codon usage bias and tRNA abundance in Drosophila. *J. Mol. Evol*, *45*, 514–523.

Mukherjee, T., Goswami, A., Chakraborty, U., Majumdar, M., Sinha, S., & Pal, N. K. (2019). A Study on Generation Time of Sensitive and Resistant *Mycobacterium tuberculosis* Isolates. *J Evolution Med Dent. Sci*, *8*, 2489–2494.

Murray, H. D., Appleman, J. A., & Gourse, R. L. (2003). Regulation of the *Escherichia coli* rrnB P2 promoter. *J Bacteriol*, *185*, 28–34.

Muto, A., & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA*, *84*, 166–169.

Nakamoto, T. (2006). A unified view of the initiation of protein synthesis. *Biochemical and Biophysical Research Communications*, *341*(3), 675–678. https://doi.org/10.1016/J.BBRC.2006.01.019

Nataraj, V., Varela, C., Javid, A., Singh, A., Besra, G. S., & Bhatt, A. (2015). Mycolic acids: Deciphering and targeting the Achilles' heel of the tubercle bacillus. In *Molecular Microbiology* (Vol. 98, Issue 1). https://doi.org/10.1111/mmi.13101

Nchioua, R., Kmiec, D., Müller, J. A., Conzelmann, C., Groß, R., Swanson, C. M., Neil, S. J. D., Stenger, S., Sauter, D., Münch, J., Sparrer, K. M. J., & Kirchhoff, F. (2020). Sars-cov-2 is restricted by zinc finger antiviral protein despite preadaptation to the low-cpg environment in humans. *MBio*, *11*(5). https://doi.org/10.1128/mBio.01930-20

Ngumbela, K. C., Ryan, K. P., Sivamurthy, R., Brockman, M. A., Gandhi, R. T., Bhardwaj, N., & Kavanagh, D. G. (2008). Quantitative Effect of Suboptimal Codon Usage on Translational Efficiency of mRNA Encoding HIV-1 gag in Intact T Cells. *PLoS ONE*, *3*, 2356.

Noller, H. F. (2012). Evolution of Protein Synthesis from an RNA World. *Cold Spring Harbor Perspectives in Biology*, *4*(4). https://doi.org/10.1101/CSHPERSPECT.A003681

Nomura, M., Gourse, R., & Baughman, G. (1984). Regulation of the synthesis of ribosomes and ribosomal components. *Annu Rev Biochem*, *53*, 75–117.

Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A General Model of Codon Bias Due to GC Mutational Bias. *PLOS ONE*, *5*(10), e13431. https://doi.org/10.1371/JOURNAL.PONE.0013431

Parker, J. (1989). Errors and alternatives in reading the universal genetic code. *Microbiol Rev*, *53*, 273–298.

Poole, E. S., Brimacombe, R., & Tate, W. P. (1997). Decoding the translational termination signal: The polypeptide chain release factor in *Escherichia coli* crosslinks to the base following the stop codon. *RNA*, *3*(9).

Poole, E. S., Brown, C. M., & Tate, W. P. (1995). The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. *The EMBO Journal*, *14*(1), 151–158. https://doi.org/10.1002/J.1460-2075.1995.TB06985.X

Poole, E. S., Major, L. L., Mannering, S. A., & Tate, W. P. (1998). Translational termination in *Escherichia coli*: Three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Research*, *26*(4). https://doi.org/10.1093/nar/26.4.954

Poulsen, L. K., Ballard, G., & Stahl, D. A. (1993). Use of rRNA fluorescence in situ hybridization for measuring the activity of single cells in young and established biofilms. *Appl Environ Microbiol*, *59*, 1354–1360.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2014). Aeromonas phages encode tRNAs for their overused codons. *International Journal of Computational Biology and Drug Design*, *7*, 168–182.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2015). *Escherichia coli* and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. *The Journal of General Virology*, *96*(Pt 5), 1169–1179. https://doi.org/10.1099/VIR.0.000050

Ran, W., & Higgs, P. G. (2012). Contributions of Speed and Accuracy to Translational Selection in Bacteria. *PLOS ONE*, *7*(12), e51652. https://doi.org/10.1371/JOURNAL.PONE.0051652

Ran, W., Kristensen, D. M., & Koonin, E. V. (2014). Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. *MBio*, *5*(2). https://doi.org/10.1128/mBio.00956-14

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

Rocha, E. P. C. (2004). Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, *14*(11), 2279–2286. https://doi.org/10.1101/GR.2896904

Rodnina, M. V. (2018). Translation in Prokaryotes. *Cold Spring Harbor Perspectives in Biology*, *10*(9), a032664. https://doi.org/10.1101/CSHPERSPECT.A032664

Roth, J. R. (1970). UGA nonsense mutations in *Salmonella typhimurium*. *J Bacteriol*, *102*, 467–475.

Ryden, S. M., & Isaksson, L. A. (1984). A temperature-sensitive mutant of *Escherichia coli* that shows enhanced misreading of UAG/A and increased efficiency for some tRNA nonsense suppressors. *Mol Gen Genet*, *193*, 38–45.

Sambrook, J. F., Fan, D. P., & Brenner, S. (1967). A strong suppressor specific for UGA. *Nature*, *214*, 452–453.

Schaechter, M., Maaloe, O., & Kjeldgaard, N. O. (1995). N.O. Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium*. *J Gen Microbiol*, *19*, 592–606.

Scharff, L. B., Childs, L., Walther, D., & Bock, R. (2011). Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genetics*, *7*(6). https://doi.org/10.1371/JOURNAL.PGEN.1002155

Sezonov, G., Joseleau-Petit, D., & D'Ari, R. (2007). *Escherichia coli* physiology in Luria-Bertani broth. *J Bacteriol*, *189*, 8746–8749.

Shan Chang, D. P., & Guan, X. L. (2021). Metabolic versatility of *Mycobacterium tuberculosis* during infection and dormancy. *Metabolites*, *11*(2). https://doi.org/10.3390/metabo11020088

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, *33*(4). https://doi.org/10.1093/nar/gki242

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. https://doi.org/10.1093/NAR/15.3.1281

Sharpe, M. E., Hauser, P. M., Sharpe, R. G., & Errington, J. (1998). *Bacillus subtilis* cell cycle as studied by fluorescence microscopy: constancy of cell length at initiation of DNA replication and evidence for active nucleoid partitioning. *J Bacteriol*, *180*, 547–555.

Shepard, C. C. (1965). Temperature optimum of *Mycobacterium leprae* in mice. *Journal of Bacteriology*, *90*(5). https://doi.org/10.1128/jb.90.5.1271-1275.1965

Shine, J., & Dalgarno, L. (1974a). Identical 3'-terminal octanucleotide sequence in 18S ribosomal ribonucleic acid from different eukaryotes. A proposed role for this sequence in the recognition of terminator codons. *Biochem J*, *141*, 609–615.

Shine, J., & Dalgarno, L. (1974b). 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 of the United States of America*, *71*(4), 1342–1346. https://doi.org/10.1073/PNAS.71.4.1342

Shine, J., & Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature*, *254*, 34–38.

Sorensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol*, *207*, 365–377.

Steitz, J. A., & Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. *Proc. Natl. Acad. Sci. U S A*, *72*, 4734–4738.

Strigini, P., & Brickman, E. (1973). Analysis of specific misreading in *Escherichia coli*. *Journal of Molecular Biology*, *75*(4). https://doi.org/10.1016/0022-2836(73)90299-4

Studer, S. M., & Joseph, S. (2006). Unfolding of mRNA Secondary Structure by the Bacterial Translation Initiation Complex. *Molecular Cell*, *22*(1). https://doi.org/10.1016/j.molcel.2006.02.014

Sun, X. Y., Yang, Q., & Xia, X. (2013). An Improved Implementation of Effective Number of Codons (Nc). *Molecular Biology and Evolution*, *30*, 191–196.

Takata, M. A., Gonçalves-Carneiro, D., Zang, T. M., Soll, S. J., York, A., Blanco-Melo, D., & Bieniasz, P. D. (2017). CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*, *550*(7674). https://doi.org/10.1038/nature24039

Taniguchi, T., & Weissmann, C. (1978). Inhibition of Qbeta RNA 70S ribosome initiation complex formation by an oligonucleotide complementary to the 3' terminal region of *E. coli* 16S ribosomal RNA. *Nature*, *275*, 770–772.

Tate, W., Greuer, B., & Brimacombe, R. (1990). Codon recognition in polypeptide chain termination: sIte directed crosslinking of termination codon to *Escherichia coli* release factor 2. *Nucleic Acids Research*, *18*(22). https://doi.org/10.1093/nar/18.22.6537

Tate, W. P., & Brown, C. M. (1992). Translational Termination: "Stop" for Protein Synthesis or "Pause" for Regulation of Gene Expression. *Biochemistry*, *31*(9). https://doi.org/10.1021/bi00124a001

Tate, W. P., & Mannering, S. A. (1996). Three, four or more: The translational stop signal at length. In *Molecular Microbiology* (Vol. 21, Issue 2). https://doi.org/10.1046/j.1365-2958.1996.6391352.x

Tate, W. P., Mansell, J. B., Mannering, S. A., Irvine, J. H., Major, L. L., & Wilson, D. N. (1999). UGA: a dual signal for "stop" and for recoding in protein synthesis. *Biochemistry (Mosc*, *64*, 1342–1353.

Tate, W. P., Poole, E. S., Dalphin, M. E., Major, L. L., Crawford, D. J. G., & Mannering, S. A. (1996). The translational stop signal: Codon with a context, or extended factor recognition element? *Biochimie*, *78*(11–12). https://doi.org/10.1016/S0300-9084(97)86716-8

Tate, W. P., Poole, E. S., Horsfield, J. A., Mannering, S. A., Brown, C. M., Moffat, J. G., Dalphin, M. E., McCaughan, K. K., Major, L. L., & Wilson, D. N. (1995). Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. *Https://Doi.Org/10.1139/O95-118*, *73*(11–12), 1095–1103. https://doi.org/10.1139/O95-118

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3645–3650. https://doi.org/10.1073/PNAS.0909910107/SUPPL_FILE/DS01.XLS

Valgepea, K., Adamberg, K., Seiman, A., & Vilu, R. (2013). *Escherichia coli* achieves faster growth by increasing catalytic and translation rates of proteins. *Molecular BioSystems*, *9*, 2344–2358.

Van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L., & Xia, X. (2011). HIV-1 Modulates the tRNA Pool to Improve Translation Efficiency. *Molecular Biology and Evolution*, *28*, 1827–1834.

Verma, A., Sampla, A. K., & Tyagi, J. S. (1999). *Mycobacterium tuberculosis* rrn promoters: differential usage and growth rate-dependent control. *J Bacteriol*, *181*, 4326–4333.

Waites, K. B., & Talkington, D. F. (2004). Mycoplasma pneumoniae and its role as a human pathogen. *Clin Microbiol Rev*, *17*, 697–728,.

Wang, H. C., Xia, X., & Hickey, D. A. (2006). Thermal adaptation of ribosomal RNA genes: a comparative study. *J. Mol. Evol*, *63*, 120–126.

Weaver, R. F. (2012). *Molecular Biology* (5th ed.). Boston.

Wei, Y., Silke, J. R., Aris, P., & Xia, X. (2020). Coronavirus genomes carry the signatures of their habitats. *PLoS ONE*, *15*(12 December 2020). https://doi.org/10.1371/journal.pone.0244025

Wei, Y., Silke, J. R., & Xia, X. (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-39369-X

Wei, Y., Wang, J., & Xia, X. (2016). Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. *Molecular Biology and Evolution*, *33*(9), 2357–2367. https://doi.org/10.1093/molbev/msw107

Wei, Y., & Xia, X. (2017). The role of +4U as an extended translation termination signal in bacteria. *Genetics*, *205*(2), 539–549. https://doi.org/10.1534/GENETICS.116.193961/-/DC1

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, *87*(1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Xia, X. (1995). Body temperature, rate of biosynthesis and evolution of genome size. *Molecular Biology and Evolution*, *12*, 834–842.

Xia, X. (1996). Maximizing transcription efficiency causes codon usage bias. *Genetics*, *144*(3). https://doi.org/10.1093/genetics/144.3.1309

Xia, X. (1998). How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics*, *149*(1). https://doi.org/10.1093/genetics/149.1.37

Xia, X. (2007). An Improved Implementation of Codon Adaptation Index. *Evolutionary Bioinformatics*, *3*, 53–58.

Xia, X. (2008). The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. *BMC Evol. Biol*, *8*, 211.

Xia, X. (2012). Rapid evolution of animal mitochondrial DNA. *Rapidly Evolving Genes and Genetic Systems*, 73–82. https://doi.org/10.1093/ACPROF:OSO/9780199642274.003.0008

Xia, X. (2013). *Comparative genomics*. Springer.

Xia, X. (2015). A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics*, *199*(2), 573–579. https://doi.org/10.1534/GENETICS.114.172106

Xia, X. (2017). DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *Journal of Heredity*, *108*, 431–437. https://doi.org/10.1093/jhered/esx033

Xia, X. (2018a). Bioinformatics and Translation Elongation. In *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics* (pp. 197–238). Springer. https://doi.org/10.1007/978-3-319-90684-3_9

Xia, X. (2018b). Bioinformatics and Translation Initiation. In *Bioinformatics and the Cell: modern computational approaches in genomics, proteomics and transcriptomics,* (pp. 173–195). Springer. https://doi.org/10.1007/978-3-319-90684-3_8

Xia, X. (2018c). Bioinformatics and Translation Termination in Bacteria. In *Bioinformatics and the Cell: modern computational approaches in genomics, proteomics and transcriptomics* (pp. 239–254). Springer.

Xia, X. (2018d). Characterizing translation efficiency. In X. Xia (Ed.), *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics* (pp. 173–206). Springer US.

Xia, X. (2018e). DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Molecular Biology and Evolution*, *35*, 1550–1552.

Xia, X. (2020). Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Molecular Biology and Evolution*, *37*(9). https://doi.org/10.1093/molbev/msaa094

Xia, X. (2021). Detailed Dissection and Critical Evaluation of the Pfizer/BioNTech and Moderna mRNA Vaccines. *Vaccines*, *9*, 734.

Xia, X. (2023a). Horizontal Gene Transfer and Drug Resistance Involving *Mycobacterium tuberculosis*. *Antibiotics*, *12*(9). https://doi.org/10.3390/antibiotics12091367

Xia, X. (2023b). Optimizing Protein Production in Therapeutic Phages against a Bacterial Pathogen, *Mycobacterium abscessus*. *Drugs and Drug Candidates 2023, Vol. 2, Pages 189-209*, *2*(1), 189–209. https://doi.org/10.3390/DDC2010012

Xia, X., & Holcik, M. (2009). Strong eukaryotic IRESs have weak secondary structure. *PLoS ONE*, *4*(1). https://doi.org/10.1371/journal.pone.0004136

Xia, X., Huang, H., Carullo, M., Betran, E., & Moriyama, E. N. (2007). Conflict between Translation Initiation and Elongation in Vertebrate Mitochondrial Genomes. *PLoS ONE*, *2*, 227.

Xia, X., Mackay, V., Yao, X., Wu, J., Miura, F., Ito, T., & Morris, D. R. (2011). Translation initiation: A regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*. *Genetics*, *189*(2). https://doi.org/10.1534/genetics.111.132068

Xia, X., Wang, H., Xie, Z., Carullo, M., Huang, H., & Hickey, D. (2006). Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. *Molecular Biology and Evolution*, *23*(7). https://doi.org/10.1093/molbev/msl012

Yang, L., Haagensen, J. A., Jelsbak, L., Johansen, H. K., Sternberg, C., Høiby, N., & Molin, S. (2008). In situ growth rates and biofilm development of Pseudomonas aeruginosa populations in chronic lung infections. *J Bacteriol*, *190*, 2767–2776.

Yap, Y. L., Zhang, X. W., & Danchin, A. (2003). Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics*, *4*. https://doi.org/10.1186/1471-2105-4-43

Yin, M., Ye, B., Jin, Y., Liu, L., Zhang, Y., Li, P., Wang, Y., Li, Y., Han, Y., & Shen, W. (2020). Changes in *Vibrio natriegens* Growth Under Simulated Microgravity. *Frontiers in Microbiology*, *11*(2040).

Zhao, J., Siddiqui, S., Shang, S., Bian, Y., Bagchi, S., He, Y., & Wang, C. R. (2015). Mycolic acid-specific T cells protect against *Mycobacterium tuberculosis* infection in a humanized transgenic mouse model. *ELife*, *4*(DECEMBER2015). https://doi.org/10.7554/eLife.08525

Zhu, M., & Dai, X. (2018). On the intrinsic constraint of bacterial growth rate: *M. tuberculosis*'s view of the protein translation capacity. *Critical Reviews in Microbiology*, *44*(4), 455–464. https://doi.org/10.1080/1040841X.2018.1425672

CHAPTER 3
Discussion/Conclusions

3.1. Impacts of Selection for Translation Efficiency on the Translation Machinery of Short-Generation and Long-Generation Bacteria

We have undertaken a comparative analysis of translation machineries in 9 bacterial species which differ substantially in generation time. Under the hypothesis that the translation subprocesses, namely initiation, elongation, and termination, are subject to stronger selection pressure for efficiency of translation in SGB such as *V. natriegens* and *E. coli* than in LGB such as *M. leprae* and *M. tuberculosis*, we engaged in comparative study of the number of rRNA and tRNA genes, start codon preference, stop codon preference, intensity of tRNA-mediated selection on codon usage, the effective number of tRNA anticodons, the secondary structure strength of mRNA near the Shine-Dalgarno sequence and start codon, the secondary structure strength of mRNA near the stop codon, and the degree of constraint exhibited in the positioning of the Shine-Dalgarno sequence. Additionally, we accounted for the influence of mutation (via genomic GC% bias) where possible, prior to drawing conclusions about selection effects. Furthermore, under a second hypothesis that highly expressed genes (HEGs) are subject to a greater intensity of selection pressure than all other CDS genes (REST genes), where applicable, we assessed differences between HEG and REST gene groups within each species.

In general, we found that our results support both hypotheses above. First, we found that SGB tend to invest more in translation machinery than LGB, such that the number of rrn operons decreases significantly with increasing generation time, and the number of tRNA genes decreases significantly with increasing generation time. Second, we found that SGB tend to prefer use of AUG as a start codon over use of any of the weaker alternative start codons (i.e. GUG, UUG,

CUG, AUA, AUC, or AUU) in comparison to LGB. Additionally, we found that genomic GC bias also has an accessory role in determining the degree of preference for AUG, and that the degree of preference for AUG is more prominent in HEGs than in REST genes. Among the start codons, AUG acts as the most efficient signal for translation initiation (Hartz et al., 1991; Hecht et al., 2017), and usage of AUG as a start codon is favored by all study species. Furthermore, of the three canonical stop codons in bacteria (i.e. UAA, UGA, and UAG), we found that SGB tend to prefer use of UAA as a stop codon in comparison to LGB. Additionally, we found that genomic GC bias plays a significant role in determining stop codon preference, explaining why the LGB study species (which all have GC-biased genomes) tend to favor usage of non-UAA stop codons, and that UAA usage is preferred more in HEGs than in REST genes. Of the stop codons, UAA exhibits the lowest rate of readthrough error (César Sánchez et al., 1998; Geller & Rich, 1980; Jorgensen et al., 1993; Meng et al., 1995; Parker, 1989; Tate et al., 1999). Third, we found that SGB tend to exhibit a greater degree of codon and tRNA anticodon adaptation in comparison to LGB in a manner that appears to be independent of mutation bias. Additionally, we found that codon optimization was more prominent in HEGs than REST genes in all study species. More specifically, SGB tend to exhibit a greater degree of codon optimization than LGB and tend to feature a smaller number of tRNAs carrying unique anticodons than LGB, and these results logically support one another. Fourth, we found that SGB genes tend to exhibit decreased secondary structure strength in the region near the translation initiation signals (i.e. SD and start codon) in comparison to LGB genes. Additionally, the evidence appears to align better with a selectionist interpretation than a selection-agnostic interpretation which invokes lack of internal base-pairing within purine-rich SD sequences as an explanation for the apparent weakening in secondary structure strength near the SD. Furthermore, genomic GC bias plays a significant role

in determining the secondary structure strength in the translation initiation region, the secondary structure strength in the translation initiation region decreases more in HEGs than in REST genes, and the positive effect of GC content on secondary structure strength appears to be greater in REST genes than in HEGs. Weakening in the secondary structure strength of the translation initiation region is essential for efficient translation of the associated gene (Calogero et al., 1988; Nakamoto, 2006; Studer & Joseph, 2006). As a final note, the pattern of secondary structure weakening in the region near the stop codon visually appears to be similar to the pattern observed in the translation initiation region. We predicted that the translation machinery would be under a higher degree of selection primarily in SGB in comparison to LGB, and secondarily in HEGs in comparison to REST genes. Overall, these results are highly consistent with the predicted effects of selection for translation efficiency on the gene and genome features which comprise the machinery for translation initiation, elongation, and termination.

3.2. Future Work

As discussed in section 2.6, the comparative analyses regarding optimization of translation machinery for translation efficiency conducted in this thesis were limited by the high degree of uncertainty involved in the measurement of generation time, the limited number of study species, and limitations in the general conceptual understanding of how the forces of selection and mutation influence optimization of translation machinery for translation efficiency in bacteria.

First, while canonical mechanisms are relatively well-understood, alternative mechanisms and the relationships between mechanisms remain poorly understood. Control of translation is

multifaceted and features extensive redundancy. Each of the key players in the translation process, including the ribosome and its subcomponents, as well as the enzymes (initiation factors, elongation factors, and release factors) that facilitate each translation subprocess, may vary in type and/or rate of synthesis, processing and modification, degradation, and recycling (Rodnina, 2018). Regulatory controls that can upregulate or downregulate any player, including lineage-specific players such as the ribosomal protein bs1 add an additional layer of complexity (Hu, 2019; Ranjan & Rodnina, 2016; Rodnina, 2018), and conditional regulation of select genes during stress extends that complexity (Oliva et al., 2015; Sawyer et al., 2021). Further, varied rates of error are associated with each coder:decoder interaction and with each proofreading and quality control step. Alternative translation mechanisms, including leaderless translation, further complicate the picture (Babitzke & O'Connor, 2017).

Of the numerous avenues which are ripe for future investigation, some include investigation into poorly understood topics such as differences in control of gene expression in SD-led, SD-independent, and leaderless genes (Sawyer et al., 2021; Scharff et al., 2011; Tuller et al., 2010), elucidation of the roles of potential confounders within the physical landscape of the cell (e.g. the association between decreased secondary structure strength in the translation initiation region could be partially attributed to the presence of poly(A) tracts preceding the start codon which enhance translation efficiency in yeast (Xia et al., 2011), as mentioned in section 2.6) and may stimulate transcription in bacteria such as *E. coli* (Aiyar et al., 1998), characterizing strategies employed for resolution of conflicting selective priorities (e.g. between optimization of translation efficiency and optimization of transcription efficiency, as alluded to in section 2.6), consequences of coder signals in "inappropriate" sequence contexts (e.g. in cases where SD-like

sequences or internal start or stop codon sequences with sequence context resembling the initiation or termination sites are present in coding regions ((Belinky et al., 2017; Hockenberry et al., 2018)) and investigation into the effects of combinations of weak and strong coder signals in translation, including which strong signals, in which context, can or cannot compensate for weak signals in terms of achieving translation efficiency (Wei et al., 2019).

Furthermore, analytical techniques similar to those used in these thesis can be used for future studies in biomanufacturing (e.g. with regards to optimization of heterologous protein production in model microorganisms such as *E. coli*), identification of therapeutic targets involving the translation machinery of pathogenic bacteria (such as *M. abscessus*, *M. tuberculosis*, and *M. leprae*), and for treatments such as phage therapy (i.e. with regards to engineering bacteriophage mRNAs for optimal expression in bacterial hosts).

To describe a potential future direction within the field of synthetic biology in more detail, a recent paper by Moger-Reischer et al. compares an engineered minimal Mycoplasma mycoides cell containing only 500 protein-coding genes to its non-minimal wildtype counterpart and shows that the minimal cell evolved to gain fitness comparable to wild-type Mycoplasma mycoides over 2000 generations, despite the first generation of the minimal cell displaying a loss in fitness of over 50% relative to wild-type (Moger-Reischer et al., 2023). Specifically, fitness was quantified using head-to-head competition assays between the first generation of the minimal cell, the 2000th generation of the minimal cell, and the wildtype reference strain of M. mycoides, and these results were corroborated by growth curves (Moger-Reischer et al., 2023).

Studies comparing use of translation signals between the first and final generations of the minimal cell as well as the wild-type M. mycoides could uncover new insights into optimization of translation as an adaptive strategy. Future work in this area, including direct replication of the analysis methods used in this thesis, has clear applications relating to tailored design of engineered host organisms for synthetic biology research.

3.3. References

Aiyar, S. E., Gourse, R. L., & Ross, W. (1998). Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase α subunit. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25). https://doi.org/10.1073/pnas.95.25.14652

Babitzke, P., & O'Connor, M. (2017). Noncanonical Translation Initiation Comes of Age. *Journal of Bacteriology*, *199*(14), 295–312. https://doi.org/10.1128/JB.00295-17

Belinky, F., Rogozin, I. B., & Koonin, E. V. (2017). Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Scientific Reports 2017 7:1*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-12619-6

Calogero, R. A., Pon, C. L., Canonaco, M. A., & Gualerzi, C. O. (1988). Selection of the mRNA translation initiation region by *Escherichia coli* ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(17). https://doi.org/10.1073/pnas.85.17.6427

César Sánchez, J., Padrón, G., Santana, H., & Herrera, L. (1998). Elimination of an HuIFN alpha 2b readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. *J Biotechnol*, *63*, 179–186.

Geller, A. I., & Rich, A. (1980). A UGA termination suppression tRNATrp active in rabbit reticulocytes. *Nature*, *283*, 41–46.

Hartz, D., McPheeters, D. S., & Gold, L. (1991). Influence of mRNA determinants on translation initiation in *Escherichia coli*. *Journal of Molecular Biology*, *218*(1), 83–97. https://doi.org/10.1016/0022-2836(91)90875-7

Hecht, A., Glasgow, J., Jaschke, P. R., Bawazer, L. A., Munson, M. S., Cochran, J. R., Endy, D., & Salit, M. (2017). Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Research*, *45*(7), 3615. https://doi.org/10.1093/NAR/GKX070

Hockenberry, A. J., Jewett, M. C., Amaral, L. A. N., & Wilke, C. O. (2018). Within-gene shine-dalgarno sequences are not selected for function. *Molecular Biology and Evolution*, *35*(10). https://doi.org/10.1093/molbev/msy150

Hu, J. (2019). *Hu, J. (2019). A systems-level view of the tRNA epitranscriptome : defining the role of tRNA abundance, stability, and modifications in the bacterial stress response.* Massachusetts Institute of Technology.

Jorgensen, F., Adamski, F. M., Tate, W. P., & Kurland, C. G. (1993). Release factor-dependent false stops are infrequent in *Escherichia coli*. *J Mol Biol*, *230*, 41–50.

Meng, S. Y., Hui, J. O., Haniu, M., & Tsai, L. B. (1995). Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. *Biochem Biophys Res Commun*, *211*, 40–48.

Moger-Reischer, R. Z., Glass, J. I., Wise, K. S., Sun, L., Bittencourt, D. M. C., Lehmkuhl, B. K., Schoolmaster, D. R., Lynch, M., & Lennon, J. T. (2023). Evolution of a minimal cell. *Nature*, *620*(7972). https://doi.org/10.1038/s41586-023-06288-x

Nakamoto, T. (2006). A unified view of the initiation of protein synthesis. *Biochemical and Biophysical Research Communications*, *341*(3), 675–678. https://doi.org/10.1016/J.BBRC.2006.01.019

Oliva, G., Sahr, T., & Buchrieser, C. (2015). Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. *FEMS Microbiology Reviews*, *39*(3), 331–349. https://doi.org/10.1093/FEMSRE/FUV022

Parker, J. (1989). Errors and alternatives in reading the universal genetic code. *Microbiol Rev*, *53*, 273–298.

Ranjan, N., & Rodnina, M. V. (2016). TRNA wobble modifications and protein homeostasis. *Translation*, *4*(1). https://doi.org/10.1080/21690731.2016.1143076

Rodnina, M. V. (2018). Translation in Prokaryotes. *Cold Spring Harbor Perspectives in Biology*, *10*(9), a032664. https://doi.org/10.1101/CSHPERSPECT.A032664

Sawyer, E. B., Phelan, J. E., Clark, T. G., & Cortes, T. (2021). A snapshot of translation in *Mycobacterium tuberculosis* during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Reports*, *34*(5), 108695. https://doi.org/10.1016/J.CELREP.2021.108695

Scharff, L. B., Childs, L., Walther, D., & Bock, R. (2011). Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genetics*, *7*(6). https://doi.org/10.1371/JOURNAL.PGEN.1002155

Studer, S. M., & Joseph, S. (2006). Unfolding of mRNA Secondary Structure by the Bacterial Translation Initiation Complex. *Molecular Cell*, *22*(1). https://doi.org/10.1016/j.molcel.2006.02.014

Tate, W. P., Mansell, J. B., Mannering, S. A., Irvine, J. H., Major, L. L., & Wilson, D. N. (1999). UGA: a dual signal for “stop” and for recoding in protein synthesis. *Biochemistry (Mosc*, *64*, 1342–1353.

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3645–3650. https://doi.org/10.1073/PNAS.0909910107/SUPPL_FILE/DS01.XLS

Wei, Y., Xia, X., & Hershberg, R. (2019). Unique Shine-Dalgarno Sequences in Cyanobacteria and Chloroplasts Reveal Evolutionary Differences in Their Translation Initiation. *Genome Biology and Evolution*, *11*(11). https://doi.org/10.1093/gbe/evz227

Xia, X., Mackay, V., Yao, X., Wu, J., Miura, F., Ito, T., & Morris, D. R. (2011). Translation initiation: A regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*. *Genetics*, *189*(2). https://doi.org/10.1534/genetics.111.132068