Microbes Carry Distinct Genomic Signatures in Adaptation to Their Translation Machinery and Host Environments

Yulong Wei

Thesis submitted to the University of Ottawa
in partial fulfilment of the requirements for the degree
Doctorate in Philosophy in Biology

Department of Biology
Faculty of Science
University of Ottawa

Thèse soumise à l'Université d'Ottawa
envers la réalization partielle des éxigeances du
Doctorat en Philosophie en Biologie

Département de Biologie
Faculté des Sciences
Université d'Ottawa

Abstract

How do bacteria grow and replicate rapidly? How do viruses and phages adapt to their host environments? Bacteria require efficient translation to grow and replicate rapidly, and translation is often rate-limited by initiation. A feature that is conserved across bacterial lineages is the Shine-Dalgarno (SD) sequence at the mRNA 5' UTR, which pairs with the anti-SD sequence located at the 3' end of mature 16S rRNA. Nonetheless, much about this interaction remains unclear. Chapter 2 reveals evolutionary differences between Cyanobacteria and chloroplast translation initiation using a new model ($D_{toStart}$) that better define optimal SD sequence and an RNA-Seq-based approach that reliably characterize the 3' end of mature 16S rRNAs.

Efficacy of translation elongation depends much on tRNA-mediated codon adaptation. In *Escherichia coli*, selection favours major codons because they are rapidly decoded by abundantly available cognate tRNAs. Nonetheless, the degree codon bias correlates with tRNA availability is unclear in many bacterial species because tRNA abundance is often inadequately approximated by gene copy numbers. To better understand tRNA-mediated codon bias, Chapter 3 describes an RNA-Seq-based approach to robustly quantify tRNA abundance. Finally, Chapter 4 evaluates the degree optimal translation initiation and elongation signals affect ribosome dynamics.

The emergence of COVID-19 pandemic poses a serious global health emergency. To establish infection during cell entry, the coronavirus Spike protein binds to the host ACE2 receptor, and a high binding potential between these two players is key to infectivity. While SARS-CoV-2 transmits efficiently in humans, it is less clear which other mammals are at risk of being infected. Chapter 5 investigates the host range of SARS-CoV-2 through comparative sequence analyses at the ACE2 receptors and the Spike proteins.

As obligate parasites, coronaviruses regularly infect host tissues that express antiviral proteins (AVPs) in abundance and must evade or adapt to the host cellular environments post-entry. Two AVPs that shape viral genomes are ZAP that binds to CpG dinucleotides to facilitate viral transcript degradation, and APOBEC3 which deaminates C into U leading to dysfunctional transcripts. Chapter 6 shows that coronavirus genomes are CpG deficient to evade ZAP and are subjected to constant C to U deamination by APOBEC3.

This thesis examines two key concepts of microbial genome evolution: 1) coevolution between gene features and the translation machinery in bacteria, and 2) adaptation of viruses to the hosts they infect. Chapters 2, 3, and 4 are aimed at improving our understanding in bacterial gene expression in the applications of transgenic biosynthesis and phage therapy. Chapters 5 and 6 are aimed at improving our understanding in the origin and evolution of SARS-CoV-2 and our ability to control the spread of infection.

Résumé

Comment les bactéries se développent-elles et se répliquent-elles rapidement ? Comment les virus et les phages s'adaptent-ils à leur environnement hôte ? Les bactéries ont besoin d'une traduction efficace pour se développer et se répliquer rapidement, et la traduction est souvent limitée par l'initiation. Une caractéristique qui est conservée dans les lignées bactériennes est la séquence Shine-Dalgarno (SD) au niveau de l'UTR 5' de l'ARNm, qui s'apparie à la séquence anti-SD située à l'extrémité 3' de l'ARNr 16S mature. Néanmoins, beaucoup de choses sur cette interaction restent floues. Le chapitre 2 révèle les différences évolutives entre les cyanobactéries et l'initiation de la traduction des chloroplastes à l'aide d'un nouveau modèle (DtoStart) qui définit mieux la séquence SD optimale et une approche basée sur l'ARN-Seq qui caractérise de manière fiable l'extrémité 3' des ARNr 16S matures.

L'efficacité de l'allongement de la traduction dépend beaucoup de l'adaptation des codons médiée par l'ARNt. Chez Escherichia coli, la sélection favorise les codons majeurs car ils sont rapidement décodés par des ARNt apparentés abondamment disponibles. Néanmoins, le degré de biais des codons en corrélation avec la disponibilité de l'ARNt n'est pas clair dans de nombreuses espèces bactériennes, car l'abondance de l'ARNt est souvent mal approximée par le nombre de copies de gènes. Pour mieux comprendre le biais de codon médié par l'ARNt, le chapitre 3 décrit une approche basée sur l'ARN-Seq pour quantifier de manière robuste l'abondance de l'ARNt. Enfin, le chapitre 4 évalue dans quelle mesure les signaux optimaux d'initiation et d'élongation de la traduction affectent la dynamique des ribosomes.

L’émergence de la pandémie de COVID-19 pose une grave urgence sanitaire mondiale. Pour établir l'infection lors de l'entrée dans la cellule, la protéine Spike du coronavirus se lie au

récepteur ACE2 de l'hôte, et un potentiel de liaison élevé entre ces deux acteurs est la clé de l'infectivité. Alors que le SARS-CoV-2 se transmet efficacement chez l'homme, il est moins clair quels autres mammifères risquent d'être infectés. Le chapitre 5 étudie la gamme d'hôtes du SARS-CoV-2 par le biais d'analyses de séquences comparatives au niveau des récepteurs ACE2 et des protéines Spike.

En tant que parasites obligatoires, les coronavirus infectent régulièrement les tissus de l'hôte qui expriment des protéines antivirales (AVP) en abondance et doivent échapper ou s'adapter aux environnements cellulaires de l'hôte après leur entrée. Deux AVP qui façonnent les génomes viraux sont ZAP qui se lie aux dinucléotides CpG pour faciliter la dégradation des transcrits viraux, et APOBEC3 qui désamine C en U conduisant à des transcrits dysfonctionnels. Le chapitre 6 montre que les génomes de coronavirus sont déficients en CpG pour échapper à la ZAP et sont soumis à une désamination C à U constante par APOBEC3.

Cette thèse examine deux concepts clés de la génomique évolutive microbienne : 1) la coévolution entre les caractéristiques des gènes et la machinerie de traduction chez les bactéries, et 2) l'adaptation des virus aux hôtes qu'ils infectent. Les chapitres 2, 3 et 4 visent à améliorer notre compréhension de l'expression des gènes bactériens dans les applications de la biosynthèse transgénique et de la phagothérapie. Les chapitres 5 et 6 visent à améliorer notre compréhension de l'origine et de l'évolution du SARS-CoV-2 et notre capacité à contrôler la propagation de l'infection.

Acknowledgements

It is with the utmost respect and gratitude that I acknowledge the support and guidance of my supervisor, Dr. Xuhua Xia. I am thankful for the freedom that he has given me to follow my research passion and for being a supplier of endless research ideas that are tailored to my interests. Over the years, I have grown more and more appreciative to his critical feedback aimed at fostering my scientific maturity. And I would be remiss if I fail to recognize his continued efforts to prepare me for my future pursuit in science, these include providing opportunities in teaching and research mentorships. I am truly grateful to have had the opportunity to pursue my Master's and PhD degree in the Xia lab, and I would like to thank him for acknowledging my small successes. May I do him proud in my pursuit of science.

I thank the many professors that I have had the pleasure of crossing paths with. The support that I have received from them has been integral in my pursuit for a PhD degree, and I am extremely appreciative of the reference letters that they have provided. I thank Dr. Alex Wong, who is a member of my thesis advisory committee for both my PhD and Master's studies. I appreciate his positive feedback and support over the years, and the accumulated commute time from Carleton in order to meet with me is something that I am indebted to. I thank Dr. Rees Kassen, who has always been supportive in helping me see the bigger picture of my research, and for keeping my TAC meetings organized. I thank Dr. Douglas Johnson, who had been indispensable to my completion of a Master's degree and who have continued to support me during my PhD studies on numerous key occasions. I thank Dr. Kristi Adamo for the great opportunity to take part in research in microbiome diversity. I thank Dr. Guy Drouin; I am indebted and grateful to his help as both my PhD qualifying examiner and thesis examiner. I thank Dr. Turlough Finan, who has

accepted to be my external thesis examiner. Last but not least, I thank Dr. Stéphane Aris-Brosou, who had been integral to my training in Bioinformatics, my completion of a Master's degree, and my early successes in PhD studies; and I am once more indebted to him for his acceptance to be my thesis examiner.

I thank all past and present lab members of the Xia lab. Juan Wang, Cailtyn Vlasschaert, and Ramandan Prabhakaran are valuable friends who provided me with much support and encoruagement in my early days as a graduate student; may they find success in their current careers. I thank all current members in the Xia lab, especially Parisa Aris and Alibek Kruglikov. I thank Parisa for her dedication, hard work, and significant contributions to our research in coronavirus. She has quickly adapted the skillsets as an independent researcher and a bright future awaits her. I thank Alibek for allowing me to take part in his research and for being a company during those late-night research at the lab.

I would not be where I am today without the love and support of my family and friends. I am grateful to my parents for the sacrifices they have made and for their never-ending support that have greatly facilitated everyday matters in which I am surprisingly inept in. I am grateful to my treasured friends for their encouragements and support outside of the lab and for being there with me in the best and worst of times. It is with special consideration that I mention Jordan (Joran) Silke, a true brother in bond and my most indispensable ally, the most brilliant student in the Xia lab, and the one other person who had made this thesis possible. Through countless challenges and adversities that we together faced, Jordan truly fits the description "for whoever sheds his blood with me today shall be my brother", if that is still a thing, albeit less literal, in this modern times.

Staying in Ottawa for my PhD was one of my best decisions in life, for I am blessed with the fortune of meeting two invaluable individuals, my supervisor Dr. Xia and my friend Jordan.

Finally, funding for this research was provided by NSERC Discovery Grant to Dr. Xuhua Xia, Ontario Graduate Scholarship (2018-2019), NSERC Alexander Graham Bell Scholarship (2019-2021), and Denis Vézina Scholarship (2020) to Yulong Wei.

Table of Contents

List of Figures and Tables

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

List of Abbreviations

A – Adenosine

C – Cytosine

G – Guanine

T – Thymine

U – Uracil

N – A, C, G, or T/U

nt – nucleotides

indel – insertion or deletion

SNP – Single Nucleotide Polymorphism

3’ tail – 3’ terminal sequence at the 16S rRNA

SD – Shine Dalgarno

UTR – untranslated region

rRNA – ribosomal RNA

mRNA – messenger RNA

tRNA – transfer RNA

cDNA – complementary DNA

ORF – Open Reading Frame

HEG – Highly expressed genes

LEG – Lowly expressed genes

SARS – Severe Acute Respiratory Syndrome

MERS – Middle East Respiratory Syndrome

CoV – Coronavirus

HIV – Human Immunodeficiency Virus

SIV – Simian Immunodeficiency Virus

MLV – Murine Leukemia Virus

MHV – Mouse Hepatitis Virus

HEV – Hepatitis E Virus

CRCoV – Canine Respiratory Coronavirus

RF1, 2, 3 – Release Factor 1, 2, 3

GTPase – Guanosine triphosphate hydrolase

ACE2 – Angiotensin Converting Enzyme 2

AVP – Antiviral proteins

ZAP – Zinc finger Antiviral Protein

APOBEC3 – Apolipoprotein B Editing Complex 3

AID – Action-Induced cytidine Deaminase

ADAR – Adenosine Deaminase Acting on RNA

ZnF1, 2, 3, 4 – Zinc Finger 1, 2, 3, 4

AlkB – Alpha-ketoglutarate-dependent dioxygenase B

NCBI – National Center for Biotechnology Information

CNCB – China National Center for Bioinformation

GEO Database – Gene Expression Omnibus Database

PaxDB – Protein Abundance Across Organisms Database

GtRNAdb – Genomic tRNA Database

SRA – Sequence Read Archive

SRX – SRA Experiment

GTEx – Genotype-Tissue Expression

DAMBE – Data Analysis in Molecular Biology and Evolution

ARSDA – Analyzing RNA-Seq Data

BLAST – Basic Local Alignment Search Tool

MAFFT – Multiple Alignment using Fast Fourier Transform

PHYML – Phylogeny reconstruction based on Maximum Likelihood principle

tAI – tRNA Adaptation Index

Fop – Frequency of Optimal Codons

CBI – Codon Bias Index

$D_{toStart}$ – Distance to Start codon

iTOL – Interactive Tree Of Life

RNA-Seq – RNA Sequencing

Ribo-Seq – Ribosome Sequencing

RSCU – Relative Synonymous Codon Usage

RTU – Relative tRNA Usage

CAI – Codon Adaptation Index

ITE – Index of Translation Elongation

FPKM – Fragments Per Kilobase of transcript per Million mapped reads

TPM – Transcripts Per Million

FIU – Fluorescence Intensity Unit

MFE – Minimum Folding Energy

RPT – Ribosome per transcript

RSCO – Relative Synonymous Codon Occupancy

Edecoding –Efficiency of decoding

PME – Proportion of mRNA expression

COD – Commonness Of Detection

ICpG – Index of CpG deficiency

MRCA – Most recent common ancestor

LoM – Lung-only Mice

List of Publications

Publications related to my Thesis Chapters:

1. Wei Y, Aris P, Farookhi H, Xia X. 2021. Predicting mammalian species at risk of being infected by SARS-CoV-2 from an ACE2 perspective. Scientific Reports 11:1702.
2. Wei Y, Silke JR, Aris P, Xia X. 2020. Coronavirus genomes carry the signatures of their habitats. PLOS ONE 15(12): e0244025.
3. Wei Y, Xia X. 2019. Unique Shine-Dalgarno Sequences in Cyanobacteria and Chloroplasts Reveal Evolutionary Differences in Their Translation Initiation. Genome Biology and Evolution 11(11): 3194-3206.
4. Wei Y, Silke JR, Xia X. 2019. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Scientific Reports 9:3184.

Other publications at the Xia lab:

1. Kruglikov A, Rakesh M, Wei Y, Xia X. 2021. Applications of protein secondary structure algorithms in SARS-CoV-2 research. Journal of Proteome Research 20:1457-1463.
2. Tehfe A, Roseshter T, Wei Y, Xia X. 2021. Does Saccharomyces cerevisiae require specific post-translational silencing against leaky translation of Hac1up? Microorganisms 9(3), 620.
3. Silke JR, Wei Y, Xia X. 2018. RNA-Seq-based analysis reveals heterogeneity in mature 16S rRNA 3' termini and extended anti-Shine-Dalgarno motifs in bacterial species. G3: Genes, Genomes, Genetics 8(12):3973-3979.
4. Wei Y, Silke JR, 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 7:17639.
5. Wei Y, Xia X. 2017. The Role of +4U as an Extended Translation Termination Signal in Bacteria. Genetics 205:439-549.
6. Wei Y, Wang J, Xia X. 2016. Coevolution between stop codon usage and release factors in bacterial species. Molecular Biology and Evolution 33:2357-2367.

CHAPTER ONE

Introduction

1.1 General introduction on bacterial translation

Rapid bacterial growth and replication depend much on efficient translation. Determining gene features that optimize translation not only advances our knowledge of molecular biology and evolution in the battle against pathogens, but also improves the efficacy of recombinant DNA in transgenic biosynthesis. To improve bacterial translation efficiency, protein-coding genes coevolve with the translation machinery. Specifically, coevolution acts on 1) initiation, between the Shine-Dalgarno (SD) sequence in the mRNA translation initiation region and the anti-SD (anti-SD) sequence at the 3' end of 16S rRNA, 2) elongation, between codon usage and tRNA availability, and 3) termination, between stop codon usage and release factor abundance. All these processes require interactions between gene features and their decoders.

Nonetheless, much about the relationships between the involved players remain unclear. While a strong SD/anti-SD pairing potential is believed to facilitate recruitment of ribosome near the start codon (Schurr, Nadir, Margalit 1993; Ma, Campbell, Karlin 2002; Starmer et al. 2006a; Lim, Furuta, Kobayashi 2012), others suggest that strong SD/anti-SD pairing may stall the ribosome during translation (Osterman et al. 2013; Li 2015; Hockenberry et al. 2017). While some suggest tRNA availability is the main driving force of codon usage bias (dos Reis, Savva, Wernisch 2004; Novoa et al. 2012), others contend that the two are weakly correlated (Sharp et al. 2005; Rojas et al. 2018). These inconsistencies can be traced to both intrinsic differences between bacterial organisms and extrinsic limitations in our approaches to pinpoint optimal interactions. For example, knowing what constitutes the full extent of the anti-SD sequence is crucial in

characterizing optimal SD/anti-SD pairs, but the 16S rRNA 3' end is often mis-characterized in bacteria. Furthermore, tRNA gene copy number as a commonly used proxy poorly estimates and poorly distinguishes tRNA content in most bacteria. More vigorous methods are required to detect, quantify, and model the complex interactions between gene features and their decoders during bacterial translation.

In subsections of Introduction 1.1, I summarize the mechanistic interactions between gene features and the translation machinery at all three translation steps, discuss the limitations in current approaches and enduring questions on the coevolution between involved players, and lay out the conceptual frameworks behind newly devised computational approaches to address those issues. Chapters 2 and 3 describe my research investigation in translation initiation and elongation, respectively, and Chapter 4 integrates those concepts to explore how presence or absence of optimal translational features affect ribosome dynamics during bacterial translation.

1.1.1 Translation initiation

Bacterial translation is often rate limited by initiation (Andersson, G Kurland 1983; Kudla et al. 2009; Duval et al. 2015) and an mRNA with efficient translation initiation possesses two features. First, the minimum requirement is that the start codon is accessible (Nakamoto 2006) and there is no stable secondary structure to embed the start codon (Osterman et al. 2013). Second, presence of a Shine-Dalgarno (SD) sequence is prominently found in the 5' untranslated region (5' UTR) of mRNAs (Nakagawa, Niimura, Gojobori 2017), which pairs with the anti-SD sequence that is located at the structure-free 3' terminus of mature 16S rRNA.

Focusing on the SD sequence, a good SD/anti-SD pairing that enhances translation initiation efficiency needs to meet two criteria (Xia 2018). The first involves SD/anti-SD pairing position because this may serve as a major determinant for a proper juxtaposition between start codon and ribosomal P site. A SD motif does not increase translation efficiency if it is not located at a proper distance away from the start codon (Hirose, Sugiura 1996; Hirose, Kusumegi, Sugiura 1998). Many researchers define a distance between the start or end of SD sequence and the start codon (Hirose, Kusumegi, Sugiura 1998; Ma, Campbell, Karlin 2002; Starmer et al. 2006b), but this is sometimes inadequate. For example, Figure 1.1a lists six genes whose proteins are abundantly produced in *Escherichia coli*, together with their SD/anti-SD pairing. One would expect that highly expressed genes would carry optimal SD sequences, and these SD sequences can take different forms and bind to different anti-SD sites on the 3' end of 16S rRNA (hereafter referred to as 3' tail). However, as illustrated in Figure 1.1b, different SD sequences can be close to or far from the start codon but have the same distance from the 3' tail to start codon. We recently proposed this improved measurement coined as $D_{toStart}$ (Prabhakaran, Chithambaram, Xia 2015; Abolbaghaei, Silke, Xia 2017; Wei, Silke, Xia 2017) because the objective of SD/anti-SD pairing is to position the start codon at the ribosomal P site to pair with the initiation tRNA. If $D_{toStart}$ is important, then we expect it to be constrained within a narrow range, which is true in *E. coli* (Figure 1.1c) (Wei, Silke, Xia 2017). Incidentally, this suggests that the use of $D_{toStart}$ to characterize the position of SD/anti-SD pairing is more informative than the traditional approach noting the distance between the SD sequence and the start codon, because the former considers the position of the ribosome relative to the start codon and the latter does not.

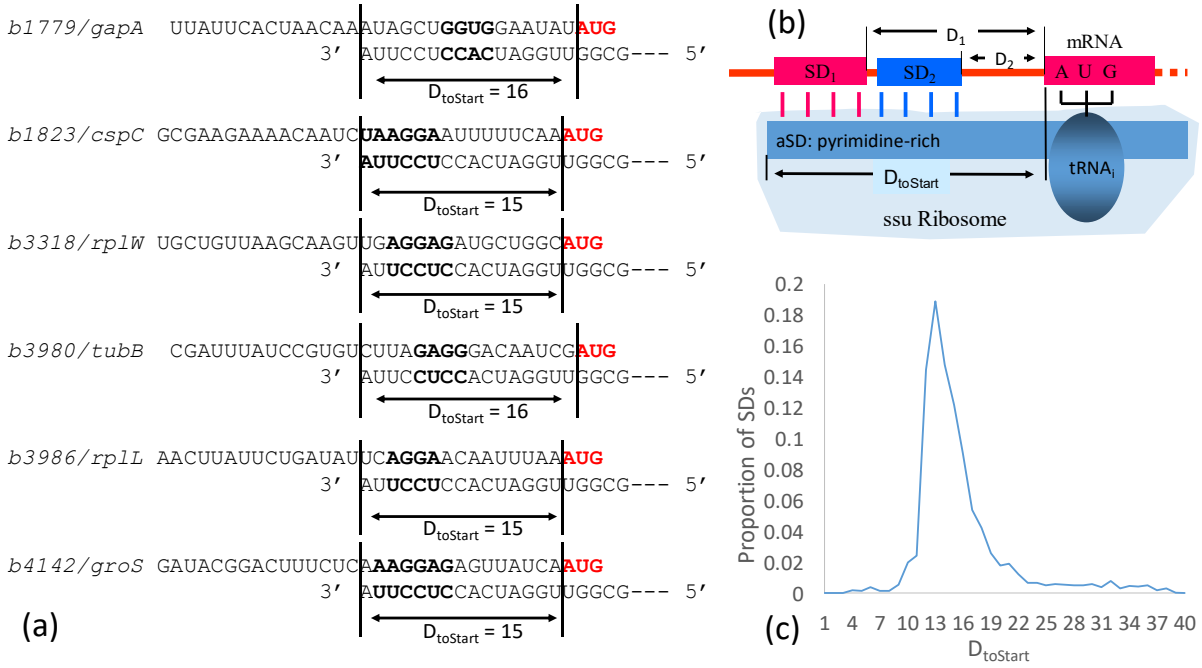


Figure 1.1. A graphic model of SD/anti-SD interactions in *E. coli*, partially redrawn from Wei, Silke, Xia (2017). (a) SD/anti-SD interaction in six genes whose proteins are abundantly produced in *E. coli*, showing that SDs (bold) can differ much from each other and in distance from start codon (red), but they have similar $D_{toStart}$. (b) Illustration of $D_{toStart}$ (number of nucleotides from the 3' end of 16S rRNA to the nucleotide immediately upstream of the start codon), showing that SDs can be close to or far from the start codon but have the same $D_{toStart}$. (c) $D_{toStart}$ is strongly constrained within a narrow range when putative SD sequences from all *E. coli* genes were considered.

The second criterion involves SD/anti-SD binding strength. Studies often considered the anti-SD to consist of the canonical CCUCCU motif because it is conserved across bacterial lineages (Nakagawa, Niimura, Gojobori 2017) and has high pairing potential (Schurr, Nadir, Margalit 1993). However, this focus on CCUCCU has excluded many other possible SD sequences. Recent studies suggest that SD/anti-SD pairing at this canonical motif may not be the most

preferred in bacteria (Wei, Silke, Xia 2017), and intermediate SD/anti-SD complementarity at flanking sites improves bacterial gene expression (Hockenberry et al. 2017; Scharff et al. 2017; Wei, Silke, Xia 2017). This is because strong SD/anti-SD pairing at CCUCCU may lead to ribosome stalling (Li, Oh, Weissman 2012; Zoschke, Bock 2018) that impedes the transition from initiation phase to elongation phase (Hockenberry et al. 2017), whereas weaker binding facilitates ribosome recruitment but does not impede ribosome movement.

It is crucial to know what constitutes the full extent of the 3' tail, as it is not only an inherent requirement to measuring SD/anti-SD pairing distance using $D_{toStart}$ but is also required to determine the complete pool of possible SD sequences in order to properly determine the relationship between motif preference and pairing affinity. Nonetheless, the 3' tail is often mis-characterized in bacteria. Take *Bacillus subtilis* for example, the 3' tail sequence was first reported as 5'-CCUCCUUUCU-3' (Shine, Dalgarno 1975; Murray, Rabinowitz 1982). Since then, alternative rDNA annotations of the *B. subtilis* 3' TAIL have emerged, including 5'-CCUCCUUUCUA-3' (NC_000964) (Barbe et al. 2009) and 5'-CCUCCUUUCUAA-3' (NZ_CP010052) which have been used in recent studies on *B. subtilis* 16S rRNA (Sohmen et al. 2015; Hockenberry et al. 2017). Discrepancies in these reported 3' tails arose because multiple exoribonucleases, including *RNase II*, *RNase R, PNPase*, *RNase PH* (Sulthana, Deutscher 2013), and *YbeY* (Jacob et al. 2013), independently participate in the maturation process of the 16S rRNA 3' terminus through unknown mechanisms. Consequently, determination of the 16S rRNA is frequently automated based on sequence similarity (Lin *et al.* 200s8; Nakagawa *et al.* 2010). However, this process is often unreliable (Starmer *et al.* 2006; Jones *et al.* 2007; Lagesen *et al.* 2007; Lin *et al.* 2008) and many such 16S ribosomal RNA sequence annotations have been discontinued in NCBI's Gene database (https://www.ncbi.nlm.nih.gov/gene/).

To determine the mature 3' tail in bacteria, we employed an RNA Sequencing mapping approach (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018), and the methodological detail of our approach is described in Chapter 2. A limitation of mapping rRNA using RNA-Seq datasets is that most experiments remove rRNAs prior to sequencing (O'Neil *et al.* 2013) in an effort to enrich target mRNAs (Choi and Hagedorn 2003). However, a test in *Lactococcus lactis* (Figure1.2) shows that ribo-depletion is often incomplete, and enough 16S rRNA reads will persist to allow for 3' tail characterization. Predicated on the availability of usable RNA-Seq datasets in NCBI's GEO (Edgar *et al.* 2002) database, we determined the 16S rRNA 3' tail in 16 bacterial species and 2 chloroplast species (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018; Wei, Xia 2019). To ensure the fidelity of our method, Table 1.1 shows that our approach recovers the 5'-GAUCACCUCCUUA-3' in *E. coli* and 5'-GAUCACCUCCUUUCU-3' in *B. subtilis* that were first reported and experimentally verified (Shine, Dalgarno 1974; Woese et al. 1975). In brief, we detected and corrected NCBI mis-annotated 16S rRNA 3' terminal sequences in 14 out of 18 species surveyed.

Table 1.1. The RNA-Seq corrected 3' tail in 16 bacterial and 2 chloroplast species. Highlighted red are species with mis-annotated 3' tail in NCBI. In blue the conserved 5'-GAUCA-3' motif, red the canonical anti-SD CCUCC, and green the extended 3' terminal sequence. Shaded nucleotide sequences are 3' tails determined through RNA-Seq mapping; unshaded and underlined nucleotides are extended and missing nucleotides in NCBI database, respectively. Results adapted and combined from Wei, Silke, Xia (2017), Silke, Wei, Xia (2018), and Wei, Xia (2019).

Species	16S 3' TAIL	NCBI accession
Escherichia coli	GAUCACCUCCUUA	NC_000913
Bacillus subtilis	GAUCACCUCCUUUCUAA	NC_000963
Listeria monocytogenes	GAUCACCUCCUUUCU	NC_003210
Streptococcus pyogenes	GAUCACCUCCUUUCU	NC_002737
Lactococcus lactis	GAUCACCUCCUUUC	NC_002662
Bacillus anthracis	GAUCACCUCC	NC_005945
Neisseria meningitidis	GAUCACCUCCUUUCU**A**	NC_003112
Clampylobacter jejuni	GAUCACCUCCUUUC	NC_002163
Deinococcus deserti	GAUCACCUCCUUUCUA	NC_012526
Mycoplasma pneumoniae	GAUCACCUCCUUUCUAAUGGAG	NC_017504
Salmonella enterica	GAUCACCUCCUUA	NC_003198
Legionella pneumophila	GAUCACCUCC	NC_002942
Desulfovibrio vulgaris	GAUCACCUCCUU	NC_002937
Leptospira interrogans	GAACACCUCCUUUUUAAGGAG	NC_005823
***Synechocystis* sp.**	GAUCACCUCCUUUAAGGG	NC_000911
Microcystis aeruginosa	GAUCACCUCCUUA	NC_010296
***Nicotiana tabacum* chloroplast**	GAUCACCUCCUUU	Z00044
***Arabidopsis thaliana* chloroplast**	GAUCACCUCCUUUCAG	NC_000932

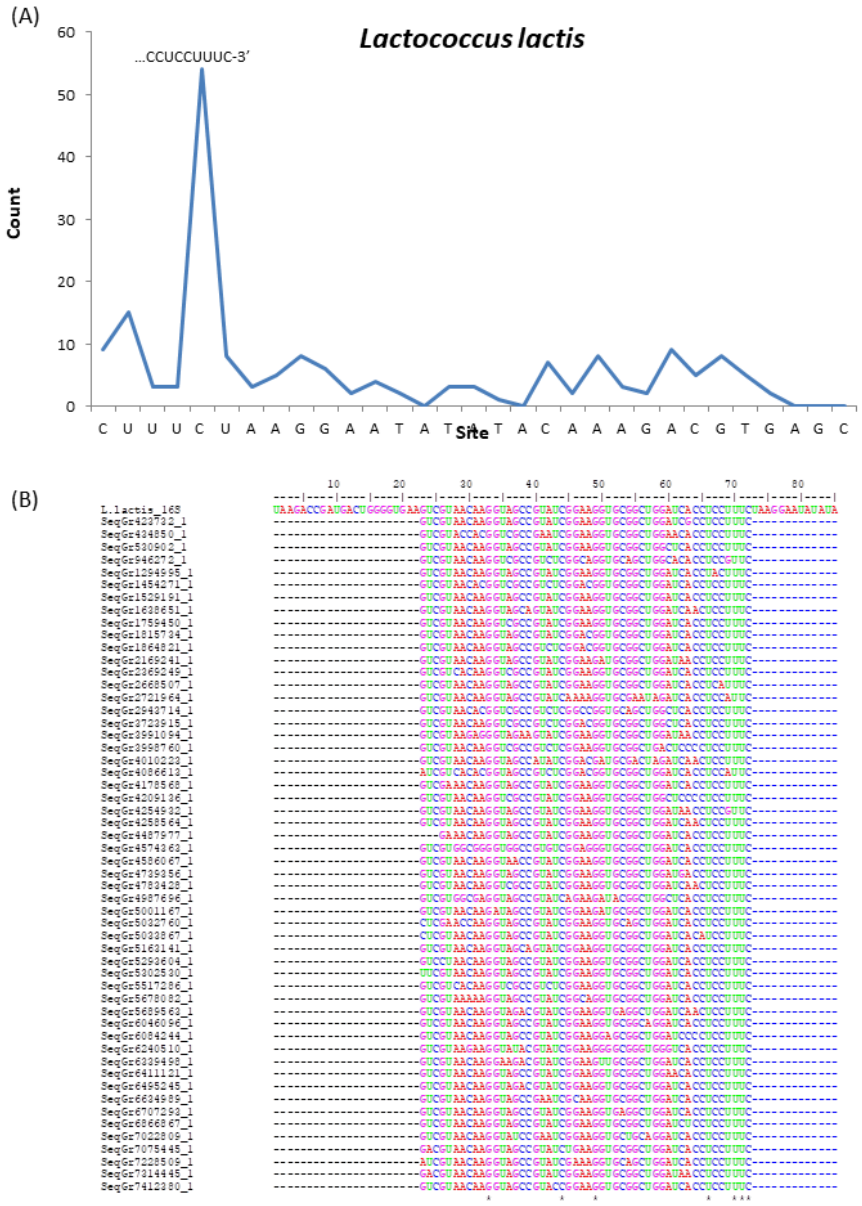


Figure 1.2. The count of RNA-Seq reads mapped to the 3' tail genomic query for *L. lactis*. (A) Mapped regions start with the last C of the canonical CCUCC motif as the first site, extended by 30 nt downstream. The single map peak represents the likely mature 3' tail in *L. lactis*: 5'-CCUCCUUUC-3'. (B) The sequence alignment between RNA-Seq reads and partial 16S rRNA genomic query in *L. lactis*. The complete length of the genomic sequence query is 205 nt long (100nt flanking the CCUCC motif on both sides). Figure adapted from Silke, Wei, Xia (2018).

Once the 3' tail (constituting the anti-SD) sequence is determined, it is important to determine the preferred binding motifs. Taking again the *B. subtilis* 5'-GAUCACCUCCUUUCU-3' as example, this 3' tail sequence extends the canonical CCUCCU motif by 4 nt. Is the 5'-UUCU-3' segment involved in pairing with SD sequences? What about the 5'-GAUCA-3' upstream of the CCUCCU motif that is also highly conserved by bacterial species (Nakagawa, Niimura, Gojobori 2017)? Assuming a putative SD sequence has no preference, then bases toward the middle of the anti-SD sequence are more predisposed to pairing with SD sequences than those towards the ends, as illustrated in Figure 1.3. Since CCUCCU constitutes the middle segment of most structure-free 3' tails, e.g., 5'-GAUCACCUCCUUA-3' (Shine, Dalgarno 1974) and 5'-GAUCACCUCCUUUCU-3' (Murray, Rabinowitz 1982) in *E. coli* and *B. subtilis*, respectively, it may be unsurprising that CCUCCU is preferred in SD pairing. Hence, to determine site preference in SD/anti-SD pairing, our $D_{toStart}$ model also calculates the observed and expected number of base pairing at each anti-SD site.

The observed value denotes the total number of times an anti-SD base is observed to pair with putative SD sequences; whereas the expected value represents the total number of times the anti-SD base is expected to pair with putative SD sequences, assuming each SD site is equally likely to be used by all SD sequences of lengths 4 nt to 12 nt. The expected number of SD/anti-SD binding at each anti-SD site is calculated independently. For example, the expected frequency at the first anti-SD site is calculated by the following equation, with N_m denoting N number of observed SD sequences of length m:

$$\sum_{m=4}^{12} \frac{N_m}{15 - m + 1}$$

while the expected frequency at the sixth anti-SD site is calculated by:

$$4 \times \frac{N_4}{12} + 5 \times \frac{N_5}{11} + 6 \times \frac{N_6}{10} + 6 \times \frac{N_7}{9} + 6 \times \frac{N_8}{8} + 6 \times \frac{N_9}{7} + 6 \times \frac{N_{10}}{6} + 6 \times \frac{N_{11}}{5} + 5 \times \frac{N_{12}}{4}.$$

These computations are implemented in DAMBE7 (Xia 2017) under the 'Analyze 5UTR' function.

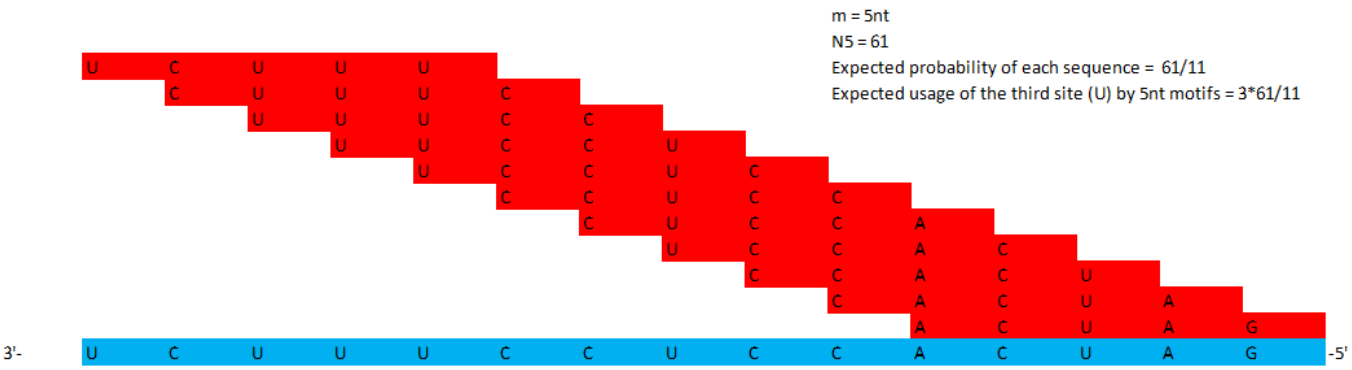


Figure 1.3. An example illustrating expected site-specific pairing at each anti-SD site by putative SD sequences (e.g. 61 SD sequences of length 5nt). Suppose each anti-SD site (blue) is equally likely to participate in SD/anti-SD binding with a SD sequence (red), assuming there is no site preference, then the expected number of anti-SD pairing is site-dependent.

An anti-SD site can be characterized as favourably selected if it pairs with putative SD sequences more frequently than expected, or has an observed to expected SD/anti-SD usage ratio (O:E) > 1. Indeed, the canonical CCUCCU motif is preferentially used in SD/anti-SD pairing. In addition, conserved regions downstream of CCUCCU also have an O:E > 1 in most bacterial species (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018), indicating that downstream bases are indeed retained because they are preferred in SD/anti-SD binding despite having weaker binding affinity than CCUCCU.

The application of above approaches is discussed in Chapter 2, where we used our $D_{toStart}$ model and RNA-Seq-based annotation for the 16S rRNA 3' terminus in a case study to reveal differences in the conservation of SD motifs between Cyanobacteria, cyanophage, and chloroplast protein-coding genes.

1.1.2 Translation elongation

The genetic code is degenerate, and amino acids are often encoded by many synonymous codons. In the bacterial codon table, most synonymous codons belong to 4-fold degeneracy families (e.g, GUN encoding Valine) or 2-fold degeneracy families (e.g., UU(U/C) encoding Phenylalanine) that differ only at the third codon site. Exceptions are Leucine and Serine encoded by six codons which can be grouped into a 2-fold family and a 4-fold family, Isoleucine encoded by a 3-fold family, and Methionine and Tryptophane each encoded by a single codon, AUG and UGG, respectively.

The term codon usage bias refers to the differential usage of synonymous codons, which can be measured by the relative synonymous codon usage (RSCU). RSCU normalizes codon frequency such that RSCU = 1 when codon usage is equal among synonymous codons. A codon is preferred when its RSCU > 1, and a codon is underused when its RSCU < 1. The general equation of RSCU is:

$$RSCU_i = \frac{\frac{codon\ i\ frequency}{\sum synonymous\ codon\ frequency}}{n}$$

Where *i* denotes a specific codon within a synonymous codon family, and *n* the number of synonymous codons within the family.

Two factors are known to affect synonymous codon usage, mutation bias (Muto, Osawa 1987; Xia 2005) and tRNA-mediated selection bias (Ikemura 1985; Bulmer 1987; Xia 1998; Carullo, Xia 2008). In terms of mutation bias, GC content affects the nucleotide composition of coding DNA regions. A species with low GC content is expected to prefer synonymous codons containing A and T over those containing C and G; for example, one expects to observe more AAA than AAG lysine codons in such a species. GC mutation bias is often measured at the third codon site (Palidwor, Perkins, Xia 2010) since it contributes the most to codon degeneracy and is therefore expected to be evolutionarily less constrained than the first and second codon sites. One can also observe strand asymmetry in nucleotide compositions in a wide spectrum of bacterial species, where the leading strand is often AG-rich and the lagging strand is often TC-rich (Rudner, Karkas, Chargaff 1968; Lobry, Sueoka 2002). AT skew is often more subtle than GC skew in bacteria, but it is still highly notable in Firmicutes species such as *Bacillus subtilis*, wherein the leading strand is A-rich and the lagging strand is T-rich (Xia 2012; Charneski, Hurst 2013).

In terms of tRNA-mediated selection bias, synonymous codons are often translated by cognate tRNAs that exist in different concentrations in the cell. It follows that for a bacterial organism to replicate rapidly, one way is to increase translation elongation efficiency such that genes preferentially encode synonymous codons that are decoded by the most abundant cognate tRNAs. This concept is conspicuous; if many tRNAs are available to translate glycine codons GGC and GGU but few tRNAs are available to translate glycine codons GGA and GGG, then translation elongation would be more efficient if genes encode more GGC and GGU. Indeed, a correlation between codon usage and tRNA availability was discovered as early as in the 1970s (Garel 1974; Chavancy et al. 1979), and this relationship holds well in *E. coli* (Ikemura 1985; Xia 1998),

especially at highly expressed genes (HEGs) including those that encode 30S and 50S ribosomal proteins (Sharp, Li 1987b; Guerdoux-Jamet et al. 1997).

Nonetheless, the general relationship between tRNA abundance and codon usage in bacteria has been the subject of debate among researchers, with some suggesting that tRNA availability is the main driving force of codon usage bias (dos Reis, Savva, Wernisch 2004; Novoa et al. 2012) and others contending that the two are weakly correlated (Sharp et al. 2005; Rojas et al. 2018). Indeed, many studies have tested the degree codon usage influences protein expression. Early studies in *E. coli* demonstrated that replacing rare codons with major codons increases protein yields (Robinson et al. 1984; Sorensen, Kurland, Pedersen 1989), wherein the term major codon was first coined by McPherson (1988) to describe synonymous codons that are more favourably selected in highly expressed genes (HEGs) than lowly expressed genes (LEGs). However, when Kudla et al. (2009) designed a synthetic library of 154 *E. coli* green fluorescent proteins that differ in synonymous codon usage, they found that elongation efficiency as estimated by the Codon Adaptation Index (CAI) (Sharp, Li 1987a) weakly influences protein expression. Despite this, more recent studies have recognized that codon choice during translation elongation affects gene expression when translation initiation is not rate-limiting (Tuller et al. 2010; Xia 2015).

To measure the degree tRNA-mediated selection shapes codon usage, tRNA availabilities have often been estimated by tRNA gene copy numbers and few experimental quantifications are available due to the difficulties in dealing with post-transcriptional modifications and in distinguishing between charged and non-charged tRNAs (discussed in more detail in Chapter 3). While there is a high correlation between relative tRNA abundance and tRNA gene copy number in the model organisms *E. coli* (Ikemura 1985), tRNA gene copy numbers is often an inadequate

approximation of tRNA abundance. Take for example the slow-growing *Leptospira interrogans* and *Mycobacterium tuberculosis*, which have only 25 and 45 annotated tRNA genes, respectively, according to the Genomic tRNA Database (GtRNAdb) (Chan, Lowe 2009). Obviously, variation in codon usage cannot be well explained by tRNA gene numbers if there is little redundancy, and the degree to which codon usage depends on tRNA abundance remains unclear in most bacterial species. It stands to reason that the coevolution between tRNA abundance and codon usage can be better characterized if tRNA abundance can be measured more accurately.

To better elucidate the relationship between tRNA abundance and codon usage, we established an RNA-Seq-based approach to quantify tRNA abundance (Wei, Silke, Xia 2019). This work is discussed in Chapter 3, and a preview of our results is shown in Figure 1.4, which illustrates the feasibility of our approach to 1) map and obtain tRNA sequence reads in bacteria despite the limited applications of RNA-Seq profiling in higher Eukaryotes where tRNAs are extensively methylated (Cozen et al. 2015; Zheng et al. 2015), and 2) precisely detect mature tRNAs that had undergone post-transcriptional addition of the 5'-CCA-3' tail (Hou 2010) when it is absent in the genomic DNA sequence.

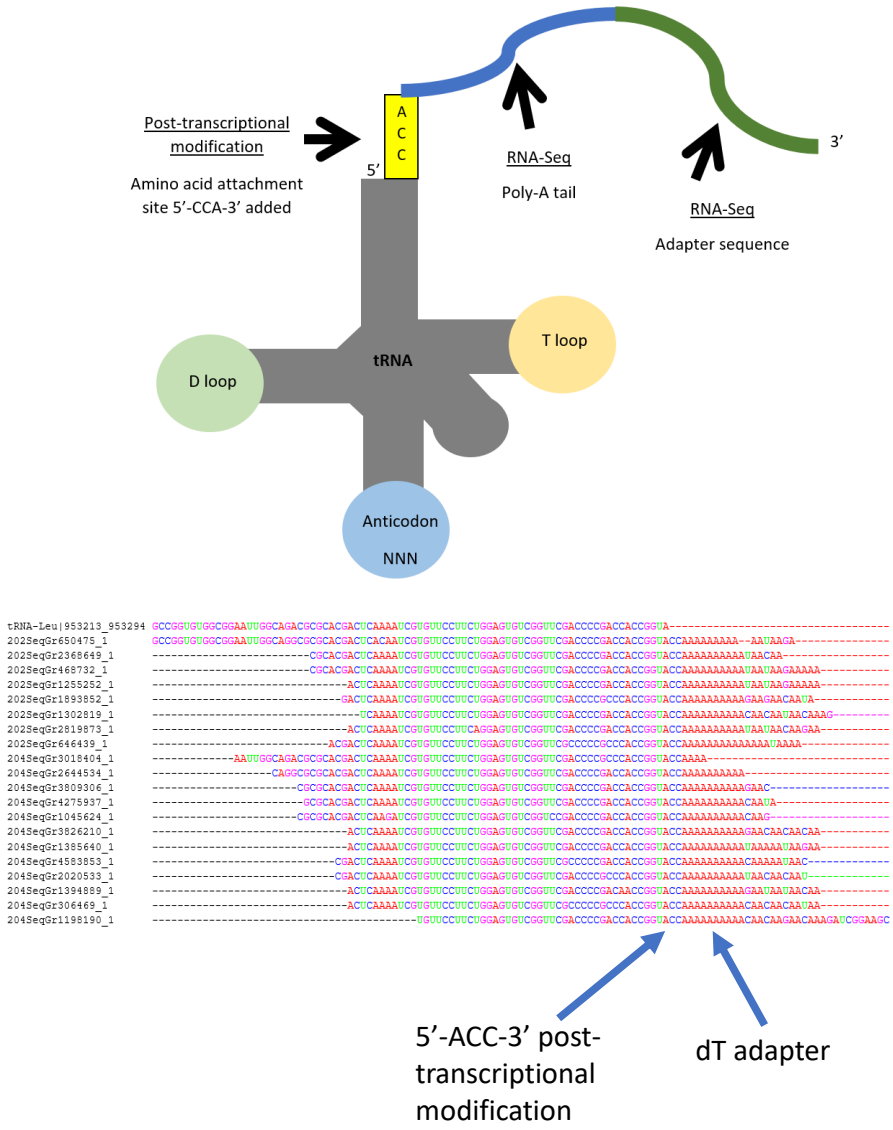


Figure 1.4. A schematic showing the feasibility of tRNA profiling using RNA-Seq data. The first entry in the alignment represents the annotated genomic sequence of a Leucine tRNA in *E. coli*, which is aligned by short RNA-Seq reads. The alignment shows that the whole tRNA sequence is sequenced in Prokaryotes albeit post-transcriptional methylation inhibits RNA sequencing in higher Eukaryotes. In addition, the appended 5'-ACC-3' tail upstream of the poly-U adapter sequence indicates that mature tRNAs can be profiled.

1.1.3 Translation termination

Three stop codons, UAA, UAG, and UGA, signal translation termination. The process takes place when a stop codon enters the 70S ribosomal A site and is recognized by a release factor. Class I release factors RF1 and RF2 are encoded by the *prfA* and *prfB* genes, respectively (Scolnick et al. 1968; Milman et al. 1969; Scolnick, Caskey 1969). RF1 recognizes stop codons UAA and UAG, whereas RF2 recognizes stop codons UAA and UGA. Class II release factor RF3 is a GTPase that facilitates the dissociation of RF1 and RF2 from the ribosome (Pavlov et al. 1997) but it is not directly involved in stop codon recognition. In comparison to studying sense codon bias, research on stop codon bias represents a simpler narrative where the effects of codon-decoder adaptation are well-traced. This section reiterates some key findings of my Master's research in order to 1) summarize coevolutionary forces at all three translation steps and 2) reinforce the concept of codon-decoder bias introduced in Chapter 1.1.2.

Just as sense codon usage depends on mutation bias and tRNA availability, stop codon usage can also be well-explained by mutation bias and relative abundances of RF1 and RF2 decoders (Supplemental Figure S1.1) (Wei, Wang, Xia 2016). The availability of protein abundance data for a number of bacterial species in PaxDB 4.0 (Wang et al. 2015) allowed us to measure the relative abundance of RF1 and RF2 and determine adaptation between stop codons and their decoders. Indeed, UAA is consistently the most preferred stop codon, especially in highly expressed genes (HEGs) relative to lowly expressed genes (LEGs), because it is decoded by both RF1 and RF2, whereas UGA usage relative to UAG usage increases significantly with relative abundance of RF2 (as measured by P_{RF2} = RF2/(RF1+RF2)) (Figure 1.5). More simply, species

with high abundance of RF1 use UAG more frequently while species with high abundance of RF2 use UGA more frequently.

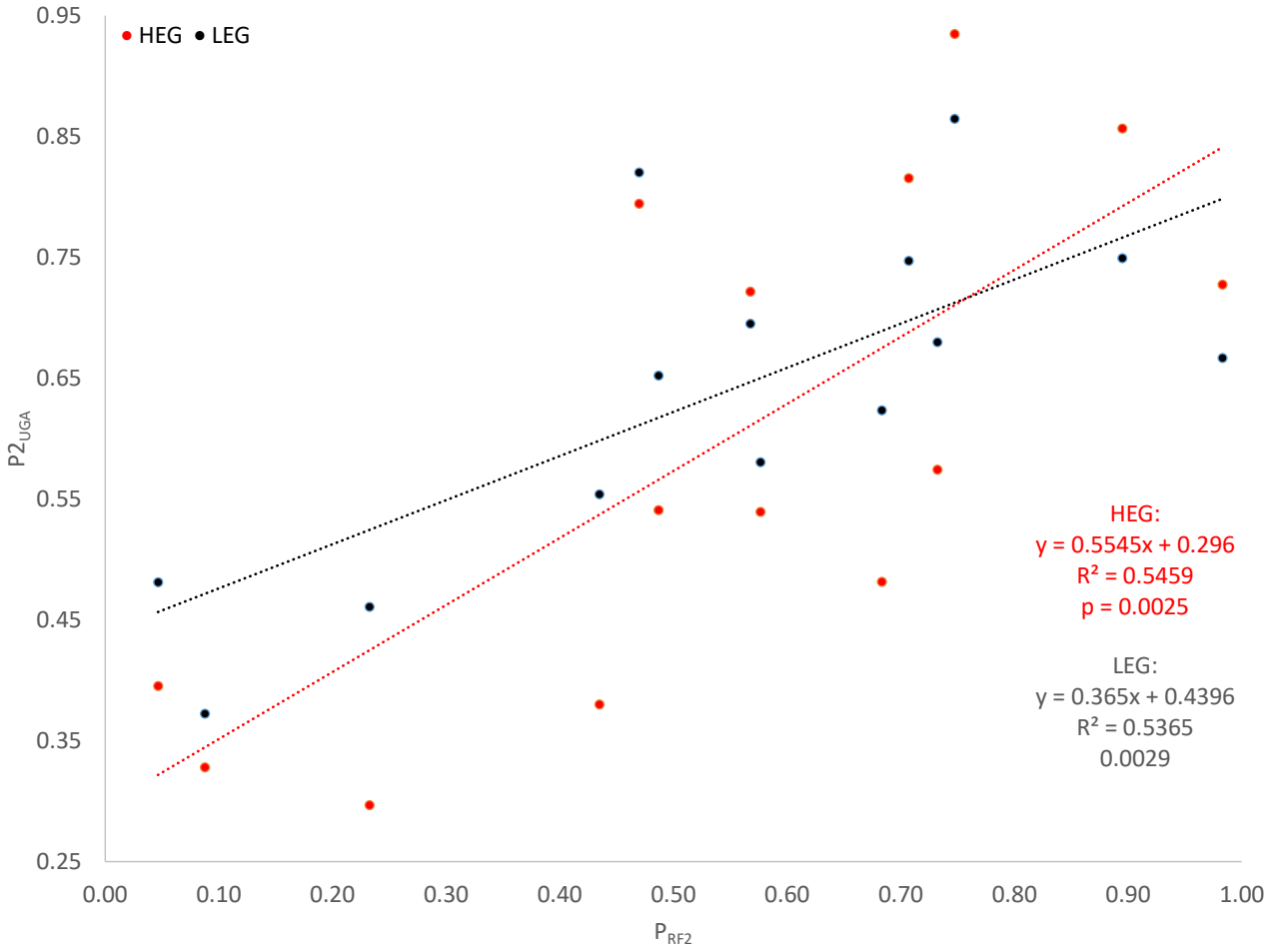


Figure 1.5. Usage of UGA relative to UAG, measured as $P2_{UGA}$ = UGA/(UGA+UAG), increases significantly with abundance of RF2 relative to RF1, measured as P_{RF2} = RF2/(RF1+RF2). Adapted from Wei, Wang, Xia (2016).

Furthermore, mutation bias as measured by genomic GC content also affects stop codon usage. Both UGA (Supplemental Figure S1.2) and RF2 (Figure 1.6) decrease in species with high AT content measured at the third codon site (AT3); this partially explains why bacterial lineages

with high AT contents such as *Mycoplasma* (Yamao et al. 1985) and *Spiroplasma* (Citti et al. 1992) often reassign UGA as a sense codon for Tryptophan.

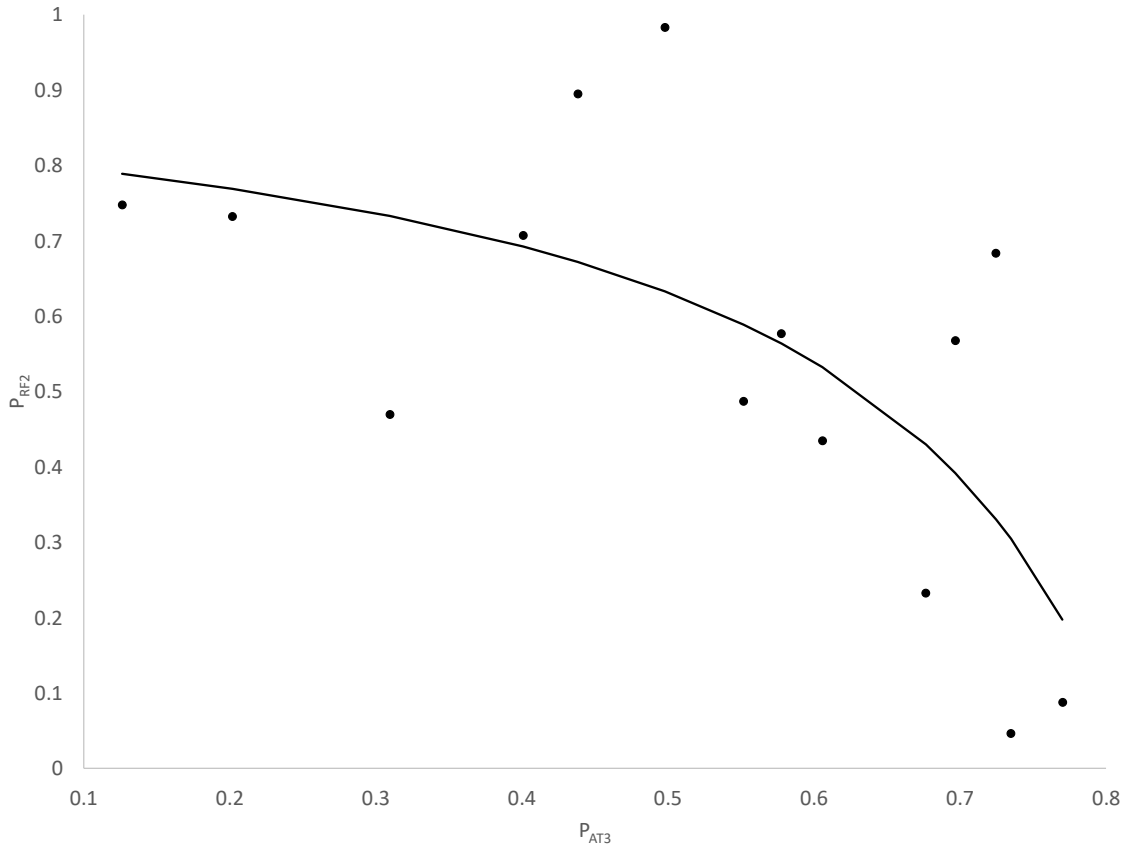


Figure 1.6. Relative abundance of RF2 decreases rapidly at high range of AT content measured by the AT proportions at the third codon site (P_{AT3}). Adapted from Wei, Wang, Xia (2016).

Nonetheless, what affects termination efficiency in bacteria remains unclear. Termination read-through occurs when a stop codon is misread as a sense codon by near-cognate tRNAs, and read-through levels vary between stop codons in both Prokaryotes and Eukaryotes (Parker 1989; Jørgensen et al. 1993; Tate et al. 1999; Dabrowski, Bukowy-Bieryllo, Zietkiewicz 2015). In addition, frameshifting can occur at the termination site, and a well-known example is the auto-regulatory translation of the *prfB* gene by programmed frameshift at a premature UGA (Craigen

et al. 1985; Craigen, Caskey 1986; Baranov, Gesteland, Atkins 2002). While UAA is the most efficient termination signal, it has been disputed whether UAG or UGA is the least efficient. Studies often found that UGA has the highest termination read-through in bacteria (Parker 1989; Jørgensen et al. 1993; Tate et al. 1999; Dabrowski, Bukowy-Bieryllo, Zietkiewicz 2015), but Povolotskaya et al. (2012) and Korkmaz et al. (2014) found that UAG is often the least used.

The complexity in termination efficiency is compounded when studies examined the interactions between release factors and termination signals using crosslinking experiments and found that presence of a +4U immediately downstream of stop codons increased binding affinity (Brown, Tate 1994; Tate et al. 1996; Poole et al. 1998). We had also showed that +4U was consistently over-represented in UAA-ending HEGs relative to LEGs, and +4U usage increased significantly in species having abundant UAA near-cognate tRNAs (Wei, Xia 2017). These results corroborated the running hypothesis that +4U is favourably selected to increase UAA termination efficiency by reducing misreading by near-cognate tRNAs (Bossi, Roth 1980; Miller, Albertini 1983; Tate et al. 1996; Poole et al. 1998). Nonetheless, surplus of +4U can be partially explained by the presence of a large number of overlapping genes with the **U**A**A**UG configuration (bold denotes stop codon of previous gene, underlined is the start codon of downstream gene) (Xia 2019).

1.2 General introduction on coronavirus genome evolution

Above I discussed that molecular evolution in microbes is partly explained by the coevolution between gene features and the translation machinery to replicate rapidly. In addition to this selection pressure, microbes evolve in adaptation to their environment such as acidity, temperature, starvation conditions. Focusing on the relationship between host and pathogens,

viruses as obligate parasites are ideal organisms to investigate how the host environments shape microbial genome evolution.

COVID-19 poses a serious global health emergency. Since its outset, the disease has prompted an immediate global response and research efforts to fight against the pandemic, and over 450,000 complete SARS-CoV-2 genomes have since been sequenced and publicly deposited (https://www.gisaid.org/, last accessed February 1, 2021). With this expansive amount of available data, the emergence of SARS-CoV-2 provides a rare window into investigating how the host cellular environment exerts both ephemeral and long-lasting selective pressures to shape the evolutionary genomics of RNA viruses.

As obligate parasites, coronaviruses such as SARS-CoV-2 evolve to carry genomic signatures in adaptation to their host-specific environments. Adaptation occurs at cell entry, where the viral Spike protein constantly evolves to bind efficiently to host ACE2 receptors. The binding potential between these two players is therefore a major determinant of coronavirus host range and provides crucial insights into the mechanism of host-switching. In addition, coronaviruses regularly infect host tissues that express antiviral proteins in abundance, and their genomes evolve to evade or adapt to the host innate immune response. Understanding the evolutionary genomics of SARS-CoV-2 will improve our ability to predict future pandemics, protect wildlife and domesticated animals, understand the origin of SARS viruses, and motivate designs for novel attenuated SARS-CoV-2 vaccines.

In subsections of Introduction 1.2, I discuss receptor binding of SARS-CoV-2 Spike protein and its application in vaccine design, how the human innate immune responses target foreign RNAs and shape viral genome evolution, and the computational frameworks to empirically

test these effects. Chapter 5 describes my research in coronavirus receptor binding, and Chapter 6 describes my research on coronavirus genome evolution under selective pressures exerted by the host innate immune response.

1.2.1 Coronavirus host-range and host-switching

How do coronaviruses evolve towards host-specificity? Coronaviruses encode a Spike protein that binds to host Angiotensin-converting enzyme 2 (ACE2) receptor to mediate cell entry. The Spike protein is encoded by a 1277 and a 1255 amino acid sequence in SARS-CoV-2 and SARS-CoV, respectively, and it is cleaved into two structural domains, S1 and S2, at a conserved 6 amino acid motif (IAYTMS; Figure 1.7) by cathepsin L in the endosome (Simmons et al. 2005; Belouzard, Chu, Whittaker 2009; Ou et al. 2020). In addition, the Spike protein in SARS-CoV-2 contains a unique furin cleavage site (PRRA) (Andersen et al. 2020; Coutard et al. 2020; Walls et al. 2020) that is absent in all known closely related viruses (Xia 2021). The S1 domain contains a signal peptide at the N terminus, an N-terminal domain, and a receptor-binding domain; whereas the S2 domain contains a fusion peptide domain, internal fusion peptide, two heptad-repeat domains, transmembrane domain, and a C-terminal domain (Xia 2021). Essentially, the S1 domain interacts with the ACE2 receptor (Tian et al. 2020; Walls et al. 2020), and the S2 domain undergoes structural rearrangements to mediate membrane fusion (Li et al. 2005).

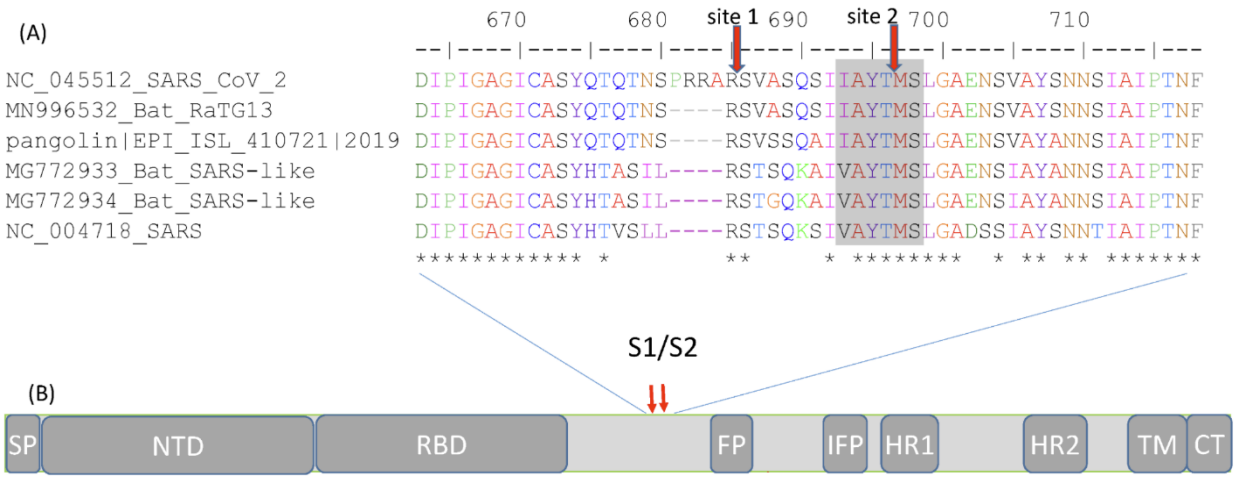


Figure 1.7. Cleavage sites at the S1/S2 boundary. (A) In grey is the 6 amino acid motif that is cleaved by cathepsin L; it is shared by the two SARS viruses and their close relatives. An additional insertion of PRRA that is unique to SARS-CoV-2 results in a new polybasic furin cleavage site. (B) Schematic domain structure of S protein and location of the two cleavage sites in SARS-CoV-2 (highlighted by the red arrows). Figure adapted from Xia (2021).

Focusing on the S1 receptor-binding domain that directly interacts with the host ACE2, differences between the 2002 SARS-CoV and the recently emerged SARS-CoV-2 at receptor contact amino acids (Lan et al. 2020; Shang et al. 2020) explain the distinct host range of the two SARS viruses (Wei et al. 2021) and why some antibodies are distinctively effective against each (Tian et al. ; Ou et al. 2020). Furthermore, a high binding potential between the Spike protein and the host ACE2 receptor is a key determinant of viral infectivity. For example, while human ACE2 binds efficiently to S proteins of SARS-CoV strain TOR2 that caused the severe 2002-2003 outbreak, the receptor does not bind efficiently to S proteins of Civet-CoV strain SZ3 isolated from palm civets or of the less severe 2003-2004 SARS-CoV strain GD (Guan et al. 2003; Marra et al. 2003; Rota et al. 2003; Song et al. 2005).

In fact, the S1 receptor-binding domain has been used extensively as an anti-SARS-CoV-2 drug in vaccine developments. This is because gaining immunity against the virus at the pre-fusion stage before the virus enters host cell is simply more effective than attempting to combat against infection post-fusion. Current vaccines such as Pfizer/BioNTech (BNT162b2) and Moderna (mRNA-1273) are mRNAs that encode a mutant Spike protein with two amino acid replacements (K986P, V987P), referred to as S-2P (Anderson et al. 2020; Jackson et al. 2020). The S-2P replacements are located at the transitional bend between HR1 and the central helix (Xia 2021) and have been previously demonstrated to strongly contribute to stabilize the translated Spike protein at the pre-fusion conformation in Betacoronavirus HKU1 (Kirchdoerfer et al. 2016) and MERS-CoV (Pallesen et al. 2017). Indeed, current SARS-CoV-2 vaccine strategies are focused on designing mRNAs that encode a modified Spike protein that is more stabilized in its pre-fusion conformation (Xia 2021).

While SARS-CoV-2 can transmit efficiently in humans, it is less clear which other mammals are at risk of being infected. Currently, SARS-CoV-2 infection has been detected in the cat (Halfmann et al. 2020), dog (Sit et al. 2020), mink (Oreshkova et al. 2020), and tiger (Wang et al. 2020). In addition, ferret (Kim et al. 2020; Shi et al. 2020), macaque (Chandrashekar et al. 2020; Rockx et al. 2020), and grivet (Woolsey et al. 2020) could be experimentally infected by SARS-CoV-2. Then, a straightforward prediction is that the ACE2 receptor of these known high-risk mammals should bear a high resemblance to human ACE2 at key Spike protein-binding sites. Moreover, two Rodentia species have been experimentally infected by SARS-CoV-2: the golden Syrian hamster (*Mesocricetus auratus*) belonging to the *Cricetidae* family, and the house mouse (*Mus musculus*) belonging to the *Muridae* family. While hamsters could be consistently infected by SARS-CoV-2 (Chan et al. 2020), wild-type mice were not susceptible to SARS-CoV-2

infection (Bao et al. 2020). It follows that these two Rodentia species should have distinct differences at the ACE2 receptor, and one expects the key binding sites at human ACE2 to be well conserved by hamster ACE2 but not by mouse ACE2.

A comparative sequence study between human ACE2 and mammalian ACE2s at key Spike protein-binding sites should therefore provide crucial empirical insights into identifying mammals at high risk of SARS-CoV-2 infection. To substantiate this narrative, one may test the prediction that ACE2 receptors of all mammals that are currently known to be infected by SARS-CoV-2 would share high similarity with human ACE2 at key Spike protein-binding sites, but that the ACE2 receptor of mouse which cannot be infected by SARS-CoV-2 would not share high similarities with human ACE2 at key S protein-binding sites.

1.2.2 Coronavirus adaptation to host innate immune responses

How do coronaviruses evade or adapt to host immune responses once they have entered host cells? SARS-CoV-2 regularly infects bronchiolar and type II alveolar epithelial cells in the lungs (Yao et al.) and enterocytes in the small intestines (Lamers et al. 2020). At the tissue level, hosts provide different cellular environments with varying levels of antiviral activity. These mammalian cells possess a variety of mechanisms to detect foreign DNA, RNA, and proteins which activate the innate and adaptive immune response against infection. Focusing on the innate immune response, two antiviral proteins (AVPs) that may contribute to the modification of viral genomes are the zinc finger antiviral protein (ZAP, gene name ZC3HAV1 in mammals) and the apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3 (APOBEC3) protein, both of which exhibit tissue-specific expressions (Fagerberg et al. 2014).

ZAP is a key component in the mammalian interferon-mediated immune response that specifically targets CpG dinucleotides in viral RNA genomes (Meagher et al. 2019) to inhibit viral replication and signal for viral genome degradation (Guo et al. 2007; Takata et al. 2017; Meagher et al. 2019; Ficarelli et al. 2020). ZAP acts against retroviruses such as HIV-1 (Zhu et al. 2011; Ficarelli et al. 2020) and single-stranded RNA viruses such as Echovirus 7 (Odon et al. 2019), Zika virus (Trus et al. 2020), and Influenza virus (Greenbaum et al. 2008). It follows that cytoplasmic ZAP activity should impose a strong CpG avoidance in RNA viruses that infect tissues abundant in ZAP. For instance, while HIV-1 infects lymph organs where ZAP is abundant (Fagerberg et al. 2014), its genome is also strongly CpG-deficient, and the viral fitness of HIV-1 diminishes as its genomic CpG content increases within a sample of patients (Theys et al. 2018). Indeed, many single-stranded RNA viruses exhibit strong CpG deficiency (Yap, Zhang, Danchin 2003; Greenbaum et al. 2008; Greenbaum, Rabadan, Levine 2009; Atkinson et al. 2014; Takata et al. 2017), but selection for CpG deficiency disappears in ZAP-deficient cells (Takata et al. 2017).

Mechanistically, ZAP binds selectively to foreign CG-rich RNA with a structurally defined CG-binding pocket containing 4 CCCH zinc fingers (ZnF1 to ZnF4) in its N-terminal RNA-binding domain (Chen et al. 2012; Meagher et al. 2019) to recruit RNA degradation factors such as the KHNYN endonuclease and the RNA exosomes (Meagher et al. 2019; Luo et al. 2020). In this interaction, both high binding affinity (Meagher et al. 2019) and CG dinucleotide targets in specific CG-rich ssRNA contexts (Luo et al. 2020) are important to ZAP activity and its ability to distinguish between self and non-self RNAs. Furthermore, ZAP-mediated RNA degradation is cumulative (Takata et al. 2017). When CpG dinucleotides were experimentally added to individual viral segments, the inhibitory effect of ZAP was weak. However, when the same CpG

dinucleotides were added to both segments, the ZAP inhibition effect was strong (Takata et al. 2017). This implies that only mRNA sequences of sufficient length would be targeted by ZAP.

Aside from ZAP, the APOBEC3 protein family (A3A, A3B, A3C, A3D, A3F, A3G, and A3H) have garnered substantial attention for their role in the antiviral immune response (Cullen 2006; Harris, Dudley 2015). Through a mechanism largely derived from HIV-1 studies, APOBEC3 enzymes have been prominently reported to restrict viral infectivity (Chiu, Greene 2008; Nabel et al. 2013; Rodriguez-Frias et al. 2013; Harris, Dudley 2015; Hayward et al. 2018) by editing C to T at the viral genomes, and HIV-1 avoids this deleterious effect by expressing Vif to target and degrade APOBEC3 enzymes (Sheehy et al. 2002; Wang, Wang 2009). Additionally, C to U RNA-editing has been demonstrated in lymphocytes, macrophages, monocytes, and natural killer cells by both A3A and A3G in response to hypoxia and interferons (Sharma et al. 2015; Sharma et al. 2016; Sharma et al. 2019), and recent studies propose that APOBEC3 enzymes may act directly to edit single-stranded RNA coronaviruses (Di Giorgio et al. 2020; Simmonds 2020; Victorovich et al. 2020). If APOBEC3 could edit RNA viruses, then this immune response could potentially restrict coronaviruses since they do not encode a Vif analogue to degrade APOBEC3 enzymes.

Unlike the cytoplasmic ZAP, APOBEC3 enzymes are mainly expressed in hematopoietic cell populations, including T cells, B cells, and myeloid cells (Koning et al. 2009). Consequently, APOBEC3 enzymes are highly expressed in lymphoid organs including the thymus, spleen, and lymph nodes (Koning et al. 2009; Refsland et al. 2010); however, APOBEC3 expression is not confined to lymphoid organs. Two studies have detected APOBEC3-encoding mRNAs from total RNA of non-lymphoid tissues (Koning et al. 2009; Refsland et al. 2010), and they suggested that

APOBEC3 enzymes are variably expressed in these tissues due to differing lymphocyte contents. For example, a considerable number of APOBEC3-expressing macrophages reside in the lung alveoli, and expectedly, both Koning et al. (2009) and Refsland et al. (2010) found the highest levels of APOBEC3 expression in lung tissues among non-lymphoid tissues.

Mechanistically, APOBEC3 enzymes have negligible catalytic activity prior to infection, but expressions of associated mRNAs are induced in response to cell stresses including hypoxia, cell crowding, and presence of interferon-α (Koning et al. 2009; Sharma et al. 2015; Sharma et al. 2019). Much about the specific protein-RNA interactions remain to be elucidated. Nevertheless, through extensive studies in binding affinity, editing activity, and secondary structural predictions, in HIV, MLV, and SLV viruses, and in *in vitro* peptide mutagenesis, APOBEC3 enzymes are known to preferentially edit C in specific motif contexts and structures. The specific motif contexts of C to U editing for each member of the APOBEC3 protein family is distinct and too lengthy to review here; those details are reviewed extensively in Chapter 6.

1.3 Thesis Objectives and Significance

How do bacterial and phage genes evolve to optimize translation under distinct host environments? How do viruses adapt to their hosts and what triggers host-switching? This thesis focuses on elucidating the principles governing microbial genome evolution. Leveraging Bioinformatics tools and evolutionary models, I perform large-scale comparative genomics, transcriptomics, and proteomics studies. My research is to understand two key thematic components of microbial genome evolution: 1) how gene features and translation machinery coevolve in bacteria, and 2) the distinct genomic signatures that coronaviruses carry in adaptation to their habitats. Answering these questions will not only advance our knowledge of microbial

molecular biology and evolution but will also motivate new designs in phage therapy and mRNA vaccines against viruses.

Rapid bacterial replication depends on efficient translation. To improve translation efficiency, protein-coding genes coevolve with the translation machinery. I study how coevolution acts on 1) translation initiation, between the Shine-Dalgarno (SD) sequence and the anti-Shine-Dalgarno (anti-SD) sequence, and on 2) translation elongation, between codon usage and tRNA abundance. The bioinformatics processes underlying these research investigations require 1) accurate and efficient approaches to handle and process large DNA-Seq and RNA-Seq datasets, and 2) methods to detect and measure coevolution between microbial gene features and their decoders. The Xia lab had recently developed the software ARSDA with aim to resolve current limitations in the processing of large sequencing datasets. Using custom software with tailor-made functions, we devised RNA-Seq-based approaches to reliably characterize and annotate 16S rRNA and quantify tRNA abundance in bacteria.

In Chapter 2, we discussed the application of the $D_{toStart}$ model described in Chapter 1.1.1 to determine optimal SD/anti-SD pairing, the RNA-Seq approach to determine 16S rRNA 3' tail, and the O:E statistic on anti-SD site usage preference, in a case study to reveal differences in preservation and function of SD sequences between Cyanobacteria and chloroplast translation initiation. We found that cyanophages well-mimic Cyanobacteria in SD usage because both have been under the same selection pressure for SD-mediated initiation. However, Chloroplasts do not retain optimal SD sequences because the need for SD-mediated initiation has been reduced in plastids as a result of host-symbiont coevolution.

In Chapter 3, we described an RNA-Seq approach to quantify tRNAs and we show that our estimator is robust with quantifications near perfectly consistent with experimentally quantified tRNA abundances. Moreover, our approach allowed us to 1) determine species-specific translationally optimal codons, 2) improve the predictive power of the tRNA adaptation index (tAI) to estimate codon adaptation, and 3) explain that tRNA availability is optimized to codon usage in fast-growing but not slow-growing species. Finally, Chapter 4 integrates research in translation initiation (Chapter 2) and elongation (Chapter 3) and apply Ribo-Seq approaches to measure how optimal translation features influence the ribosome dynamics during bacterial translation.

The emergence of COVID-19 pandemic poses a serious global health emergency. As obligate parasites, coronaviruses evolve to carry genomic signatures in adaptation to their host-specific environments. To understand the evolutionary genomics of SARS-CoV-2, I investigate 1) host-specificity during cell entry and 2) how coronaviruses evade or adapt to host immune responses once they have entered host cells. The aim of these investigations is to understand virus-host adaptation in contribution to the global research efforts to battle against the current pandemic.

In Chapter 5, we tested the hypothesis that mammals at high risk of SARS-CoV-2 infection should have human-like ACE2 receptors at key Spike protein-binding sites. Furthermore, if high-risk mammals share human-like ACE2 receptors, they may provide the adequate environments to host species-specific coronaviruses that express SARS-like Spike proteins at key ACE2-binding sites. We performed large-scale comparative sequence analyses to determine which coronaviruses carry SARS-like Spike proteins and which mammalian species carry human-like ACE2 receptors. This investigation aims to provide insights to improve our ability to predict and control future

pandemics, to manage and protect wildlife and domesticated animals, and to understand the origin and evolution of SARS viruses.

In Chapter 6, we tested the hypothesis that APOBEC3 and ZAP exert selective pressure to shape coronavirus genome evolution. Coronaviruses regularly infect host tissues that express antiviral proteins (AVPs) in abundance. It follows that the genome of an infecting virus should trend towards reduced CpG content to evade ZAP and increased C to U mutations due to APOBEC3 editing. In a large-scale comparative genomics study, we detected changes to the genome compositions of tissue-specific coronaviruses such as SARS-CoV-2 over their infection time course in distinct cellular environments with different levels of AVP expressions. We showed that coronaviruses that regularly infect tissues with abundant AVPs have CpG-deficient and U-rich genomes; whereas those that do not infect tissues with abundant AVPs do not share these sequence hallmarks. Understanding the evolutionary pressures exerted by host immune systems onto the genomes of infecting viruses will improve our ability to control the spread of infection, and we expect these findings will improve the designs for attenuated vaccines against coronaviruses such as SARS-CoV-2.

1.4 References

Abolbaghaei, A, JR Silke, X Xia. 2017. How Changes in Anti-SD Sequences Would Affect SD Sequences in *Escherichia coli* and *Bacillus subtilis*. G3 (Bethesda).

Andersen, KG, A Rambaut, WI Lipkin, EC Holmes, RF Garry. 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26:450-452.

Anderson, EJ, NG Rouphael, AT Widge, et al. 2020. Safety and Immunogenicity of SARS-CoV-2 mRNA-1273 Vaccine in Older Adults. New England Journal of Medicine 383:2427-2438.

Andersson, DI, C G Kurland. 1983. RAM ribosomes are defective proofreaders. Molecular & general genetics 191:378-381.

Atkinson, NJ, J Witteveldt, DJ Evans, P Simmonds. 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 Res 42:4527-4545.
Bao, L, W Deng, B Huang, et al. 2020. The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. Nature 583:830-833.
Baranov, PV, RF Gesteland, JF Atkins. 2002. Release factor 2 frameshifting sites in different bacteria. EMBO reports 3:373-377.
Barbe, V, S Cruveiller, F Kunst, et al. 2009. From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. Microbiology 155:1758-1775.
Belouzard, S, VC Chu, GR Whittaker. 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proceedings of the National Academy of Sciences 106:5871.
Bossi, L, JR Roth. 1980. The influence of codon context on genetic code translation. Nature 286:123-127.
Brown, CM, WP Tate. 1994. Direct recognition of mRNA stop signals by Escherichia coli polypeptide chain release factor two. J Biol Chem 269:33164-33170.
Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728-730.
Carullo, M, X Xia. 2008. An Extensive Study of Mutation and Selection on the Wobble Nucleotide in tRNA Anticodons in Fungal Mitochondrial Genomes. Journal of Molecular Evolution 66:484.
Chan, JF-W, AJ Zhang, S Yuan, et al. 2020. Simulation of the clinical and pathological manifestations of Coronavirus Disease 2019 (COVID-19) in golden Syrian hamster model: implications for disease pathogenesis and transmissibility. Clin. Infect. Dis.
Chan, PP, TM Lowe. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37:4.
Chandrashekar, A, J Liu, AJ Martinot, et al. 2020. SARS-CoV-2 infection protects against rechallenge in rhesus macaques. Science 369:812-817.
Charneski, CA, LD Hurst. 2013. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol 11:12.
Chavancy, G, A Chevallier, A Fournier, JP Garel. 1979. Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie 61:71-78.
Chen, S, Y Xu, K Zhang, X Wang, J Sun, G Gao, Y Liu. 2012. Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA. Nat Struct Mol Biol 19:430-435.
Chiu, Y-L, WC Greene. 2008. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. Annual Review of Immunology 26:317-353.
Citti, C, L Maréchal-Drouard, C Saillard, JH Weil, JM Bové. 1992. Spiroplasma citri UGG and UGA tryptophan codons: sequence of the two tryptophanyl-tRNAs and organization of the corresponding genes. J Bacteriol 174:6471-6478.
Coutard, B, C Valle, X de Lamballerie, B Canard, NG Seidah, E Decroly. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research 176:104742.
Cozen, AE, E Quartley, AD Holmes, E Hrabeta-Robinson, EM Phizicky, TM Lowe. 2015. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Methods 12:879-884.

Craigen, WJ, CT Caskey. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature 322:273-275.
Craigen, WJ, RG Cook, WP Tate, CT Caskey. 1985. Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. Proc Natl Acad Sci U S A 82:3616-3620.
Cullen, BR. 2006. Role and Mechanism of Action of the APOBEC3 Family of Antiretroviral Resistance Factors. Journal of Virology 80:1067.
Dabrowski, M, Z Bukowy-Bieryllo, E Zietkiewicz. 2015. Translational readthrough potential of natural termination codons in eucaryotes--The impact of RNA sequence. RNA Biol 12:950-958.
Di Giorgio, S, F Martignano, MG Torcia, G Mattiuz, SG Conticello. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances:eabb5813.
dos Reis, M, R Savva, L Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036-5044.
Duval, M, A Simonetti, I Caldelari, S Marzi. 2015. Multiple ways to regulate translation initiation in bacteria: Mechanisms, regulatory circuits, dynamics. Biochimie 114:18-29.
Fagerberg, L, BM Hallström, P Oksvold, et al. 2014. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. Molecular & Cellular Proteomics 13:397.
Ficarelli, M, I Antzin-Anduetza, R Hugh-White, AE Firth, H Sertkaya, H Wilson, SJD Neil, R Schulz, CM Swanson. 2020. CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. J Virol 94:01337-01319.
Garel, JP. 1974. Functional adaptation of tRNA population. J Theor Biol 43:211-225.
Greenbaum, BD, AJ Levine, G Bhanot, R Rabadan. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4:1000079.
Greenbaum, BD, R Rabadan, AJ Levine. 2009. Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. PLoS One 4:0005969.
Guan, Y, BJ Zheng, YQ He, et al. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302:276-278.
Guerdoux-Jamet, P, A Hénaut, P Nitschké, JL Risler, A Danchin. 1997. Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res 4:257-265.
Guo, X, J Ma, J Sun, G Gao. 2007. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. Proceedings of the National Academy of Sciences 104:151.
Halfmann, PJ, M Hatta, S Chiba, et al. 2020. Transmission of SARS-CoV-2 in Domestic Cats. N. Engl. J. Med. 383:592-594.
Harris, RS, JP Dudley. 2015. APOBECs and virus restriction. Virology 480:131-145.
Hayward, JA, M Tachedjian, J Cui, AZ Cheng, A Johnson, ML Baker, RS Harris, L-F Wang, G Tachedjian. 2018. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. Molecular Biology and Evolution 35:1626-1637.

Hirose, T, T Kusumegi, M Sugiura. 1998. Translation of tobacco chloroplast rps14 mRNA depends on a Shine-Dalgarno-like sequence in the 5'-untranslated region but not on internal RNA editing in the coding region. FEBS Lett 430:257-260.
Hirose, T, M Sugiura. 1996. Cis-acting elements and trans-acting factors for accurate translation of chloroplast psbA mRNAs: development of an in vitro translation system from tobacco chloroplasts. Embo J 15:1687-1695.
Hockenberry, AJ, AR Pah, MC Jewett, LAN Amaral. 2017. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. Open Biology 7.
Hou, Y-M. 2010. CCA addition to tRNA: implications for tRNA quality control. IUBMB life 62:251-260.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13-34.
Jackson, LA, EJ Anderson, NG Rouphael, et al. 2020. An mRNA Vaccine against SARS-CoV-2 — Preliminary Report. New England Journal of Medicine 383:1920-1931.
Jacob, AI, C Kohrer, BW Davies, UL RajBhandary, GC Walker. 2013. Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. Mol Cell 49:427-438.
Jørgensen, F, FM Adamski, WP Tate, CG Kurland. 1993. Release factor-dependent false stops are infrequent in Escherichia coli. J Mol Biol 230:41-50.
Kim, Y-I, S-G Kim, S-M Kim, et al. 2020. Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. Cell host microbe: Elsevier Inc. p. 704-709.e702.
Kirchdoerfer, RN, CA Cottrell, N Wang, J Pallesen, HM Yassine, HL Turner, KS Corbett, BS Graham, JS McLellan, AB Ward. 2016. Pre-fusion structure of a human coronavirus spike protein. Nature 531:118-121.
Koning, FA, EN Newman, EY Kim, KJ Kunstman, SM Wolinsky, MH Malim. 2009. Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. J Virol 83:9474-9485.
Korkmaz, G, M Holm, T Wiens, S Sanyal. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. The Journal of biological chemistry 289:30334-30342.
Kudla, G, AW Murray, D Tollervey, JB Plotkin. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324:255-258.
Lamers, MM, J Beumer, J van der Vaart, et al. 2020. SARS-CoV-2 productively infects human gut enterocytes. Science 369:50-54.
Lan, J, J Ge, J Yu, et al. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581:215-220.
Li, G-W. 2015. How do bacteria tune translation efficiency? Current opinion in microbiology 24:66-71.
Li, G-W, E Oh, JS Weissman. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538-541.
Li, W, C Zhang, J Sui, et al. 2005. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. Embo. J. 24:1634-1643.
Lim, K, Y Furuta, I Kobayashi. 2012. Large variations in bacterial ribosomal RNA genes. Mol Biol Evol 29:2937-2948.

Lobry, JR, N Sueoka. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biology 3:research0058.0051.
Luo, X, X Wang, Y Gao, J Zhu, S Liu, G Gao, P Gao. 2020. Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. Cell Rep 30:46-52.
Ma, J, A Campbell, S Karlin. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733-5745.
Marra, MA, SJ Jones, CR Astell, et al. 2003. The Genome sequence of the SARS-associated coronavirus. Science 300:1399-1404.
McPherson, DT. 1988. Codon preference reflects mistranslational constraints: a proposal. Nucleic Acids Res 16:4111-4120.
Meagher, JL, M Takata, D Gonçalves-Carneiro, et al. 2019. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. Proc Natl Acad Sci U S A 116:24303-24309.
Miller, JH, AM Albertini. 1983. Effects of surrounding sequence on the suppression of nonsense codons. J Mol Biol 164:59-71.
Milman, G, J Goldstein, E Scolnick, T Caskey. 1969. Peptide chain termination. 3. Stimulation of in vitro termination. Proc Natl Acad Sci U S A 63:183-190.
Murray, CL, JC Rabinowitz. 1982. Nucleotide sequences of transcription and translation initiation regions in Bacillus phage phi 29 early genes. J Biol Chem 257:1053-1062.
Muto, A, S Osawa. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84:166-169.
Nabel, CS, JW Lee, LC Wang, RM Kohli. 2013. Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. Proceedings of the National Academy of Sciences:201306345.
Nakagawa, S, Y Niimura, T Gojobori. 2017. Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. Nucleic Acids Res 45:3922-3931.
Nakamoto, T. 2006. A unified view of the initiation of protein synthesis. Biochem Biophys Res Commun 341:675-678.
Novoa, EM, M Pavon-Eternod, T Pan, L Ribas de Pouplana. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149:202-213.
Odon, V, JJ Fros, N Goonawardane, I Dietrich, A Ibrahim, K Alshaikhahmed, D Nguyen, P Simmonds. 2019. The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. Nucleic acids research 47:8061-8083.
Oreshkova, N, RJ Molenaar, S Vreman, et al. 2020. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. Euro. Surveill. 25:1560-7917.
Osterman, IA, SA Evfratov, PV Sergiev, OA Dontsova. 2013. Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic acids research 41:474-486.
Ou, X, Y Liu, X Lei, et al. 2020. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature Communications 11:1620.
Palidwor, GA, TJ Perkins, X Xia. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5:0013431.

Pallesen, J, N Wang, KS Corbett, et al. 2017. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proceedings of the National Academy of Sciences 114:E7348.

Parker, J. 1989. Errors and alternatives in reading the universal genetic code. Microbiological reviews 53:273-298.

Pavlov, MY, DV Freistroffer, V Heurgué-Hamard, RH Buckingham, M Ehrenberg. 1997. Release factor RF3 abolishes competition between release factor RF1 and ribosome recycling factor (RRF) for a ribosome binding site. J Mol Biol 273:389-401.

Poole, ES, LL Major, SA Mannering, WP Tate. 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 26:954-960.

Povolotskaya, IS, FA Kondrashov, A Ledda, PK Vlasov. 2012. Stop codons in bacteria are not selectively equivalent. Biology Direct 7:30.

Prabhakaran, R, S Chithambaram, X Xia. 2015. *Escherichia coli* and *Staphylococcus* phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. J Gen Virol 96:1169-1179.

Refsland, EW, MD Stenglein, K Shindo, JS Albin, WL Brown, RS Harris. 2010. Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. Nucleic Acids Res 38:4274-4284.

Robinson, M, R Lilley, S Little, JS Emtage, G Yarranton, P Stephens, A Millican, M Eaton, G Humphreys. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic acids research 12:6663-6671.

Rockx, B, T Kuiken, S Herfst, et al. 2020. Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. Science 368:1012-1015.

Rodriguez-Frias, F, M Buti, D Tabernero, M Homs. 2013. Quasispecies structure, cornerstone of hepatitis B virus infection: mass sequencing approach. World journal of gastroenterology 19:6995-7023.

Rojas, J, G Castillo, LE Leiva, S Elgamal, O Orellana, M Ibba, A Katz. 2018. Codon usage revisited: Lack of correlation between codon usage and the number of tRNA genes in enterobacteria. Biochemical and Biophysical Research Communications 502:450-455.

Rota, PA, MS Oberste, SS Monroe, et al. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300:1394-1399.

Rudner, R, JD Karkas, E Chargaff. 1968. Separation of B. subtilis DNA into complementary strands, I. Biological properties. Proc Natl Acad Sci U S A 60:630-635.

Scharff, LB, M Ehrnthaler, M Janowski, LH Childs, C Hasse, J Gremmels, S Ruf, R Zoschke, R Bock. 2017. Shine-Dalgarno Sequences Play an Essential Role in the Translation of Plastid mRNAs in Tobacco. Plant Cell 29:3085-3101.

Schurr, T, E Nadir, H Margalit. 1993. Identification and characterization of E.coli ribosomal binding sites by free energy computation. Nucleic Acids Res 21:4019-4023.

Scolnick, E, R Tompkins, T Caskey, M Nirenberg. 1968. Release factors differing in specificity for terminator codons. Proceedings of the National Academy of Sciences of the United States of America 61:768-774.

Scolnick, EM, CT Caskey. 1969. Peptide chain termination. V. The role of release factors in mRNA terminator codon recognition. Proceedings of the National Academy of Sciences of the United States of America 64:1235-1241.

Shang, J, G Ye, K Shi, Y Wan, C Luo, H Aihara, Q Geng, A Auerbach, F Li. 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 30:020-2179.
Sharma, S, SK Patnaik, RT Taggart, BE Baysal. 2016. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Sci Rep 6.
Sharma, S, SK Patnaik, R Thomas Taggart, ED Kannisto, SM Enriquez, P Gollnick, BE Baysal. 2015. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nature Communications 6:6881.
Sharma, S, J Wang, E Alqassim, S Portwood, E Cortes Gomez, O Maguire, PH Basse, ES Wang, BH Segal, BE Baysal. 2019. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol 20:019-1651.
Sharp, PM, E Bailes, RJ Grocock, JF Peden, RE Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141-1153.
Sharp, PM, WH Li. 1987a. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281-1295.
Sharp, PM, WH Li. 1987b. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research 15:1281-1295.
Sheehy, AM, NC Gaddis, JD Choi, MH Malim. 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418:646-650.
Shi, J, Z Wen, G Zhong, et al. 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. Science 368:1016-1020.
Shine, J, L Dalgarno. 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 71:1342-1346.
Shine, J, L Dalgarno. 1975. Determinant of cistron specificity in bacterial ribosomes. Nature 254:34-38.
Silke, JR, Y Wei, X Xia. 2018. RNA-Seq-Based Analysis Reveals Heterogeneity in Mature 16S rRNA 3' Termini and Extended Anti-Shine-Dalgarno Motifs in Bacterial Species. G3 8:3973-3979.
Simmonds, P. 2020. Rampant C to U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. bioRxiv : the preprint server for biology:2020.2005.2001.072330.
Simmons, G, DN Gosalia, AJ Rennekamp, JD Reeves, SL Diamond, P Bates. 2005. Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. Proceedings of the National Academy of Sciences of the United States of America 102:11876.
Sit, THC, CJ Brackman, SM Ip, et al. 2020. Infection of dogs with SARS-CoV-2. Nature.
Sohmen, D, S Chiba, N Shimokawa-Chiba, CA Innis, O Berninghausen, R Beckmann, K Ito, DN Wilson. 2015. Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. 6:6941.
Song, HD, CC Tu, GW Zhang, et al. 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. USA 102:2430-2435.
Sorensen, MA, CG Kurland, S Pedersen. 1989. Codon usage determines translation rate in Escherichia coli. J Mol Biol 207:365-377.
Starmer, J, A Stomp, M Vouk, D Bitzer. 2006a. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. PLoS Comput Biol 2:19.

Starmer, J, A Stomp, M Vouk, D Bitzer. 2006b. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. PLoS Comput Biol 2:e57.
Sulthana, S, MP Deutscher. 2013. Multiple exoribonucleases catalyze maturation of the 3' terminus of 16S ribosomal RNA (rRNA). J Biol Chem 288:12574-12579.
Takata, MA, D Gonçalves-Carneiro, TM Zang, SJ Soll, A York, D Blanco-Melo, PD Bieniasz. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124-127.
Tate, WP, JB Mansell, SA Mannering, JH Irvine, LL Major, DN Wilson. 1999. UGA: a dual signal for 'stop' and for recoding in protein synthesis. Biochemistry 64:1342-1353.
Tate, WP, ES Poole, ME Dalphin, LL Major, DJG Crawford, SA Mannering. 1996. The translational stop signal: Codon with a context, or extended factor recognition element? Biochimie 78:945-952.
Theys, K, AF Feder, M Gelbart, M Hartl, A Stern, PS Pennings. 2018. Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV. PLoS genetics 14:e1007420-e1007420.
Tian, X, C Li, A Huang, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody: Emerg Microbes Infect. 2020 Feb 17;9(1):382-385. doi: 10.1080/22221751.2020.1729069. eCollection 2020.
Tian, X, C Li, A Huang, et al. 2020. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. Emerg. Microbes. Infect. p. 382-385.
Trus, I, D Udenze, N Berube, C Wheler, M-J Martel, V Gerdts, U Karniychuk. 2020. CpG-Recoding in Zika Virus Genome Causes Host-Age-Dependent Attenuation of Infection With Protection Against Lethal Heterologous Challenge in Mice. Frontiers in immunology 10:3077-3077.
Tuller, T, YY Waldman, M Kupiec, E Ruppin. 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 107:3645-3650.
Victorovich, KV, G Rajanish, KT Aleksandrovna, KS Krishna, SA Nicolaevich, PV Vitoldovich. 2020. Translation-associated mutational U-pressure in the first ORF of SARS-CoV-2 and other coronaviruses. bioRxiv : the preprint server for biology:2020.2005.2005.078238.
Walls, AC, YJ Park, MA Tortorici, A Wall, AT McGuire, D Veesler. 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181:281-292.
Wang, L, PK Mitchell, PP Calle, et al. 2020. Complete Genome Sequence of SARS-CoV-2 in a Tiger from a U.S. Zoological Collection. In: S Roux, editor. Microbiol. Resour. Announc. p. e00468-00420.
Wang, M, CJ Herrmann, M Simonovic, D Szklarczyk, C von Mering. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163-3168.
Wang, SM, CT Wang. 2009. APOBEC3G cytidine deaminase association with coronavirus nucleocapsid protein. Virology 388:112-120.
Wei, Y, P Aris, H Farookhi, X Xia. 2021. Predicting mammalian species at risk of being infected by SARS-CoV-2 from an ACE2 perspective. Scientific Reports 11:1702.
Wei, Y, JR Silke, X Xia. 2017. Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in Escherichia coli and Bacillus subtilis using RNA-Seq data. Sci Rep 7:017-17918.

Wei, Y, JR Silke, X Xia. 2019. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci Rep 9:3184.
Wei, Y, J Wang, X Xia. 2016. Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. Mol Biol Evol 33:2357-2367.
Wei, Y, X Xia. 2017. The Role of +4U as an Extended Translation Termination Signal in Bacteria. Genetics 205:539.
Wei, Y, X Xia. 2019. Unique Shine-Dalgarno Sequences in Cyanobacteria and Chloroplasts Reveal Evolutionary Differences in Their Translation Initiation. Genome Biol Evol 11:3194-3206.
Woese, CR, GE Fox, L Zablen, T Uchida, L Bonen, K Pechman, BJ Lewis, D Stahl. 1975. Conservation of primary structure in 16S ribosomal RNA. Nature 254:83-86.
Woolsey, C, V Borisevich, AN Prasad, et al. 2020. Establishment of an African green monkey model for COVID-19. bioRxiv : the preprint server for biology: Cold Spring Harbor Laboratory.
Xia, X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37-44.
Xia, X. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345:13-20.
Xia, X. 2012. DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. Current genomics 13:16-27.
Xia, X. 2015. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199:573-579.
Xia, X. 2017. DAMBE6: New tools for microbial genomics, phylogenetics and molecular evolution. J Hered.
Xia, X. 2018. "Bioinformatics and translation initiation" in Bioinformatics and the Cell: Springer, Cham.
Xia, X. 2019. Optimizing Phage Translation Initiation. OBM Genetics 3(4).
Xia, X. 2021. Domains and Functions of Spike Protein in SARS-Cov-2 in the Context of Vaccine Design. Viruses 13.
Yamao, F, A Muto, Y Kawauchi, M Iwami, S Iwagami, Y Azumi, S Osawa. 1985. UGA is read as tryptophan in Mycoplasma capricolum. Proceedings of the National Academy of Sciences of the United States of America 82:2306-2309.
Yao, XH, ZC He, TY Li, et al. Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient: Cell Res. 2020 Jun;30(6):541-543. doi: 10.1038/s41422-020-0318-5. Epub 2020 Apr 28.
Yap, YL, XW Zhang, A Danchin. 2003. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinformatics 4:1471-2105.
Zheng, G, Y Qin, WC Clark, Q Dai, C Yi, C He, AM Lambowitz, T Pan. 2015. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods 12:835-837.
Zhu, Y, G Chen, F Lv, X Wang, X Ji, Y Xu, J Sun, L Wu, Y-T Zheng, G Gao. 2011. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. Proceedings of the National Academy of Sciences 108:15834.
Zoschke, R, R Bock. 2018. Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. Plant Cell 30:745-770.

CHAPTER TWO

Unique Shine-Dalgarno sequences in Cyanobacteria and chloroplasts reveal evolutionary differences in their translation initiation

This chapter is a modified version of the publication below, in that 2.3 Introduction is partly removed, reorganized, and rewritten into Chapter 1.1.1:

Wei, Y. and Xia, X. 2019. Unique Shine-Dalgarno sequences in Cyanobacteria and chloroplasts reveal evolutionary differences in their translation initiation. Genome Biology and Evolution 11(11):3194-3206.

2.1 Author Contributions

Yulong Wei[1], 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.

Y.W. and X.X. designed the study. Y.W. collected and analyzed the data. Y.W. wrote the manuscript, and X.X. provided guidance and editing. X.X. developed the computational programs. Y.W. made all Figures, Tables, Supplemental Figures, and Supplemental Files. All authors reviewed the manuscript. X.X. supervised the project. This work is published (PMID: 31621842)

2.2 Abstract

Microorganisms require efficient translation to grow and replicate rapidly, and translation is often rate-limited by initiation. A prominent feature that facilitates translation initiation in bacteria is the Shine-Dalgarno (SD) sequence. However, there is much debate over its conservation in Cyanobacteria and in chloroplasts which presumably originated from endosymbiosis of ancient Cyanobacteria. Elucidating the utilization of SD sequences in Cyanobacteria and in chloroplasts is therefore important to understand whether 1) SD role in Cyanobacterial translation has been reduced prior to chloroplast endosymbiosis or 2) translation in Cyanobacteria and in plastid has been subjected to different evolutionary pressures. To test these alternatives, we employed genomic, proteomic, and transcriptomic data to trace differences in SD usage between *Synechocystis* species, *Microcystis aeruginosa*, cyanophages, *Nicotiana tabacum* chloroplast, and *Arabidopsis thaliana* chloroplast. We corrected their mis-annotated 16S rRNA 3' terminus using an RNA-Seq-based approach, and we determined their genes' SD/anti-SD location constraints using an improved measurement $D_{toStart}$. We found that genes in cyanophages well mimic genes in Cyanobacteria in SD usage because both have been under the same selection pressure for SD-mediated initiation. Whereas genes in chloroplasts lost this similarity because the need for SD-facilitated initiation has been reduced in plastids having much reduced genome size and different ribosomal proteins as a result of host-symbiont coevolution. Consequently, SD sequence increases protein expression in Cyanobacteria but not in chloroplasts, and only Cyanobacterial genes compensate for a lack of SD sequence by having weaker secondary structures at the 5' UTR. These results suggest different evolutionary pressures operate on translation initiation in Cyanobacteria and in chloroplast.

2.3 Introduction

As detailed in Chapter 1.1.1, bacterial translation is often rate limited by initiation (Andersson, G Kurland 1983; Kudla et al. 2009; Duval et al. 2015) and an efficient translation initiation possesses two features. First, the minimum requirement is that the start codon is accessible (Nakamoto 2006) and there is no stable secondary structure to embed the start codon (Osterman et al. 2013). Second, presence of a Shine-Dalgarno (SD) sequence is prominent at the 5' untranslated region (5' UTR) upstream of start codon in prokaryotes (Nakagawa, Niimura, Gojobori 2017). The SD sequence facilitates recruitment of ribosome to start codon by pairing with the anti-SD sequence located at the 3' terminus of mature 16S rRNA (hereafter referred as 3' tail) (Shine, Dalgarno 1974).

Distinctions between SD-facilitated genes and SD-independent genes in model organisms such as *Escherichia coli* and *Bacillus subtilis* reveal the importance of SD mechanism in bacterial translation initiation: genes with a well-positioned SD/anti-SD pair are more efficiently translated in bacteria (Schurr, Nadir, Margalit 1993; Osterman et al. 2013; Abolbaghaei, Silke, Xia 2017) and in their bacteriophages (Prabhakaran, Chithambaram, Xia 2015) than other genes; while translation in SD-independent genes are more reliant on reduced secondary structure stability at the initiation region (Nakagawa, Niimura, Gojobori 2017; Scharff et al. 2017) to compensate for the lack of SD sequence.

2.3.1 Controversies on SD function in Cyanobacteria and chloroplasts

Nonetheless, there have been many debates over the conservation and function of SD sequence in Cyanobacteria and in chloroplasts. Translation initiation is important to

Cyanobacterial and chloroplast gene expression as there is no stable secondary structure to embed start codons in these species (Nakamoto 2006; Scharff et al. 2011). However, studies suggest that SD motifs are weakly conserved in Cyanobacteria and in plant chloroplasts that presumably originated from early endosymbiosis of ancient Cyanobacterium for several reasons. First, translation initiation often relies on S1 protein in Cyanobacteria (Mutsuda, Sugiura 2006) and on an additional trans-acting factor in tobacco chloroplast (Plader, Sugiura 2003; Hirose, Sugiura 2004). This may explain why presence of SD sequence is reduced in both Cyanobacteria (Ma, Campbell, Karlin 2002; Starmer et al. 2006; Nakagawa et al. 2010) and plastids (Scharff et al. 2017). Second, SD/anti-SD pairing distance to start codon was found to be weakly constrained in both *Synechocystis* sp. (Sazuka, Ohara 1996; Hirosawa, Sazuka, Yada 1997; Ma, Campbell, Karlin 2002) and tobacco chloroplast (Sugiura, Hirose, Sugita 1998; Plader, Sugiura 2003). Nevertheless, these studies had focused on SD interactions with the canonical anti-SD motif CCUCCU and had measured pairing distance from end of SD sequence to start codon, both are problematic, as discussed in Chapter 1.1.1.

Consequently, there are conflicting views on the importance of SD-facilitated initiation in both Cyanobacteria and chloroplasts. While some studies showed that a well-positioned SD sequence increased plastid gene expression (Ye et al. 2001; Oey et al. 2009), others found that mutating such SD sequences had no detectable effect on plastid gene translation initiation (Fargo et al. 1998; Nickelsen et al. 1999), and still others showed that weak SD/anti-SD complementarity increased plastid translation initiation (Scharff et al. 2017). In one case, the presence of SD sequence even reduced gene expression for one mRNA (Plader, Sugiura 2003). Hence, the importance of SD mechanism in chloroplast translation remains unclear, and this uncertainty raises

skepticism over SD function in Cyanobacteria because chloroplasts presumably originated from Cyanobacteria.

If SD-facilitated initiation is unimportant in Cyanobacteria and chloroplasts, then the selection constraining SD usage and SD/anti-SD pairing was most likely lost or weakened prior to chloroplast endosymbiosis. On the contrary, if SD-facilitated initiation is important in Cyanobacteria but not important in chloroplasts, then SD usage and locational constraints would manifest in the former but not in the latter. This would also lead to the prediction that cyanophages, which depends on Cyanobacterial translation machinery for translation, should exhibit SD usage and SD/anti-SD pairing mimicking its Cyanobacterial hosts (similar to phages that infect *E. coli* and *B. subtilis*). This scenario would imply that SD function had been reduced in chloroplasts after having differentially evolved under host-symbiont coevolution away from Cyanobacteria.

2.3.2 An integrated approach sheds light on evolutionary differences between Cyanobacteria and chloroplast translation initiation

To test the predictions outlined above, we performed an integrated study using publicly available transcriptomic, genomic, and proteomic data of Cyanobacteria (*Microcystis aeruginosa* and *Synechocystis* sp.), chloroplasts (*Nicotiana tabacum* chloroplast and *Arabidopsis thaliana* chloroplast), and cyanophages (S-*SSM6a* and *S-SSM6b*) to determine SD usage and locational constraints and measure gene expression in presence or absence of a SD signal.

An issue that confounds the determination of functional SD sequences is the difficulty in reliably characterizing the 3' terminus of mature 16S rRNA in bacteria (Lin et al. 2008; Wei, Silke, Xia 2017) and in plastids (Gallaher et al. 2018). The annotated 3' tails in *S.* sp., *M. aeruginosa* and

chloroplasts were determined by automated computational predictions based on sequence homology against other bacterial species, and these predictions are often erroneous (Jones, Brown, Baumann 2007; Lin et al. 2008). These errors are problematic for two reasons: 1) determination of all potential SD sequences requires knowing the full extent of anti-SD sequence which constitutes the 3' tail, and 2) identity of the 3' tail is also required to correctly measure SD/anti-SD locational constraints by $D_{toStart}$.

Uncertainty in 3' tail determination arises because at least four RNases (RNase II, RNase R, PNPase, and YbeY) independently participates in the 3' maturation process of the 16S rRNA with unknown mechanisms (Sulthana, Deutscher 2013), and it is unclear whether they cleave off nucleotides on the 3' end of the pre-16S rRNA until the same length. Thus, the mature 16S rRNA may have a variety of end points as opposed to a single deterministic 3' tail. Indeed, we have previously observed a degree of heterogeneity in the 3' tail in five bacterial species (Silke, Wei, Xia 2018).

Recently, we devised an RNA-Seq-based approach (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018) to determine the most prominent 3' tail in the rRNA pool, which led to the correction of 3' tail mis-annotations in 12 bacterial species, including *S.* sp. To test the fidelity of our method, we have previously shown that the most prominent 3' tail sequence determined by our RNA-Seq-based approach (Wei, Silke, Xia 2017) matches the mature 16S rRNA 3' terminus that was first experimentally validated in *E. coli* and in *B. subtilis* (Shine, Dalgarno 1974; Shine, Dalgarno 1975). To our knowledge, similar experimental procedures have not been done in *S.* sp., *M. aeruginosa*, and chloroplasts. Hence, we applied our RNA-Seq-based approach to correct the potentially mis-annotated 3' tail in these species.

Our results suggest differences between chloroplast and Cyanobacterial translation initiation. SD motif usage and $D_{toStart}$ are strongly and similarly constrained in both Cyanobacteria and cyanophages, consistent with our prediction. In addition, SD-facilitated genes had significantly higher protein expression in comparison to other genes in both *S.* sp. and *M. aeruginosa*. Yet, in both chloroplasts $D_{toStart}$ was loosely constrained and their SD usage differed from those in Cyanobacteria. Furthermore, there was no significant difference in protein expression between SD-facilitated and other genes in chloroplasts. Moreover, secondary structures at the 5' UTR were significantly weaker in SD-independent genes to compensate for the lack of SD sequence in Cyanobacteria but not in chloroplasts. While cyanophages mimic Cyanobacteria in SD usage because both have been under the same selection pressure for SD usage, chloroplasts had lost this similarity likely because the need for SD-facilitated initiation was reduced in plastids during host-symbiont coevolution.

2.4 Materials and Methods

2.4.1 Processing genomic, proteomic, and RNA-Seq data

The annotated genome of *Synechocystis* sp. PCC 6803 (NC_017277.1), *Microcystis aeruginosa* NIES-843 (NC_010296.1), *Nicotiana tabacum* chloroplast (Z00044.2), *Arabidopsis thaliana* chloroplast (NC_000932.1) and chloroplast genes from *A. thaliana* chromosomes 1-5: (NC_003070.8, 1.8, 4.8, 5.8, 6.8), cyanophages *S-SSM6a* (HQ317391.1) and *S-SSM6b* (HQ316603.1) were retrieved in Genbank format from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). *S.* sp. protein abundances (in ppm) from Wegener et al. (2010) and *M. aeruginosa* integrated protein abundances (in ppm) were retrieved from PaxDb 4.0 (Wang et al. 2015), and protein IDs were matched to NCBI gene IDs.

N. tabacum chloroplast protein abundances (in unique peptide counts) were retrieved from Wu, Yan (2018). *A. thaliana* chloroplast protein abundances (in total spectral count) were retrieved from Ferro et al. (2010); here, we only considered chloroplasts proteins previously detected by Zybailov et al. (2008).

High-throughput RNA Sequencing datasets of *S.* sp. PCC6803 (PRJNA381210: SRX2694285, 6, 7, 8; PRJNA473849: SRX4145044, 5, 6), *M. aeruginosa* (strain KW, PRJNA421714: SRX3459379, 80; strain PCC 7806: PRJNA427104: SRX3501057, 8, 9), *N. tabacum* (PRJNA526208: SRX5495062, 75; PRJNA299013: SRX1342044, 5) and *A. thaliana* (PRJNA432917: SRX3647857, 58, 59, 60; PRJNA354600: SRX2367970) were retrieved from GEO Datasets in FASTQ format. All selected datasets are control runs for wild-type species, with the exception of *M. aeruginosa* KW (PRJNA421714) where S- and W- morphotype strains were selected. We chose RNA-Seq runs from plant species because there are no records of chloroplast-only RNA-Seq datasets in GEO Datasets, and because these experimental designs were tailored to study plastid transcripts (Pasoreck et al. 2016; Chang et al. 2017; Armarego-Marriott et al. 2019; Zhao, Huang, Chory 2019).

These FASTQ datasets were then processed to eliminate low-quality reads and adapter sequences. We used Trimmomatic 0.38 (Bolger, Lohse, Usadel 2014) to remove poor quality sequences with average Phred scores < 20 (1% probability of base calling errors) (Ewing, Green 1998). Next, we used CutAdapt 1.17 (Martin 2011) to trim off flanking adapter sequences as listed in GEO Dataset's '*Construction protocol'* or in '*Publicatio*ns' (Supplemental File S1), with 10% error rate and retaining reads ≥ 25nt to mitigate bias in expression levels (Williams et al. 2016). Then, processed FASTQ datasets were converted into FASTA+ format using ARSDA 1.1 (Xia

2017) (http://dambe.bio.uottawa.ca/Include/software.aspx). This step groups identical reads under a new unique ID indicating the number of identical copies (SeqID_# of copies), reducing data size without loss of information.

2.4.2 Characterizing the 3' tail using RNA-Seq data

To determine the 3' tail, we mapped reads from each FASTA+ file onto the 16S rDNA query sequence. The FASTA+ files were converted into BLAST databases using the '*Create BLAST DB*' function in ARSDA. The first NCBI annotated 16S rRNA sequence was extracted from .gbk file in all species, and the BLAST query sequence was selected as the CCUCC core anti-SD motif plus 100 nt upstream and downstream (205nt total query length). The query sequence was searched against the BLAST databases using BLASTn (Altschul et al. 1990) implemented in DAMBE7 (Xia 2018) with the following specifications: E-value cutoff of 10^{-10}, ungapped alignment, minimum match length of 25.

The most prominent 3' tails were characterized based on map abundances that met two criteria: 1) they contained the core CCUCC and map within or close to 5'-GAUCACCUCCUU(U or A)-3', and 2) they were the highest peaks with mapped counts appreciably greater than counts at sites further downstream (greater than the combined count numbers of at-least 5 downstream bases). The basis for the first criterion was that CCUCC is conserved in 16S rRNA among bacterial species due to its presumed essential role in SD/anti-SD pairing, and that the genomic sequence 5'-GAUCACCUCCUU(U or A)-3' is conserved across 277 prokaryote species (Nakagawa et al. 2010). The premise for the second criterion was that intermediate sequences between the mature and pre-16S rRNA are rapidly cleaved by RNases in *E. coli* (Cangelosi, Brabant 1997; Sulthana,

Deutscher 2013), hence we expect to observe few map counts immediately downstream of the mature 3' tail.

2.4.3 Determining functional Shine-Dalgarno sequences based on pairing potential and location

The most prominent 3' tails are 5'-GAUCACCUCCUUU-3' in *S.* sp., 5'-GAUCACCUCCUU-3' in *M. aeruginosa*, and 5'-GAUCACCUCC-3' in both *N. tabacum* chloroplast and *A. thaliana* chloroplast. These sequences were used as the complementary anti-SD sequence to identify ≥ 4 nt complementary SD sequences in non-pseudo genes. Additionally, *S.* sp. and *M. aeruginosa* 3' tails were each used to identify SD sequences in cyanophages. To identify putative SD sequences, we used DAMBE following the method used in previous studies (Prabhakaran, Chithambaram, Xia 2015; Abolbaghaei, Silke, Xia 2017; Wei, Silke, Xia 2017): 30 nt upstream of start codon of coding-DNA sequences were extracted and matched against the 16S rRNA 3' tail with '*Analyzing 5'UTR*', with minimum SD length = 4, maximum SD length = length of 3' tail. DAMBE outputs the distance from the 16S rRNA 3' end to the start codon denoted as $D_{toStart}$. Notably, this 30 nt encompassed the -2 to -29 and -5 to -19 sites relative to start codon where most SD-like AGGAGG sequences were found in chloroplast (Hirose, Kusumegi, Sugiura 1998) and in *S.* sp. (Sazuka, Ohara 1996), respectively. The total number of putative SD sequences determined were 2209 in *S.* sp., 3860 in *M. aeruginosa*, 62 in *N. tabacum* chloroplast, and 50 in *A. thaliana* chloroplast.

In addition, DAMBE outputs the site-specific observed and expected numbers of pairing at anti-SD sequence (hereafter designated as O_i and E_i, respectively, where i refers to site i in the anti-SD sequence). If there is no selection at site i then O_i is expected to be equal to E_i (Wei, Silke,

Xia 2017), and site i is preferred if number of observed pairing was higher than that of expected (a O_i:E_i ratio of > 1). To test whether O_i:E_i > 1 is significant, we applied the Central Limit Theorem for proportions. We defined the total sample size n as $\sum O_i$, the sum of O_i at all i anti-SD sites; then the sample proportion of observed counts at each site was $p_i = \frac{O_i}{n}$, and the expected proportion at each site was $\widehat{p_i} = \frac{E_i}{n}$. Because the sample size n was sufficiently large (in our case n = 8709 in *S.* sp., 17373 in *M. aeruginosa*, 283 in *N. tabacum* chloroplast, and 314 in *A. thaliana* chloroplast), by CTL p_i approximates a normal distribution $N(\widehat{p_i}, \sqrt{\frac{\widehat{p_i}(1-\widehat{p_i})}{n}})$ and the Z-score $Z_i = \frac{p_i - \widehat{p_i}}{\sqrt{\frac{\widehat{p_i}(1-\widehat{p_i})}{n}}}$. Because we were only interested in testing whether O_i:E_i > 1 is significant, we took an upper-tailed hypothesis test with H_1: $p_i > \widehat{p_i}$ and H_0: $p_i = \widehat{p_i}$, and we rejected the null hypothesis if $p < 0.05$ and $Z_i \geq 1.645$.

To examine SD role in gene expression, we defined genes with non-zero protein abundances as SD-facilitated if the determined putative SD sequences were located within the preferred $D_{toStart}$ ranges of 11-21 nt in *S.* sp., 11-20 nt in *M. aeruginosa*, 8-16 nt in *S-SSM6a*, 10-21 nt in *S-SSM6b*, 8-16 nt in *N. tabacum* chloroplast, and 10-21 nt in *A. thaliana* chloroplast. These optimal ranges were determined because they encompassed the confined $D_{toStart}$ peaks in Cyanobacteria and cyanophages and constituted the most abundantly mapped region in chloroplasts (illustrated in Figure 2.2, % of SDs they encompass described in Supplemental File S2). Whereas all other genes with non-zero protein abundances were defined as SD-independent genes.

2.4.4 Measuring translation efficiency with protein per transcript

To measure protein per transcript, mRNA abundance was profiled in ARSDA. A query file containing coding DNA sequences of all SD-facilitated and SD-independent genes was built in FASTA format. This query sequence was searched against BLAST databases (of RNA-Seq reads) using ARSDA's *'Gene expression from BLAST Database'*, with the following specifications: E-value = 10^{-10}, ungapped alignment, minimum match length = 25, Max target hits = 1000000. ARSDA outputs the normalized map abundances in Fragments per Kilobase of transcript per Million mapped reads (FPKM). For each gene, the averaged FPKM values were calculated from FPKM values obtained in all selected RNA-Seq replicates within the same BioProject. Finally, protein per transcript was calculated as $\frac{protein\ abundance}{average\ mRNA\ FPKM}$.

2.4.5 Determining secondary structure stability at the translation initiation region

In SD-facilitated and SD-independent genes, we measured folding energy of a 40 nt sequence at the 5' UTR following the approach by Tuller et al. (2010) and by Kudla, et al. (2009). In particular, Tuller et al. (2010) measured the folding energy of a 40 nt 5' UTR sequence immediately upstream of the start codon. We followed this approach and designated a 40 nt 5' UTR sequence as the translation initiation region since our objective was to compare the accessibility of the 5' UTR having or lacking a SD sequence. To measure folding energy, we used Minimum Folding Energy (MFE, kcal/mol) via Vienna RNA Folding Library (Hofacker 2003) implemented in DAMBE, with options: no lonely pair, Window size = 40, StepSize = 1. Importantly, this 40 nt encompassed all putative SD sequences that we have determined.

2.5 Results

2.5.1 The 3' tail of mature 16S rRNA is required to determine preferred anti-SD motifs

We characterized the most prominent 3' tail of mature 16S rRNA in *S.* sp. (5'-GAUCACCUCCUUU-3'), *M. aeruginosa* (5'-GAUCACCUCCUU-3'), *N. tabacum* chloroplast (5'-GAUCACCUCC-3'), and *A. thaliana* chloroplast 5'-GAUCACCUCC-3') (Figure 2.1; See Materials and Methods for criteria used in 3' tail determination) and corrected the mis-annotated (in red) *S.* sp. 5'-GAUCACCUCCUUUAAGGG-3' (NC_017277.1), *M. aeruginosa* 5'-GAUCACCUCCUUA-3' (NC_010296.1), *N. tabacum* chloroplast 5'-GAUCACCUCCUUU-3' (Z00044.2), and *A. thaliana* chloroplast 5'-GAUCACCUCCUUU-3' (NC_000932.1). Indeed, automated annotations of 16S rRNA are often erroneous (Starmer et al. 2006; Jones, Brown, Baumann 2007; Lagesen et al. 2007; Lin et al. 2008). Notably, a secondary 3' tail peak was observed from both independent RNA-Seq studies in *S.* sp. (5'-GAUCACCUCCU-3') and in *A. thaliana* chloroplasts (5'-GAUCACCUCCUUUUCAG-3'). Similarly, we have previously reported a degree of heterogeneity in the 3' tail in four other bacterial species (Silke, Wei, Xia 2018).

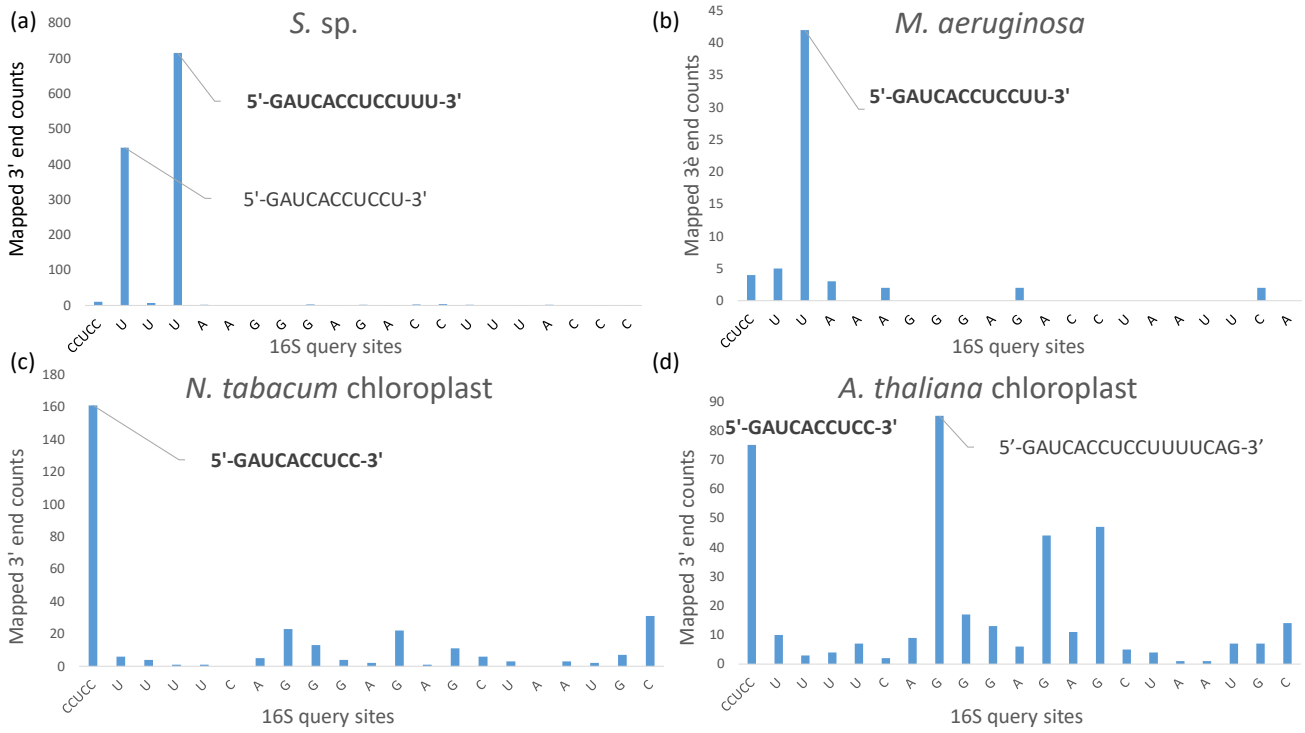


Figure 2.1. Site-specific counts of RNA-Seq read 3' ends mapped to the 16S rDNA query sequence in (a) *S.* sp., (b) *M. aeruginosa*, (c) *N. tabacum* chloroplast, and (d) *A. thaliana* chloroplast. The first site is the CCUCC motif, followed by 20 downstream sites. The identity of the most prominent 3' tail that meets the two identification criteria is annotated in bold. A secondary peak is observed and annotated in *S.* sp. and *A. thaliana* chloroplast.

Next, we showed that terminal bases in the most prominent 3' tail are just as important in Cyanobacteria as they are in *E. coli* or *B. subtilis* because they are often more preferred in SD/anti-SD pairing than the canonical CCUCC (or CCUCCU) motif (Wei, Silke, Xia 2017). We defined an anti-SD site as preferred if the number of times the site was observed in pairing with putative SD sequence was significantly greater than expected (O:E > 1, $p < 0.05$, upper-tailed, See Materials and Methods). Here, we designated the characterized 3' tails (annotated in bold in Figure 2.1) as the anti-SD sequences. Table 2.1 shows that the preferred anti-SD sequence is 5'-UCCUUU-3' in

S. sp., 5'-UCCUU-3' in *M. aeruginosa*. However, a preferred anti-SD motif with ≥ 4 nt could not be obtained from *N. tabacum* chloroplast or *A. thaliana* chloroplast.

Table 2.1. Site-specific pairing preference between anti-SD and putative SD sequences in Cyanobacterial and chloroplast non-pseudo genes. O:E denotes the ratio of observed to expected pairs. In red are preferred anti-SD sites with significant O:E > 1 ($p < 0.05$, upper-tailed test).

anti-SD site	O:E *S.* sp.	O:E *M. aeruginosa*	O:E *N. tabacum* chloroplast	O:E *A. thaliana* chloroplast
G	0.97	1.03	0.71	1.45
A	0.98	1.13	1.22	1.32
U	0.79	0.92	0.88	1.54
C	0.69	0.79	0.74	1.27
A	0.72	0.78	0.76	1.05
C	0.68	0.65	0.69	0.85
C	0.94	0.87	0.97	0.59
U	1.11	1.11	1.19	0.67
C	1.15	1.21	1.73	0.77
C	1.35	1.30	2.75	0.73
U	1.36	1.58		
U	1.33	1.54		
U	1.63			

2.5.2 SD usage and SD/anti-SD location constraints are similar between Cyanobacteria and cyanophages but differ in chloroplasts

SD location constraints and motif usage differ between Cyanobacteria and chloroplasts. Figure 2.2a and 2.3b highlight that $D_{toStart}$ is similarly confined in *S.* sp. and *M. aeruginosa* but differ between Cyanobacteria and chloroplasts. In contrast, Figure 2.2c and 2.2d show that $D_{toStart}$ constraints in Cyanobacteria are well mimicked by cyanophages *S-SSM6a* and *S-SSM6b*. In addition, we found that usage of SD motifs located at optimal $D_{toStart}$ (Figure 2.3; See Materials and Methods for optimal $D_{toStart}$ identification) were similar between *S.* sp. and *M. aeruginosa* and they were well mimicked by cyanophages; however, SD motif usage varied drastically in

chloroplasts. A test for ordinal association using Kendall's tau-b statistic with adjustment for tied ranks showed SD motif usage is highly correlated between *S.* sp. and *M. aeruginosa* ($\underline{\tau_b}$ = 0.606, p < 0.00001, two-tailed) and cyanophages (*S-SSM6a*: $\underline{\tau_b}$ = 0.812, p < 0.00001, two-tailed; *S-SSM6b*: $\underline{\tau_b}$ = 0.809, p < 0.00001, two-tailed) but only moderately correlated between *S.* sp. and chloroplasts (*N. tabacum* chloroplast: $\underline{\tau_b}$ = 423, p = 0.00057, two-tailed; *A. thaliana* chloroplast: $\underline{\tau_b}$ = 0.339, p = 0.00482, two-tailed). In both Figures 2.3 and 2.4, we considered all putative SD sequences (2209 in *S.* sp., 3860 in *M. aeruginosa*, 62 in *N. tabacum* chloroplast, 50 in *A. thaliana* chloroplast, 297 in *S-SSM6a*, and 201 in *S-SSM6b*).

Figure 2.1 illustrates a secondary 3' tail peak in *S.* sp. (5'-GAUCACCUCCU-3') and in *A. thaliana* chloroplast (5'-GAUCACCUCCUUUUCAG-3'). We have additionally determined putative SD sequences using these alterative 3' tails and analyzed $D_{toStart}$ constraints and SD motif usage at optimal $D_{toStart}$ (Supplemental Figure S2.1). Similarly, Figure S2.1a and S2.1b highlight that $D_{toStart}$ constraints resemble among Cyanobacteria and cyanophages, whereas chloroplast SD sequences lost this resemblance. Additionally, Figure S2.1c shows that SD motif usage is again comparable between *S.* sp. and *M. aeruginosa* (τ_b = 0.621, p < 0.00001, two tailed) even though AAGGAG, AAGG, and AAGGA could not be determined in *S.* sp. when the shorter anti-SD 5'-GAUCACCUCCU-3' was used; whereas similarities in SD motif usage are again weaker between Cyanobacteria and chloroplasts (*N. tabacum* chloroplast: τ_b = 0.580, p < 0.00001, two tailed; A. thaliana chloroplast: τ_b = 0.169, p = 0.132, two-tailed).

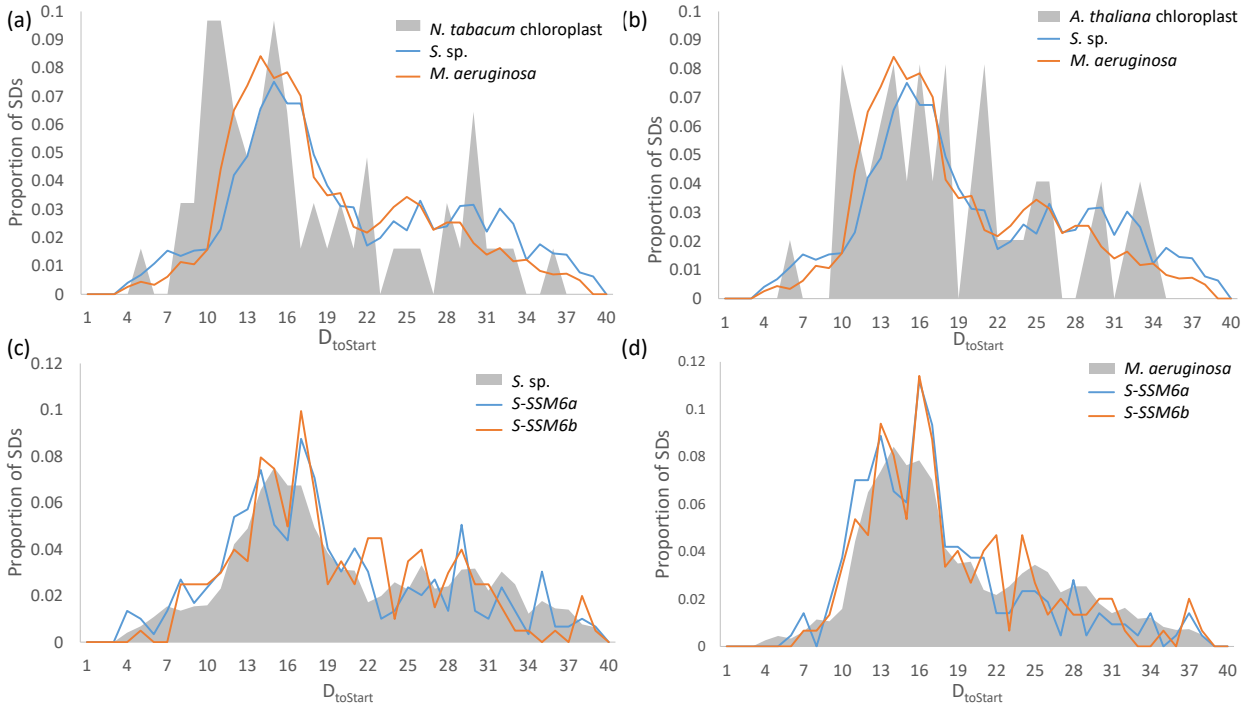


Figure 2.2. Comparisons of $D_{toStart}$ between Cyanobacteria, cyanophages, and chloroplasts. $D_{toStart}$ constraints differ between Cyanobacteria (*S.* sp. and *M. aeruginosa*) and (a) *N. tabacum* chloroplast and (b) *A. thaliana* chloroplast. Whereas $D_{toStart}$ constraints in (c) *S.* sp. and in (d) *M. aeruginosa* are well mimicked by cyanophages (*S-SSM6a* and *S-SSM6b*). Phage $D_{toStart}$ in (c) was obtained using *S.* sp. anti-SD 5'-GAUCACCUCCUUU-3', and in (d) was obtained using *M. aeruginosa* anti-SD 5'-GAUCACCUCCUU-3'. Shaded grey are $D_{toStart}$ constraints for outgroup species subjected to comparison.

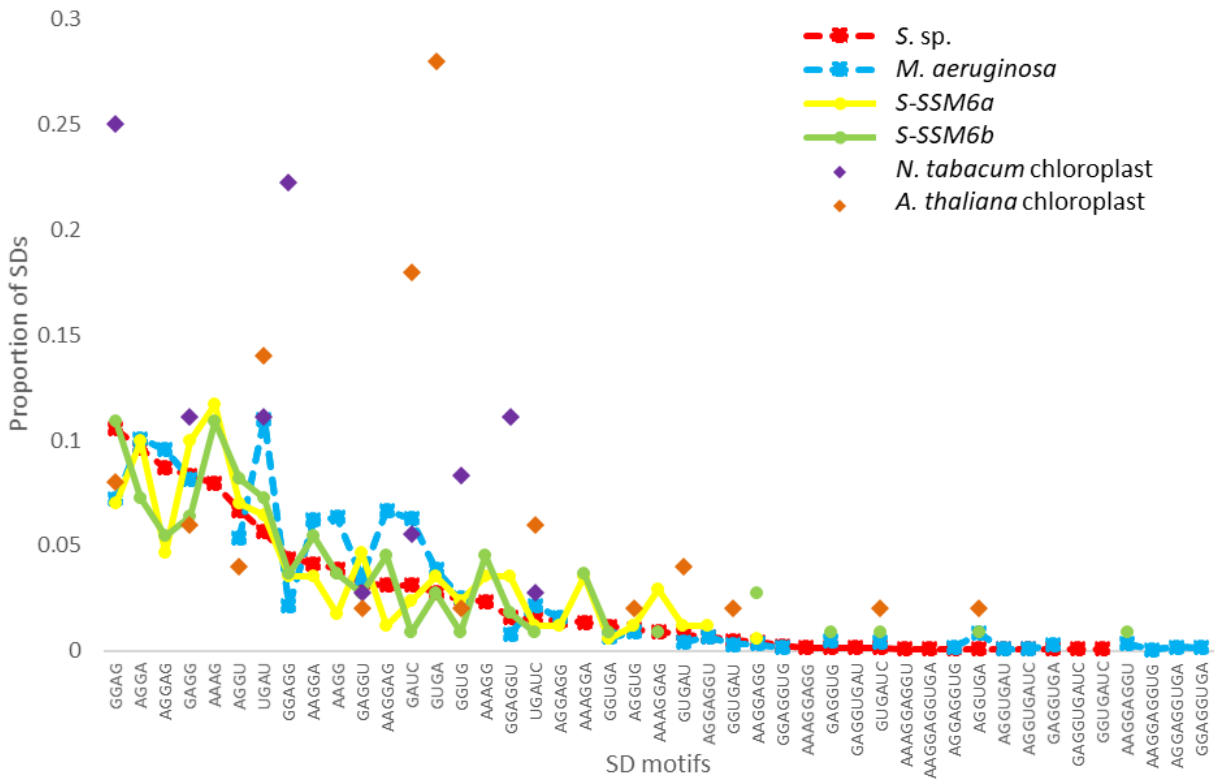


Figure 2.3. Comparison in SD motif usage between Cyanobacteria, cyanophages, and chloroplasts. Usage of SD motifs at optimal $D_{toStart}$ is comparable between Cyanobacteria *S.* sp. (red), *M. aeruginosa* (blue), cyanophage *S-SSM6a* (yellow), and cyanophage *S-SSM6b* (green), but varies appreciably in *N. tabacum* chloroplast (purple) and in *A. thaliana* chloroplast (orange). All data points presented have non-zero SD proportions.

2.5.3 SD-facilitated initiation increases Cyanobacterial gene expression but weakly affects chloroplast gene expression

Results above show differences in SD usage between Cyanobacteria and chloroplasts. To examine the effect of SD on gene expression, we measured how a well-positioned SD sequence (with optimal $D_{toStart}$) affects Cyanobacterial and chloroplast protein abundances. As expected, we observed a significant increase in protein abundance of SD-facilitated genes over other genes in *S.*

sp. (Figure 2.4a, Wilcoxon rank sum test with continuity correction: $p < 0.00001$) and in *M. aeruginosa* (Figure 2.4b, Wilcoxon rank sum test with continuity correction: $p < 0.00001$), but difference in protein abundances was not significant in *N. tabacum* chloroplast (Figure 2.4c, Wilcoxon rank sum test with continuity correction: $p = 0.429$) or in *A. thaliana* chloroplast (Figure 2.4d, Wilcoxon rank sum test with continuity correction: $p = 0.460$).

To further contrast translation efficiency between SD-facilitated and SD-independent genes, we also calculated protein per transcript (protein abundance/ average mRNA FPKM; Supplemental Figure S2.2) which normalized protein abundance to mRNA transcript abundance. For every species, we had retrieved RNA-Seq datasets from two independent studies (See Materials and Methods); therefore, two protein per transcript values were calculated for each gene. Figure S2.2 shows that protein per transcript values are significantly higher in SD-facilitated genes than SD-independent genes in *S.* sp (Wilcoxon rank sum test with continuity correction: $p < 0.0001$ in both panels), but protein per transcript values do not differ significantly between the two gene sets in *M. aeruginosa* (Wilcoxon rank sum test with continuity correction: $p = 0.201$ and 0.132), *N. tabacum* chloroplast (Wilcoxon rank sum test with continuity correction: $p = 0.931$ and 0.900), or *A. thaliana* chloroplast (Wilcoxon rank sum test with continuity correction: $p = 0.210$ and 0.135).

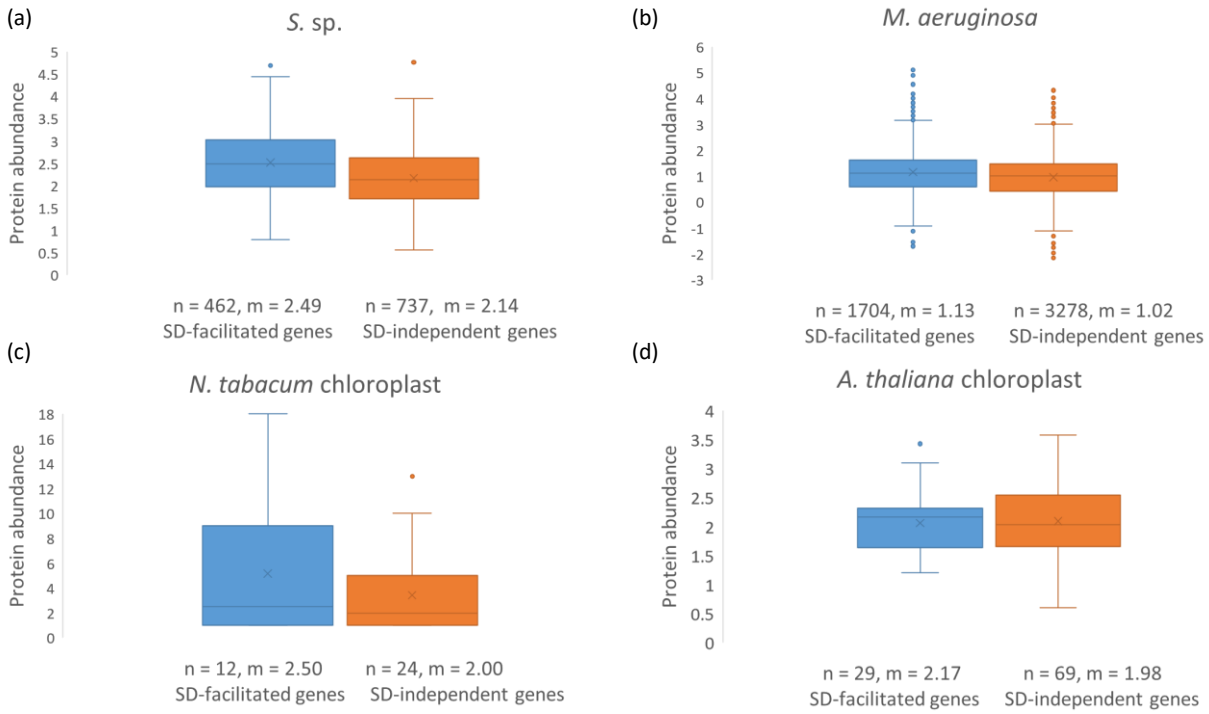


Figure 2.4. Contrasts in protein abundances between SD-facilitated genes (SD/anti-SD pair ≥ 4 nt and SD within optimal $D_{toStart}$) and SD-independent genes. Protein abundances in Cyanobacteria (a, b) were measured by log(ppm) retrieved from PaxDB database, in *N. tabacum* chloroplast (c) were measured by number of unique peptides retrieved from Wu, Yan (2018), and in *A. thaliana* chloroplast (d) were measured by log(total spectral counts) retrieved from Ferro et al. (2010) and curated as plastid proteins by Zybailov et al. (2008). All genes are non-pseudo with protein abundance > 0. For each gene set, the number of genes (n) and median (m) are denoted under the boxplots.

2.5.4 Translation initiation region in SD-independent genes is compensated with weaker secondary structure stability in Cyanobacteria but not in chloroplasts

In Figure 2.5 we compared the accessibility of ribosome binding sites between SD-facilitated and SD-independent genes in Cyanobacteria and chloroplasts. We defined a translation

initiation region to consist of 40 nt immediately upstream of the start codon because it encompassed all putative SD sequences in SD-facilitated genes. We observed a significant decrease in secondary structure stability (smaller -kcal/mol) in SD-independent genes over SD-facilitated genes in *S.* sp. (Wilcoxon rank sum test with continuity correction: $p = 0.00018$) and in *M. aeruginosa* (Wilcoxon rank sum test with continuity correction: $p < 0.00001$), but not in *N. tabacum* chloroplast (Wilcoxon rank sum test with continuity correction: $p = 0.158$) or in *A. thaliana* chloroplast (Wilcoxon rank sum test with continuity correction: $p = 0.840$).

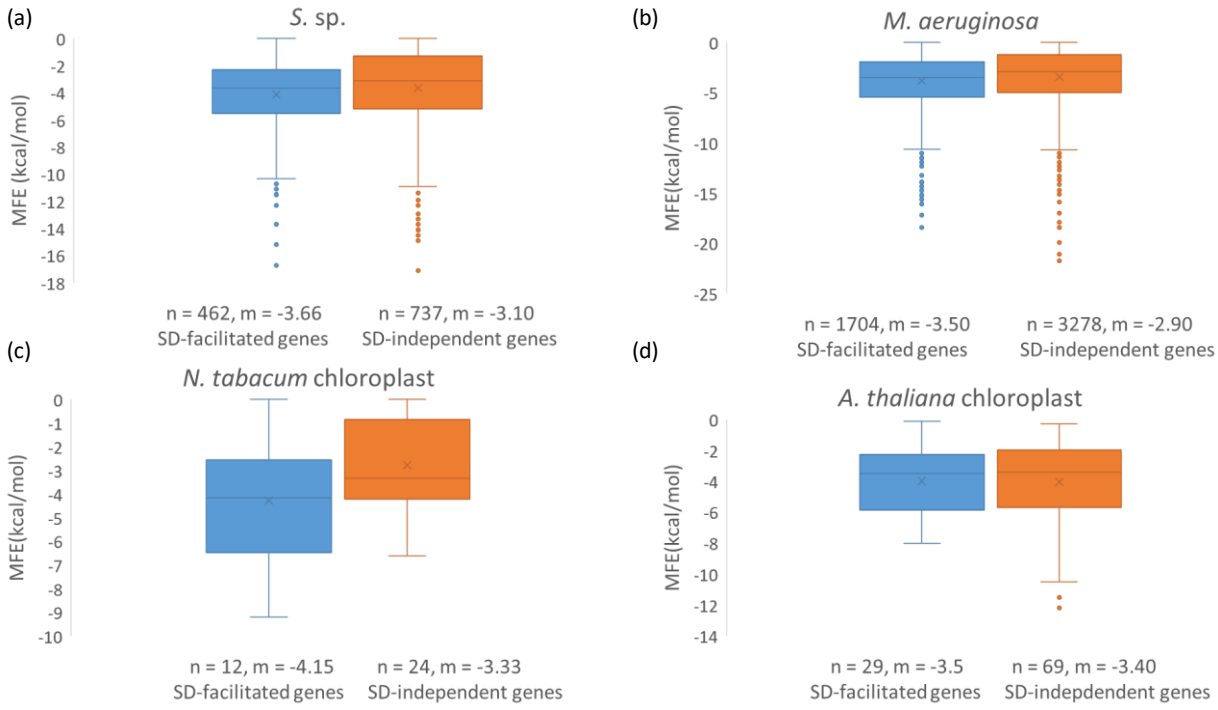


Figure 2.5. Contrasts in MFE (Kcal/mol) at the 40 nt initiation region immediately upstream of the start codon between SD-facilitated genes (SD/anti-SD pair ≥ 4 nt and optimal $D_{toStart}$) and SD-independent genes in (a) *S.* sp., (b) *M. aeruginosa*, (c) *N. tabacum* chloroplast, and (d) *A. thaliana* chloroplast. All genes are non-pseudo with non-zero protein abundances. For each gene set, the number of genes (n) and median (m) are denoted under the boxplots.

Figure 2.6 illustrates an example of the evolutionary advantage for SD-independent genes to have weakened secondary structure stability at translation initiation region in *S.* sp. Among the five genes with most abundant proteins in *S.* sp. characterized by Wegener et al. (2010), four genes exhibit properly positioned SD/anti-SD base-pairing interactions, but one ssl2598/*psbH*, whose protein happens to be the most abundant of all genes, cannot form SD/anti-SD interaction with a $D_{toStart}$ near the peak shown in Figure 2.2a. One would predict that ssl2598/*psbH* would need to have reduced secondary structure stability near the translation initiation region to compensate for the lack of a well-positioned SD/anti-SD. We calculated minimum folding energy (MFE) for the 40 nt immediately upstream of the start codon of these five genes. The result is consistent with the expected compensation effect between SD/anti-SD and secondary structure stability. The MFE (kcal/mol) is −10.95 for sll1577/*cpcB*, −5.81 for slr2067/*apcA*, −7.78 for sll1578/*cpcA*, and −7.08 for slr1986/*apcB*, but only −2.51 for ssl2598/*psbH*, with more negative MFEs corresponding to stronger secondary structures.

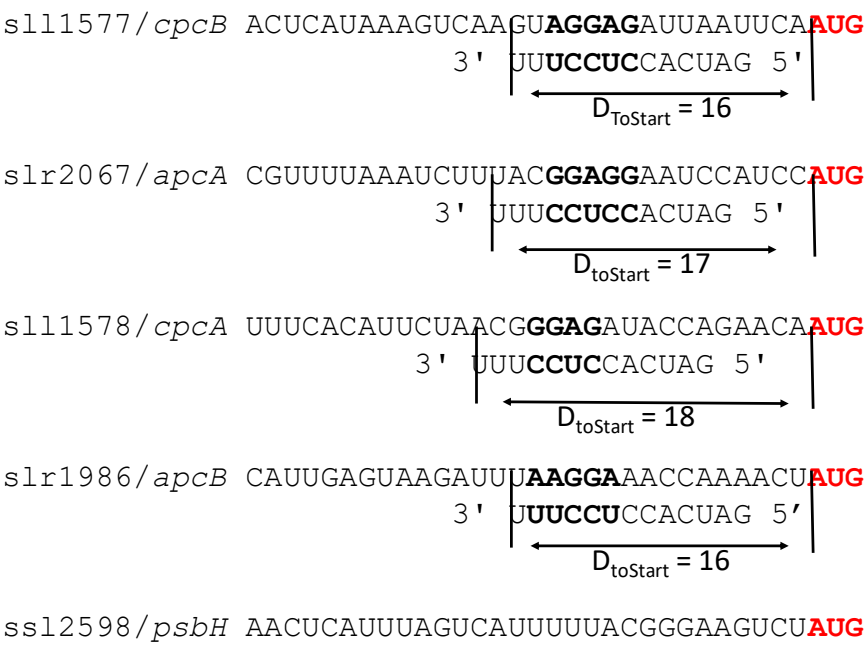


Figure 2.6. SD/anti-SD interactions can be found between mRNA and 3' tail of 16S rRNA in four of the five *S.* sp. genes with the most abundant proteins, but not in ssl2598/*psbH*. In red is the start codon, in bold the SD/anti-SD pair.

2.6 Discussion

There have been many debates over the conservation and function of SD sequence in Cyanobacteria and in chloroplasts. Elucidating SD role in Cyanobacterial and in plastid translation initiation is crucial to understand how these species may have differentially evolved. One of the first indications of differences between bacteria and chloroplast translation came from isolation of chloroplast ribosomal proteins that have no homologues in *E. coli* (Plader, Sugiura 2003). Because chloroplasts are presumably derived from Cyanobacteria through endosymbiosis, this difference between chloroplast and *E. coli* may be traced to Cyanobacteria. Therefore, contrasts in SD usage

and function between chloroplasts and Cyanobacteria have two distinct implications. If the location of SD sequence is poorly constrained and SD sequences weakly influence gene expression in both Cyanobacteria chloroplasts, then SD role in translation initiation may have already been weakened in Cyanobacteria prior to chloroplast endosymbiosis. However, if features of functional SD sequence in *E. coli* are retained in Cyanobacteria and are mimicked by cyanophages, then the SD mechanism is crucial for Cyanobacterial translation initiation and chloroplast SD function may have been weakened as a result of host-symbiont coevolution after symbiosis.

Knowing the full extent of 3' tail is required to identify the complete pool of functional SD motifs. We had previously devised an RNA-Seq-based approach to correct the mis-annotated 3' tail in several bacterial species (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018). Here, we rectified the mis-annotated 3' tails in *S.* sp., *M. aeruginosa*, *N. tabacum* chloroplast, and *A. thaliana* chloroplast. Using an O:E metric we showed that the preferred anti-SD motif is 5'-CCUCCUUU-3' in *S.* sp. and 5'-UCCUU-3' in *M. aeruginosa*, but a preferred anti-SD motif could not be determined in chloroplasts (Table 2.1). Like in *E. coli* and *B. subtilis*, the 3' terminal sequences extending downstream of the canonical CCUCCU are preferred in SD/anti-SD pairing in Cyanobacteria. This finding corroborates recent studies that suggested intermediate SD/anti-SD complementarity increases bacterial gene expression (Hockenberry et al. 2017; Scharff et al. 2017); whereas strong complementarity at CCUCCU may lead to ribosome stalling in *E. coli* (Li, Oh, Weissman 2012). Additionally, SD/anti-SD pairing position is crucial to SD function, and characterizing the most prominent 3' tail have allowed us to accurately determine SD/anti-SD location constraints using $D_{toStart}$ and to find functional SD sequences with proper juxtapositioning between start codon and ribosomal P site.

Our results provided three lines of evidence to suggest that the SD mechanism is important for Cyanobacterial translation initiation. First, we showed that SD/anti-SD pairing location, as measured by $D_{toStart}$, is well constrained in Cyanobacteria (Figure 2.2a). This constraint was similarly found for SD usage in *E. coli* and *B. subtilis* (Wei, Silke, Xia 2017). Second, both SD/anti-SD location constraints and SD motif usage in Cyanobacteria were well mimicked by cyanophages (Figures 2.3c, 2.3d, 2.4a). These similarities indicate host-phage specificity and emphasizes the importance of SD-facilitated initiation in Cyanobacteria because the same patterns were also observed in *E. coli* and coliphages (Prabhakaran, Chithambaram, Xia 2015). Third, SD-facilitated genes had higher protein abundance whereas SD-independent genes compensated by having weaker secondary structure stability at the initiation region (Figures 2.5, 2.6). This result is consistent with an previous study (Nakamoto 2006; Scharff et al. 2011) that showed weaker secondary structure stability at the 5' UTR in genes lacking a SD sequence in Proteobacteria and in Cyanobacteria. Because recruitment of ribosome is slower in absence of SD sequence, it is evolutionarily more advantageous for SD-independent genes in bacteria to adopt a structure-less initiation region in order to facilitate ribosome loading (Nakamoto 2006; Scharff et al. 2011). Indeed, SD sequence play an important role in Cyanobacterial gene expression.

In contrast, there were notable differences in SD usage between chloroplasts and Cyanobacteria (Figure 2.3). In addition, chloroplast $D_{toStart}$ was weakly constrained in comparison to Cyanobacteria (Figure 2.2). These results imply weakened SD function in chloroplasts in support for the theory that the mechanisms of translation initiation are subjected to different evolutionary pressure between bacteria and plastids. Resultantly, chloroplast SD-facilitated genes did not have significantly increased protein expressions (Figure 2.4), and other genes did not compensate by having weaker secondary structures at the 5' UTR (Figure 2.5). A caveat of this

study is the small number of unique proteins retrievable from chloroplasts: only 38 in *N. tabacum* chloroplast from Wu, Yan (2018) and 98 in *A.* thaliana chloroplast from Ferro et al. (2010) that were previously detected as plastid proteins by Zybailov et al. (2008). This may be attributed to the fact that chloroplast genomes have shrank significantly due to host-endosymbiont coevolution (Timmis et al. 2004) and about 120 genes remain in present-day green plant chloroplasts (Li, Oh, Weissman 2012; Zoschke, Bock 2018). This may entail that remaining chloroplast genes would carry SD sequences that resemble optimal ones used in Cyanobacteria, should the SD mechanism remains crucial to chloroplast translation initiation; however, this was not the case.

To examine SD role in translation efficiency, we additionally contrasted protein per transcript values between SD-facilitated genes and SD-independent genes. Similar to Figure 2.4, we found that protein per transcript levels were significantly higher in SD-facilitated genes than SD-independent genes in *S.* sp. but not in chloroplasts (Figure S2.2). Surprisingly, we observed no significant difference in protein per transcript between the two gene sets in *M. aeruginosa*. A plausible explanation for this result is that *M. aeruginosa* gene coverage in mRNA profiling is poor (88.8% coverage from PRJNA421714 and 76.4% coverage from PRJNA427104 for 4982 genes of interest, as opposed to > 95% coverage for all other species). In fact, for PRJNA427104, the authors' provided FPKM file has about 47% gene coverage. A plausible explanation for this poor gene coverage is that annotated genes in *M. aeruginosa* (NC_010286.1) are putative and determined by automated predictions based on sequence homology against other cyanobacterial species (Kaneko et al. 2007). A degree of mis-annotations is to be expected from automated predictions (Schnoes et al. 2009). Nevertheless, Figure S2.2 shows that SD-facilitated initiation modestly increases translation efficiency in the Cyanobacteria *S.* sp. but not in chloroplasts.

An alternative 3’ tail was found in *S.* sp. (5’-GAUCACCUCCU-3’) and *A. thaliana* chloroplast (5’-GAUCACCUCCUUUUCAG-3’). Similarly, we have previously reported 3’ tail heterogeneity in four other bacterial species (Silke, Wei, Xia 2018). These results suggest that in some species the mature 16S rRNA molecules may have a variety of end points rather than a single deterministic 3’ end, since the pre-16S rRNA 3’ end is independently processed by four RNases: RNase II, RNase R, PNPase, and YbeY (Sulthana, Deutscher 2013). Alternatively, a secondary peak extending far downstream from the conserved genomic sequence 5’-GAUCACCUCCUU(U or A) (Nakagawa et al. 2010) may constitute the pre-16S rRNA, which is accumulated (Sulthana, Deutscher 2013) because the localization of RNases to this precursor sequence is a rate limiting step during 3’ maturation. Nonetheless, we determined putative SD sequences using this alternative 3’ tail in *S.* sp. and *A. thaliana* chloroplast and investigated $D_{toStart}$ constraints and SD motif usage (Supplemental Figure S2.1). Again, we found that both $D_{toStart}$ and SD motif usage were similar between *S.* sp. and *M. aeruginosa*, but they differed notably in *A. thaliana* chloroplast.

Our results suggest that the SD mechanism is crucial to Cyanobacterial gene expression, and cyanophages mimic Cyanobacteria in SD sequence usage since it is evolutionarily advantageous for bacteriophages to take up features that increase host gene expression. On the contrary, chloroplasts lost this similarity because the need for SD-facilitated initiation is reduced in plastids having much reduced genome size (Timmis et al. 2004), different ribosomal proteins (Plader, Sugiura 2003), and other possible mechanisms of translation initiation (Plader, Sugiura 2003; Hirose, Sugiura 2004) as a result of host-symbiont coevolution. Nonetheless, because the most recent common ancestor (MRCA) of chloroplast is unknown, our findings does not exclude the possibility that the importance of SD/anti-SD base-pairing might have already been decreased in MRCA.

2.7 Data availability

Supplemental figures for Chapter 2 are available in Appendix B. Supplemental files are available at https://academic.oup.com/gbe/article/11/11/3194/5588849#173567624: Supplemental File S1 contains RNA-Seq mapped reads and 3' tail characterization for all species; S2 contains the identification of SD motifs, $D_{toStart}$, secondary structure stability, summarized descriptions for Materials and Methods and statistical tests; S3 contains the protein abundance data, mRNA FPKMs, and protein per transcript values.

2.8 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to X.X. [RGPIN/2018-03878], and NSERC Doctoral Scholarship to Y.W. [PGSD/2019-535291]. We thank Jordan R. Silk for comments.

2.9 References

Abolbaghaei, A, JR Silke, X Xia. 2017. How Changes in Anti-SD Sequences Would Affect SD Sequences in Escherichia coli and Bacillus subtilis. G3 (Bethesda).

Altschul, SF, W Gish, W Miller, EW Myers, DJ Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Andersson, DI, C G Kurland. 1983. RAM ribosomes are defective proofreaders. Molecular & general genetics 191:378-381.

Armarego-Marriott, T, Ł Kowalewska, A Burgos, et al. 2019. Highly Resolved Systems Biology to Dissect the Etioplast-to-Chloroplast Transition in Tobacco Leaves. Plant physiology 180:654-681.

Bolger, AM, M Lohse, B Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Cangelosi, GA, WH Brabant. 1997. Depletion of pre-16S rRNA in starved Escherichia coli cells. J Bacteriol 179:4457-4463.

Chang, SH, S Lee, TY Um, JK Kim, Y Do Choi, G Jang. 2017. pTAC10, a Key Subunit of Plastid-Encoded RNA Polymerase, Promotes Chloroplast Development. Plant physiology 174:435-449.

Duval, M, A Simonetti, I Caldelari, S Marzi. 2015. Multiple ways to regulate translation initiation in bacteria: Mechanisms, regulatory circuits, dynamics. Biochimie 114:18-29.
Ewing, B, P Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194.
Fargo, DC, M Zhang, NW Gillham, JE Boynton. 1998. Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. Mol Gen Genet 257:271-282.
Ferro, M, S Brugiere, D Salvi, et al. 2010. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol Cell Proteomics 9:1063-1084.
Gallaher, SD, ST Fitz-Gibbon, D Strenkert, SO Purvine, M Pellegrini, SS Merchant. 2018. High-throughput sequencing of the chloroplast and mitochondrion of Chlamydomonas reinhardtii to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. Plant J 93:545-565.
Hirosawa, M, T Sazuka, T Yada. 1997. Prediction of translation initiation sites on the genome of Synechocystis sp. strain PCC6803 by Hidden Markov model. DNA Res 4:179-184.
Hirose, T, T Kusumegi, M Sugiura. 1998. Translation of tobacco chloroplast rps14 mRNA depends on a Shine-Dalgarno-like sequence in the 5'-untranslated region but not on internal RNA editing in the coding region. FEBS Lett 430:257-260.
Hirose, T, M Sugiura. 2004. Functional Shine-Dalgarno-like sequences for translational initiation of chloroplast mRNAs. Plant Cell Physiol 45:114-117.
Hockenberry, AJ, AR Pah, MC Jewett, LAN Amaral. 2017. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. Open Biology 7.
Hofacker, IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res 31:3429-3431.
Jones, CE, AL Brown, U Baumann. 2007. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 8:170.
Kaneko, T, N Nakajima, S Okamoto, et al. 2007. Complete genomic structure of the bloom-forming toxic cyanobacterium Microcystis aeruginosa NIES-843. DNA research : an international journal for rapid publication of reports on genes and genomes 14:247-256.
Kudla, G, AW Murray, D Tollervey, JB Plotkin. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324:255-258.
Lagesen, K, P Hallin, EA Rodland, HH Staerfeldt, T Rognes, DW Ussery. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100-3108.
Li, G-W, E Oh, JS Weissman. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538-541.
Lin, YH, BC Chang, PW Chiang, SL Tang. 2008. Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes. Gene 416:44-47.
Ma, J, A Campbell, S Karlin. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733-5745.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

Mutsuda, M, M Sugiura. 2006. Translation initiation of cyanobacterial rbcS mRNAs requires the 38-kDa ribosomal protein S1 but not the Shine-Dalgarno sequence: development of a cyanobacterial in vitro translation system. J Biol Chem 281:38314-38321.

Nakagawa, S, Y Niimura, T Gojobori. 2017. Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. Nucleic Acids Res 45:3922-3931.

Nakagawa, S, Y Niimura, K-i Miura, T Gojobori. 2010. Dynamic evolution of translation initiation mechanisms in prokaryotes. Proceedings of the National Academy of Sciences 107:6382-6387.

Nakamoto, T. 2006. A unified view of the initiation of protein synthesis. Biochem Biophys Res Commun 341:675-678.

Nickelsen, J, M Fleischmann, E Boudreau, M Rahire, JD Rochaix. 1999. Identification of cis-acting RNA leader elements required for chloroplast psbD gene expression in Chlamydomonas. Plant Cell 11:957-970.

Oey, M, M Lohse, B Kreikemeyer, R Bock. 2009. Exhaustion of the chloroplast protein synthesis capacity by massive expression of a highly stable protein antibiotic. Plant J 57:436-445.

Osterman, IA, SA Evfratov, PV Sergiev, OA Dontsova. 2013. Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic acids research 41:474-486.

Pasoreck, EK, J Su, IM Silverman, SJ Gosai, BD Gregory, JS Yuan, H Daniell. 2016. Terpene metabolic engineering via nuclear or chloroplast genomes profoundly and globally impacts off-target pathways through metabolite signalling. Plant Biotechnol J 14:1862-1875.

Plader, W, M Sugiura. 2003. The Shine-Dalgarno-like sequence is a negative regulatory element for translation of tobacco chloroplast rps2 mRNA: An additional mechanism for translational control in chloroplasts. The Plant journal 34:377-382.

Prabhakaran, R, S Chithambaram, X Xia. 2015. Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. J Gen Virol 96:1169-1179.

Sazuka, T, O Ohara. 1996. Sequence features surrounding the translation initiation sites assigned on the genome sequence of Synechocystis sp. strain PCC6803 by amino-terminal protein sequencing. DNA Res 3:225-232.

Scharff, LB, L Childs, D Walther, R Bock. 2011. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. PLoS Genet 7:23.

Scharff, LB, M Ehrnthaler, M Janowski, LH Childs, C Hasse, J Gremmels, S Ruf, R Zoschke, R Bock. 2017. Shine-Dalgarno Sequences Play an Essential Role in the Translation of Plastid mRNAs in Tobacco. Plant Cell 29:3085-3101.

Schnoes, AM, SD Brown, I Dodevski, PC Babbitt. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:11.

Schurr, T, E Nadir, H Margalit. 1993. Identification and characterization of E.coli ribosomal binding sites by free energy computation. Nucleic Acids Res 21:4019-4023.

Shine, J, L Dalgarno. 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 71:1342-1346.

Shine, J, L Dalgarno. 1975. Determinant of cistron specificity in bacterial ribosomes. Nature 254:34-38.

Silke, JR, Y Wei, X Xia. 2018. RNA-Seq-Based Analysis Reveals Heterogeneity in Mature 16S rRNA 3' Termini and Extended Anti-Shine-Dalgarno Motifs in Bacterial Species. G3 8:3973-3979.
Starmer, J, A Stomp, M Vouk, D Bitzer. 2006. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. PLoS Comput Biol 2:e57.
Sugiura, M, T Hirose, M Sugita. 1998. Evolution and mechanism of translation in chloroplasts. Annu Rev Genet 32:437-459.
Sulthana, S, MP Deutscher. 2013. Multiple exoribonucleases catalyze maturation of the 3' terminus of 16S ribosomal RNA (rRNA). J Biol Chem 288:12574-12579.
Timmis, JN, MA Ayliffe, CY Huang, W Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-135.
Tuller, T, YY Waldman, M Kupiec, E Ruppin. 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 107:3645-3650.
Wang, M, CJ Herrmann, M Simonovic, D Szklarczyk, C von Mering. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163-3168.
Wegener, KM, AK Singh, JM Jacobs, et al. 2010. Global proteomics reveal an atypical strategy for carbon/nitrogen assimilation by a cyanobacterium under diverse environmental perturbations. Mol Cell Proteomics 9:2678-2689.
Wei, Y, JR Silke, X Xia. 2017. Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in Escherichia coli and Bacillus subtilis using RNA-Seq data. Sci Rep 7:017-17918.
Williams, CR, A Baccarella, JZ Parrish, CC Kim. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17:103.
Wu, W, Y Yan. 2018. Chloroplast proteome analysis of Nicotiana tabacum overexpressing TERF1 under drought stress condition. Botanical Studies 59:26.
Xia, X. 2017. ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. G3: Genes|Genomes|Genetics 7:3839-3848.
Xia, X. 2018. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol Biol Evol 35:1550-1552.
Ye, GN, PT Hajdukiewicz, D Broyles, D Rodriguez, CW Xu, N Nehra, JM Staub. 2001. Plastid-expressed 5-enolpyruvylshikimate-3-phosphate synthase genes provide high level glyphosate tolerance in tobacco. Plant J 25:261-270.
Zhao, X, J Huang, J Chory. 2019. GUN1 interacts with MORF2 to regulate plastid RNA editing during retrograde signaling. Proc Natl Acad Sci U S A 116:10162-10167.
Zoschke, R, R Bock. 2018. Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. Plant Cell 30:745-770.
Zybailov, B, H Rutschow, G Friso, A Rudella, O Emanuelsson, Q Sun, KJ van Wijk. 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS One 3:0001994.

CHAPTER THREE

An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria

This chapter is a modified version of the publication below, in that 3.3 Introduction is partly removed, reorganized, and rewritten into Chapter 1.1.2:

Wei, Y., Silk, JR., Xia, X. 2019. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Scientific Reports. 9:3184

3.1 Author Contributions

Yulong Wei[1], Jordan R. Silke[1], 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.

Y.W. and X.X. designed the study. Y.W and J.R.S. wrote the manuscript text, and X.X. provided guidance and editing. Y.W. and J.R.S. collected and analyzed the data. X.X. developed the computer programs. Y.W. made Figures 3.1, 3.2, 3.3, S3.3, S3.4, S3.5, Tables 3.1, 3.2, S3.1, and Supplemental File S2 and File S3. J.R.S. made Figure S3.1, S3.2, Table 3.3, and Supplemental File S1. All authors reviewed the manuscript. X.X. supervised the project. Y.W. and J.R.S. contributed equally to this work.

3.2 Abstract

The degree to which codon usage can be explained by tRNA abundance in bacterial species is often inadequate, partly because relative tRNA abundance is often poorly approximated by tRNA copy numbers. To better understand the coevolution between tRNA abundance and codon usage, we provide a better estimate of tRNA abundance by profiling tRNA mapped reads (tRNA tpm) using publicly available RNA Sequencing data. To emphasize the feasibility of our approach, we demonstrate that tRNA tpm is consistent with tRNA abundances derived from RNA fingerprinting experiments in *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica*. Furthermore, we do not observe an appreciable reduction in tRNA sequencing efficiency due to post-transcriptional methylations in seven bacteria studied. To determine optimal codons, we calculate codon usage in highly and lowly expressed genes measured by protein per transcript. We found that tRNA tpm is sensitive to identify more translationally optimal codons than gene copy number and early tRNA fingerprinting abundances. Additionally, tRNA tpm improves the predictive power of tRNA adaptation index to explain codon preference.

3.3 Introduction

As discussed in Chapter 1.1.2, codon optimization is important to researchers seeking to improve protein production. Codon preference, especially in highly expressed genes (HEGs), is largely attributed to tRNA availability. This effect is coined as the tRNA-mediated codon usage bias and has been broadly observed in a variety of organisms, including the gram-negative *Escherichia coli* and *Salmonella enterica* serovar typhimurium (Ikemura 1985), the gram-positive *Bacillus subtilis* (Kanaya et al. 1999), eukaryotes such as yeast (Ikemura 1985; Xia 1998), a variety of fungal and invertebrate mitochondrial genomes (Carullo, Xia 2008; Xia 2008), viruses such as HIV (van Weringh et al. 2011), and bacteriophages (Chithambaram, Prabhakaran, Xia 2014a; Prabhakaran, Chithambaram, Xia 2015).

To incorporate tRNA availability in the identification of translationally optimal codons, many codon usage indices use tRNA gene copy as proxy. These include the Codon Bias Index (CBI) (Bennetzen, Hall 1982), Frequency of Optimal Codons (F_{op}) (Ikemura 1985), and tRNA Adaptation Index (tAI) (dos Reis, Savva, Wernisch 2004). All three of these indices define a translationally optimal codon as one that corresponds to the most abundant isoacceptor tRNA, with CBI additionally incorporating gene expression information.

3.3.1 Gene copy number is a poor estimator of tRNA abundance

The use of tRNA gene copy is often undesirable. This is exemplified in *B. subtilis* (dos Reis, Savva, Wernisch 2004), in which tAI fails to predict codon usage although codon usage correlates strongly with experimental tRNA abundance (Kanaya et al. 1999) and conforms well to the selection-mutation-drift theory (Bulmer 1991; Sharp et al. 2005). Moreover, while tRNA copy

number is correlated with codon usage in many fast-growing bacterial species having genomes that contain high numbers of tRNA gene copies, slow-growing species exhibit little tRNA gene redundancy with many tRNA genes existing as a single copy. Resultantly, codon usage in slow-growing species is poorly predicted by tRNA gene copy number (Rocha 2004). For example, *Leptospira interrogans* and *Mycobacterium tuberculosis* have only 25 and 45 annotated tRNA genes, respectively, according to the Genomic tRNA Database (GtRNAdb) (Chan, Lowe 2009). Obviously, variation in codon usage cannot be well explained by tRNA gene redundancy if there is little variation in tRNA gene copy number. It stands to reason that the coevolutionary relationship between tRNA abundance and codon usage can be better characterized if tRNA abundance can be measured more accurately.

3.3.2 A new RNA-Seq approach to estimate tRNA abundance

To accurately estimate tRNA abundance in bacteria, RNA-Seq profiling shows promise as a better alternative. RNA-Seq data are publicly available for many bacteria, and reads can be mapped to tRNA genes deposited in GtRNAdb and quantified in transcripts per kilobase million (tpm) using kallisto (Bray et al. 2016). To improve mapping efficacy, we have recently developed a new tool to process RNA-Seq data, ARSDA (Xia 2017), that stores identical reads as single entries to drastically reduce data storage and computation time for analyzing large RNA-Seq datasets relative to previous methods (Leinonen, Sugawara, Shumway 2011; Kodama, Shumway, Leinonen 2012).

It is worth mentioning that in eukaryotes, tRNA sequences generated via standard Illumina protocols are of limited use in gene profiling due to the well-known issue of post-transcriptional methylation that occurs at numerous tRNA sites (e.g., N^1-methyladenosine (m^1A), N^3-

methylcytosine (m^3C) and N^1-methylguanosine (m^1G) structural modifications) to hinder cDNA reverse transcription (Cozen et al. 2015; Zheng et al. 2015). In order to generate full length eukaryotes tRNA reads, both DM-tRNA-seq (Zheng et al. 2015) and ARM-seq (Cozen et al. 2015) approaches use the *E. coli* derived AlkB demethylase enzyme to efficiently removed tRNA methylation (Falnes et al. 2004) before RNA sequencing. As such, one may reason that AlkB treatments may not be necessary to remove tRNA methylations in bacteria that naturally encode their own AlkB homologs. Besides *E. coli* (Falnes, Johansen, Seeberg 2002; Falnes et al. 2004; Cozen et al. 2015; Zheng et al. 2015), several lines of evidence suggest AlkB homologs are present in other bacterial species (van den Born et al. 2009; Schulz et al. 2010; Wang et al. 2010), such as *B. subtilis* (Gao et al. 2016), *S. enterica*, *Mycobacterium tuberculosis* (van den Born et al. 2009; Nie et al. 2014), *Synechocystis* species (van den Born et al. 2009; Cassier-Chauvat, Veaudor, Chauvat 2016), and species in *Leptospira* (van den Born et al. 2009; Nie et al. 2014) and *Bacteroides* (van den Born et al. 2008) genera. Regardless, a key point of our investigation is determining whether bacterial tRNA sequences generated from standard Illumina protocols can be reliably used for abundance profiling.

We surveyed seven bacterial species (*E. coli*, *B. subtilis*, and *S. enterica*, *B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* sp.). To our knowledge, our results are the first to show that tRNA quantification by RNA-Seq data is well correlated with early tRNA abundances derived from RNA fingerprinting experiments (i.e., separating radiolabelled RNA by 2D gel electrophoresis followed by quantification of radioactivity) reported previously in *E. coli*, *S. enterica* and *B. subtilis* (Ikemura 1985; Dong, Nilsson, Kurland 1996; Xia 1998; Kanaya et al. 1999). Despite the challenges associated with tRNA sequencing in yeast (Pang et al. 2014; Cozen et al. 2015) and mammals (Cozen et al. 2015; Zheng et al. 2015; Loher, Telonis, Rigoutsos

2018), our results suggest that tRNA methylation may not appreciably influence tRNA sequencing efficiency in the bacterial species studied herein. We devised an integrated approach to show that tRNA tpm better predicts translationally optimal codons in *E. coli* than F_{op} and improves the predictive power of tAI to explain codon preference. We found that the dependence of codon preference on tRNA availability is not always stronger in fast-growing species, and optimal codons can be well explained by tRNA availability in certain slow-growing species. Conversely, tRNA availability is better optimized to codon usage in highly expressed genes of fast-growing than slow-growing species.

3.4 Materials and Methods

3.4.1 Processing genomic, proteomic, and RNA-seq data

We retrieved the annotated genomes for three fast growing species (*E. coli*, *B. subtilis*, and *S. enterica*) and four slow-growing species (*B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* sp.) in GenBank format from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). These species were selected because their protein abundance data are available in PaxDb (Wang et al. 2015), their growth rates are described on the basis of generation time (bacteria with > 2.5 hour generation times are considered slow growing and all those with lower generation times are fast growing) (Rocha 2004), and their RNA-Seq data are available (GEO Datasets). In addition, all documented tRNA methyltransferase genes in each species were retrieved from GenBank annotations (Supplemental File S1).

Protein abundance data of the seven surveyed species were retrieved from PaxDb 4.0 (Wang et al. 2015) and protein IDs were associated with Gene IDs retrieved from the genomes using DAMBE 7 (Xia 2018c). The integrated protein abundance dataset was taken when available.

RNA-Seq runs of wildtype species were fetched from GEO DataSets (https://www.ncbi.nlm.nih.gov/gds/) in FASTQ format. In total, we employed 14 publicly available RNA-Seq datasets, two for each of the seven species studied. FASTQ files were converted to FASTQ+ format using ARSDA 1.1 (Xia 2017) in order to reduce file sizes by grouping identical reads under a single ID while retaining the copy number for each read (in the format of ID_copy#). The FASTQ+ datasets were then processed to remove flanking adapter sequences and to purge low quality reads. For experiments that use the oligo(dT)-adapter primer for cDNA synthesis, RNA fragments are first poly-adenylated at the 3' end. In these cases, we use CutAdapt 1.17 (Martin 2011) and set to recognize "AAAAA". We also used CutAdapt to recognize and remove all possible adapters in experiments that used custom adapters, with 10% mismatch error rate. When the adapters conformed to standard Illumina protocols, we simply used the "ILLUMINACLIP" function built into Trimmomatic 0.38 (Bolger, Lohse, Usadel 2014) to trim the relevant library of adapters from reads (ILLUMINACLIP:<adapters.fa>:2:20:6). After adapters were trimmed, we retained reads that were a minimum of 25 nt long (-m 25 in CutAdapt or MINLEN:25 in Trimmomatic) to mitigate bias in expression levels (Williams et al. 2016). Lastly, we filtered trimmed reads to remove poor quality sequences with average Phred scores lower than 30 (0.1% probability of a base calling error) (Ewing et al. 1998).

3.4.2 Mapping RNA-Seq read to tRNAs and mRNAs

We retrieved all tRNA sequences from each bacterial genome deposited in the Genomic tRNA Database (GtRNAdb 2.0) (Chan, Lowe 2016). Predicted pseudo-tRNAs and those with unspecified anticodons were removed, and identical tRNA sequences were grouped under the same ID that indicate the number of identical copies. Since mature tRNAs are post-transcriptionally modified to have 5'-CCA-3' appended to their 3' end, we manually added 5'-CCA-3' to sequences lacking this motif (Zheng et al. 2015).

For each species, the tRNA FASTA file was indexed and RNA-Seq reads from processed FASTQ+ files were pseudo-aligned to each tRNA index using kallisto v0.44.0 (Bray et al. 2016) to quantify tRNA tpm. The tRNA pseudo-alignments were sorted and site-specific depth values were generated for each tRNA using the sorted pseudo-alignments via 'sort' and 'depth' commands from SAMtools (Li et al. 2009), respectively.

Similarly, for each species, all non-pseudo and non-hypothetical coding DNA sequences with non-zero PaxDb protein abundance values were retrieved using DAMBE in FASTA format and indexed. Then, RNA-Seq reads from processed FASTQ+ files were pseudo-aligned to the mRNA index using kallisto to quantify mRNA tpm values.

3.4.3 Determining highly and lowly translated genes by protein per transcript

In each species except *B. subtilis*, protein per transcript (ppm/tpm) was measured twice using both RNA-Seq datasets. In *B. subtilis*, protein per transcript was obtained with only SRX515181 dataset, because SRX2804667 MiSeq experimental protocol effectively removes large mRNA transcripts to specifically study tRNAs (Ernst et al. 2018). From each species, genes

with top and bottom 30% ppm/tpm values were selected (Supplemental File S3, Table S3.1). A gene is considered highly translated (highly expressed gene, HEG) if the gene ID is found in both top 30% gene sets, and a gene is considered lowly translated (lowly expressed gene, LEG) if the gene ID is found in both bottom 30% gene sets. We also determined for presence of ribosomal protein (30S and 50S subunit) genes in all gene sets. We observed a great number of ribosomal protein genes in the top 30% gene sets and nearly none in the bottom 30% gene sets (Supplemental Table S3.1). This verifies our approach since ribosomal protein genes are highly translated (Sharp, Li 1987; Rocha 2004).

3.4.4 Computing relative synonymous codon usage and tRNA usage metrics

Relative synonymous codon usage (RSCU) (Sharp, Li 1986) were computed for each species by loading HEGs and LEGs (Supplemental File S3) into DAMBE and selecting "Seq. Analysis" > "Codon Usage" > "Relative synonymous codon usage". DAMBE's implementation of RSCU automatically splits 6-fold degenerate codon families into a 2-fold and 4-fold degenerate family to separate synonymous codons that differ at the first nucleotide position.

To acquire relative tRNA usage for each synonymous codon, we adapted the RSCU formula to calculate RTU (1) in the same way that Relative tRNA Gene frequency was calculated by Novoa et al. (2012) using tRNA gene copy number:

$$RTU_i = \frac{tpm\ of\ tRNA_i}{\frac{\sum_i^n tRNA_i}{n}} \qquad (1)$$

Where *i* is any codon within a 2 or 4-fold degenerate codon family, and *n* is the total number of codons in the synonymous group. RSCU and RTU are both calculated by breaking 6-fold codon

families (Arginine, Leucine, and Serine) into 2 and 4-fold groups (e.g., 4-fold CGN (arg) and 2-fold AGN (arg)). Methionine and Tryptophan have been omitted from RSCU and RTU calculation because they are encoded by a single codon (with RSCU and RTU = 1). For all regression analyses, correlation coefficients (R^2) are calculated by taking the square of Pearson's correlation r.

3.4.5 Computation of tAI and correlation *S*

We first calculated tAI using the original formulation of the model (dos Reis, Wernisch, Savva 2003; dos Reis, Savva, Wernisch 2004) which considers the copy number of all isoacceptor tRNAs for each codon via the tAI R package version 0.2 (https://github.com/mariodosreis/tai) for all species surveyed in this study. We additionally computed a modified version of tAI (tAI') useing the summed tpm values associated with codon-specific isoacceptor tRNAs in lieu of tRNA copy number. Rather than using all annotated DNA coding sequences in these calculations, as was done originally (dos Reis, Savva, Wernisch 2004) , we only considered genes with non-zero protein abundance values from the integrated datasets stored in PaxDb. The tAI and tAI′ values for each species were then plotted against the effective number of codons, corrected for silent substitutions at the third codon position (f[GC3s] - Nc), to determine the *S* and *S'* correlation coefficients, respectively.

3.5 Results

3.5.1 tRNA abundances profiled via RNA-Seq mappings are consistent with experimental estimates

To accurately profile tRNA transcripts in tpm, we processed RNA-Seq data to remove adapters and low-quality sequences (See Materials and Methods for more detail) and quantified

reads mapped to all unique tRNA sequences (Supplemental File S1) in tRNA tpm. To demonstrate the fidelity of tRNA tpm in bacteria, we compared these values with RNA fingerprinting abundances (Figure 3.1, Supplemental File S2) previously reported in *E. coli* (Ikemura 1985; Dong, Nilsson, Kurland 1996), *S. enterica* (Ikemura 1985), and *B. subtilis* (Kanaya et al. 1999). In all cases, tRNA tpm correlates with RNA fingerprinting abundance (Figure 3.1: $R^2 > 0.4$, $P < 0.05$).

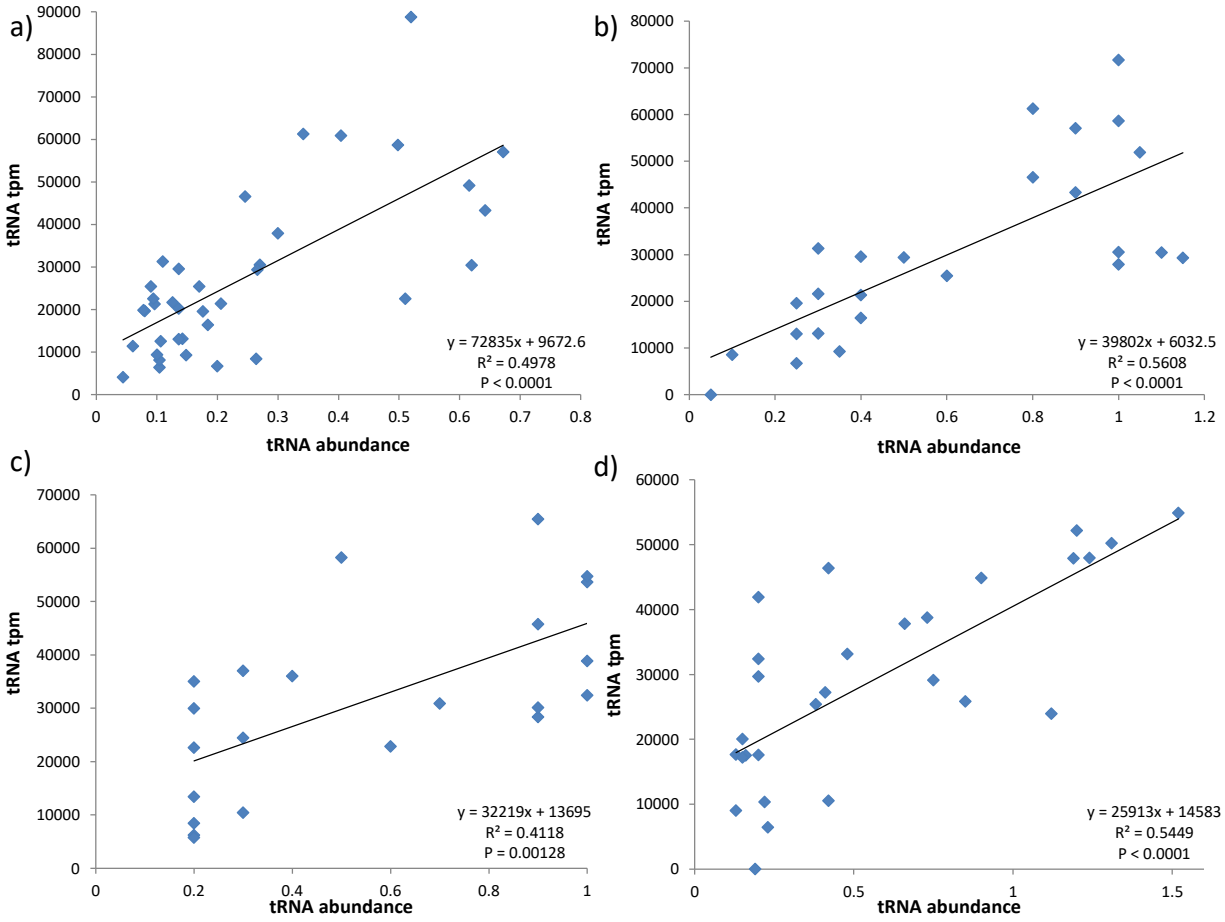


Figure 3.1. Comparison between tRNA tpm and fingerprinting abundance in *E. coli* (a and b), *S. enterica* (c), and *B. subtilis* (d). The X-axis shows experimental tRNA abundances: in a) averaged *E. coli* tRNA abundances across five growth phases retrieved from Dong, Nilsson, Kurland (1996), in b) and c) *E. coli* and *S. enterica* tRNA abundances retrieved from Ikemura (1985), and in d) *B. subtilis* tRNA abundances retrieved from Kanaya et al. (1999).

3.5.2 Documented tRNA methylation does not appreciably affect tRNA sequencing in bacteria studied

To investigate the potential effects of site-specific tRNA methylation on standard RNA-Seq experiments in bacteria, we visualized the RNA-Seq read depths of all seven species studied. In *E. coli* (Figure 3.2), read depths before and after documented tRNA methylation sites (18, 32, 34, 37, 46, and 54) in GenBank annotation (NC_000913, Supplemental File S1) do not vary substantially, and we observe no partial tRNA mappings (Figure 3.2a-d) contrary to the "hard-stops" previously described in both yeast and human tRNA sequencing in the absence of demethylation treatment (Cozen et al. 2015). Additionally, tpm values of sets of tRNAs that can be potentially modified at five or all six documented methylation sites do not differ substantially from tpm values of set of tRNAs that can be potentially modified at four or less sites (Figure 3.2d-e; two-tailed Student's t-test with unequal variance: $P = 0.477$). To distinguish sets of tRNAs by levels of methylation, we defined tRNAs that can potentially be methylated at >4 sites as heavily methylated with respect to other tRNAs. This cut-off was chosen because it divided the 50 unique tRNA sequences into two subsets of roughly equal size. Similarly, hard-stops were not observed at documented methylated sites in all six other bacteria studied (Supplemental File S1).

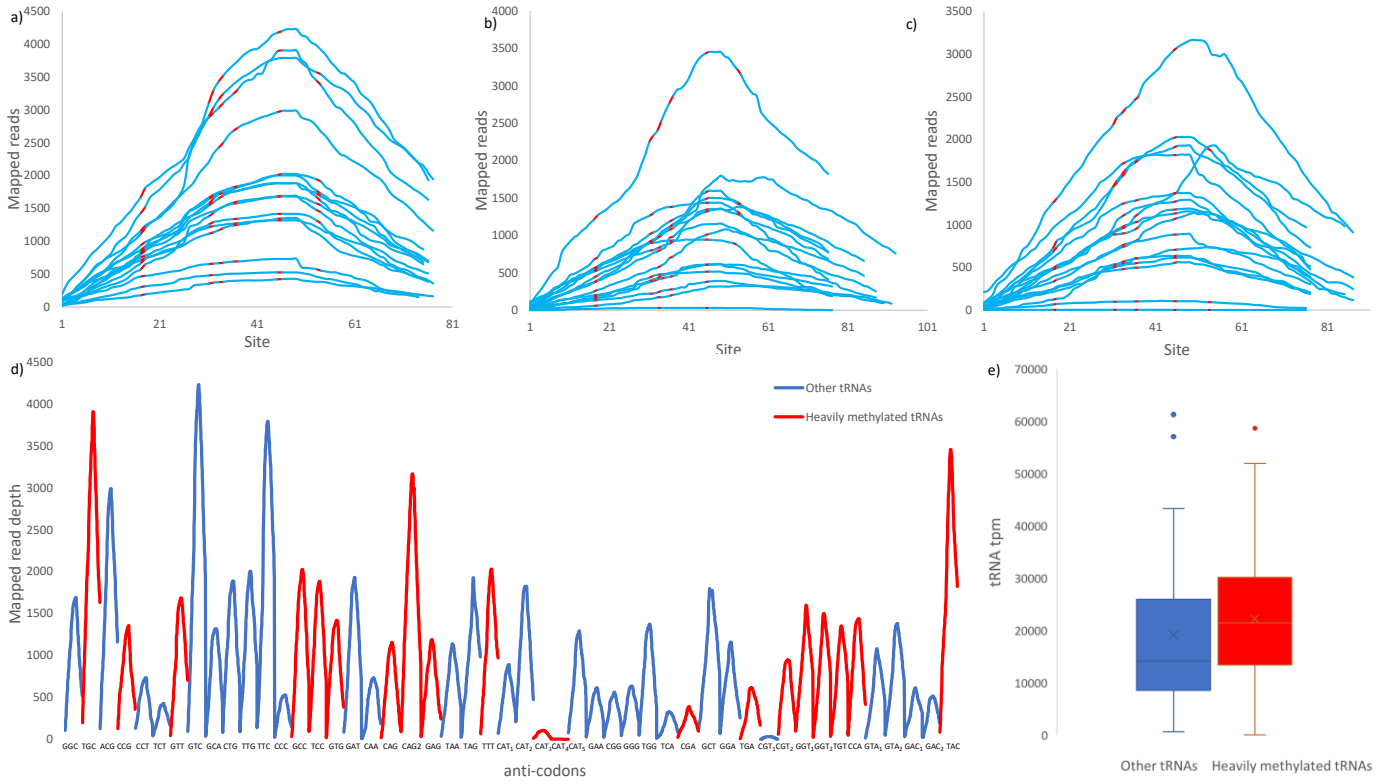


Figure 3.2. RNA-Seq read map for all 50 unique tRNA sequences in *E. coli*, split in three sets in panels a), b) and c). Each line represents the read depth at entire tRNA sequence region, with sites susceptible to methylation (m^2G18, m^2C32 or m^2U32, m^5U34 or m^2C34 or cmo^5U34, m^6A37, m^7G46 and m^5U54) highlighted red. In d) the distribution of mapped reads across the entire length of each tRNA sequence, with tRNAs that are potentially methylated at >4 sites (heavily methylated) highlighted in red and all others highlighted in blue. In e) the distribution of total tRNA tpm in sets of heavily methylated and other tRNAs.

3.5.3 An improved estimation of codon preference using tRNA tpm

To investigate how well tRNA tpm explains codon preference across bacterial lineages, we first designed an approach to determine translationally optimal codons. We used two criteria to define what constitutes a translationally optimal codon: 1) it has the highest relative synonymous

codon usage (RSCU) (Sharp, Li 1986) in HEGs, and 2) its RSCU in HEGs is greater than that in lowly expressed genes (LEGs). These two criteria combine the specifications of optimal codons defined by two previous studies: the first criteria is used by Novoa et al. (2012) and the second is adapted from Rocha (2004) (HEGs vs. others). These criteria are also consistent with those used in the calculation of the Codon Adaptation Index (Sharp, Li 1987) and Index of Translation Elongation (Xia 2015). By our definition, there can be only one codon that is most translationally optimal within each synonymous group, consistent with the original definition of optimal codon determined by F_{op} (Ikemura 1985), and the characterized translationally optimal codons are expected to increase elongation efficiency relative to others. The availability of protein abundance data in parts per million (ppm) for species studied herein and mRNA transcript abundance (in tpm) determined using kallisto (Bray et al. 2016) enabled us to calculate protein per transcript (ppm/tpm) in the identification of HEGs and LEGs (see Materials and Methods for more detail).

Next, we determined readable isoacceptor tRNA abundance. We considered all cognate and near-cognate interactions, as well as those enabled by anticodon modifications. Most studies (Ikemura 1985; Dong, Nilsson, Kurland 1996; Kanaya et al. 1999; dos Reis, Savva, Wernisch 2004) considered anticodon-codon cognate (e.g., $tRNA^{Arg}_{UGC}$ reading GCA) and near-cognate (e.g., $tRNA^{Arg}_{UGC}$ reading GCG) pairings; while some (Rocha 2004; Novoa et al. 2012) also considered pairings allowed due to anticodon modifications (e.g., $tRNA^{Arg}_{UmGC}$ reading GCC and GCU) which increased the predictive influence of tRNA content on codon usage (Novoa et al. 2012). In order to explain RSCU, we adapted the idea of Relative tRNA Gene frequency from Novoa et al. (2012) to derive Relative tRNA Usage (RTU) (see Materials and Methods for more detail). By considering tRNA abundance for synonymous codons, RTU improved tRNA tpm as an estimator of tRNA abundance (Figure 3.3). In particular, the correlation between tRNA tpm and average

tRNA abundance (Dong, Nilsson, Kurland 1996) in *E. coli* (Figure 3.1a; $R^2 = 0.498$, $P < 0.0001$) increased if we consider their RTU values (Figure 3.3a: $R^2 = 0.646$, $P < 0.0001$). Furthermore, both RTU values correlate with codon usage in HEGs (Figure 3.3b).

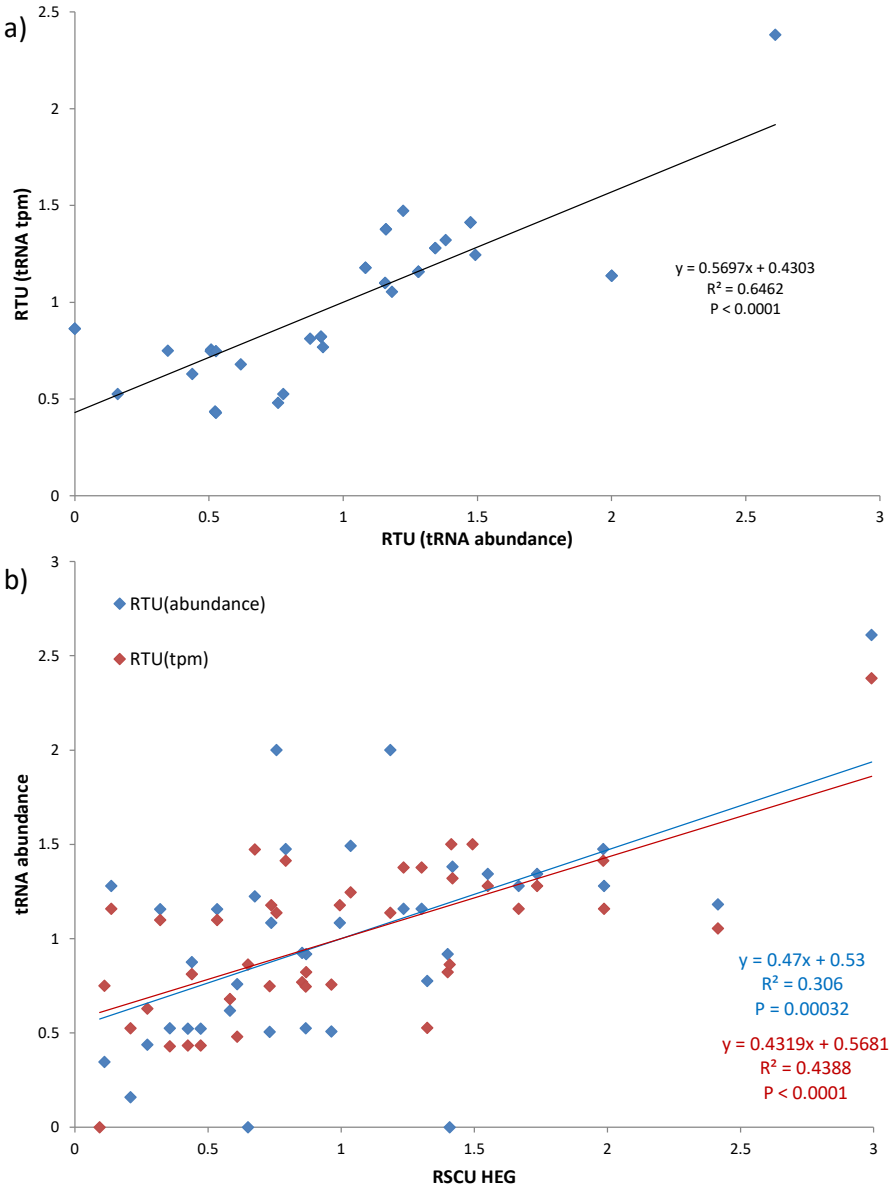


Figure 3.3. a) *E. coli* tRNA tpm and averaged tRNA abundance from RNA fingerprinting experiments (Dong, Nilsson, Kurland 1996) well correlates when both are measured in RTU, and b) both RTU correlates with RSCU in *E. coli* HEGs.

With translationally optimal codons and tRNA RTU established, we determined whether translationally optimal codons are indeed codons that have the highest tRNA availability. Here, we adapted the four rules of codon-anticodon constraint (Ikemura 1985). Rule one states that codon usage is constrained by tRNA availability (Ikemura, Ozeki 1983). Rules two to four focus on specific base pairing efficiencies and they describe that cognate codon-anticodon pairs are generally more efficient and preferable relative to near-cognate wobble pairs (Nishimura 1978; Weissenbach, Dirheimer 1978; Bennetzen, Hall 1982; Grosjean, Fiers 1982; Ikemura 1982). Hence, synonymous codons are first ranked by highest RTU, and then ranked by highest near-cognate tRNA abundances (Supplemental File S2). Supplemental Figure S3.1 provides a flowchart explaining the approach to predict translationally optimal codons by tRNA availability. For example, applying this two-step identification approach for Threonine codons in *E. coli*, we first selected ACC and ACU since they both have the highest RTU and are both readable by the same tRNAs ($tRNA^{Thr}_{GGU}$ and $tRNA^{Thr}_{UGU}$) due to anticodon modification by ADATs (Novoa et al. 2012). Next, we ranked codon preference by cognate tRNA abundance: tRNA tpm of the cognate $tRNA^{Thr}_{GGU}$ for ACC is 46537.7, and ACU is not decoded by a cognate tRNA. Hence, the predicted optimal codon based on tRNA availability and pairing constraints is ACC for Threonine. On the other hand, ACC is also the translationally optimal codon for Threonine based on usage criteria: 1) it has the highest RSCU in HEGs (2.111), and 2) its RSCU HEG is higher than RSCU LEG (1.468). We extended these approaches to predict translationally optimal codons in all seven species and checked whether predicted optimal codons based on tRNA availability meets predicted optimal codons based on usage in genes (Table 3.1).

Table 3.1. Predicted translationally optimal codons, based on RSCU criteria, in seven bacterial species. Highlighted red are translationally optimal codons that also have the highest tRNA availability estimated by RTU.

Amino acid[1]	Synonymous codons	*E. coli*	*S. enterica*	*B. subtilis*	*B. thetaiotaomicron*	*Synechocystis* sp.	*M. tuberculosis*	*L. interrogans*
Ala	GCA, GCC, GCG, GCU	-[3]	-	GCA[a]	GCU	GCC	GCC	GCA
Cys	UGC, UGU	UGC[b]	UGC	UGC	UGU	-	UGC	-
Asp	GAC, GAU	-	-	-	GAU	-	GAC	-
Glu	GAA,GAG	GAA[ab]	GAA	GAA[a]	GAA	GAA	GAG	GAA
Phe	UUC,UUU	UUC[ab]	UUC	-	UUC	-	UUC	-
Gly	GGN	GGC[ab]	GGC	GGC	GGU	GGU	GGC*	-
His	CAC, CAU	CAC[b]	-	-	-	-	CAC	-
Ile	AUA, AUC, AUU	AUC[ab]	AUC	-	AUC	-	AUC	-
Lys	AAA, AAG	AAA[ab]	AAA	AAA[a]	AAA	-	AAG	-
Leu 2-fold[2]	UUA, UUG	UUG[b]	UUG	UUA[a]	-	UUG	UUG	-
Leu 4-fold	CUA, CUC, CUG, CUU	CUG[a]	CUG	CUU	-	CUG*	CUG	-
Asn	AAC, AAU	AAC[ab]	AAC	-	-	-	AAC	-
Pro	CCA, CCC, CCG, CCU	CCG*[ab]	CCG	-	-	CCC*	CCG	CCU
Gln	CAA, CAG	CAG[ab]	CAG	CAA[a]	-	-	CAG	-
Arg 2-fold	AGA, AGG	AGA[a]	AGA	AGA	AGA	AGG	AGG	AGA
Arg 4-fold	CGA, CGC, CGG, CGU	CGU[a]	CGU	CGC	CGU	CGG*	CGC	CGU
Serine 2-fold	AGC, AGU	AGC[b]	AGC	AGC	-	-	AGC	-
Serine 4-fold	UCA, UCC, UCG, UCU	UCU	UCC	UCA	UCU	UCC	-	UCU
Thr	ACA, ACC, ACG, ACU	ACC[ab]	ACC	ACA[a]	ACU	ACC	ACC	ACU
Val	GUA, GUC, GUG, GUU	-		GUU	GUA	-	-	
Tyr	UAC, UAU	-		-	-	-	UAC	

[1] - Two amino acids (Met and Trp) are omitted because they are each encoded by a single codon.
[2] - The 6-fold degenerate codon families (Leu, Arg, and Ser) are broken into 2 and 4-fold families because of differences in the first codon base.
[3] - An optimal codon cannot be determined by our definition (e.g., RSCU HEG < RSCU LEG violates the second criterion).
* - Translationally optimal codons that have highest RTU values calculated from only one of the two RNA-Seq data analyzed.
[a] – Translationally optimal codons determined by F_{op} in Ikemura (1985) and Kanaya et al. (1999).
[b] – Translationally optimal codons identified using average tRNA abundance from RNA fingerprinting approach, retrieved from Dong, Nilsson, Kurland (1996).

Many studies have demonstrated that bacterial tRNA abundance fluctuates due to growth phase and culture conditions such as temperature and growth media (Dong, Nilsson, Kurland 1996; Dittmar et al. 2004; Chen, Texada 2006; Avcilar-Kucukgoze et al. 2016). Hence, for each species, we have acquired tRNA tpm values from two independent, but experimentally consistent (i.e., all strains are wildtype, and all cultures are taken during log-phase growth) RNA-Seq datasets (Table 3.2) that were prepared for sequencing on the same platform (Illumina) to verify consistency in predicting translationally optimal codons using tRNA tpm (Table 3.1).

Table 3.2. Seven bacterial species are studied herein since protein abundance, growth rate information, and RNA-Seq data are publicly available.

Species	Strain	Growth Rate[1]	NCBI Accession	Experiment ID*
Bacteroides thetaiotaomicron	VPI-5482	Slow	NC_004663	SRX020805, SRX860738
Bacillus subtilis	168	Fast	NC_000964	SRX515181, SRX2804667
Escherichia coli	K-12	Fast	NC_000913	SRX515174, SRX669653
Leptospira interrogans	Fiocruz L1-130	Slow	AE016823	SRX2448246, SRX405952
Mycobacterium tuberculosis	H37Rv	Slow	NC_000962	SRX1372108, SRX4374910
Salmonella enterica	LT2	Fast	NC_003197	SRX1638989, SRX1258668
***Synechocystis* sp.**	PCC 6803	Slow	NC_017277	SRX347145, SRX4145044

[1] – Information on species growth-rate are retrieved from Rocha (2004).
* - All selected datasets have matching species strains between protein abundance, GtRNAdb, NCBI and RNA-Seq data, with the exception of *L. interrogans* (SRX2448246) due to the lack of a second RNA-Seq data in GEO Datasets for strain Fiocruz L1-130. All cultures in RNA-Seq experiments were isolated during log phase of growth. The first listed SRX dataset was selected for all analyses, and both were used for Table 3.1 and Figure S3.4.

3.5.4 Implementing tRNA tpm in tAI calculation improves the non-parametric *S* correlation

We calculated tAI using gene copy number and tRNA tpm (See Materials and Methods for more detail, Supplemental Figure S3.2), and Table 3.3 shows the non-parametric regression *S* which reflects the correlation between tAI values and effective number of codons (Wright 1990) corrected for silent substitutions (dos Reis, Savva, Wernisch 2004). For six out of seven species

studied, *S'* calculated with tRNA tpm was higher than *S* calculated using tRNA gene copy number (Table 3). Indeed, tAI' values calculated using tRNA tpm better correlated with codon usage than tAI values obtained from tRNA gene copy number (two-tailed Student's t-test with unequal variance: $P < 0.0001$; Supplemental Figure S3.2) in all species except *L. interrogans*. Note that *S* values calculated herein use non-hypothetical and non-pseudo genes, whereas those originally calculated (S_0) (dos Reis, Savva, Wernisch 2004) use all coding DNA sequences, although value ranks stayed consistent (Table 3.3).

Table 3.3. Non-parametric regression *S* correlations between tAI values and effective number of codons.

Species	S_0	*S*	*S'*
E. coli	0.70[1]	0.61[2]	0.71[3]
S. enterica	0.63	0.59	0.69
B. thetaiotaomicron	0.55	0.4	0.62
***Synechocystis* sp.**	0.38	0.27	0.5
B. subtilis	-0.01	0.18	0.33
M. tuberculosis	-0.04	0.1	0.13
L. interrogans	N/A	0.23	0.2

[1] – *S* correlations retrieved from dos Reis, Savva, Wernisch (2004), calculated using all coding DNA sequences.
[2] – *S* correlations calculated using tRNA gene copy number, using genes having non-zero protein abundances.
[3] – *S* correlations calculated using tRNA tpm, using genes having non-zero protein abundances.

3.5.5 Abundance of tRNA depends on codon usage in fast-growing species.

Above we investigated whether tRNA abundance explains codon usage. Codon usage preferences can conversely drive up tRNA content (Bulmer 1987; Rocha 2004). For each tRNA species (distinguished by anticodon), we defined its readable codon usage as the sum usage of its cognate and near-cognate codons (e.g., the readable codon usage of $tRNA^{Ala}_{GGC}$ = GCC usage + GCU usage). Indeed, readable codon usage in HEGs better explained tRNA tpm in fast-growing

species than in slow-growing species. More specifically, readable codon usage better correlated with tRNA tpm in *E. coli*, *S. enterica*, and *B. subtilis*, than in *B. thetaiotaomicron*, *Synechocystis* sp., *M. tuberculosis*, and *L. interrogans* (Supplemental Figure S3.3).

3.6 Discussion

Two approaches can be taken to characterize the effect of tRNA on codon usage. The first is to test whether gain or loss of tRNA genes predict changes in codon usage (Xia et al. 2007; Xia 2012; Xia 2018a). In tunicates and bivalves, an additional $tRNA^{Met/UAU}$ gene is present in the mitochondrial genome. One would expect that the additional $tRNA^{Met/UAU}$ would favor increased usage of AUA codons, and this expectation is empirically substantiated (Xia et al. 2007; Xia 2012). One may also reason that, if a bacteriophage encodes many tRNA genes in its own genome, especially when these tRNAs are rare in the host, then the phage codon usage will be less dependent on the host tRNA pool. This expectation is also consistent with empirical evidence (Chithambaram, Prabhakaran, Xia 2014b; Prabhakaran, Chithambaram, Xia 2014). The second approach is to quantify the within-species association between codon usage and tRNA abundance which is often approximated by tRNA gene copy number, but this proxy has two key shortcomings. First, we do not know if it is generally true that tRNA gene copy number well correlate tRNA abundance. Second, some bacterial species, such as *M. tuberculosis*, have only a single tRNA gene for all anticodons. In such cases, we cannot use tRNA gene copy number as a proxy of abundance because of its lack of variability. Thus, accurate quantification of tRNA abundance is crucial for detecting codon adaptation.

Our RNA-Seq-based analysis shows that snapshots of bacterial tRNA pools can be captured using RNA-Seq profiling, in spite of claims that standard Illumina protocols ineffectively

sequence full length tRNAs in eukaryotes (Pang et al. 2014; Zheng et al. 2015; Loher, Telonis, Rigoutsos 2018) due to a number of modifications that increase the stability of tRNA secondary structure. While some tRNA modifications are shared among the three kingdoms of life, others are kingdom specific (Hori 2014). Nonetheless, tRNA post-transcriptional methylation is extensive in Eubacteria (Hori 2014), and we acknowledge that tRNA tpm values may underestimate tRNA abundances since tRNA is highly structured and difficult to denature. Despite this, we do not observe any drastic read count drops or hard-stops in mapped reads at or flanking documented methylated sites in the seven species studied here, nor do we see partially mapped tRNA regions (Figure 3.2a-c; Supplemental File S1) that were commonly observed in AlkB-untreated tRNA sequencing data in Eukaryotes (Cozen et al. 2015). A caveat is the presence of a hard-stop in most tRNAs at site 50 for *B. subtilis* (SRX2804667), *Synechocystis* sp. (SRX4145044) and *L. interrogans* (SRX2448246) (Supplemental File S1). However, there is no documented methyltransferase activity acting at this site for these species, and the drop in mapped reads is only observed in one of the two SRX datasets retrieved for each species and could be explained by length specifications in RNA Sequencing (typically 50). Nonetheless, we acknowledge that tRNA sequencing may be potentially enhanced with demethylase treatments (Cozen et al. 2015; Zheng et al. 2015) that should be employed in future tRNA-Seq studies in bacteria.

Because eukaryotic tRNA read mapping abundances are considerably higher in demethylated samples than untreated samples (Cozen et al. 2015), we expected that our read mapping abundances may be similarly impacted by tRNA methylation. In light of this, we considered all annotated methylation sites in bacterial tRNAs, even though they may not be methylated at all times due to structural constraints. Surprisingly, our observations reveal that read mapping abundances of *E. coli* tRNAs that are potentially heavily methylated do not differ from

those that are susceptible to methylation at fewer than five methylation sites (Figure 3.2e). This suggests that the requirement for demethylase treatment prior to tRNA sequencing in bacterial species with functional AlkB demethylase homologs may be more relaxed, especially since demethylation treatments in current eukaryotic tRNA sequencing approaches are bacterial AlkB-facilitated (Cozen et al. 2015; Zheng et al. 2015).

We determined the most translationally optimal codon in 21 codon groups (Table 3.1) based on RSCU differences between HEG and LEG subsets. These subsets of genes are established based on protein per transcript (Supplemental Table S3.1, File S3), with the aim of providing a more accurate estimate of translation efficiency from protein abundance by decoupling rates of transcription. However, species-specific translationally optimal codons cannot always be established because the two described criteria are violated in some groups (e.g., the codon with the highest RSCU is more over-represented in LEGs than HEGs). In these cases, there is no evidence that the most abundantly used synonymous codon would contribute to increase translation efficiency. In particular, *L. interrogans* represents a slow-growing species wherein codon optimization is very poor (only seven translationally optimal codons can be characterized in 21 codon groups).

Our tRNA quantification approach better predicts translationally optimal codons over F_{op} (Ikemura 1985). In *E. coli*, synonymous codons with the highest tRNA tpm (ranked by highest cognate tRNA TPU followed by highest near-cognate tRNA TPU) match 15 out of 17 translationally optimal codons, but 13 out of 17 when we replace tRNA tpm with averaged RNA fingerprinting abundance retrieved from Dong, Nilsson, Kurland (1996) (Table 3.1). In contrast, F_{op} determines 12 such translationally optimal codons (Ikemura 1985) of which AGA (Arg) is the

only codon that our method does not predict to be optimal (Table 3.1). Additionally, all translationally optimal codons determined herein are consistent with optimal codons determined by F_{op}, except in the Serine 4-fold family where F_{op} predicts UCC whereas we predict UCU to be optimal. In the case of *B. subtilis*, both tRNA tpm and F_{op} determine six translationally optimal codons (Table 3.1). It is worth mentioning that F_{op} determines 16 optimal codons in *B. subtilis* (Kanaya et al. 1999), but most may not be translationally optimal. For example, CCA (Pro) was determined to be an optimal codon, but CCG (Pro) is substantially more preferred than CCA (Pro) (RSCU in HEGs are 1.735 and 0.782, respectively).

Nonetheless, bacterial tRNA abundance does not fully explain the variation in codon usage (Supplemental Figure S3.4). First, codon preference cannot always be inferred reliably from tRNA abundance. For example, Inosine is expected to pair best with C and U, but less with A, presumably because of the bulky I/A pairing involving two purines (Xia 2018b). Second, what matters in translation elongation is the availability of charged tRNAs. It is difficult to determine the level of charged tRNAs, and researchers typically would use transcriptionally determined tRNAs or even the number of tRNA genes in the genome as a proxy of charged tRNAs. Similarly, the limitation of our RNA-Seq approach is that it does not reflect the abundance of charged tRNA (Elf et al. 2003). Lastly, other factors such as mutation bias (Muto, Osawa 1987; Osawa et al. 1988; Duret 2002; Sharp et al. 2005; Yang, Nielsen 2008) may exert more pressure on codon usage in certain species.

One important implementation of tRNA tpm is in the calculation of tAI. Our results (Table 3.3, Supplemental Figure S3.2) show that tAI' (tAI calculated using tRNA tpm) better correlates effective number of codons than tAI (calculated using tRNA gene copy number) for all species

studied except *L. interrogans*. Considering *S* and S_0 calculated using tRNA gene copy numbers, their differences are likely due to our usage of the subset of non-hypothetical and non-pseudo genes that have non-zero protein abundance values (Wang et al. 2015) (*S*) whereas all DNA coding sequences (including hypothetical and pseudo genes) were used in the original calculation (dos Reis, Savva, Wernisch 2004) (S_0). Additionally, both the GtRNAdb (Chan, Lowe 2009; Chan, Lowe 2016) and DNA coding sequences (GenBank annotations) have been continuously curated since tAI release in 2004. Lastly, only wobble base pairings were considered in the original introduction of tAI (dos Reis, Wernisch, Savva 2003; dos Reis, Savva, Wernisch 2004); whereas we have also considered possible anticodon modifications (Rocha 2004; Novoa et al. 2012) that further relax codon pairing. These differences improve the calculation of the *S* correlation using tRNA copy numbers, notably in *B. subtilis* and *M. tuberculosis*. In contrast, the negative S_0 correlation originally calculated for *B. subtilis* was a major shortcoming of the tAI method (dos Reis, Savva, Wernisch 2004) and was criticized (Sharp et al. 2005) for suggesting a lack of selective pressure exerted by tRNA abundance on codon preference in this species.

In the case of *M. tuberculosis*, the *S* correlations are consistently the lowest (Table 3.3), yet we identified 17 out of 19 translationally optimal codons using tRNA tpm (Table 3.1). Our method recaptures the "weak but significant codon usage preference" previous reported (Andersson, Sharp 1996; Sharp et al. 2005) in this slow-growing species, and show that the degree to which tRNA availability explains optimal codon usage is species-specific and does not always depend on growth-rate.

Conversely, codon preference also drives up tRNA availability (Bulmer 1987; Rocha 2004), more so in fast-growing (*E. coli*, *B. subtilis* and *S. enterica*) than in slow-growing species (*B.*

thetaiotaomicron, *L. interrogans*, *M. tuberculosis* and *Synechocystis* sp.) (Supplemental Figure S3.3). This result supports the theory that tRNA translation machinery is better optimized to codon usage in fast-growing than slow-growing species (Rocha 2004; Higgs, Ran 2008). Indeed, duplicating tRNA genes may be an effective way to elevate transcript abundance in species that grow and replicate rapidly (Ikemura 1985; Dong, Nilsson, Kurland 1996; Xia 1998; Kanaya et al. 1999), but not in slow-growing species (Supplemental Figure S3.5).

We studied the coevolution between codon usage and tRNA abundance in three fast-growing species (*E. coli*, *S. enterica*, and *B. subtilis*) and four slow-growing species (*B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* sp.). Our findings indicate that tRNA quantification by tpm offers better predictions of translationally optimal codons over F_{op} in *E. coli*, and improves the calculation of tAI to better reflect codon preference in all species studied except *L. interrogans*. The usage of translationally optimal codons can be well explained by relative tRNA tpm in *E. coli* and *S. enterica*; however, both tRNA tpm and RNA fingerprinting abundances (Kanaya et al. 1999) offer weaker explanations for codon preference in *B. subtilis*. The influence of tRNA availability on codon bias is not always stronger in fast-growing species, and optimal codons can be well explained by tRNA content in certain slow-growing species such as *M. tuberculosis*.

3.7 Data availability

Supplemental figures for Chapter 3 are available in Appendix C. Supplemental files are available at https://www.nature.com/articles/s41598-019-39369-x: Supplemental File S1 contains RNA-Seq read depths and tRNA methylation profile. Supplemental File S2 contains identifications of translationally optimal codons and tRNA abundances (gene copy, tpm and fingerprinting data) and

tRNA quantification approaches. Supplemental File S3 contains protein per transcript data, protein abundance information, identified translationally HEGs and LEGs.

3.8 Acknowledgements

This work was supported by the Discovery Grant of Natural Science and Engineering Research Council of Canada to X.X. (NSERC, RGPIN/2018-03878), and the Ontario Graduate Scholarship 2018-2019 to Y.W.

3.9 References

Andersson, GE, PM Sharp. 1996. Codon usage in the Mycobacterium tuberculosis complex. Microbiology 142:915-925.

Avcilar-Kucukgoze, I, A Bartholomaus, JA Cordero Varela, RF Kaml, P Neubauer, N Budisa, Z Ignatova. 2016. Discharging tRNAs: a tug of war between translation and detoxification in Escherichia coli. Nucleic Acids Res 44:8324-8334.

Bennetzen, JL, BD Hall. 1982. Codon selection in yeast. J Biol Chem 257:3026-3031.

Bolger, AM, M Lohse, B Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bray, NL, H Pimentel, P Melsted, L Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525-527.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728-730.

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

Carullo, M, X Xia. 2008. An Extensive Study of Mutation and Selection on the Wobble Nucleotide in tRNA Anticodons in Fungal Mitochondrial Genomes. Journal of Molecular Evolution 66:484.

Cassier-Chauvat, C, T Veaudor, F Chauvat. 2016. Comparative Genomics of DNA Recombination and Repair in Cyanobacteria: Biotechnological Implications. Frontiers in microbiology 7:1809-1809.

Chan, PP, TM Lowe. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37:4.

Chan, PP, TM Lowe. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research 44:D184-D189.

Chen, D, DE Texada. 2006. Low-usage codons and rare codons of Escherichia coli. Gene Therapy and Molecular Biology 10:1.

Chithambaram, S, R Prabhakaran, X Xia. 2014a. Differential codon adaptation between dsDNA and ssDNA phages in Escherichia coli. Mol Biol Evol 31:1606-1617.

Chithambaram, S, R Prabhakaran, X Xia. 2014b. Differential Codon Adaptation between dsDNA and ssDNA Phages in Escherichia coli. Molecular Biology and Evolution 31:1606-1617.
Cozen, AE, E Quartley, AD Holmes, E Hrabeta-Robinson, EM Phizicky, TM Lowe. 2015. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Methods 12:879-884.
Dittmar, KA, EM Mobley, AJ Radek, T Pan. 2004. Exploring the regulation of tRNA distribution on the genomic scale. J Mol Biol 337:31-47.
Dong, H, L Nilsson, CG Kurland. 1996. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. Journal of Molecular Biology 260:649-663.
dos Reis, M, R Savva, L Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036-5044.
dos Reis, M, L Wernisch, R Savva. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res 31:6976-6985.
Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640-649.
Elf, J, D Nilsson, T Tenson, M Ehrenberg. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300:1718-1722.
Ernst, FGM, L Erber, J Sammler, F Juhling, H Betat, M Morl. 2018. Cold adaptation of tRNA nucleotidyltransferases: A tradeoff in activity, stability and fidelity. RNA Biol 15:144-155.
Ewing, B, L Hillier, MC Wendl, P Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175-185.
Falnes, PØ, M Bjørås, PA Aas, O Sundheim, E Seeberg. 2004. Substrate specificities of bacterial and human AlkB proteins. Nucleic acids research 32:3456-3461.
Falnes, PO, RF Johansen, E Seeberg. 2002. AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli. Nature 419:178-182.
Gao, P, G Li, Y Li, Y Li, H Tian, Y Wang, J Zhou, T Ma. 2016. An Exogenous Surfactant-Producing Bacillus subtilis Facilitates Indigenous Microbial Enhanced Oil Recovery. Frontiers in microbiology 7:186-186.
Grosjean, H, W Fiers. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199-209.
Higgs, PG, W Ran. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25:2279-2291.
Hori, H. 2014. Methylated nucleosides in tRNA and tRNA methyltransferases. Frontiers in genetics 5:144-144.
Ikemura, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573-597.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13-34.
Ikemura, T, H Ozeki. 1983. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harb Symp Quant Biol 2:1087-1097.

Kanaya, S, Y Yamada, Y Kudo, T Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143-155.

Kodama, Y, M Shumway, R Leinonen. 2012. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40:18.

Leinonen, R, H Sugawara, M Shumway. 2011. The sequence read archive. Nucleic Acids Res 39:9.

Li, H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Loher, P, AG Telonis, I Rigoutsos. 2018. Accurate Profiling and Quantification of tRNA Fragments from RNA-Seq Data: A Vade Mecum for MINTmap. Methods Mol Biol:7339-7332_7316.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

Muto, A, S Osawa. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84:166-169.

Nie, Y, CQ Chi, H Fang, JL Liang, SL Lu, GL Lai, YQ Tang, XL Wu. 2014. Diverse alkane hydroxylase genes in microorganisms and environments. Sci Rep 4.

Nishimura, S. 1978. Modified nucleosides and isoaccepting tRNA. . Cambridge, Mass: MIT Press.

Novoa, EM, M Pavon-Eternod, T Pan, L Ribas de Pouplana. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149:202-213.

Osawa, S, T Ohama, F Yamao, A Muto, TH Jukes, H Ozeki, K Umesono. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci U S A 85:1124-1128.

Pang, YL, R Abo, SS Levine, PC Dedon. 2014. Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. Nucleic Acids Res 42:27.

Prabhakaran, R, S Chithambaram, X Xia. 2014. Aeromonas phages encode tRNAs for their overused codons. Int J Comput Biol Drug Des 7:168-182.

Prabhakaran, R, S Chithambaram, X Xia. 2015. Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. J Gen Virol 96:1169-1179.

Rocha, EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Research 14:2279-2286.

Schulz, S, A Perez-de-Mora, M Engel, JC Munch, M Schloter. 2010. A comparative study of most probable number (MPN)-PCR vs. real-time-PCR for the measurement of abundance and assessment of diversity of alkB homologous genes in soil. J Microbiol Methods 80:295-298.

Sharp, PM, E Bailes, RJ Grocock, JF Peden, RE Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141-1153.

Sharp, PM, WH Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28-38.

Sharp, PM, WH Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research 15:1281-1295.

van den Born, E, A Bekkelund, MN Moen, MV Omelchenko, A Klungland, PO Falnes. 2009. Bioinformatics and functional analysis define four distinct groups of AlkB DNA-dioxygenases in bacteria. Nucleic Acids Res 37:7124-7136.
van den Born, E, MV Omelchenko, A Bekkelund, V Leihne, EV Koonin, VV Dolja, PØ Falnes. 2008. Viral AlkB proteins repair RNA damage by oxidative demethylation. Nucleic acids research 36:5451-5461.
van Weringh, A, M Ragonnet-Cronin, E Pranckeviciene, M Pavon-Eternod, L Kleiman, X Xia. 2011. HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol 28:1827-1834.
Wang, L, W Wang, Q Lai, Z Shao. 2010. Gene diversity of CYP153A and AlkB alkane hydroxylases in oil-degrading bacteria isolated from the Atlantic Ocean. Environ Microbiol 12:1230-1242.
Wang, M, CJ Herrmann, M Simonovic, D Szklarczyk, C von Mering. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163-3168.
Weissenbach, J, G Dirheimer. 1978. Pairing properties of the methylester of 5-carboxymethyl uridine in the wobble position of yeast tRNAArg3. Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis 518:530-534.
Williams, CR, A Baccarella, JZ Parrish, CC Kim. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17:103.
Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87:23-29.
Xia, X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37-44.
Xia, X. 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evolutionary Biology 8:211.
Xia, X. 2012. Rapid evolution of animal mitochondria. In: RS Singh, J Xu, RJ Kulathinal, editors. Evolution in the fast lane: Rapidly evolving genes and genetic systems. Oxford: Oxford University Press. p. 73-82
Xia, X. 2015. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199:573-579.
Xia, X. 2017. ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. G3: Genes|Genomes|Genetics 7:3839-3848.
Xia, X. 2018a. Bioinformatics and Translation Elongation. Bioinformatics and the Cell: Springer, Cham. p. 197-238.
Xia, X. 2018b. "Bioinformatics and Translation Elongation" in Bioinformatics and the Cell: Springer, Cham.
Xia, X. 2018c. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol Biol Evol 35:1550-1552.
Xia, X, H Huang, M Carullo, E Betran, EN Moriyama. 2007. Conflict between Translation Initiation and Elongation in Vertebrate Mitochondrial Genomes. PLoS One 2:e227.
Yang, Z, R Nielsen. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25:568-579.
Zheng, G, Y Qin, WC Clark, Q Dai, C Yi, C He, AM Lambowitz, T Pan. 2015. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods 12:835-837.

CHAPTER FOUR

Centrality of efficient translation initiation and elongation in *Escherichia coli* and *Bacillus subtilis* gene expression

This chapter is being prepared for submission.

4.1 Author Contributions

Yulong Wei[1], 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, University of Ottawa, Ontario, Canada K1H 8M5.

Y.W. designed the study. Y.W. collected and analyzed the data and wrote the manuscript. X.X. developed the computational programs. Y.W. made all Figures, Tables, and Supplemental Figures. All authors reviewed the manuscript. X.X. supervised the project.

4.2 Abstract

In bacteria, sequence features that increase translation initiation and elongation efficiencies improve ribosome loading and ribosome decoding rates, respectively. Consequently, efficient translation should enable a consistent flux of ribosome decoders along the mRNA, whereas poor translation efficiencies may increase ribosomal stalling. Here, we employed publicly available genomic, transcriptomic, and proteomic data to exhibit these effects in *Escherichia coli* and *Bacillus subtilis*. Our Ribo-Seq approach showed that ribosome loading rate is precisely affected by structure stability of initiation region and by presence of optimal SD sequences. Additionally, we provide an approach to measure ribosome occupancy and show that while minor codons having few tRNA availabilities frequently stall at the ribosomal A site, major codons having high tRNA availabilities are rapidly decoded. Lastly, we measured how these changes in ribosome dynamics affect protein expression, and we showed that decoder abundance better predicts protein expression in genes where optimal translation features are well-retained. While our results are consistent with the conventional wisdom that initiation is often rate-limiting in bacterial translation, they also emphasize that elongation efficiency strongly influence gene expression even when initiation is poor.

4.3 Introduction

Rate of protein production in bacteria relies much on efficient translation initiation and elongation. As discussed in Chapter 2, ribosome loading during translation initiation is often facilitated by the presence of a Shine-Dalgarno (SD) sequence and the absence of a strong secondary structure flanking the start codon. As discussed in Chapter 3, rate of decoding during translation elongation increases with use of codons that have readily available cognate tRNAs. While many have perceived initiation as the rate-limiting step in translation (Jacques, Dreyfus 1990; Kudla et al. 2009; Xia 2015), others suggest elongation efficiency is the main factor affecting protein expression (Nie, Wu, Zhang 2006; Welch et al. 2009; Tuller et al. 2010; Guimaraes, Rocha, Arkin 2014).

4.3.1 Sequence features affecting translation initiation and their roles in protein expression

Two gene features may be essential for efficient translation initiation in bacteria (Xia 2018a). The minimum requirement is that the start codon is accessible (Nakamoto 2006) and there is no stable secondary structure to embed the start codon (Scharff et al. 2011; Osterman et al. 2013). As a result, the first 5 – 10 codon sites in bacterial coding DNA sequences (CDS) prefer AT-rich codons that are often rare elsewhere due to selection to reduce secondary structure stability at initiation region (Bentele et al. 2013). Second, a purine-rich Shine-Dalgarno (SD) sequence (Shine, Dalgarno 1974; Shine, Dalgarno 1975; Nakagawa et al. 2010; Nakagawa, Niimura, Gojobori 2017) is prominently found in the 5' untranslated region (5' UTR) upstream of bacterial CDS (Nakagawa et al. 2010). In SD-mediated translation initiation, the ≥ 4 nt SD sequence interacts with the anti-SD sequence, located at the 3' end of the 16S rRNA, to facilitate ribosome docking near the start codon. An effective SD sequence requires both optimal SD/anti-SD pairing location (Ma,

Campbell, Karlin 2002; Starmer et al. 2006; Abolbaghaei, Silke, Xia 2017) and optimal base complementarity (Schurr, Nadir, Margalit 1993; Hockenberry et al. 2017; Wei, Silke, Xia 2017).

While it is commonly accepted that start codon accessibility is crucial to translation, the importance of the SD sequence is less well-established. Despite studies suggesting SD/anti-SD pairing is important to protein production in *Escherichia coli* and *Bacillus subtilis* (Band, Henner 1984; Osterman et al. 2013), others found that protein expression is weakly affected by the presence of SD sequence in *E. coli* and *Desulfovibrio vulgaris* (Nie, Wu, Zhang 2006; Guimaraes, Rocha, Arkin 2014), and others still showed a lack of SD influence in *E. coli* protein expression (Vimberg et al. 2007; Li et al. 2014).

4.3.2 Sequence features affecting translation elongation and their role in protein expression

Translation elongation efficiency depends much on synonymous codon usage. In *E. coli* and *Salmonella enterica*, usage of major (or optimal) codons are preferred over synonymous alternatives (minor codons) because major codons have the highest tRNA availabilities (Ikemura 1985; Dong, Nilsson, Kurland 1996; Xia 1998; Kanaya et al. 1999; Wei, Silke, Xia 2019), and one expects codon usage bias to influence gene expression. Indeed, highly expressed genes prefer major codons, whereas lowly expressed genes overuse minor codons in comparison (Sharp, Li 1987; Xia 2015).

Many studies have tested the degree that codon usage influences protein expression. Kudla et al. (2009) designed a synthetic library of 154 *E. coli* green fluorescent proteins that differ in synonymous codon usage and found that elongation efficiency as measured by the Codon Adaptation Index (CAI) (Sharp, Li 1987) weakly influences protein expression. A later reanalysis by Tuller et al. (2010) showed that protein expression is jointly influenced by translation initiation

and elongation. Using the recently formulated Index of Translation Elongation (ITE) which extends CAI by accounting for mutation bias, Xia (2015) observed an increased correlation between elongation efficiency and protein production. In brief, all three studies suggest that elongation efficiency as measured by codon usage correlates with protein expression when translation initiation is efficient. These results are congruous with early studies concluding that optimizing codon usage may serve primarily to increase global ribosomal availability but not highly increase gene-specific expression (Liljenstrom, von Heijne 1987; Bulmer 1991; Andersson, Sharp 1996; Kudla et al. 2009). Nonetheless, others have shown that modifying codon usage in *E. coli* did affect rate of translation (Robinson et al. 1984; Varenne et al. 1984; Sorensen, Kurland, Pedersen 1989; Komar, Lesnik, Reiss 1999; Ben-Yehezkel et al. 2015), and protein expression in species such as *E. coli* and *D. vulgaris* is highly influenced by translation elongation efficiency (Nie, Wu, Zhang 2006; Guimaraes, Rocha, Arkin 2014).

4.3.3 The degree that optimal translation features are conserved is species-specific

While translation initiation is perceived by many as the rate-limiting step in bacterial translation (Jacques, Dreyfus 1990; Kudla et al. 2009; Xia 2015), others have argued that translation elongation is the major factor influencing bacterial protein expression (Nie, Wu, Zhang 2006; Welch et al. 2009; Tuller et al. 2010; Guimaraes, Rocha, Arkin 2014). Despite debates over centrality of translation initiation and elongation in bacterial gene expression, what may be important to consider is that conservation of features optimizing translation varies in bacteria. Proportionally more genes are SD-facilitated in *B. subtilis* than in *E. coli* (Abolbaghaei, Silke, Xia 2017; Wei, Silke, Xia 2017), and gene expression is more sensitive to SD sequence modification in *B. subtilis* than in *E. coli* (Band, Henner 1984). Additionally, the extent that codon bias is tRNA-mediated differs between species. Based on Frequency of Optimal Codons (Ikemura 1985; Kanaya

et al. 1999), RNA-Seq quantification (Wei, Silke, Xia 2019), and tRNA Adaptation Index (tAI) (dos Reis, Savva, Wernisch 2004), tRNA availability well explains codon bias in *E. coli* but not in *B. subtilis*. It is reasonable to assume that the influence translation initiation and elongation exert over protein expression varies between bacterial species.

4.3.4 Explaining gene expression using an integrative approach involving genomics, transcriptomics, and proteomics data

As mentioned previously, translation initiation efficiency affects ribosome loading rate, whereas translation elongation efficiency affects ribosome decoding rate. Both processes influence protein production by affecting ribosome flux (Rodnina 2016). Indeed, a recent computational model by Shaham, Tuller (2017) predicted that both protein expression and ribosome dynamics change with rates of translation initiation and elongation in *E. coli*. To estimate translation initiation efficiency, we determined for presence of optimal SD sequence in the 5' UTR and secondary structure stability at the initiation region embedding the start codon. To estimate elongation efficiency, we measured codon adaptation using the index ITE and we contrasted decoding efficiencies between major and minor codons by measuring their ribosome occupancies. Summarily, to trace how translation variably influences gene expression in *E. coli* and *B. subtilis*, we devised ribosome profiling approaches using publicly available genomic, transcriptomic, and proteomic data.

Ribosome sequencing datasets contain ribosome protected mRNA fragments which are snapshots of ribosomal footprints during translation. We demonstrated several applications of Ribosome Sequencing (Ribo-Seq) data to illustrate the following known effects: 1) both translation initiation features, secondary structure stability and SD/anti-SD pairing, affect

ribosome loading efficiency, and 2) during elongation, codon decoding is stalled when a minor codon enter the ribosomal A site but decoding is efficient when a major codon is present. To reveal how these effects influence gene expression, we showed that the degree ribosome abundance correlates protein expression highly depends on elongation efficiency but weakly depends on initiation efficiency. This suggests that the constancy of ribosome flux during translation elongation is crucial to protein production rate. Our findings are congruous with the conventional wisdom that initiation is often rate-limiting in bacterial translation, but they do not support the prediction that elongation efficiency weakly influences gene expression when initiation is poor.

4.4 Materials and Methods

4.4.1 Processing genomic, proteomic, and transcriptomic data

The genomes of *E. coli* K-12 MG1655 (NC_000913.3) and *B. subtilis* 168 (NC_000964.3) were retrieved in GenBank format from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov) and CDS were extracted using DAMBE 7.0 (Xia 2018b). For each species, the integrated protein abundance data were retrieved from PaxDb 4.0 (Wang et al. 2015) and protein IDs were matched to NCBI gene IDs via UniProt (Consortium 2018).

Publicly deposited raw Ribo-Seq FASTQ datasets (of log-phase wild-type *E. coli*: SRX119208, SRX4615480, SRX461548082, and of log-phase wild-type *B. subtilis*: SRX119212) were fetched from GEO Datasets (https://www.ncbi.nlm.nih.gov/gds/). These datasets were retrieved because their associated studies have experimentally verified the ribosomal A site at the ribosome protected reads (discussed in section 4.4.5). In addition, raw FASTQ datasets from two RNA-Seq experiments (log-phase wild-type *E. coli*: SRX515174 and log-phase wild-type *B. subtilis*: SRX515181) were retrieved from GEO Datasets.

Raw FASTQ datasets were processed to trim off adapter sequences and to remove sequence reads with poor quality. First, we trimmed off experiment-specific adapter sequences as described in '*Construction Protocols*' using CutAdapt (Martin 2011) with 10% error rate and retained reads with lengths ≥ 25 to mitigate bias in expression levels (Williams et al. 2016). Then, we used Trimmomatic 0.38 (Bolger, Lohse, Usadel 2014) to remove poor quality sequences with average Phred scores lower than 20 (1% probability of base calling errors) (Ewing et al. 1998). Each processed FASTQ file was converted into a FASTA+ file (FASTA format) and a FASTQ+ file (FASTQ format) using ARSDA1.1 (Xia 2017); these sequence storage formats group identical reads under a single unique ID indicating the number of copies (SeqID_# of copies), this reduced file sizes without loss of information.

4.4.2 Transcript profiling using Ribo- and RNA-Seq data

To perform transcript profiling in ARSDA, a BLASTn database was created for each FASTA+ file in ARSDA, and a query file was made to contain all non-pseudo CDS. For comparison, the FASTA+ files were also inputted in ARSDA and FASTQ+ files were inputted in Kallisto v0.44.0 (Bray et al. 2016) for transcript profiling. Gene expression in Fragments Per Kilobase per Million (FPKM) was calculated via 'Gene expression from BLAST database' in ARSDA with options: E-Value = 10^{-10}, Match length ≥ 18, Strand = both, and 'Ungapped' alignment. To profile map abundance using Kallisto, we created an index file for all CDS in FASTA format to pseudo-align with each FASTQ+ file, and transcript per million (tpm) values were calculated. Additionally, RPFdb 2.0 (Li et al. 2018) deposited map abundances (in FPKM computed by Bowtie2) for each datasets studied herein, these FPKM values were retrieved for comparison. We compared relative read abundances between these three alignment approaches to ensure fidelity of ribosome abundance measurements using any approach (comparison between

any two alignment approaches: *E. coli*: $R^2 > 0.91$, $P < 0.0001$; *B. subtilis*: $R^2 > 0.84$, $P < 0.0001$, all comparisons were between log(FPKM) and log(tpm), correlation coefficients R^2 are Pearson's correlation r squared), and FPKMs from BLASTn in ARSDA were used for downstream analyses.

4.4.3 Determining features affecting translation initiation efficiency

To determine the translation initiation efficiency of CDS, we scanned for presence of a SD sequence and measured secondary structure stability at the translation initiation region. To detect putative SD sequences, we used 16S rRNA 3' terminal sequences (*E. coli*: 5'-GAUCACCUCCUUA-3', *B. subtilis*: 5'-GAUCACCUCCUUUCU-3') first reported by Shine and Dalgarno (Shine, Dalgarno 1974; Shine, Dalgarno 1975) as the complementary anti-SD sequence to match against the 5' untranslated region, following the approach used in previous studies (Prabhakaran, Chithambaram, Xia 2015; Abolbaghaei, Silke, Xia 2017; Wei, Silke, Xia 2017): 30 nt upstream of start codon of all CDSs were extracted and matched against the anti-SD sequence with 'Analyzing 5′UTR' in DAMBE, with minimum SD length = 4 nt and maximum SD length = 12.

For each putative SD sequence, DAMBE outputted a SD/anti-SD pairing location coined as $D_{toStart}$. $D_{toStart}$ measures the distance from the 3' end of 16S rRNA to the start codon, not between the putative SD sequence to start codon. We proposed $D_{toStart}$ under the assumption that this distance is functionally constrained to properly juxtapose the ribosomal P-site near the start codon (Prabhakaran, Chithambaram, Xia 2015; Wei, Silke, Xia 2017) in Chapter 1. The need of $D_{toStart}$ arose because the conventional SD location measurement (the distance from SD motif to start codon) will sometimes inaccurately assign one SD sequence as more optimal than another even-though their effective distance from the free 16S rRNA 3' tail sequence to the start codon is the

same. In other words, two SD motifs may be located at different distances from the start codon but have the same $D_{toStart}$. Thus, genes are considered as SD-facilitated if the putative >4 nt SD sequence is located within the highly confined $D_{toStart}$ peak ranging from 15 to 21 in *B. subtilis* and 10 to 21 in *E. coli* (Supplemental Figure S4.1).

We obtained a 60 nt translation initiation region consisting of 30 nt upstream of the start codon, the start codon, and 27 nt downstream of the start codon. This region embeds the SD sequence located within 30 nt upstream of the start codon (Shine, Dalgarno 1975; Dontsova, Kopylov, Brimacombe 1991; Nakagawa, Niimura, Gojobori 2017) and the first 10 codons that are AT-rich due to selection disfavouring strong mRNA secondary structures (Bentele et al. 2013). mRNA secondary structure stability at this translation initiation region was calculated by Minimum Folding Energy (MFE, kcal/mol) via the Vienna RNA Folding Library (Hofacker 2003) implemented in DAMBE (with options: no lonely pair, Window size = 40, and StepSize = 1), and secondary structure stability was determined as the most negative MFE (kcal/mol) among 40 nt windows.

4.4.4 Determining features affecting translation elongation efficiency

To determine translation elongation efficiency, we measured codon adaptation in genes using Index of Translation Elongation ITE (Xia 2015) and determined optimal codons by tRNA availability. Using ITE implemented in DAMBE, we measured codon adaptation in CDS with respect to two reference gene sets: 1) highly expressed ribosomal protein genes and 2) lowly expressed genes with the lowest non-zero protein abundances, with the optional 'Break 8-fold and 6-fold families into 2' to consider differences in cognate tRNA types among synonymous codons. ITE was used instead of Codon Adaptation Index (CAI) (Sharp, Li 1987) because the former

considers codon usages in both highly and lowly expressed reference genes whereas the latter only considers codon usage in highly expressed genes. This extension accounts for selection bias in codon usage in highly expressed genes and for mutation bias in codon usage affecting non-highly expressed genes (Xia 2015) (as such, CAI is a special case of ITE when there is no codon usage bias in non-highly expressed genes).

Using a RNA-Seq profiling approach, we had previously identified which synonymous codons are major and which are minor in a number of bacterial species, including *E. coli* and *B. subtilis* (Wei, Silke, Xia 2019). These quantifications are consistent with early reports using experimental RNA fingerprinting methods (Ikemura 1985; Dong, Nilsson, Kurland 1996; Kanaya et al. 1999). In brief, we defined major codons by two criteria: 1) they have the highest cognate tRNA availabilities in comparison to synonymous alternatives, and 2) but its RSCU in HEGs is greater than its RSCU in lowly expressed genes (LEGs). The second criterion is a combination of optimal codon specifications used in two previous studies: the first half is used by Novoa et al. (2012) and the second half is adapted from Rocha (2004) (which compared HEGs vs. other genes). These criteria are also consistent with those used in the calculation of codon adaptation in the ITE index. By our approach, there can be at most one major codon that is translationally optimal within a synonymous group, consistent with the original definition of optimal codon determined by F_{op} (Ikemura 1985).

4.4.5 Distinguishing efficiently translated genes from poorly translated genes in *E. coli* and *B. subtilis*

We characterized efficiently and poorly translated genes based on the presence of gene features affecting initiation or elongation. To differentiate genes by initiation efficiency, all non-pesudo CDSs were determined as SD-facilitated (those having SD within optimal $D_{toStart}$) and SD-

independent (all others). Next, we ranked genes by highest to lowest secondary structure stabilities (most negative to least negative MFE in kcal/mol) at the 60 nt initiation region. Together, we considered 30% of SD-facilitated genes with the weakest secondary structure stabilities as those having efficient translation initiaiton, and 30% of SD-independent genes with the strongest secondary structure stabilities as those having poor translation initiation. Similar to the model proposed by Salis et al. (2009), our criteria for rate of translaiton initiation were 1) folding energy at the translation initiation region and 2) presence of SD sequence with optimal physical distance between 16S rRNA binding site and the start codon. Unlike in Salis et al. (2009), we did not determine SD based on the highest SD/anti-SD pairing affinity, as more recent studies suggest intermediate pairing affinities may be optimal for translaiton initiation (Hockenberry et al. 2017; Wei, Silke, Xia 2017). Translation elongation efficiencies were estimated by codon adaptation calculated by ITE. In a given group of genes, 30% of genes with highest ITE values were considered to have optimized codon usage and good translation elongation efficiency, whereas 30% of genes with lowest ITE values were considered to have poorly optimized codon usage and poor translation elongation efficiency.

In total, we obtained four sets of *E. coli* and *B. subtilis* genes based on estimated translation efficiencies: 1) efficiently translated genes with high initiation and elongation efficiencies ($Gene_{HI,HE}$), 2) genes with high initiation but low elongation efficiencies ($Gene_{HI,LE}$), 3) genes with low initiation but high elongation efficiencies ($Gene_{LI,HE}$), and 4) poorly translated genes with low initiation and elongation efficiencies ($Gene_{LI,LE}$).

4.4.6 Measuring site-specific ribosome abundance in ribosomes per transcript

Ungapped BLAST alignments in ARSDA (with options E-Value = 10^{-10}, Match length ≥ 18, Strand = both, and 'Ungapped' alignment) outputs the number of reads and mapping locations at each query gene to enable us to measure site-specific map counts at each gene. Then, three steps were taken to adjust ribosome abundance measurements for site comparisons between two sets of genes with equal elongation efficiencies (e.g., $Gene_{HI,LE}$ vs. $Gene_{LI,LE}$). First, at nucleotide site *i* in a given gene *j*, we measured ribosome $RPKM_{i,j}$ as the number of Ribo-Seq reads mapped at site i in gene j/(length of gene j/1000×total number of reads/1000000) to normalize for differences in gene lengths and differences in total number of Ribo-Seq reads mapped to gene sets (e.g., $Gene_{HI,LE}$ vs. $Gene_{LI,LE}$). Next, we measured the total ribosome abundance at each site *i* (ribosome $RPKM_i$) by summing $RPKM_{i,j}$ for all *j* genes in a set. Finally, to consider for differences in mRNA transcript abundances between $Gene_{HI,:E}$ and $Gene_{LI,:E}$, we calculated site-specific ribosomes per transcript (RPT_i) as (ribosome $RPKM_i$)/(sum of mRNA FPKM in all j genes).

4.4.7 Measuring decoding efficiencies at major and minor codons

To determine codon occupancies at the ribosomal A site, we use Ribo-Seq datasets SRX119208 for *E. coli* and SRX119212 for *B. subtilis* where the ribosomal A site was experimentally approximated at the center of ribosome footprints (Li, Oh, Weissman 2012). Additionally, we retrieved *E. coli* Ribo-Seq dataset SRX4615480, where the P-site location was experimentally approximated at 15 nt (hence the A site 12nt) away from the 3' end of ribosome protected reads (Woolstenhulme et al. 2015; Mohammad, Green, Buskirk 2019). Ungapped BLAST alignment in ARSDA outputs the reads' mapping locations, this enabled us to determine the codon at the assigned ribosomal A sites and at flanking ribosomal sites (e.g., A+3, A+2, A+1, A, P, E, E-1, and E-2). For a few mapped reads, the center is located right between two codons

(e.g, Leu-CUA|Ile-AUC, where | is the location of the center position), these reads were omitted because the ribosomal A site codon cannot be determined.

Codon occupancy is estimated as the total number of times that a codon is observed at a ribosome binding site when all Ribo-Seq reads were considered. We next calculated relative codon occupancy at each ribosome binding site and coined the measurement as relative synonymous codon occupancy (RSCO, equation 1), for 59 codons encoding 18 common amino acids (the single codon families Methionine encoded by ATG and Tryptophan encoded by TGG were omitted). $RSCO_i$ is calculated using the same formula as RSCU, where i is a ribosome binding site (e.g., *i* = A, P and E, etc.).

$$RSCO_i = \frac{codon\ occupancy\ at\ site\ i}{\frac{\sum_1^n synonymous\ codon\ \ occupancy\ at\ i-site}{n}} \qquad (1)$$

The efficiency of decoding for a given codon *j* ($E_{decoding,\ j}$) was calculated as in equation (2), where $RSCO_{average,j}$ is the averaged RSCOs at sites A+3, A+2, A+1, P, E, E-1, E-2, and E-3 (excluding site A) for codon j and $RSCO_{A,j}$ is the RSCO at site A for codon j.

$$E_{decoding,j} = \frac{RSCO_{average,j}}{RSCO_{A,j}} \qquad (2)$$

The rationale for $E_{decoding}$ is as follows: if a codon is decoded inefficiently and ribosome stalling occurs, then its ribosomal A site occupancy should be greater than its P, E, and flanking site occupancies (or $RSCO_A > RSCO_P$, $RSCO_E$, etc.) and $E_{decoding} < 1$. A similar comparison approach between sites A, P and E were illustrated in Mohammad, Green, Buskirk (2019) to show ribosome stalling at Serine codons in yeast. On the other hand, if a codon is rapidly decoded, then we would observe its A site occupancy to be lower than P and E site occupancies (or $RSCO_A < RSCO_P$ and $RSCO_E$, etc.) and $E_{decoding} > 1$. For completeness, relative synonymous codon occupancies depend

not only on its decoding rate but also its usage in CDSs relative to synonymous alternatives. Nonetheless, if we were to normalize RSCO for relative codon usage as RSCO/RSCU, where RSCU measures the codon's relative usage in CDS, then $E_{decoding,j}$ would still be calculated in the same way since $[RSCO_{average,j}/RSCU_j]/[RSCO_{A,j}/RSCU_j] = RSCO_{average,j}/RSCO_{A,j}$.

4.5 Results

4.5.1 Translation efficiencies affect ribosome abundance in *E. coli* and *B. subtilis*

We expected that changes in both translation initiation and elongation efficiencies will influence the overall number of ribosome units on mRNA transcripts. To test this prediction, we contrasted ribosome abundances between $Gene_{HI,HE}$, $Gene_{HI,LE}$, $Gene_{LI,HE}$, and $Gene_{LI,LE}$ in *E. coli* and *B. subtilis*. Note that number of $Gene_{HI}$ is much higher than number of $Gene_{LI}$ (Figure 4.1) since most genes in both *E. coli* and *B. subtilis* have a SD sequence in their 5' UTR. Since genes also differ in mRNA transcript abundance, Figure 4.1 shows the log transformed ribosomes per transcript (logRPT) abundance calculated as log(Ribosome FPKM/mRNA FPKM).

As expected, efficiently translated genes in *E. coli* with high initiation and elongation efficiencies ($Gene_{HI,HE}$) have the highest logRPT (Figure 4.1a) (See Materials and Methods for logRPT calculation). Reducing either initiation efficiency or elongation efficiency significantly decreased logRPT ($Gene_{HI,HE}$ median (m) = -0.2056 vs. $Gene_{HI,LE}$ m = -0.9755 and $Gene_{LI,HE}$ m = -0.8622, all $P < 0.0001$ based on Wilcoxon rank sum test with continuity correction), and poorly translated genes with both poor initiation and elongation efficiencies have the least logRPT ($Gene_{HI,HE}$ m = -0.2056 vs. $Gene_{LI,LE}$ m = -1.3149, $P < 0.0001$). These patterns were consistently observed when we analyzed Ribo-Seq datasets from a different study (Supplemental Figure S4.2).

Similarly, efficiently translated genes in *B. subtilis* also have the highest logRPT (Figure 4.1b). In comparison, logRPT decreased significantly when either initiation or elongation efficiencies decreases ($Gene_{HI,HE}$ m = -0.0867 vs. $Gene_{HI,LE}$ m = -0.4683, $P < 0.0001$; $Gene_{HI,HE}$ m = -0.0867 vs. $Gene_{LI,HE}$ m = -0.4448, $P = 0.0201$), and logRPT was lowest when both initiation and elongation decrease ($Gene_{HI,HE}$ m = -0.0867 vs $Gene_{LI,LE}$ m = -0.8026, $P < 0.0001$). These results illustrate that both translation initiation and elongation efficiencies affect the overall number of translation units on mRNA transcripts in bacteria.

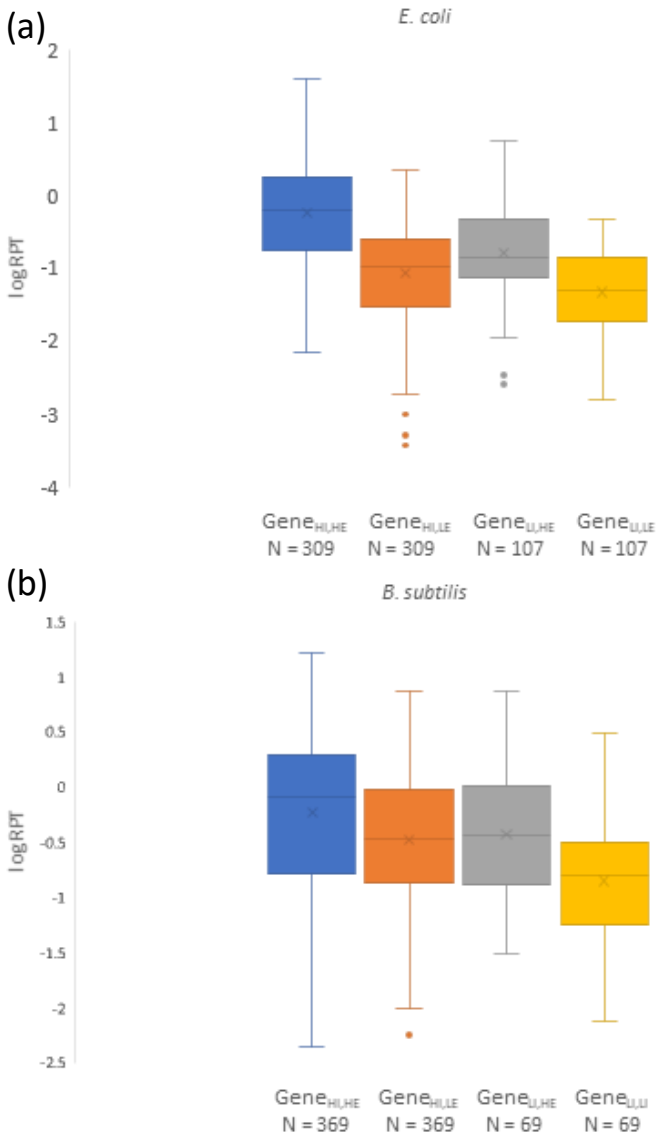


Figure 4.1. Ribosomes per transcript (logRPT) in genes that differ in initiation or elongation efficiencies in (a) *E. coli* and (b) *B. subtilis*. logRPT was calculated as log(ribosome FPKM/RNA FPKM), using Ribo-Seq datasets SRX119208 and SRX119212 and RNA-Seq datasets SRX515174 and SRX515181, for *E. coli* and *B. subtilis* respectively. FPKM values were calculated in ARSDA. N denotes the number of genes in each gene set.

4.5.2 Visualizing effects of SD sequence and secondary structure stability in ribosome loading

We determined genes as having efficient translation initiation by presence of an optimal SD sequence and an easily accessible initiation region encompassing the start codon. We expect these features to substantially increase ribosome loading efficiency. To test this prediction, we compared site-specific RPT read depth at the 60 nt initiation region, extended by 30 nt upstream and 90 nt downstream (180 nt total), between $Gene_{HI,LE}$ and $Gene_{LI,LE}$. Because the number of genes in $Gene_{HI,LE}$ was greater than the number of genes in $Gene_{LI,LE}$, we randomly selected genes from the set of $Gene_{HI,LE}$ so that, for each species, the two sets have equal number of genes (107 in *E. coli* and 69 in *B. subtilis*). Additionally, we chose to analyze sets of genes having poor translation elongation to examine the independent effect of translation initiation efficiency. To verify that these genes only differed in initiation efficiency and not in elongation efficiency, we compared codon usage between the two gene sets and Supplemental Figure S4.3 shows that RSCUs almost perfectly correlate (*E. coli*: $R^2 = 0.9323$, $P < 0.0001$; *B. subtilis*: $R^2 = 0.9521$, $P < 0.0001$).

Indeed, translation initiation efficiency affects ribosome loading. Figure 4.2 shows a notable rise of RPT in $Gene_{HI,LE}$ that aligns with the beginning of optimal SD sequence location as measured by $D_{toStart}$ (region in yellow). Additionally, site-specific RPTs remained much higher in $Gene_{HI,LE}$ than $Gene_{LI,LE}$ until the first five codons in both *E. coli* and *B. subtilis*.

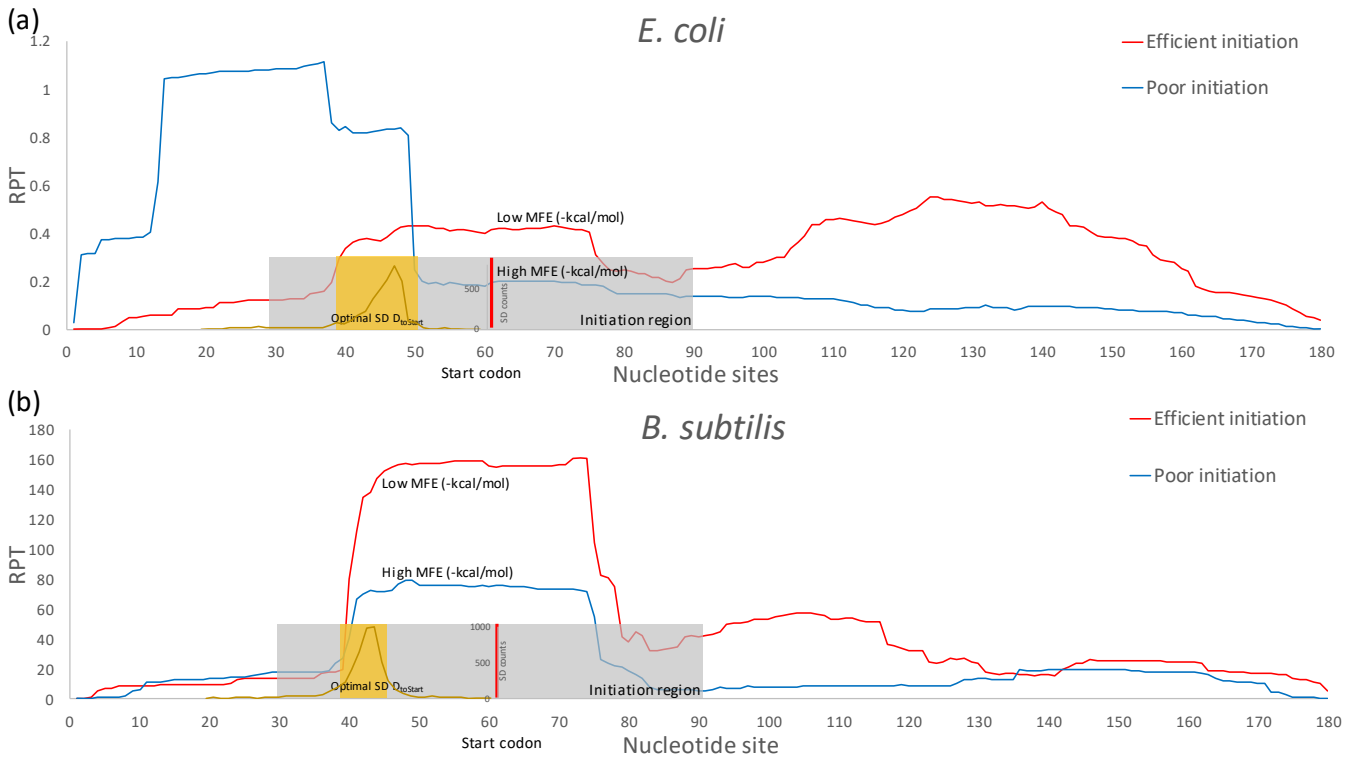


Figure 4.2. Site-specific ribosomes per transcript (RPT) at the translation initiation region (shaded grey) and flanking sites in (a) *E. coli* and (b) *B. subtilis* genes with efficient initiation ($Gene_{HI,LE}$, red, N = 107 and 69, respectively) and genes with poor initiation ($Gene_{LI,LE}$, blue, N = 107 and 69, respectively). All gene sets contain non-pseudo and non-hypothetical genes with protein abundance > 0. The start codon is located at nucleotide sites 61, 62 and 63. In yellow is the Shine-Dalgarno $D_{toStart}$ distribution, and shaded yellow is the optimal $D_{toStart}$ region. The Y-axes for RPT and SD counts are not drawn to scale. Retrieved Ribo-Seq datasets are SRX119208 (*E.* coli) and SRX119212 (*B. subtilis*) and RNA-Seq datasets are SRX515174 (*E. coli*) and SRX515181 (*B. subtilis*).

4.5.3 Estimate decoding efficiency among synonymous codons

To measure a codon's decoding efficiency as $E_{decoding}$ (See materials and Methods), we compared its ribosomal A site occupancy relative to its P, E and flanking site occupancies with the underlying assumption that a notable increase in occupancy at the A site suggests ribosome stalling (Mohammad, Green, Buskirk 2019). As major codons have high tRNA availabilities and minor codons have few tRNA availabilities, we expect to observe ribosome stalling for minor codons but not for major codons.

At each ribosome binding site and flanking site, we took the average of RSCOs (See Materials and Methods for detail) for all major codons (16 in *E. coli* and six in *B. subtilis*) and all minor codons (all others). As expected, Figure 4.3 shows that $RSCO_A$ is the lowest for major codons but the highest for minor codon in both *E. coli* and *B. subtilis*. This suggests that major codons are decoded rapidly when it enters the ribosomal A site, whereas decoding is stalled when minor codons enter the A site. Additionally, we repeated the analysis using the *E. coli* Ribo-Seq dataset SRX4615480 where the experimentally verified ribosomal A site was determined by the 3' end assignment strategy (Mohammad, Green, Buskirk 2019) instead of the center assignment strategy (data from SRX119208 in Figure 4.3, See Materials and Methods). However, we did not observe convincing evidence of A site stalling in minor codons using dataset SRX4615480 (Supplemental Figure S4.4).

Here, we compared site RSCOs for major and minor codons independently, as RSCO is generally higher in major codons than minor codons across all ribosome binding sites (Figure 4.3) because usage of major codons are preferred over minor codons in CDS (Table 4.1). Furthermore, Table 4.1 shows that decoding is efficient ($E_{decoding} > 1$) in 12 out of 15 (80%) major codons in *E. coli* and in 5 out of 6 (83.3%) major codons in *B. subtilis* (See Discussion for limitations in

classifying major codons in *B. subtilis*). In comparison, $E_{decoding} > 1$ in only 18 out of 44 (40.9%) minor codons in *E. coli* and 24 out of 53 (45.3%) minor codons in *B. subtilis*. Indeed, tRNA-mediated codon bias affect ribosome dynamics; while genes with high ITE ($Gene_{HE)}$ prefer major codons that are decoded efficiently, genes with poor ITE ($Gene_{LE}$) overused minor codons that are often decoded slowly (Figure 4.3, Table 4.1).

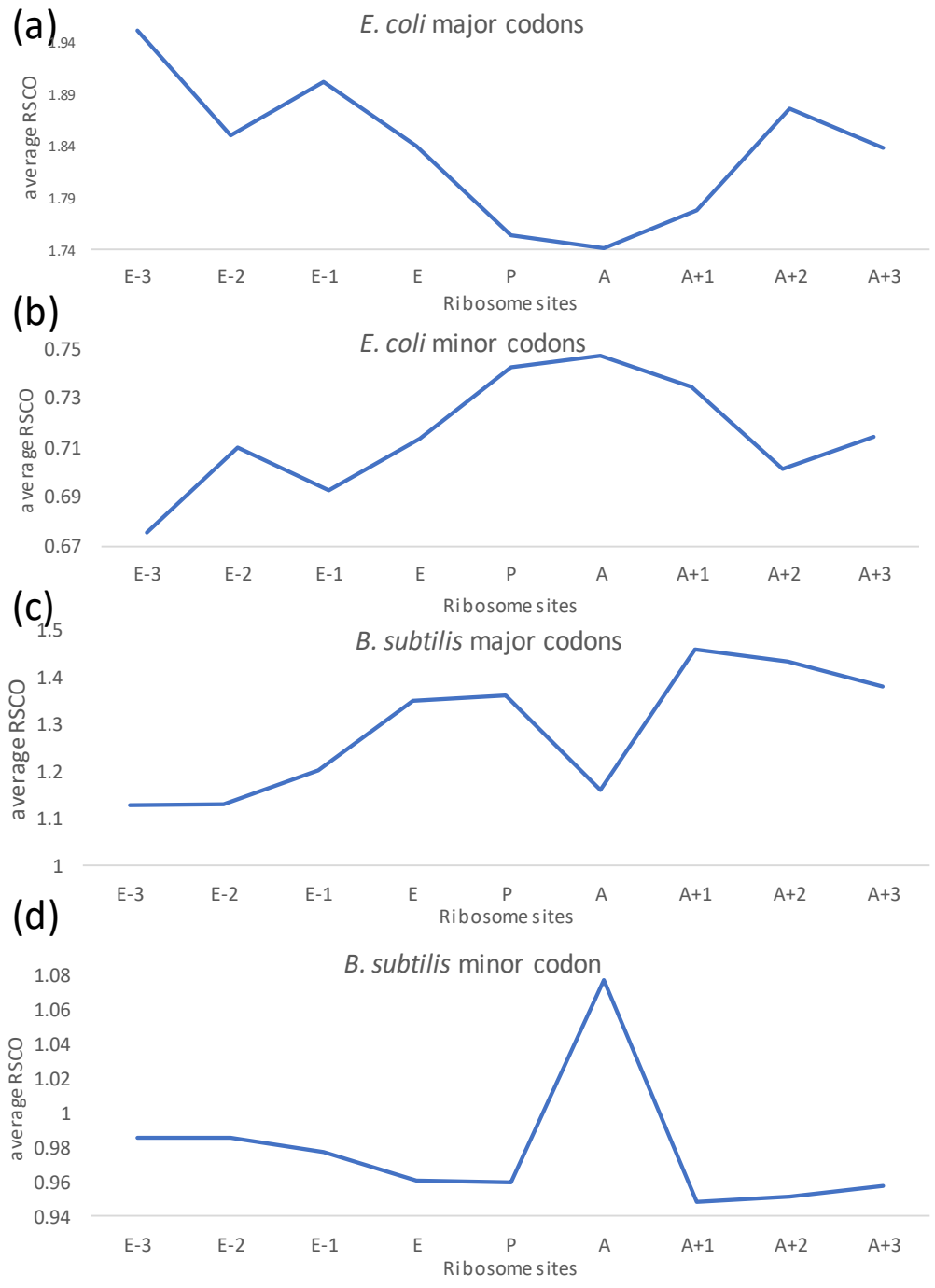


Figure 4.3. The average site-specific RSCO calculated considering entire sequences of all non-pseudo genes, at 15 major codons and 44 minor codons in *E. coli* (from Ribo-Seq dataset SRX119208) and at 6 major codons and 53 minor codons in *B. subtilis* (from Ribo-Seq dataset SRX119212).

Table 4.1. The relative synonymous codon usage (RSCU) for all 61 codons and 3 stop codons in *E. coli* and *B. subtilis* genes with high elongation efficiency ($Gene_{HE}$, N = 416 and 438, respectively) and genes with poor elongation efficiency ($Gene_{LE}$, N = 416 and 438, respectively), and codon's decoding efficiency relative to synonymous alternatives ($E_{decoding}$) in 59 sense codons. Highlighted red are values associated with 15 major codons in *E. coli* and six major codons in *B. subtilis* identified previously (Wei, Silke, Xia 2019).

		E. coli			***B. subtilis***		
A.A.[1]	**Codon**	**RSCU $Gene_{HE}$**	**RSCU $Gene_{LE}$**	**$E_{decoding}$**	**RSCU $Gene_{HE}$**	**RSCU $Gene_{LE}$**	**$E_{decoding}$**
STOP[2]	UGA	0.568	0.909	-	0.398	0.863	-
STOP	UAG	0.079	0.346	-	0.398	0.39	-
STOP	UAA	2.353	1.745	-	2.204	1.747	-
A	GCU	0.765	0.601	0.747	1.24	0.866	0.743
A	GCG	1.396	1.382	1.252	0.913	1.144	1.099
A	GCC	0.999	1.122	1.280	0.57	0.942	1.541
A	GCA	0.839	0.896	1.098	1.278	1.048	0.793
C	UGU	0.797	1.003	0.627	1.016	0.815	0.534
C	UGC	1.203	0.997	1.277	0.984	1.185	1.819
D	GAU	1.12	1.377	0.889	1.281	1.284	1.382
D	GAC	0.88	0.623	1.094	0.719	0.716	0.677
E	GAG	0.54	0.678	1.058	0.538	0.699	1.120
E	GAA	1.46	1.322	0.977	1.462	1.301	1.130
F	UUU	0.869	1.329	0.990	1.199	1.428	1.214
F	UUC	1.131	0.671	1.005	0.801	0.572	0.824
G	GGU	1.629	1.171	0.919	1.011	0.607	1.057
G	GGG	0.362	0.74	1.257	0.44	0.76	0.505

G	GGC	1.809	1.439	1.121	1.279	1.42	2.102
G	GGA	0.2	0.65	0.738	1.271	1.213	0.459
H	CAC	1.069	0.694	1.06	0.746	0.61	0.636
H	CAU	0.931	1.306	0.881	1.254	1.39	1.679
I	AUU	1.343	1.68	0.811	1.601	1.501	1.389
I	AUA	0.07	0.346	0.860	0.301	0.402	0.277
I	AUC	1.587	0.975	1.1111	1.098	1.097	1.334
K	AAA	1.562	1.524	0.955	1.494	1.35	1.399
K	AAG	0.438	0.476	1.129	0.506	0.65	0.623
L	CUA	0.102	0.298	0.830	0.457	0.247	0.435
L	CUC	0.466	0.614	0.698	0.542	0.775	0.563
L	CUG	3.049	2.364	1.053	1.105	1.658	2.112
L	CUU	0.383	0.724	0.741	1.895	1.32	0.630
L	UUA	0.938	1.107	0.824	1.267	1.02	0.945
L	UUG	1.062	0.893	1.161	0.733	0.98	1.058
M	AUG	1	1	-	1	1	-
N	AAC	1.367	0.853	1.045	0.965	0.814	1.062
N	AAU	0.633	1.147	0.810	1.035	1.186	0.973
P	CCA	0.732	0.827	0.586	0.956	0.633	0.780
P	CCC	0.256	0.684	0.520	0.203	0.439	0.301
P	CCU	0.483	0.786	0.972	1.486	0.983	0.830
P	CCG	2.529	1.703	1.214	1.355	1.945	1.7645
Q	CAA	0.563	0.83	0.862	1.24	0.912	0.459
Q	CAG	1.437	1.17	1.030	0.76	1.088	3.120
R	AGA	1.458	1.307	1.138	1.712	1.369	1.189
R	AGG	0.542	0.693	0.745	0.288	0.631	0.944

R	CGA	0.114	0.442	0.828	0.503	0.638	0.231
R	CGC	1.712	1.597	1.249	1.436	1.217	2.682
R	CGG	0.19	0.637	0.967	0.412	1.269	0.785
R	CGU	1.983	1.323	0.908	1.648	0.877	1.221
S	AGC	1.488	1.149	1.108	1.266	1.433	1.279
S	AGU	0.512	0.851	0.709	0.734	0.567	0.727
S	UCA	0.557	1.119	1.042	1.534	1.344	0.587
S	UCC	1.304	0.821	1.346	0.553	0.878	1.381
S	UCG	0.823	1.172	1.529	0.329	0.729	1.166
S	UCU	1.316	0.888	0.748	1.584	1.049	1.277
T	ACC	2.086	1.424	1.024	0.395	0.742	1.530
T	ACA	0.314	0.718	1.371	1.85	1.56	0.528
T	ACG	0.799	1.266	0.789	0.819	1.217	1.192
T	ACU	0.802	0.593	1.013	0.937	0.481	0.627
V	GUU	1.183	0.951	0.852	1.359	0.961	0.867
V	GUG	1.426	1.498	1.132	0.794	1.205	2.060
V	GUC	0.746	0.947	1.086	0.806	1.161	0.847
V	GUA	0.645	0.604	1.139	1.041	0.673	0.638
W	UGG	1	1	-	1	1	-
Y	UAC	1.073	0.709	1.104	0.753	0.651	0.881
Y	UAU	0.927	1.291	0.826	1.247	1.349	1.103

[1] – Amino acid (A.A.).

[2] – Stop codon (STOP).

4.5.4 The degree ribosome per transcript abundance correlates protein expression depends more on translation elongation efficiency than translation initiation efficiency

Above results showed that both translation initiation and elongation affect ribosome dynamics and lead to the following predictions: 1) when translation is efficient, ribosome abundance should well correlate protein abundance as ribosomes flow in absence of traffic jams, and 2) when translation is poor, ribosome abundance do not well correlate protein abundance as ribosome traffic variably stalls. To test these predictions, we measured the degree ribosome per transcript abundance (in logRPT) correlates protein expression when translation initiation and elongation efficiencies are altered.

Figure 4.4 shows that *E. coli* logRPT indeed well correlates with protein abundance in genes having efficient translation initiation and elongation ($Gene_{HI,HE}$: $R^2 = 0.4182$, $P < 0.0001$), but logRPT poorly correlates protein abundance when both initiation and elongation efficiencies are poor ($Gene_{LI,LE}$: $R^2 = 0.1867$, $P < 0.0001$). Additionally, compared to efficiently translated genes, the correlation between logRPT and protein abundance was considerably reduced in genes having poor elongation efficiency even when initiation was kept efficient ($Gene_{HI,LE}$: $R^2 = 0.0935$, $P < 0.0001$). In comparison, reducing initiation efficiency but keeping efficient elongation moderately reduced the correlation between logRPT and protein abundance ($Gene_{LI,HE}$: $R^2 = 0.2932$, $P < 0.0001$). Similarly, these patterns were observed when we calculated logRPT by employing Ribo-Seq datasets from a different study (SRX4615480 and SRX4615482, Supplemental Figure S4.5).

As for *B. subtilis*, Figure 4.5 shows a modest correlation between logRPT and protein abundance in efficiently translated genes ($Gene_{HI, HE}$: $R^2 = 0.2946$, $P < 0.0001$) and a poor correlation between these two variables in poorly translated genes ($Gene_{LI,LE}$: $R^2 = 0.1207$, P =

0.0116). In comparison to efficiently translated genes, the correlation was considerably reduced in genes having poor elongation efficiency but good initiation efficiency ($Gene_{HI,LE}$: $R^2 = 0.0953$, $P < 0.0001$), and the correlation was moderately reduced in genes having poor initiation efficiency but good elongation efficiency ($Gene_{LI,HE}$: $R^2 = 0.1922$, $P < 0.0001$).

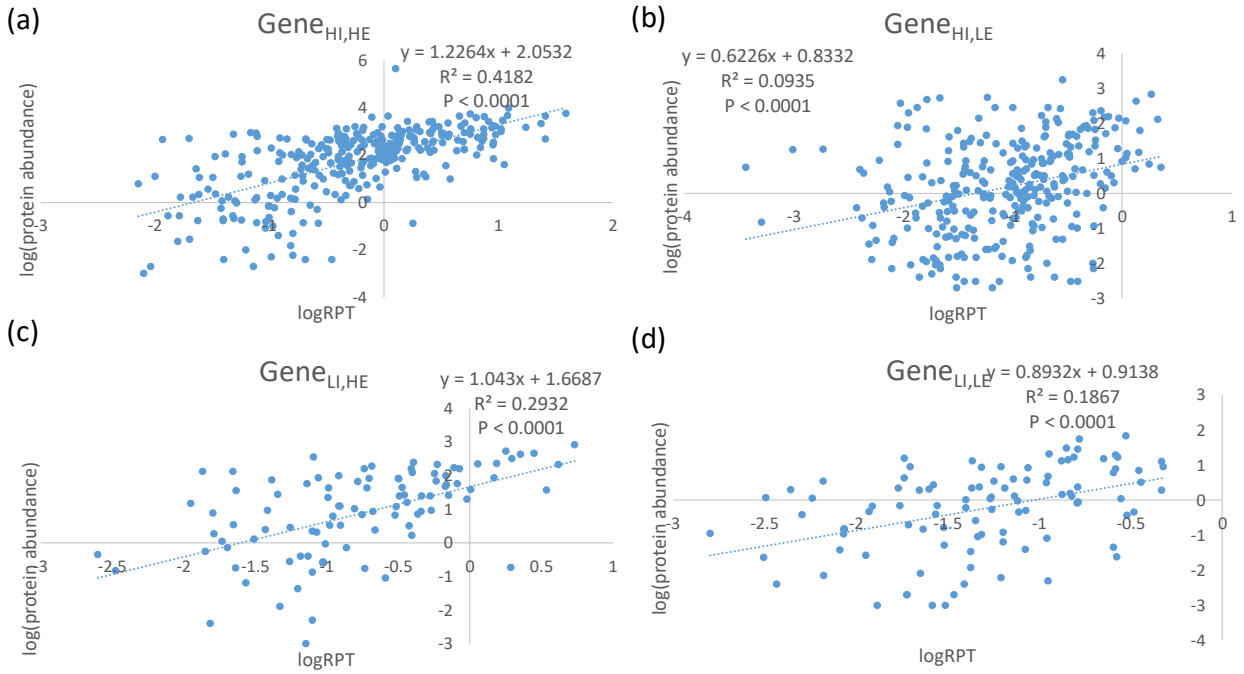


Figure 4.4. Relationship between *E. coli* log protein abundance (Integrated protein abundance from PaxDB) and logRPT (Ribo-Seq dataset: SRX119208, RNA-Seq dataset: SRX515174), in four sets of genes separated by translation initiation and elongation efficiencies. All gene sets contain non-pseudo and non-hypothetical genes with protein abundance > 0. See Figure 4.1 for number of genes in each set.

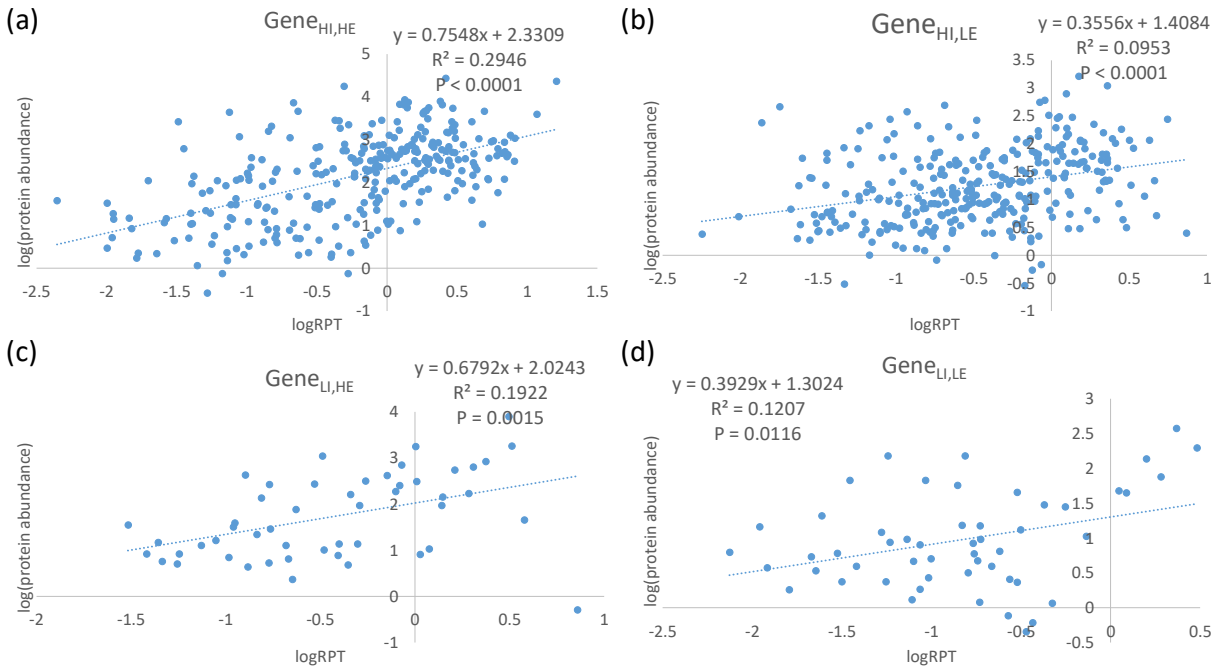


Figure 4.5. Relationship between *B. subtilis* log protein abundance (Integrated protein abundance from PaxDB) and logRPT (Ribo-Seq dataset: SRX119212, RNA-Seq dataset: SRX515181) in four sets of genes separated by translation initiation and elongation efficiencies. All gene sets contain non-pseudo and non-hypothetical genes with protein abundance > 0. See Figure 4.1 for number of genes in each set.

4.6 Discussion

It is commonly accepted that initiation is rate-limiting in bacterial translation (Jacques, Dreyfus 1990; Kudla et al. 2009; Xia 2015); this implies that increasing elongation efficiency may not strongly impact protein production if translation is bottlenecked by poor initiation. Nevertheless, previous studies have debated over the centrality of translation initiation and elongation in bacterial gene expression. The availability of Ribosome Sequencing data prompted us to investigate the influence each translation subprocess exerts on gene expression in *E. coli* and *B. subtilis*. To this end, we traced how changes in translation initiation and elongation efficiencies

affect ribosome dynamics along the mRNA and then measured how this change in decoder flow affects protein expression.

In both *E. coli* and *B. subtilis*, we found high ribosome occupancy near the beginning of start codon (Figure 4.2), consistent with early ribosome profiling studies (Ingolia et al. 2009; Subramaniam, Zid, O'Shea 2014). To compare ribosome loading efficiencies between $Gene_{HI,LE}$ and $Gene_{LI,LE}$ that differ only in translation initiation efficiency, we examined site specific RPT. In both bacteria, our approach detected a notable rise in RPT in $Gene_{HI}$ over $Gene_{LI}$, beginning from the SD sequence location until the first five codons (Figure 4.2). This result is indicative of an increase in ribosome loading efficiency in genes having good translation initiation features: 1) the presence of an SD with optimal location accurately measured by $D_{toStart}$, and 2) reduced secondary structure stability up until the first five codon sites, consistent with a previously report (Bentele et al. 2013) that the first 5-10 codons are AT-rich due to selection to reduce secondary structure stability in bacteria. Hence, differences in ribosome loading rate at least partly explains why both whole-sequence RPT (Figure 4.1) and site-specific RPT (Figure 4.2) are greater in $Gene_{HI}$ than $Gene_{LI}$. Lastly, it is worth noting that in *E. coli* $Gene_{LI,LE}$ (Figure 4.2a), there was high traffic of ribosomes upstream of CDS which drops sharply around site 50 (-10 nt relative to start codon); this may indicate that poor translation initiation has bottlenecked ribosome flow.

Early studies showed that modifying codon usage affects rate of translation (Robinson et al. 1984; Varenne et al. 1984; Sorensen, Kurland, Pedersen 1989; Komar, Lesnik, Reiss 1999). Yet, recent ribosome profiling studies were unable to observe a significant correlation between ribosome stalling and usage of codon encoding different amino acids (Li, Oh, Weissman 2012; Charneski, Hurst 2013). These results may not be surprising because amino acid usage are variable (Akashi, Gojobori 2002; Paul, Malakar, Chakraborty 2018) and decoding efficiencies differ among

amino acids (e.g.; charged vs. uncharged) (Trylska et al. 2004; Lu, Deutsch 2008; Charneski, Hurst 2013; Dana, Tuller 2014a).

To better compare codon decoding efficiencies, we contrasted ribosome occupancies among only synonymous codons (RSCO, Equation (1)). A similar use of Ribo-seq data was presented in Shaham, Tuller (2017), which described an in-depth TASEP-based (Shaw, Zia, Lee 2003) computational model to estimate rates of translation elongation in *E. coli* based on ribosomal density, traffic jam, and decoding rate. To estimate codon-specific decoding time (decoding efficiency) they normalized the distribution of ribosome footprint counts for any given codon *j* in windows of 50 codons (Dana, Tuller 2014b). The difference between our approaches is that we determined decoding efficiency at the resolution of single ribosome binding sites.

Our approach indicates that minor codons having few tRNA availabilities are indeed decoded slowly at the A site, whereas major codons having high tRNA availabilities tend to be decoded rapidly at the A site (Figure 4.3). These effects are congruous with a previous observation of increased Leucine codon occupancy at the A site when the availability of charged Leucine tRNAs decreases (Subramaniam, Zid, O'Shea 2014). Codon decoding efficiency as measured by $E_{decoding}$ tends to be higher in major codons than minor codons in both *E. coli* and *B. subtilis* (Table 4.1). Moreover, $Gene_{LE}$ with poor codon adaptation as measured by ITE overused minor codons whereas $Gene_{HE}$ having high ITE preferred major codons (Table 4.1). Together, these factors explain why whole-sequence ribosome abundance is significantly reduced in $Gene_{LE}$ in comparison to $Gene_{HE}$ (Figure 4.1). Nonetheless, it is worth mentioning that while stalling events create a drop-off in footprint abundance downstream of the stalled sites, abundance of ribosome footprints at the stalled site would increase. How these two trends balance acts as an additional

layer of complexity that determines whether overall ribosomes per transcript decrease in genes with poor elongation.

A limitation in our analysis for codon decoding efficiency is that a very limited number of major codons were identified in *B. subtilis*. Using an RNA-Seq-based approach, we had previously determined 15 major codons in Chapter 3 (Table 4.1) in *E. coli* and all identifications were consistent with those determined by F_{op}. In *B. subtilis* however, our approach had only identified six major codons, the reason being many codons with abundant tRNA availabilities did not have preferred usage over synonymous alternatives in CDS, even for highly expressed genes. Alternatively, using RNA fingerprinting experiments, Kanaya et al. (1999) determined 16 optimal codons in *B. subtilis*, but many may not be translationally optimal because more than one major codon was determined for many amino acids. Additionally, their identification of major codons did not consider codon preference. For example, CCA (Pro) was determined to be an optimal codon (Kanaya et al. 1999), but CCG (Pro) is substantially more preferred than CCA (Pro) (RSCU in highly expressed genes are 1.735 and 0.782, respectively). Furthermore, tRNA copy number was an especially poor estimator in *B. subtilis* (dos Reis, Savva, Wernisch 2004) and tAI (based on tRNA gene copy) failed to predict codon usage in this specific species. Hence, while our RNA-Seq-based approach was able to identify a limited number of major codons in *B. subtilis*, our identifications are likely to be more accurate than alternative methods.

Finally, we measured how well ribosomes per transcript correlates with protein expression when efficacy of translation subprocesses change. In both *E. coli* and *B. subtilis*, RPT best correlated protein abundance in efficiently translated genes. More specifically, in both *E. coli* and *B. subtilis*, the degree RPT correlates protein abundance depended heavily on translation elongation efficiency but not on translation initiation efficiency (Figure 4.4, 4.5). A caveat is that

a degree of noise pertains to this association since we did not consider protein degradation rates due to the scarcity of available data (Goldberg 1972; Maurizi 1992; Cameron, Collins 2014).

In summary, our results are consistent with the conventional wisdom that translation initiation is often rate-limiting in bacteria because rate of ribosome recruitment is bottlenecked by initiation efficiency. When translation initiation is efficient, elongation becomes rate-limiting and poor elongation leads to ribosome stalling that results in ribosome traffic jam (Shaham, Tuller 2017) and dissociation (Subramaniam, Zid, O'Shea 2014). Nevertheless, our results do not support the prediction that translation elongation efficiency would weakly influence gene expression when initiation is poor. Rather, they emphasize the importance of elongation efficiency as measured by codon adaptation in predicting gene expression.

4.7 Data availability

Supplemental figures for Chapter 4 are available in Appendix D. Data are available upon request.

4.8 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to X.X. [RGPIN/2018-03878], and NSERC Doctoral Scholarship to Y.W. [CGSD/2019-535291]. We thank Jordan R. Silk for comments.

4.9 References

Abolbaghaei, A, JR Silke, X Xia. 2017. How Changes in Anti-SD Sequences Would Affect SD Sequences in Escherichia coli and Bacillus subtilis. G3 (Bethesda).

Akashi, H, T Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proceedings of the National Academy of Sciences 99:3695-3700.

Andersson, GE, PM Sharp. 1996. Codon usage in the Mycobacterium tuberculosis complex. Microbiology 142:915-925.

Band, L, DJ Henner. 1984. Bacillus subtilis requires a "stringent" Shine-Dalgarno region for gene expression. Dna 3:17-21.
Ben-Yehezkel, T, S Atar, H Zur, et al. 2015. Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. RNA Biol 12:972-984.
Bentele, K, P Saffert, R Rauscher, Z Ignatova, N Bluthgen. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9:32.
Bolger, AM, M Lohse, B Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
Bray, NL, H Pimentel, P Melsted, L Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525-527.
Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-907.
Cameron, DE, JJ Collins. 2014. Tunable protein degradation in bacteria. Nature Biotechnology 32:1276.
Charneski, CA, LD Hurst. 2013. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol 11:12.
Consortium, TU. 2018. UniProt: a worldwide hub of protein knowledge. Nucleic acids research 47:D506-D515.
Dana, A, T Tuller. 2014a. The effect of tRNA levels on decoding times of mRNA codons. Nucleic acids research 42:9171-9181.
Dana, A, T Tuller. 2014b. Properties and determinants of codon decoding time distributions. BMC Genomics 15:1471-2164.
Dong, H, L Nilsson, CG Kurland. 1996. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. Journal of Molecular Biology 260:649-663.
Dontsova, O, A Kopylov, R Brimacombe. 1991. The location of mRNA in the ribosomal 30S initiation complex; site-directed cross-linking of mRNA analogues carrying several photo-reactive labels simultaneously on either side of the AUG start codon. Embo J 10:2613-2620.
dos Reis, M, R Savva, L Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036-5044.
Ewing, B, L Hillier, MC Wendl, P Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.
Goldberg, AL. 1972. Degradation of abnormal proteins in Escherichia coli (protein breakdown-protein structure-mistranslation-amino acid analogs-puromycin). Proceedings of the National Academy of Sciences of the United States of America 69:422-426.
Guimaraes, JC, M Rocha, AP Arkin. 2014. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. Nucleic acids research 42:4791-4799.
Hockenberry, AJ, AR Pah, MC Jewett, LAN Amaral. 2017. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. Open Biology 7.
Hofacker, IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Research 31:3429-3431.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13-34.
Ingolia, NT, S Ghaemmaghami, JR Newman, JS Weissman. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218-223.

Jacques, N, M Dreyfus. 1990. Translation initiation in Escherichia coli: old and new questions. Mol Microbiol 4:1063-1067.
Kanaya, S, Y Yamada, Y Kudo, T Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143-155.
Komar, AA, T Lesnik, C Reiss. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462:387-391.
Kudla, G, AW Murray, D Tollervey, JB Plotkin. 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324:255-258.
Li, G-W, E Oh, JS Weissman. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538-541.
Li, GW, D Burkhardt, C Gross, JS Weissman. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157:624-635.
Li, H, L Yang, Y Wang, H Wang, Z Xie, L Chen. 2018. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic acids research 47:D230-D234.
Liljenstrom, H, G von Heijne. 1987. Translation rate modification by preferential codon usage: intragenic position effects. J Theor Biol 124:43-55.
Lu, J, C Deutsch. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 384:73-86.
Ma, J, A Campbell, S Karlin. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733-5745.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.
Maurizi, MR. 1992. Proteases and protein degradation in Escherichia coli. Experientia 48:178-201.
Mohammad, F, R Green, AR Buskirk. 2019. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. Elife 6:42591.
Nakagawa, S, Y Niimura, T Gojobori. 2017. Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. Nucleic Acids Res 45:3922-3931.
Nakagawa, S, Y Niimura, K-i Miura, T Gojobori. 2010. Dynamic evolution of translation initiation mechanisms in prokaryotes. Proceedings of the National Academy of Sciences 107:6382-6387.
Nakamoto, T. 2006. A unified view of the initiation of protein synthesis.
Nie, L, G Wu, W Zhang. 2006. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. Genetics 174:2229-2243.
Novoa, EM, M Pavon-Eternod, T Pan, L Ribas de Pouplana. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149:202-213.
Osterman, IA, SA Evfratov, PV Sergiev, OA Dontsova. 2013. Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic acids research 41:474-486.
Paul, P, AK Malakar, S Chakraborty. 2018. Codon usage and amino acid usage influence genes expression level. Genetica 146:53-63.

Prabhakaran, R, S Chithambaram, X Xia. 2015. Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. Journal of General Virology 96:1169-1179.
Robinson, M, R Lilley, S Little, JS Emtage, G Yarranton, P Stephens, A Millican, M Eaton, G Humphreys. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic acids research 12:6663-6671.
Rocha, EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Research 14:2279-2286.
Rodnina, MV. 2016. The ribosome in action: Tuning of translational efficiency and protein folding. Protein science : a publication of the Protein Society 25:1390-1406.
Scharff, LB, L Childs, D Walther, R Bock. 2011. Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites.
Schurr, T, E Nadir, H Margalit. 1993. Identification and characterization of E.coli ribosomal binding sites by free energy computation. Nucleic Acids Res 21:4019-4023.
Shaham, G, T Tuller. 2017. Genome scale analysis of Escherichia coli with a comprehensive prokaryotic sequence-based biophysical model of translation initiation and elongation. DNA Research 25:195-205.
Sharp, PM, WH Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28-38.
Sharp, PM, WH Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281-1295.
Shaw, LB, RK Zia, KH Lee. 2003. Totally asymmetric exclusion process with extended objects: a model for protein synthesis. Phys Rev E Stat Nonlin Soft Matter Phys 68:18.
Shine, J, L Dalgarno. 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 71:1342-1346.
Shine, J, L Dalgarno. 1975. Determinant of cistron specificity in bacterial ribosomes. Nature 254:34-38.
Sorensen, MA, CG Kurland, S Pedersen. 1989. Codon usage determines translation rate in Escherichia coli. J Mol Biol 207:365-377.
Starmer, J, A Stomp, M Vouk, D Bitzer. 2006. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. PLoS Comput Biol 2:19.
Subramaniam, AR, BM Zid, EK O'Shea. 2014. An integrated approach reveals regulatory controls on bacterial translation elongation. Cell 159:1200-1211.
Trylska, J, R Konecny, F Tama, CL Brooks, 3rd, JA McCammon. 2004. Ribosome motions modulate electrostatic properties. Biopolymers 74:423-431.
Tuller, T, YY Waldman, M Kupiec, E Ruppin. 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 107:3645-3650.
Varenne, S, J Buc, R Lloubes, C Lazdunski. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 180:549-576.
Vimberg, V, A Tats, M Remm, T Tenson. 2007. Translation initiation region sequence preferences in Escherichia coli. BMC Mol Biol 8:1471-2199.

Wang, M, CJ Herrmann, M Simonovic, D Szklarczyk, C von Mering. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163-3168.
Wei, Y, JR Silke, X Xia. 2017. Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in Escherichia coli and Bacillus subtilis using RNA-Seq data. Sci Rep 7:017-17918.
Wei, Y, JR Silke, X Xia. 2019. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci Rep 9:3184.
Welch, M, S Govindarajan, JE Ness, A Villalobos, A Gurney, J Minshull, C Gustafsson. 2009. Design Parameters to Control Synthetic Gene Expression in Escherichia coli. PLoS One 4:e7002.
Williams, CR, A Baccarella, JZ Parrish, CC Kim. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17:103-103.
Woolstenhulme, CJ, NR Guydosh, R Green, AR Buskirk. 2015. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep 11:13-21.
Xia, X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37-44.
Xia, X. 2015. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199:573-579.
Xia, X. 2017. ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. G3: Genes|Genomes|Genetics 7:3839-3848.
Xia, X. 2018a. Bioinformatics and the cell: modern computational apporaches in genomics, proteomics and transcriptomics: Springer.
Xia, X. 2018b. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol Biol Evol 35:1550-1552.

CHAPTER FIVE

Predicting mammalian species at risk of being infected by SARS-CoV-2 from an ACE2 perspective

This chapter is published:

Wei, Y., Aris, P., Farookhi, H., Xia, X. 2021. Predicting mammalian species at risk of being infected by SARS-CoV-2 from an ACE2 perspective. Scientific Reports 11:1702.

5.1 Author Contributions

Yulong Wei[1], Parisa Aris[1], Heba Farookhi[1], 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.

Y.W. and X.X. designed the study. Y.W. wrote the manuscript, and X.X. provided guidance and editing. Y.W., P.A., and H.F. collected and analyzed the data. Y.W. made Figures 5.1, 5.2, 5.3, 5.4, and Table 5.1. P.A. made Figures S5.1, S5.2, S5.3. Y.W. and H.F. contributed equally in making Supplemental File S1. All authors reviewed the manuscript. X.X. supervised the study.

5.2 Abstract

SARS-CoV-2 can transmit efficiently in humans, but it is less clear which other mammals are at risk of being infected. SARS-CoV-2 encodes a Spike (S) protein that binds to human ACE2 receptor to mediate cell entry. A species with a human-like ACE2 receptor could therefore be at risk of being infected by SARS-CoV-2. We compared amino acid sequences between 132 mammalian *ACE2* genes and between 17 coronavirus S protein gene. We showed that while global similarities reflected by whole *ACE2* gene alignments are poor predictors of high-risk mammals, local similarities at key S protein-binding sites highlight several high-risk mammals that share good *ACE2* homology with human. Bats are likely reservoirs of SARS-CoV-2, but there are other high-risk mammals that share better *ACE2* homologies with human. Both SARS-CoV-2 and SARS-CoV are closely related to the bat coronavirus. Yet, among host-specific coronaviruses infecting high-risk mammals, key ACE2-binding sites on S proteins share highest similarities between SARS-CoV-2 and Pangolin-CoV and between SARS-CoV and Civet-CoV. These results suggest that direct coronavirus transmission from bat to human is unlikely, and that adaptation of a bat SARS-like coronavirus in high-risk intermediate hosts could have allowed it to acquire distinct high binding potential between S protein and human-like ACE2 receptors.

5.3 Introduction

The *Betacoronavirus* SARS-CoV-2 poses a serious global health emergency. Since its emergence in Wuhan city, Hubei province of China in December 2019, the viral outbreak has resulted in over 20 million confirmed cases of COVID-19 worldwide (https://www.who.int/emergencies/diseases/novel-coronavirus-2019, last accessed August 25, 2020). While it is evident that SARS-CoV-2 can transmit efficiently from person to person, it is less clear which other mammalian species are at high risk of being infected. The answer to this question is important to 1) improve our ability to predict and control future pandemics, and 2) manage and protect wildlife and domesticated animals.

5.3.1 Mammals at high risk of SARS-CoV-2 infection should have human-like ACE2 receptors with conserved key S protein-binding sites

Both SARS-CoV and SARS-CoV-2 genomes encode a Spike (S) protein that binds to human Angiotensin-converting enzyme 2 (ACE2) receptor to mediate viral entry into the host cell (Li et al. 2005a; Li et al. 2005b; Cui, Li, Shi 2019; Lan et al. 2020; Shang et al. 2020; Zhou et al. 2020). Mechanistically, two S protein domains are involved in coronavirus infection in mammalian cells that express ACE2: the S1 domain interacts with the ACE2 receptor (Tian et al. 2020; Walls et al. 2020), and the S2 domain undergoes structural rearrangements to mediate membrane fusion (Li et al. 2005b). Interacting with the S1 domain, the S protein-binding sites on ACE2 receptor are primarily located in the α-helix 1 and β-sheet 5 domains (Li et al. 2005b). The efficacy of the interaction between the S protein and the ACE2 receptor is a good predictor of the severity of coronavirus infection (Hamming et al. 2004; Nie et al. 2004; Li et al. 2005b). For example, a potential source of the SARS-CoV outbreak was the masked palm civets (Guan et al.

2003; Song et al. 2005). Palm civet ACE2 receptors bind efficiently to S proteins of Civet-CoV strain SZ3 isolated from infected palm civets, to S proteins of SARS-CoV strain TOR2 that caused the severe 2002-2003 outbreak, and to S proteins of SARS-CoV strain GD that caused the mild 2003-2004 outbreak (Guan et al. 2003; Marra et al. 2003; Rota et al. 2003; Song et al. 2005). Whereas human ACE2 receptors bind efficiently to S proteins of the severe SARS-CoV strain TOR2 but do not bind efficiently to S proteins of Civet-CoV strain SZ3 or of the less severe SARS-CoV strain GD (Li et al. 2005b). Indeed, the binding potential between viral S protein and host ACE2 receptor is a key determinant of viral infectivity.

The binding potential between viral S protein and host ACE2 receptor is attributed to several key binding sites. Differences at key S protein-binding sites between mammalian ACE2 receptors explained why SARS-CoV efficiently infected humans and palm civets but not rats, and introducing point mutations into the *AEC2* gene variably affected the binding potential between ACE2 receptor and SARS-CoV S protein in these species (Li et al. 2005b). For example, experimentally mutating rat His353 to human Lys353 turned the rat ACE2 receptor from one that poorly binds S protein into one that is efficient for binding, introducing amino acid residues 82 to 84 from human *ACE2* into rat *ACE2* also led to an increase in S protein-binding potential, but replacing human *ACE2* Met82 to rat *ACE2* Asn82 partially inhibited S protein-binding. Furthermore, mutating human *AEC2* at Lys31, Tyr41, Asp355, and Arg357 also decreases receptor binding potential with S protein. On the contrary, changes to other human *ACE2* sites such as Gly354 (corresponding to Asp354 in palm civet) and amino acid residues 90-93 that potentially interact with S protein residue 479 do not affect the efficacy of S protein-binding[10]. In brief, gene mutation experiments (Li et al. 2005a; Li et al. 2005b) showed that the conservation of Lys31 and Tyr41 on α-helix 1, residues 82-84 in the vicinity of α-helix 3, and residues 353-357 in β-sheet 5

in human ACE2 receptor are crucial for SARS-CoV infection because their replacements weakened the binding potential between SARS-CoV S protein and ACE2 receptor.

To showcase the importance of contrasting between mammalian *ACE2* genes at key S protein-binding sites, Figure 5.1 shows comparisons at *ACE2* genes in a sample of mammalian species in two ways: the global similarities among whole *ACE2* gene alignments as reflected by the mammalian phylogenetic relationships on the left, and the local similarities at key *ACE2* sites that are involved in SARS-CoV S protein-binding (Li et al. 2005a; Li et al. 2005b) in the table on the right. The phylogenetic relationship shows that human *ACE2* shares higher global similarity with *ACE2* of species from Primates and Rodentia orders than with *ACE2* of species from Carnivora, Artiodactyla, and Chiroptera orders. This emphasizes that a phylogenetic relationship based on global *ACE2* gene comparisons is not a good predictor of mammals at high risk of being infected by SARS-CoV for three reasons. First, it does not identify the *Rhinolophus* bats as a potential reservoir for the progenitor virus of SARS-CoV as previously postulated (Liu et al. 2020). Second, it does not reveal masked palm civets as a potential intermediate host (Guan et al. 2003; Song et al. 2005). Third, it misidentifies the rat as a high-risk mammal while experimental evidence suggests SARS-CoV poorly infects the rat (Li et al. 2005b). In contrast, the table of comparisons at key SARS-CoV S protein-binding sites on *ACE2* genes show that species in the Carnivora, Artiodactyla, and Chiroptera orders share highest local similarities with human. These high-risk mammals include the *Rhinolophus* bat and masked palm civets, even though they share less global *ACE2* similarity with human than Primates and Rodentia species. Hence, comparing key *ACE2* sites may provide insights into the host range of viral infection.

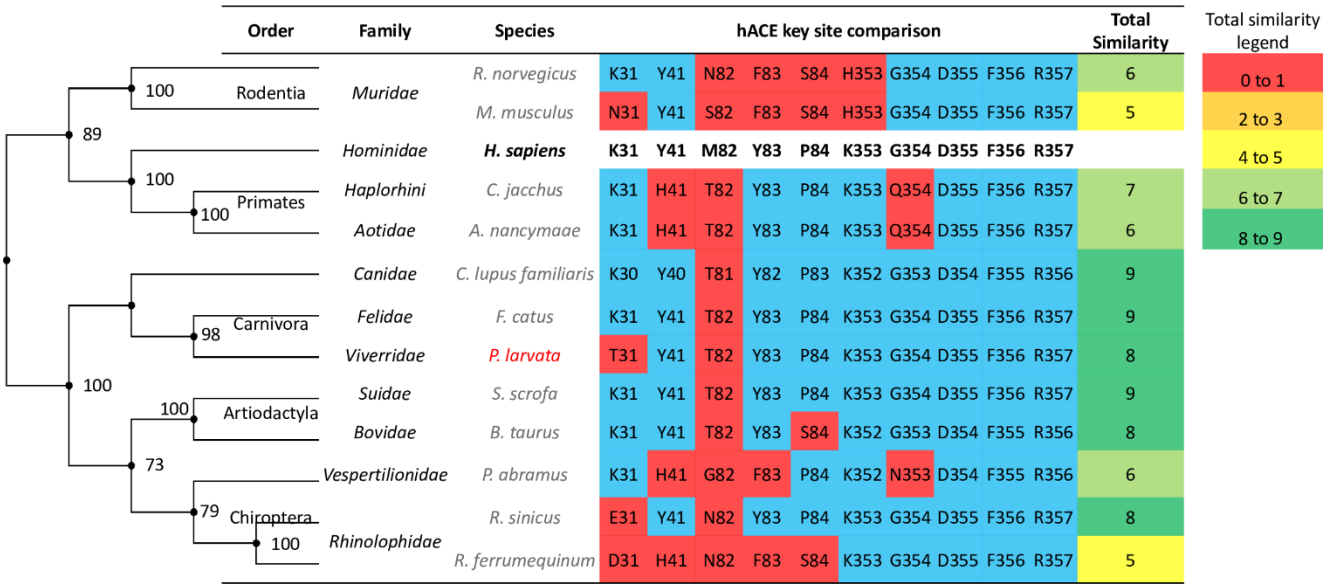


Figure 5.1. A sample comparison between mammalian *ACE2* genes. Global similarities represented by whole *ACE2* gene comparisons (the phylogenetic tree on the left) poorly predict mammals at high risk of being infected by SARS-CoV; whereas local similarities at key human *ACE2* sites (the table on the right) show that species in the Carnivora, Artiodactyla and Chiroptera orders are at high risk of SARS-CoV infection. The tips of the phylogenetic tree are aligned with the corresponding species listed in the table. Species in red is the masked palm civet, a possible intermediate host of SARS-CoV. Species in black bold is human, and all other species are in grey. Amino acids highlighted in blue and red are matching and mis-matching sites, respectively, between mammalian *ACE2* and human *ACE2* at key SARS-CoV S protein binding sites. "Total Similarity" is calculated as the sum of matching amino acids between mammalian and human *ACE2* at key binding sites. A complete comparison at key *ACE2* sites involved in SARS-CoV binding between human and 131 mammals is shown in Supplemental Figure S5.3.

The above observations allow for the prediction that high-risk mammals of the newly emerging SARS-CoV-2 should have human-like ACE2 receptors at key SARS-CoV-2 S protein-

binding sites. Current research on SARS-COV-2 infection in animals is limited. Nonetheless, SARS-CoV-2 infection has been detected in the cat (Halfmann et al. 2020), dog (Sit et al. 2020), mink (Oreshkova et al. 2020), and tiger (Wang et al. 2020). In addition, ferret (Kim et al. 2020; Shi et al. 2020), macaque (Chandrashekar et al. 2020; Rockx et al. 2020), and grivet (Woolsey et al. 2020) could be experimentally infected by SARS-CoV-2. We therefore expect the *ACE2* genes of these known high-risk mammals to bear a high resemblance to human *ACE2* at key S protein-binding sites. Moreover, two Rodentia species have been experimentally infected by SARS-CoV-2: the golden Syrian hamster (*Mesocricetus auratus*) belonging to the *Cricetidae* family, and the house mouse (*Mus musculus*) belonging to the *Muridae* family. While hamsters could be consistently infected by SARS-CoV-2 (Chan et al. 2020), wild-type mice were not susceptible to SARS-CoV-2 infection (Bao et al. 2020). It follows that these two Rodentia species should have distinct differences at the ACE2 receptor, and we expect key *ACE2* sites in human to be well conserved by hamster *ACE2* but not by mouse *ACE2*. Sequence comparisons between human ACE2 receptor and mammalian ACE2 receptors at key S protein-binding sites may provide crucial insights into the identification of mammals at high risk of SARS-CoV-2 infection.

Human ACE2 receptors bind to S proteins on SARS-CoV and on SARS-CoV-2 with overall structural similarity; this is likely constrained by the structure of ACE2 (Lan et al. 2020). However, the S protein-coding genes in SARS-CoV and SARS-CoV-2 share only about 75% homology (Gralinski, Menachery 2020; Zhou et al. 2020). This dissimilarity may contribute to a difference in viral infectivity and host range between the two SARS viruses. Indeed, based on two recent X-ray crystallography experiments (Lan et al. 2020; Shang et al. 2020), different key residues on human ACE2 receptor are structurally involved in binding to S proteins of SARS-CoV-2 and SARS-CoV. Table 5.1 summarizes all experimentally identified site-specific

interactions at the ACE2 receptor-S protein interface (Lan et al. 2020; Shang et al. 2020). Four key human ACE2 sites bind to S protein of SARS-CoV-2 but they do not interact with S protein of SARS-CoV: ACE2 Ser19 forms a hydrogen bond with S protein Ala475, ACE2 Asp30 forms a hydrogen bond with S protein Lys417, ACE2 Glu35 forms a hydrogen bond with S protein Gln493, and ACE2 Arg393 forms a hydrogen bond with S protein Tyr505. Furthermore, there are 11 other key human ACE2 sites (Gln24, Lys31, His34, Glu37, Asp38, Tyr41, Gln42, Leu79, Met82, Tyr83, and Lys353) that form either hydrogen bonds or salt bridges with S proteins of both SARS-CoV-2 and SARS-CoV, but they interact with different S protein residues. For example, ACE2 Gln24 forms a hydrogen bond with SARS-CoV-2 S protein at Leu472 but with SARS-CoV S protein at Asn473, and ACE2 Tyr41 forms a hydrogen bond with SARS-CoV-2 S protein at Thr500 and Asn501 but with SARS-CoV S protein at Thr486 and Thr487. The extended data in Lan et al. (2020) also highlights ACE2 sites Thr27, Phe28, Asn330, Gly354 and Asp355 to be in proximity with SARS-CoV-2 S protein. However, these residues were excluded from analysis because no chemical bond evidence was shown at these sites. In brief, we analyzed all key interacting residues (listed in Table 5.1) whose structures were shown in structural diagrams and whose chemical bond evidences were reported, by one or both X-ray crystallography experiments (Lan et al. 2020; Shang et al. 2020) with structure resolution cut-off < 3Å and interface distance cut-off < 5Å. Nonetheless, it is worth noting that future gene mutation studies should be performed to verify whether all 15 key ACE2 sites are essential for SARS-CoV-2 infection.

Table 5.1. Contacting sites at the interface of ACE2-S protein of SARS-CoV-2 (column 1, 2) and of SARS-CoV (column 3, 4). Data are retrieved from X-ray crystallography experiments by Shang et al. (2020) and Lan et al. (2020). Each row specifies the site-specific contacts formed between human ACE2 and S protein (e.g., first row: ACE2 S19 binds SARS-CoV-2 S protein A475). In

red are key ACE2 sites that distinctly interact with SARS-CoV-2 or SARS-CoV S protein sites. In blue are ACE2 sites that interact with both SARS-CoV-2 and SARS-CoV S proteins at different ACE2-binding sites. The structural and chemical bond evidence for all listed interactions were described in Shang et al. (2020) and Lan et al. (2020).

Human ACE2	SARS-CoV-2	Human ACE2	SARS-CoV
S19[1]	A475[1]	Q24[2]	N473[2]
Q24[2]	N487[2]	K31[1,2]	Y442[1,2]
Q24[2], M82[1,2], L79[1,2], Y83[1,2]	F486[1,2]	H34[2]	L443[2], N479[2]
D30[2]	K417[2]	E37[2]	Y491[1,2]
K31[1,2]	L455[1,2], Q493[2]	D38[1,2]	Y436[2], Y484[2]
H34[2], E35[1,2]	F456[2], Q493[1,2]	Y41[2]	Y484[2], T486[2], T487[2]
E37[2], R393[2]	Y505[2]	Q42[2]	Y436[2], Y484[2]
D38[1,2]	Y449[2], Q498[2]	L79[2], M82[1,2]	L472[1,2],
Y41[2]	Q498[2], T500[2], N501[2]	Y83[2]	Y475[2], N473[2]
Q42[2]	G446[2], Y449[2], Q498[2]	Q325[2], E329[2]	R426[2]
Y83[2]	Y489[2], N487[2]	N330[2]	T486[2]
K353[1,2]	N501[1], G502[2]	K353[1,2]	T487[1], G488[2]

[1] Interacting residues retrieved from Shang et al. (2020)

[2] Interacting residues retrieved from Lan et al. (2020)

5.3.2 Species-specific coronaviruses infecting high-risk mammals may express SARS-like S proteins with conserved key ACE2-binding sites

Next, we wish to know which mammalian-specific coronaviruses are similar to SARS-CoV-2 and to SARS-CoV in infectivity. We discussed above that a high-risk mammal that is potentially capable of carrying SARS viruses should express a human-like ACE2 receptor. It follows that because the mammal possesses a cellular environment expressing human-like ACE2 receptors, its host-specific coronavirus may be SARS-like at the spike protein due to adaptation. Just as mammals at high risk of SARS-CoV-2 infection should have human-like ACE2 receptors

at key S protein-binding sites**,** we expect SARS-like mammalian coronaviruses to express SARS-like S proteins at key ACE2-binding sites. Table 5.1 lists 15 key SARS-CoV-2 S protein sites and 13 key SARS-CoV S protein sites that were deemed as essential to binding with human ACE2 (Lan et al. 2020; Shang et al. 2020). Determining similarities at key ACE2-binding sites among S proteins may help identify candidate coronaviruses that are SARS-like in infectivity and shed light on the zoonosis of SARS-CoV-2.

To expand on the above point, many recent studies point to the bat *Betacoronavirus* RaTG13 as a close relative of SARS-CoV-2 (Andersen et al. 2020; Shang et al. 2020; Tang et al. 2020). SARS-CoV-2 S protein contains a unique Gly482, Val483, Glu484, and Gly485 four-residue motif in the binding ridge that may facilitate contact with human ACE2 N-terminal helix. Indeed, bat RaTG13 also contains a similar four-residue motif (Shang et al. 2020). Moreover, residues Leu455 and Asn501 on SARS-CoV-2 S protein are conserved by RaTG13 S protein (Shang et al. 2020); these sites contribute favorably to human ACE2-binding as mutations at these sites reduced binding potential. These observations may explain why RaTG13 could use human ACE2 as its receptor (Shang et al. 2020). Nonetheless, many residues in RaTG13 S protein are not fine-tuned for binding human ACE2 (Lan et al. 2020; Shang et al. 2020). Changing bat RaTG13 S protein at residues Lys486 and Tyr493 to SARS-CoV-2 S protein residues Phe486 and Gln493, respectively, enhanced human ACE2 recognition (Shang et al. 2020). Besides bat RaTG13, the pangolin *Betacoronavirus* Pangolin-CoV also shares high sequence similarity with SARS-CoV-2 (Andersen et al. 2020), and it also contains the 482-485 four-residue motif in its S protein (Shang et al. 2020). Furthermore, key ACE2-binding sites such as Leu455, Phe486, Gln493, and Asn501 are conserved between SARS-CoV-2 and Pangolin-CoV S proteins. Indeed, both bats and pangolins have been proposed to be potential intermediate hosts for SARS-CoV-2, and host-

specific coronaviruses of these two mammals bear a resemblance to SARS-CoV-2 at key S protein sites (Xiao et al. 2020).

Our investigation considered 132 *ACE2* genes from mammals across 19 orders and 17 mammalian-specific coronaviruses infecting high-risk species. Key *ACE2* sites are strongly conserved among Primates. Among other mammals, key binding sites on human *ACE2* are conserved by selected species of the Chiroptera, Artiodactyla, Rodentia, Carnivora, Perissodactyla, Pholidota, Lagomorpha, Proboscidea, and Sirenia orders. Among these high-risk mammals, 12 species are known to be infected by host-specific coronaviruses. We found that key S protein sites in SARS-CoV are most conserved by Civet-CoV, whereas key S protein sites in SARS-CoV-2 are most conserved by Pangolin-CoV. Both SARS viruses also share several key S protein sites with bat RaTG13 but not with other mammalian-specific coronaviruses. Together, our results reinforce the current hypothesis that the progenitor of both SARS-CoV-2 and SARS-CoV are likely of bat origin. The palm civet and pangolin may have served as distinct intermediate hosts that facilitated the adaptation of SARS viruses to bind human ACE2 receptors prior to zoonosis, because 1) both mammals express human-like ACE2 receptors at key S protein-binding sites and 2) their host-specific coronaviruses encode S proteins that share highest similarities with S proteins of SARS viruses at key ACE2-binding sites.

5.4 Materials and Methods

5.4.1 Retrieving and processing 132 mammalian *ACE2* genes and 17 coronavirus S protein-coding genes

The nucleotide sequences of 266 *ACE2* gene variants from 132 mammalian species were retrieved from the National Center for Biotechnology Information (NCBI) Nucleotide Database (https://www.ncbi.nlm.nih.gov/). The NCBI Nucleotide Database was queried for records containing "ACE2" as gene name and "Mammalia" as taxonomic class, excluding whole-genome and chromosome-wide results. Next, each record was searched for /product= "angiotensin-converting enzyme 2", and all others were removed. For each *ACE2* gene entry, only the coding DNA sequence region was extracted in FASTA format. The coding DNA sequences were translated from nucleotides into amino acids using DAMBE7 (Xia 2018) and verified with annotated amino acid sequences from NCBI GenBank files. Then, for sequence files, gene IDs were renamed as follow: ACE2_NCBI gene accession ID_Species name (Supplemental File S1). Similarly, the amino acid sequences of S proteins encoded by 17 host-specific coronaviruses infecting 12 mammalian species were retrieved from NCBI and extracted in FASTA format.

Multiple *ACE2* isoforms were retrieved for some species, but only one *ACE2* gene per species was selected for analysis. Many isoforms differ only in the 5' and 3' UTRs. For human, *ACE2* transcript variant 2 (NM_021804.3) contains 19 exons, whereas transcript variant 1 (NM_001371415.1) contains 18 exons, although both variants encode the same protein. Furthermore, some isoforms were experimentally validated, while others were predictions by automated computational approaches. We selected experimentally verified variants when available, but when only predicted variants are available, we picked one among those whose 1) amino acid

identities are conserved by most other variants, and 2) sequence is not truncated (selected *ACE2* variants are listed in Supplemental File S1). Similarly, one representative coronavirus was picked out of several available strains (e.g., strain Urbani was picked for SARS-CoV).

Next, amino acid sequences of *ACE2* genes and of S protein-coding genes were aligned with MAFFT (Katoh, Asimenos, Toh 2009) with the slow but accurate G-INS-i option. We then extracted the location and identity of amino acids that aligned to key human *ACE2* sites and to key S protein-coding gene sites in SARS-CoV-2 and SARS-CoV listed in Table 5.1. Mammalian *ACE2* match-mismatch heat-maps were then generated, and a total similarity score (the total number of matching amino acid identities between human and mammals at key *ACE2* sites) was calculated for each mammal. Similarly, coronavirus S protein match-mismatch heat-maps were generated (one against SARS-CoV-2 S protein and another against SARS-CoV S protein), and the total similarity score was calculated for each mammalian-specific coronavirus.

5.4.2 Phylogenetic reconstruction based on 132 *ACE2* genes and 16 S protein-coding genes

Three phylogenetic trees were constructed using MAFFT G-INS-i aligned amino acid sequences with the maximum-likelihood-based PHYML approach (Guindon et al. 2010): one tree for aligned *ACE2* genes from 132 mammalian species (bootstrap = 500, model = JTT + G + I + F), another tree for *ACE2* genes from 13 sample mammalian species (bootstrap =100, model = JTT + G + I + F), and a third tree for 16 mammalian-specific coronavirus S proteins (bootstrap = 100, model = WAG + G + I + F). All were constructed using the PHYML model implemented in DAMBE. The tree improvement option "-s" was set to "BEST" (best of NNI and SPR search). The "-o" option was set to "tlr" which optimizes the topology, branch lengths and rate parameters.

Tree figure illustrations were made using the Interactive Tree Of Life (iTOL) v4 (Letunic, Bork 2019).

5.5 Results

5.5.1 Key binding sites on the human ACE2 receptor are most conserved among Primates species and variably conserved in selected species belonging to eight other mammalian orders

We investigated 132 species belongs to 19 mammalian orders with available *ACE2* gene records. Among Primates, key sites on human *ACE2* are perfectly conserved by species belonging to the *Hominidae* and *Cercopithecoidae* families and highly conserved by other species (Supplemental Figure S5.1). Among the other 18 orders, key *ACE2* sites are highly conserved in selected species belonging to eight orders: Artiodactyla, Chiroptera, Carnivora, Rodentia, Lagomorpha, Perissodactyla, Proboscidea, and Sirenia (Supplemental Figure S5.1). More specifically, in the Artiodactyla order, species belonging to the *Bovidae*, *Monodontidae*, *Phocoenidae*, and *Physeteridae* families share very high similarities with human at key *ACE2* sites. In the Chiroptera order, species belonging to the *Pteropodidae* and *Rhinolophidae* share high similarities with human at key *ACE2* sites. In the Carnivora order, species belonging to the *Canidae*, *Felidae*, and *Ursidae* families share high similarities with human at key *ACE2* sites. In the Rodentia order, all species (e.g., belonging to *Cricetidae*, *Sciuridae*) share high similarities with human at key *ACE2* sites except species in the *Muridae* family. In the other four orders, the specific species that share high similarities with human at key *ACE2* sites are *Ochotona princeps* and *Oryctolagus cuniculus* (order Lagomorpha), *Ceratotherium simum* (order Perissodactyla), *Loxodonta Africana* (order Proboscidea), and *Trichechus manatus latirostris* (order Sirenia).

Similar to the SARS-CoV example displayed in Figure 5.1, while local *ACE2* site comparisons suggest that select species among nine orders could be at high risk to SARS-CoV-2 infection (Supplemental Figure S5.1), global similarities among whole *ACE2* gene alignments (Supplemental Figure S5.2) grouped 132 species by order and did not point out any species as high-risk.

Figure 5.2 showcases key *ACE2* site comparisons among a sample of mammals. These species were selected based on the following three criteria: 1) they all belong to the nine orders that contain high-risk mammals (Primates, Artiodactyla, Chiroptera, Carnivora, Rodentia, Lagomorpha, Perissodactyla, Proboscidea, and Sirenia), 2) all species that have been experimentally identified as susceptible to SARS-CoV-2 infection were selected (underlined), and 3) some species were selected because they are known to be infected by their own host-specific coronaviruses (in bold). In addition, the pangolin (*Manis javanica*) belonging to the Pholidota order shares medium similarity with human at key *ACE2* sites; the species was added to Figure 5.2 because it is a possible intermediate host of SARS-CoV-2 (Xiao et al. 2020).

Order	Family	Species	Host coronavirus	ACE2 binding hotspots similarities with hACE2															Total similarity
	Hominidae	***Homo sapiens***	SARS-CoV-2, SARS-CoV, MERS	S19	Q24	D30	K31	H34	E35	E37	D38	Y41	Q42	L79	M82	Y83	K353	R393	
Primates	*Cercopithecidae*	*Macaca mulatta*	n/a	S19	Q24	D30	K31	H34	E35	E37	D38	Y41	Q42	L79	M82	Y83	K353	R393	15/15
		Macaca fascicularis	n/a	S19	Q24	D30	K31	H34	E35	E37	D38	Y41	Q42	L79	M82	Y83	K353	R393	15/15
		Cercopithecus aethiops	n/a	S19	Q24	D30	K31	H34	E35	E37	D38	Y41	Q42	L79	M82	Y83	K353	R393	15/15
	Mustelidae	*Mustela putorius furo*	n/a	S19	L24	E30	K31	Y34	E35	E37	E38	Y41	Q42	H79	T82	Y83	K353	R393	9/15
	Canidae	***Canis lupus familiaris***	Canine CRCoV	S19	L23	E29	K30	Y33	E34	E36	E37	Y40	Q41	L78	T81	Y82	K352	R392	10/15
Carnivora	*Felidae*	***Felis catus***	Feline-CoV	S19	L24	E30	K31	H34	E35	E37	E38	Y41	Q42	L79	T82	Y83	K353	R393	11/15
		Panthera tigris	n/a	S11	L16	E22	K23	H26	E27	E29	E30	Y33	Q34	L71	T74	Y75	K345	R385	11/15
	Viverridae	***Paguma larvata***	Civet-SARS-CoV 007/2004	S19	L24	E30	T31	Y34	E35	Q37	E38	Y41	Q42	L79	T82	Y83	K353	R393	8/15
Rodentia	*Cricetidae*	*Mesocricetus auratus*	n/a	S19	Q24	D30	K31	Q34	E35	E37	D38	Y41	Q42	L79	N82	Y83	K353	R393	13/15
	Muridae	***Mus musculus***	MHV	S19	N24	N30	N31	Q34	E35	E37	D38	Y41	Q42	T79	S82	F83	H353	R393	7/15
	Bovidae	***Bos taurus***	Bovine-CoV	S19	Q24	E30	K31	H34	E35	E37	D38	Y41	Q42	M79	T82	Y83	K352	R392	12/15
Artiodactyla	*Camelidae*	***Camelus ferus***	Camel HKU23	S19	L24	E30	E31	H34	E35	E37	D38	Y41	Q42	T79	T82	Y83	K353	R393	10/15
	Suidae	***Sus scrofa***	Swine PHEV	S19	L24	E30	K31	L34	E35	E37	D38	Y41	Q42	I79	T82	Y83	K353	R393	10/15
		Rhinolophus ferrumequinum	Bat SARS-like-CoV ATO98145	S19	L24	D30	D31	S34	E35	E37	N38	H41	Q42	L79	N82	F83	K353	R393	8/15
Chiroptera	*Rhinolophidae*	***Rhinolophus sinicus***	BtRs-BetaCoV/YN2013	S19	R24	D30	E31	S34	E35	E37	N38	Y41	Q42	L79	N82	Y83	K353	R393	10/15
		Rhinolophus affinis	Bat RaTG13	S19	R24	D30	N31	H34	E35	E37	D38	Y41	Q42	L79	N82	Y83	K353	R393	12/15
Lagomorpha	*Leporidae*	***Oryctolagus cuniculus***	Rabbit HKU14	S19	L24	E30	K31	Q34	E35	E37	D38	Y41	Q42	L79	T82	Y83	K353	R393	11/15
Perissodactyla	*Equidae*	***Equus caballus***	Equine CoV	S19	L24	E30	K31	S34	E35	E37	E38	H41	Q42	L79	T82	Y83	K353	R393	9/15
	Rhinocerotidae	*Ceratotherium simum*	n/a	S19	L24	E30	K31	P34	E35	E37	D38	Y41	Q42	L79	T82	Y83	K353	R393	11/15
Sirenia	*Trichechidae*	*Trichechus manatus latirostris*	n/a	S19	L24	D30	T31	Q34	E35	E37	D38	Y41	Q42	L79	N82	F83	K348	R388	10/15
Proboscidea	*Elephantidae*	*Loxodonta africana*	n/a	S19	L24	D30	T31	Q34	E35	E37	D38	Y41	Q42	L79	D82	F83	K348	R388	10/15
Pholidota	*Manidae*	***Manis javanica***	Pangolin-CoV	S19	E24	E30	K31	S34	E35	E37	E38	Y41	Q42	I79	N82	Y83	K353	R393	9/15

Total similarity
5 or less
6 ot 7
8 to 9
10 to 11
12 to 13
14 to 15

Figure 5.2. Comparisons at key *ACE2* sites between human and a sample of mammals belong to ten orders. Species that have been experimentally identified as susceptible to SARS-CoV-2 infection are underlined. Species that can be infected by their own host-specific coronaviruses are in black bold, and other species are in grey. Highlighted blue and red are matching and mis-matching amino acid residues, respectively, between mammalian *ACE2* and human *ACE2* at key SARS-CoV-2 S protein binding sites. Underlined amino acids are conserved among species known to be infected by SARS-CoV-2. Total similarity designates the total number of matching amino acid residues with respect to human *ACE2*. The total similarity score for mammalian *ACE2* in blue highlights perfect site

similarities (14 to 15 matching sites), in green highlights high similarity (12 to 13 matching sites), in light green highlights medium-high similarity (10 to 11 matching sites), in yellow highlights medium similarity (8 to 9 matching sites), in orange highlights medium-low similarity (6 to 7 matching sites), and in red highlights low similarity (5 or less matching sites). A complete comparison of key *ACE2* sites involved in SARS-CoV-2 binding between human and 131 mammals is shown in Supplemental Figure S5.1.

All mammals currently known to be susceptible to SARS-CoV-2 infection share similarities with human at key *ACE2* sites (Figure 5.2). These include cat (*Felis catus*) (Halfmann et al. 2020), dog (*Canis lupus familiaris*) (Sit et al. 2020), and tiger (*Panthera tigris*) (Wang et al. 2020) where SARS-CoV-2 infections were detected, and ferret (*Mustela putorius furo*) (Kim et al. 2020; Shi et al. 2020), macaque (*Macaca mulatta* (Chandrashekar et al. 2020) and *Macaca fascicularis* (Rockx et al. 2020)), and grivet (*Cercopithecus aethiops*) (Woolsey et al. 2020) that were experimentally infected by SARS-CoV-2. One additional mammal that is known to be infected by SARS-CoV-2 is the mink (*Mustela lutreola*) (Oreshkova et al. 2020), a relative of the ferret. However, the mink was not included in Figure 5.2 because there were no *ACE2* gene records for this species in the NCBI gene database (last accessed April 15, 2020). Indeed, the *ACE2* genes in all these species share high similarity with human *ACE2* at key binding sites, except for ferret, which shares medium similarity. Wild-type mouse (*Mus musculus*) is not susceptible to SARS-CoV-2 infection (Bao et al. 2020), and expectedly, mouse *ACE2* shares poor similarity with human *ACE2* at key binding sites (Figure 5.2).

Among mammals that are known to be infected by SARS-CoV-2, nine key human *ACE2* sites (Ser19, Lys31, Glu35, Glu37, Tyr41, Gln42, Tyr83, Lys353 and Arg393) are conserved by their mammalian *ACE2* genes, but the other six key sites (Gln24, Asp30, His34, Asp38, Leu79, and Met82) are not well conserved (Figure 5.2). One may reason that only these nine conserved ACE2 sites could be important in S protein-binding. Nonetheless, mis-match amino acids at the six non-conserved sites all share similar physiochemical properties with the human *ACE2* residue (the replacements resulted in small Graham's distance D (Grantham 1974), which implies similar composition, polarity, and molecular volume between the two amino acids). Aspartic acid at sites 30 and 38 are replaced by Glutamic acid (Grantham's distance D = 45), Glutamine at site 24 is

replaced by Lysine (D = 53), Histidine at site 34 is replaced by Tyrosine (D = 83), and Methionine at site 82 is replaced by Threonine (D = 81). Nonetheless, future *ACE2* gene mutation studies are required to better elucidate which key *ACE2* sites among the 15 contacting residues are essential for SARS-CoV-2 infection.

5.5.2 Key ACE2-binding sites on SARS-CoV-2 and SARS-CoV S proteins are distinctly conserved by mammalian coronaviruses

Above results highlighted several mammals at high risk of SARS-CoV-2 infection. Figure 5.2 also highlights 12 species-specific coronaviruses that infect high-risk mammals (belonging to the Artiodactyla, Canivora, Chiroptera, Perissodactyla, Pholidota, and Lagomorpha orders) and one MHV coronavirus that infects the low-risk mice, but mammals of the Proboscidea and Sirenia orders do not have records of species-specific coronaviruses. Some of these coronaviruses may have SARS-like infectivity because they infect mammals having human-like ACE2 receptors. We thus investigated whether key S protein sites in SARS-CoV-2 and SARS-CoV are conserved by mammalian-specific coronaviruses.

Figure 5.3 shows distinct differences in the conservation of key SARS-CoV-2 and SARS-CoV S protein sites by mammalian-specific coronaviruses. Between the two SARS viruses, key S protein sites in SARS-CoV-2 are weakly conserved by SARS-CoV (Figure 5.3a) and vice versa (Figure 5.3b). In addition, key S protein sites in both SARS viruses share a medium degree of similarity with S protein sites in bat coronavirus RaTG13 isolated from *Rhinolophus affinis*, but they share poor similarities with S protein sites in two other bat coronaviruses, isolated from *Rhinolopus sinicus* and *Rhinolopus ferrumequinum*. Importantly, key S protein sites in SARS-CoV-2 share highest similarity with S protein sites of Pangolin-CoV strain Guangdong (GD) that

was sequenced with high coverage but share lower similarity with S protein sites in Pangolin-CoV strain Guangxi (GX) that was flagged as poorly sequenced by GISAID (Figure 5.3a). Indeed, the unique 482-485 domain in SARS-CoV-2 (highlighted yellow, Figure 5.3a) is perfectly conserved by Pangolin-CoV GD, partially conserved by bat RaTG13 and Pangolin-CoV GX, and not conserved by any other coronaviruses surveyed. In contrast, key S protein sites in SARS-CoV share highest similarity with S protein sites in Civet-CoV but share medium similarity with S protein sites in Pangolin-CoVs (Figure 5.3b). As for other mammalian-specific coronaviruses, including the human MERS-CoV with a presumed camel origin, their S proteins share little similarities with SARS-CoV-2 and SARS-CoV at key sites. Together, these findings imply that the bat coronavirus RaTG13 is SARS-like in infectivity, but SARS-CoV-2 more closely resembles Pangolin-CoV and SARS-CoV more closely resembles Civet-CoV.

Global similarities among whole S protein-coding amino acid sequences (Figure 5.4) also showed that SARS viruses are distinctly closely related to mammalian coronaviruses: SARS-CoV-2 closely relates to Pangolin-CoV and bat RaTG13 (Figure 5.4a), and SARS-CoV closely relates to Civet-CoV (Figure 5.4b). However, while SARS-CoV-2 is more similar to bat RaTG13 than to Pangolin-CoV in terms of global S protein similarity (Figure 5.4a), key S protein sites are more conserved between SARS-CoV-2 and Pangolin-CoV (Figure 5.3a). The notion that comparing local but not global S protein similarities is crucial in determining SARS-like coronaviruses was therefore consistent with the notion that comparing local but not global *ACE2* similarities is crucial in determining mammals at high risk of infection by SARS viruses (Figure 5.1, 5.2).

a)

Genus	S protein accession	Infected species	Aligned Spike protein key sites																				Total Similarity
			K417	G446	Y449	L455	F456	A475	G482	V483	E484	G485	F486	N487	Y489	Q493	Q498	T500	N501	G502	Y505		
Betacoronavirus	YP_009724390 (SARS-Cov-2 Wuhan-Hu-1)	Human (*Homo sapiens*)																				
Betacoronavirus	EPI_ISL_410721 (Pangolin-CoV GD)	Pangolin (*Manis javanica*) Guangdong	R413	G442	Y445	L451	F452	A471	G478	V479	E480	G481	F482	N483	Y485	Q489	H494	T496	N497	G498	Y501	13/15	
Betacoronavirus	QIQ54048.1 (PangolinCoV GX_P2V)	Pangolin (*Manis javanica*) Guangxi	V417	G446	Y449	L455	F456	A475	G482	Q483	V484	G485	L486	N487	Y489	E493	H498	T500	T501	G502	Y505	10/15	
Betacoronavirus	QHR63300.2 (RaTG13)	Bat (*Rhinolophus affinis*)	K417	G446	F449	L455	F456	A475	G482	Q483	T484	G485	L486	N487	Y489	Y493	Y498	T500	D501	G502	H505	9/15	
Betacoronavirus	AAP13441.1 (SARS Urbani)	Human (*Homo sapiens*)	V404	T433	Y436	Y442	L443	P462	P469	-	P470	A471	L472	N473	Y475	N479	Y484	T486	T487	G488	Y491	6/15	
Betacoronavirus	AAU04646.1 (Civet SARS CoV 007)	Palm civet (*Paguma larvata*)	V404	T433	Y436	Y442	L443	S462	P469	-	P470	A471	P472	N473	Y475	R479	Y484	T486	S487	G488	Y491	6/15	
Betacoronavirus	QBM11748.1 (MERS)	Human (*Homo sapiens*)	P463	T492	-	P494	L495	L517	S528	D539	Y540	Y541	R542	K543	L545	E549	L554	A556	S557	G558	V561	6/15	
Betacoronavirus	ATO98145.1 (Bat SARS-like-CoV)	Bat (*Rhinolophus ferrumequinum*)	V401	-	-	S434	H435	-	-	-	-	-		N452	V454	S458	N463	N465	V466	P467	Y470	2 /15	
Betacoronavirus	AIA62330.1 (BtRs-BetaCoV/YN2013)	Bat (*Rhinolophus sinicus*)	V400	-	-	S433	H434	-	-	-	-	-		N451	V453	S457	N462	N464	V465	P466	Y469	2 /15	
Betacoronavirus	ALA50080.1 (HKU23)	Camel (*Camelus dromedarius*)	Y412	S441	W448	F454	T455	F480	L491	T519	Y520	Y521	L522	T523	Y525	T540	Q583	Q585	A586	F587	W590	2 /15	
Betacoronavirus	AVZ61129.1 (Bovine CoV)	Cattle (*Bos taurus*)	Y412	S441	W448	F454	T455	F480	L491	T519	N520	Y521	L522	T523	Y525	T540	Q580	Q582	A583	F584	W587	2 /15	
Betacoronavirus	YP_005454245.1 (HKU14)	Rabbit (*Oryctolagus cuniculus*)	Y411	S440	W447	F453	T454	F479	L490	T518	N519	Y520	L521	T522	Y524	L539	Q579	Q581	A582	F583	W586	2/15	
Betacoronavirus	BAJ52885.1 (Equine CoV)	Horse (*Equus caballus*)	Y412	T441	W448	F454	N455	F479	N490	T518	S519	Y520	R521	T522	F524	N539	P580	N582	A583	F584	W587	0/15	
Betacoronavirus	ACN89763.1 (murine CoV MHV-JHM.IA)	Rat (*Mus musculus*)	F410	N439	W446	F452	N453	F472	Q483	T505	I506	H507	R508	E509	S511	D540	A581	D583	S584	F585	W588	0/15	
Betacoronavirus	AVV64341.1 (PHEV)	Swine (*Sus scrofa*)	Y412	T441	W448	F454	N455	F474	T485	T501	T502	V503	R504	K505	F507	T526	K566	Q568	A569	F570	W573	0/15	
Alphacoronavirus	AFW97360.1 (CRCoV)	Canine (*Lupus familiaris*)	Y412	S441	W448	F454	T455	F480	L491	T519	N520	Y521	L522	T523	Y525	L540	Q580	K582	A583	F584	W587	2 /15	
Alphacoronavirus	AFH58021.1 (Feline CoV)	Feline (*Felis catus*)	A458	S495	N502	M540	K541	I570	H580	I591	F592	N593	Q594	E595	T597	D601	K639	D641	V642	A643	T646	0/15	

Total similarity

5 or less
6 to 7
8 to 9
10 to 11
12 to 13

b)

Genus	S protein accession	Infected species	Aligned Spike protein key sites													Total Similarity
			R426	Y436	Y442	L443	L472	N473	Y475	N479	Y484	T486	T487	G488	Y491	
Betacoronavirus	AAP13441.1 (SARS Urbani)	Human (*Homo sapiens*)														
Betacoronavirus	AAU04646.1 (Civet SARS CoV 007)	Palm civet (*Paguma larvata*)	R426	Y436	Y442	L443	P472	N473	Y475	R479	Y484	T486	S487	G488	Y491	10/13
Betacoronavirus	QIQ54048.1 (PangolinCoV GX_P2V)	Pangolin (*Manis javanica*) Guangxi	V439	Y449	L455	F456	L486	N487	Y489	E493	H498	T500	T501	G502	Y505	8/13
Betacoronavirus	YP_009724390 (SARS-Cov-2 Wuhan-Hu-1)	Human (*Homo sapiens*)	N439	Y449	L455	F456	F486	N487	Y489	Q493	Q498	T500	N501	G502	Y505	6/13
Betacoronavirus	EPI_ISL_410721 (Pangolin-CoV GD)	Pangolin (*Manis javanica*) Guangdong	N435	Y445	L451	F452	F482	N483	Y485	Q489	H494	T496	N497	G498	Y501	6/13
Betacoronavirus	QHR63300.2 (RaTG13)	Bat (*Rhinolophus affinis*)	K439	F449	L455	F456	L486	N487	Y489	Y493	Y498	T500	D501	G502	H505	6/13
Betacoronavirus	AIA62330.1 (BtRs-BetaCoV/YN2013)	Bat (*Rhinolophus sinicus*)	A422	-	S433	H434		N451	V453	S457	N462	N464	V465	P466	Y469	2/13
Betacoronavirus	ATO98145.1 (Bat SARS-like-CoV)	Bat (*Rhinolophus ferrumequinum*)	A423	-	S434	H435		N452	V454	S458	N463	N465	V466	P467	Y470	2/13
Betacoronavirus	AVZ61129.1 (Bovine CoV)	Cattle (*Bos taurus*)	P434	W448	F454	T455	L522	T523	Y525	T540	Q580	Q582	A583	F584	W587	2/13
Betacoronavirus	QBM11748.1 (MERS)	Human (*Homo sapiens*)	P485	-	P494	L495	R542	K543	L545	E549	L554	A556	S557	G558	V561	2/13
Betacoronavirus	ALA50080.1 (HKU23)	Camel (*Camelus dromedarius*)	P434	W448	F454	T455	L522	T523	Y525	T540	Q583	Q585	A586	F587	W590	2/13
Betacoronavirus	YP_005454245.1 (HKU14)	Rabbit (*Oryctolagus cuniculus*)	P433	W447	F453	T454	L521	T522	Y524	L539	Q579	Q581	A582	F583	W586	2/13
Betacoronavirus	BAJ52885.1 (Equine CoV)	Horse (*Equus caballus*)	G434	W448	F454	N455	R521	T522	F524	N539	P580	N582	A583	F584	W587	1/13
Betacoronavirus	AVV64341.1 (PHEV)	Swine (*Sus scrofa*)	P434	W448	F454	N455	R504	K505	F507	T526	K566	Q568	A569	F570	W573	0/13
Betacoronavirus	ACN89763.1 (murine CoV MHV-JHM.IA)	Rat (*Mus musculus*)	P432	W446	F452	N453	R508	E509	S511	D540	A581	D583	S584	F585	W588	0/13
Alphacoronavirus	AFW97360.1 (CRCoV)	Canine (*Lupus familiaris*)	P434	W448	F454	T455	L522	T523	Y525	L540	Q580	K582	A583	F584	W587	2/13
Alphacoronavirus	AFH58021.1 (Feline CoV)	Feline (*Felis catus*)	H488	N502	M540	K541	Q594	E595	T597	D601	K639	D641	V642	A643	T646	0/13

Total similarity

3 or less
4 to 5
5 to 7
8 to 9
10 to 11

Figure 5.3. Amino acid comparisons at key S protein sites among 17 mammalian-specific coronaviruses. a) displays the conservation of key ACE2-binding sites in SARS-CoV-2 S protein by 16 mammalian-specific S proteins. b) displays the conservation of key ACE2-

binding sites in SARS-CoV S protein by 16 mammalian-specific S proteins. Highlighted blue and red are matching and mis-matching amino acid residues, respectively, between mammalian coronaviruses and the two human SARS coronaviruses at key S protein sites. Highlighted yellow is the 482-485 GVEG motif found in SARS-CoV-2. Total similarity designates the total number of matching amino acid residues with respect to SARS-CoV-2 or SARS-CoV, and scores in green highlights high similarity, in light green highlights medium-high similarity, in yellow highlights medium similarity, in orange highlights medium-low similarity, and in red highlights low similarity.

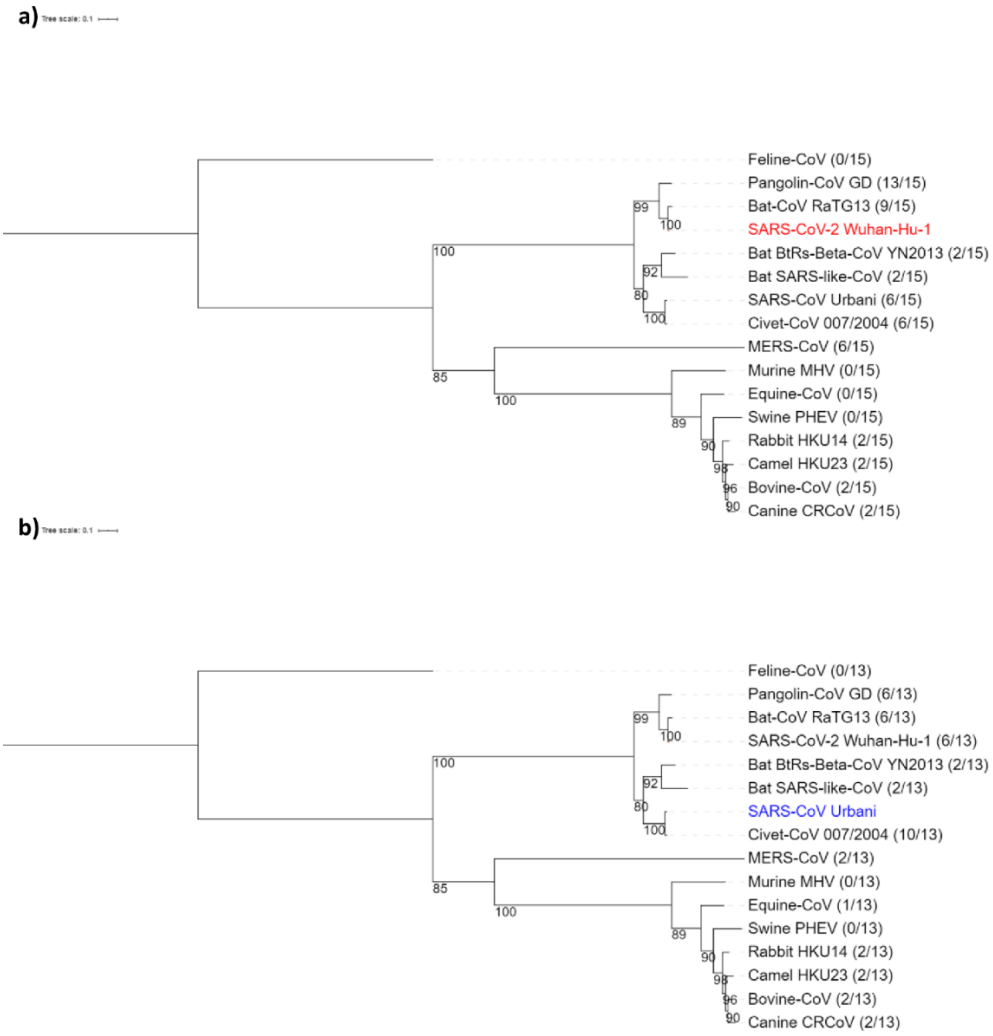


Figure 5.4. Phylogenetic reconstruction of 16 MAFFT G-INS-i aligned coronavirus S protein-coding genes. Infected mammalian species belong to the Artiodactyla, Canivora, Chiroptera, Lagomorpha, Perissodactyla, Primates, Pholidota and Rodentia orders. The phylogenetic tree is constructed using the maximum-likelihood-based PHYML approach, with best model by AIC criteria = WAG + G + I + F, and Bootstrap = 100. Pangolin-CoV GX strain was excluded in favor of Pangolin-CoV GD strain having better sequence coverage. The ratio appended to each coronavirus indicates the total number of amino acid matches when coronavirus S protein gene was compared to a) SARS-CoV-2 S protein gene (in red) at its 15 key ACE2-binding sites and to b) SARS-CoV S protein gene (in blue) at its 13 key ACE-2 binding sites.

5.6 Discussion

SARS-CoV-2 can transmit efficiently in humans, but it is less clear which other mammals are at risk of being infected. We performed comparative gene analyses to trace differences at the *ACE2* gene of 132 mammalian species belonging to 19 orders. Mammalian *ACE2* genes were compared in two ways, one was global similarities reflected by the phylogenetic relationship from whole *ACE2* sequence alignments, and the other was local comparisons at key human *ACE2* sites. While global similarities (Supplemental Figure S5.2) were not good predictors of mammals at high-risk of being infected by SARS-CoV-2, local similarities highlighted several high-risk mammals belonging to nine out of 19 orders surveyed (Figure 5.2, Supplemental Figure S5.1: Primates, Artiodactyla, Canivora, Chiroptera, Lagomorpha, Perissodactyla, Proboscidea, Rodentia and Sirenia).

Species currently known to be susceptible to SARS-CoV-2 infection are indeed high-risk mammals that share high similarities with human at key *ACE2* sites (Figure 5.2). For example, while golden Syrian hamster could be consistently infected by SARS-CoV-2 (Chan et al. 2020) and its *ACE2* gene shares high local similarity with human *ACE2* gene, wild-type mouse could not be infected by SARS-CoV-2 (Bao et al. 2020) and its *ACE2* gene expectedly shares poor local similarity with human *ACE2* gene. However, these differences between the two Rodentia species could not be distinguished from global *ACE2* sequence similarities (Supplemental Figure S5.2). Among other susceptible mammals (Figure 5.2), confirmed cases of SARS-CoV-2 infection have been reported for domesticated cats and dogs across several U.S. states (https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/sa_one_health/sars-cov-2-animals-us, last accessed August 25, 2020). These findings may prompt future investigations to perform

SARS-CoV-2 screening in other high-risk mammals, especially other domesticated animals living in proximity with humans such as pigs and cattle. Indeed, the pig was previously predicted to be susceptible to SARS-CoV-2 infection based on computational models of ACE2 structures (Wan et al. 2020), but SARS-CoV-2 infection in pig has yet to be detected (Schlottau et al. 2020; Shi et al. 2020).

We also analyzed which host-specific coronaviruses infecting high-risk mammals could be SARS-like in infectivity. To this end, we measured global similarities at whole S protein-coding genes and local similarities at key S protein sites among 17 coronaviruses (infecting human, 12 high-risk mammals, and the low-risk mouse). We showed that key S protein sites in the two SARS viruses are modestly conserved by one bat coronavirus RaTG13. More importantly, key S protein sites in SARS-CoV and SARS-CoV-2 are distinctly most conserved by Civet-CoV and Pangolin-CoV GD, respectively (Figure 5.3). Indeed, a recent study (Andersen et al. 2020) had also found that SARS-CoV-2 is more similar to Pangolin-CoV GD than to bat RaTG13 or to SARS-CoV at the S1 binding domain of the S protein. However, based on global similarities, the S protein of SARS-CoV-2 is more closely related to S protein of bat RaTG13 than to S protein of Pangolin-CoV (Figure 5.4), similar to what others have shown (Lan et al. 2020; Zhang, Wu, Zhang 2020). Hence, similarities in coronavirus infectivity can be better determined by local than global S protein comparisons. Nevertheless, both Figures 5.3 and 5.4 suggest that the two SARS viruses are of bat origin but have since distinctly evolved.

Our results corroborate the current hypothesis on the origin and evolution of SARS viruses. The progenitors of SARS-CoV-2 and SARS-CoV are likely to have a bat coronavirus origin (Cui, Li, Shi 2019; Perlman 2020), and the bat serves as a potential reservoir for these viruses. However,

transmissions of SARS viruses from bats to humans are unlikely, and the viruses may had undergone adaptation in distinct intermediate hosts, namely the palm civet for SARS-CoV (Guan et al. 2003; Li et al. 2005b) and the pangolin for SARS-CoV-2 (Lam et al. 2020; Xiao et al. 2020; Zhang, Wu, Zhang 2020). Indeed, the *ACE2* genes in both palm civet and pangolin share similarities with human *ACE2* gene at key S protein-binding sites. Additionally, key S protein sites are most conserved between Civet-CoV and SARS-CoV, and between Pangolin-CoV and SARS-CoV-2. It is plausible that, prior to zoonotic transmission, evolution of progenitor SARS virus in distinct intermediate hosts allowed it to adapt high binding potential between viral S protein and human-like ACE2 receptor and led to differences between SARS-CoV-2 and SARS-CoV at the S protein.

5.7 Data availability

Supplemental Figures for Chapter 5 are available in Appendix E. Supplemental files are available at https://www.nature.com/articles/s41598-020-80573-x: Supplemental file S1 contains data for mammalian *ACE2* and coronavirus S protein gene accessions, the selected species used for phylogenetic reconstruction, and key site comparisons between 132 mammalian species at aligned *ACE2* genes and between 17 coronaviruses at aligned S protein-coding genes.

5.8 Acknowledgements

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to X.X. [RGPIN/2018-03878], and NSERC Doctoral Scholarship to Y.W. [CGSD/2019-535291].

5.9 References

Andersen, KG, A Rambaut, WI Lipkin, EC Holmes, RF Garry. 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26:450-452.
Bao, L, W Deng, B Huang, et al. 2020. The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. Nature 583:830-833.
Chan, JF-W, AJ Zhang, S Yuan, et al. 2020. Simulation of the clinical and pathological manifestations of Coronavirus Disease 2019 (COVID-19) in golden Syrian hamster model: implications for disease pathogenesis and transmissibility. Clin. Infect. Dis.
Chandrashekar, A, J Liu, AJ Martinot, et al. 2020. SARS-CoV-2 infection protects against rechallenge in rhesus macaques. Science 369:812-817.
Cui, J, F Li, ZL Shi. 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17:181-192.
Gralinski, LE, VD Menachery. 2020. Return of the Coronavirus: 2019-nCoV. Viruses 12:135.
Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.
Guan, Y, BJ Zheng, YQ He, et al. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302:276-278.
Guindon, S, JF Dufayard, V Lefort, M Anisimova, W Hordijk, O Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307-321.
Halfmann, PJ, M Hatta, S Chiba, et al. 2020. Transmission of SARS-CoV-2 in Domestic Cats. N. Engl. J. Med. 383:592-594.
Hamming, I, W Timens, ML Bulthuis, AT Lely, G Navis, H van Goor. 2004. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. J. Pathol. 203:631-637.
Katoh, K, G Asimenos, H Toh. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. 537:39-64.
Kim, Y-I, S-G Kim, S-M Kim, et al. 2020. Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. Cell Host Microbe: Elsevier Inc. p. 704-709.e702.
Lam, TT-Y, MH-H Shum, H-C Zhu, et al. 2020. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. Nature 583:282-285.
Lan, J, J Ge, J Yu, et al. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581:215-220.
Letunic, I, P Bork. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. p. W256-W259.
Li, F, W Li, M Farzan, SC Harrison. 2005a. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309:1864-1868.
Li, W, C Zhang, J Sui, et al. 2005b. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. Embo. J. 24:1634-1643.
Liu, P, JZ Jiang, XF Wan, et al. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog. 16.
Marra, MA, SJ Jones, CR Astell, et al. 2003. The Genome sequence of the SARS-associated coronavirus. Science 300:1399-1404.
Nie, Y, P Wang, X Shi, et al. 2004. Highly infectious SARS-CoV pseudotyped virus reveals the cell tropism and its correlation with receptor expression. Biochem. Biophys. Res. Commun. 321:994-1000.

Oreshkova, N, RJ Molenaar, S Vreman, et al. 2020. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. Euro. Surveill. 25:1560-7917.
Perlman, S. 2020. Another Decade, Another Coronavirus. N. Engl. J. Med. 382:760-762.
Rockx, B, T Kuiken, S Herfst, et al. 2020. Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. Science 368:1012-1015.
Rota, PA, MS Oberste, SS Monroe, et al. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300:1394-1399.
Schlottau, K, M Rissmann, A Graaf, et al. 2020. SARS-CoV-2 in fruit bats, ferrets, pigs, and chickens: an experimental transmission study. The Lancet Microbe.
Shang, J, G Ye, K Shi, Y Wan, C Luo, H Aihara, Q Geng, A Auerbach, F Li. 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 30:020-2179.
Shi, J, Z Wen, G Zhong, et al. 2020. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. Science 368:1016-1020.
Sit, THC, CJ Brackman, SM Ip, et al. 2020. Infection of dogs with SARS-CoV-2. Nature.
Song, HD, CC Tu, GW Zhang, et al. 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. USA 102:2430-2435.
Tang, X, C Wu, X Li, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. Nat. Sci. Rev. p. 1012-1023.
Tian, X, C Li, A Huang, et al. 2020. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. Emerg. Microbes. Infect. p. 382-385.
Walls, AC, YJ Park, MA Tortorici, A Wall, AT McGuire, D Veesler. 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181:281-292.
Wan, Y, J Shang, R Graham, RS Baric, F Li. 2020. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. In: T Gallagher, editor. J. Virol. p. e00127-00120.
Wang, L, PK Mitchell, PP Calle, et al. 2020. Complete Genome Sequence of SARS-CoV-2 in a Tiger from a U.S. Zoological Collection. In: S Roux, editor. Microbiol. Resour. Announc. p. e00468-00420.
Woolsey, C, V Borisevich, AN Prasad, et al. 2020. Establishment of an African green monkey model for COVID-19. bioRxiv : the preprint server for biology: Cold Spring Harbor Laboratory.
Xia, X. 2018. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol. Biol. Evol. 35:1550-1552.
Xiao, K, J Zhai, Y Feng, et al. 2020. Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. bioRxiv. p. 2020.2002.2017.951335.
Zhang, T, Q Wu, Z Zhang. 2020. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr. Biol. 30:1578.
Zhou, P, X-L Yang, X-G Wang, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273.

CHAPTER SIX

Coronavirus genomes carry the signatures of their habitats

This chapter is a modified version of the publication below, which won OISB PhD Paper of the Year 2020. Section 6.3 Introduction is partly removed, reorganized, and rewritten into Chapter 1.2.2:

Wei, Y., Silke, J.R., Aris, P., Xia, X. 2020. Coronavirus genomes carry the signatures of their habitats. Plos One 15(12):e0244025.

6.1 Author Contributions

Yulong Wei[1], Jordan R. Silke[1], Parisa Aris[1], 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.

Y.W. and X.X. designed the study. Y.W wrote the manuscript, J.R.S contributed substantially to manuscript editing, and X.X. provided guidance and editing. Y.W., J.R.S, P.A. collected and analyzed the data. X.X. developed the computational programs. Y.W. made Figures 6.1, 6.2, 6.4, 6.5, 6.6, 6.7, 6.8, S6.4, S6.6, S6.7, S6.8, S6.9, S6.10, S6.11, Tables 6.2, 6.3, 6.4, S6.1, and all Supplemental Files. J.R.S. made Figure 6.3, S6.3, S6.5, Table 6.1, and LoM datasets in Supplemental File S1. P.A. made Figures S6.1 and S6.2. All authors reviewed the manuscript. X.X. supervised the study. Y.W. and J.R.S contributed equally to this work. This work is published (PMID: 33351847)

6.2 Abstract

Coronaviruses such as SARS-CoV-2 regularly infect host tissues that express antiviral proteins (AVPs) in abundance. Understanding how they evolve to adapt or evade host immune responses is important in the effort to control the spread of infection. Two AVPs that may shape viral genomes are the zinc finger antiviral protein (ZAP) and the apolipoprotein B mRNA editing enzyme-catalytic polypeptide-like 3 (APOBEC3). The former binds to CpG dinucleotides to facilitate the degradation of viral transcripts while the latter frequently deaminates C into U residues which could generate notable viral sequence variations. We tested the hypothesis that both APOBEC3 and ZAP impose selective pressures that shape the genome of an infecting coronavirus. Our investigation considered a comprehensive number of publicly available genomes for seven coronaviruses (SARS-CoV-2, SARS-CoV, and MERS infecting *Homo sapiens*, Bovine CoV infecting *Bos taurus*, MHV infecting *Mus musculus*, HEV infecting *Sus scrofa*, and CRCoV infecting *Canis lupus familiaris*). We show that coronaviruses that regularly infect tissues with abundant AVPs have CpG-deficient and U-rich genomes, whereas those that do not infect tissues with abundant AVPs do not share these sequence hallmarks. Among the coronaviruses surveyed herein, CpG is most deficient in SARS-CoV-2 and a temporal analysis showed a marked increase in C to U mutations over four months of SARS-CoV-2 genome evolution. Furthermore, the preferred motifs in which these C to U mutations occurred are the same as those subjected to APOBEC3 editing in HIV-1. These results suggest that both ZAP and APOBEC3 shape the SARS-CoV-2 genome: ZAP imposes a strong CpG avoidance, and APOBEC3 constantly edits C to U. Evolutionary pressures exerted by host immune systems onto viral genomes may motivate novel strategies for SARS-CoV-2 vaccine development.

6.3 Introduction

The COVID-19 pandemic is a serious global health emergency. Understanding how coronaviruses adapt or evade tissue-specific host immune responses is important in the effort to control the spread of infection and to facilitate vaccine-development strategies. As obligate parasites, coronaviruses evolve in mammalian hosts and carry genomic signatures shaped by their host-specific environments. For instance, SARS-CoV-2 regularly infects bronchiolar and type II alveolar epithelial cells in the lungs (Yao et al.) and enterocytes in the small intestines (Lamers et al. 2020). At the tissue level, hosts provide different cellular environments with varying levels of antiviral activity. Two antiviral proteins (AVPs) that may contribute to the modification of viral genomes are the zinc finger antiviral protein (ZAP, gene name ZC3HAV1 in mammals) and the apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3 (APOBEC3) protein, both of which exhibit tissue-specific expressions (Fagerberg et al. 2014).

6.3.1 ZAP is endogenously expressed in tissues and relies on motif-specific transcript targeting to mediate antiviral activity

As described in Chapter 1.2.2, ZAP is a key component in the mammalian interferon-mediated immune response that specifically targets CpG dinucleotides in viral RNA genomes (Meagher et al. 2019) to inhibit viral replication and signal for viral genome degradation (Guo et al. 2007; Takata et al. 2017; Meagher et al. 2019; Ficarelli et al. 2020). Consequently, retroviruses such as HIV-1 (Zhu et al. 2011; Ficarelli et al. 2020) and single-stranded RNA viruses such as Ecovirus 7 (Odon et al. 2019), Zika virus (Trus et al. 2020), and Influenza virus (Greenbaum et al. 2008) all exhibit strong CpG deficiency. In addition, recent studies have shown strong CpG deficiency in the SARS-CoV-2 genome in comparison to SARS-CoV and MERS-CoV genomes,

and they suggested that SARS-CoV-2 could be adapted to evade the ZAP antiviral defense (Nchioua et al. 2020; Xia 2020). Indeed, endogenous ZAP activity has been shown to restrict SARS-CoV-2 replication in human lung cell lines expressing ZAP in abundance (Nchioua et al. 2020).

Furthermore, ZAP may prefer to target CG dinucleotides in specific CG-rich ssRNA contexts. Based on an *in vitro* crystal structure study that examined the binding affinity between the mouse ZAP zinc-finger motif and a variety of CG-rich ssRNA sequences (Luo et al. 2020), the mouse ZAP-preferred motifs were determined to be CN_XGNCG, where N_X is a spacer sequence of length 4 nt to 8 nt. However, to date, a human ZAP-preferred consensus motif remains to be determined (Meagher et al. 2019).

6.3.2 APOBEC3 is variably expressed in tissue-specific cell lines, and its catalytic activity is induced in response to cell stress

Aside from ZAP, the APOBEC3 cytidine deaminase enzymes have also garnered substantial attention for their role in the antiviral immune response (Cullen 2006; Harris, Dudley 2015). As discussed in Chapter 1.2.2, APOBEC3 enzymes are mainly expressed in hematopoietic cell populations but are not confined to lymphoid organs. Among tested APOBEC3 RNAs, A3A, A3B, A3F, and A3G, but not A3H, were detected in the human mammary epithelial cells (Okeoma et al. 2010), and transcripts of both A3G and A3F were detected in human lung epithelial cells (Pauli et al. 2009). In particular, A3G transcripts were upregulated in response to viral infection in human lung epithelial cells. Presence of APOBEC3 enzymes in non-lymphoid tissues is largely attributed to the presence of tissue-residing lymphocytes and their expression levels have been

observed to vary at the tissue level (Koning et al. 2009; Refsland et al. 2010). Although, it remains unclear whether APOBEC3 expression in tissue cells plays a role in restricting viral infections,

APOBEC3 enzymes have negligible catalytic activity prior to infection, but expressions of associated mRNAs are induced in response to cell stresses including hypoxia, cell crowding, and presence of interferon-α (Koning et al. 2009; Sharma et al. 2015; Sharma et al. 2019). Through a mechanism largely derived from HIV-1 studies, APOBEC3 enzymes have been prominently reported to restrict viral infectivity (Chiu, Greene 2008; Nabel et al. 2013; Rodriguez-Frias et al. 2013; Harris, Dudley 2015; Hayward et al. 2018) by editing C to T at the viral genomes, and HIV-1 avoids this deleterious effect by expressing Vif to target and degrade APOBEC3 enzymes (Sheehy et al. 2002; Wang, Wang 2009).

Additionally, C to U RNA-editing has been demonstrated in lymphocytes, macrophages, monocytes, and natural killer cells by both A3A and A3G in response to hypoxia and interferons (Sharma et al. 2015; Sharma et al. 2016; Sharma et al. 2019), and recent studies propose that APOBEC3 enzymes may act directly to edit single-stranded RNA coronaviruses (Di Giorgio et al. 2020; Simmonds 2020; Victorovich et al. 2020). This notion is supported by lines of evidence showing that A3C, A3F, and A3H may inhibit HCoV-NL63 coronavirus infection in humans (Milewska et al. 2018). Indeed, many C to U mutations have been detected in SARS-CoV-2 genomes (Di Giorgio et al. 2020; Jiang 2020; Simmonds 2020; Victorovich et al. 2020), and if RNA-editing by APOBEC3 is involved, then this immune response could potentially restrict SARS-CoV-2 because coronaviruses do not encode a Vif analogue to degrade APOBEC3 enzymes.

6.3.3 APOBEC3 editing is preferred at specific motif contexts and is influenced by substrates' higher-order structure

APOBEC3 enzymes prefer to edit C in specific motif contexts. For instance, in HIV-1 (Hultquist et al. 2011; Ebrahimi, Alinejad-Rokny, Davenport 2014), MLV (Harris et al. 2003; Bishop et al. 2004; Langlois et al. 2005), and SIV (Dang et al. 2006), A3G tends to deaminate C mostly in the context of 5' C<u>C</u> (underlined is site subjected to C to T editing), whereas all other APOBEC3 paralogues deaminate C in the context of 5' T<u>C</u> (Hultquist et al. 2011; Feng et al. 2014; Chen, MacCarthy 2017; Martinez et al. 2019). In HIV, these edits cause 5' GG to 5' AG and 5' GA to 5' AA hypermutations on the positive DNA strand (Hultquist et al. 2011; Ebrahimi, Alinejad-Rokny, Davenport 2014) to potentially disrupt protein function (Sadler et al. 2010; Gillick et al. 2013). However, not all 5' T<u>C</u> and 5' C<u>C</u> are deaminated with equal efficiency because the identities of the -2 and +1 nucleotides flanking the 5'N<u>C</u> are important to APOBEC3 target selection (Bishop et al. 2004; Liddament et al. 2004; Desimmie et al. 2016; Wan, Nagata, Katahira 2018; McDaniel et al. 2020).

In addition to motif preference, the structural configuration of the substrates bound to the APOBEC3 zinc center may also influence APOBEC3 editing activity (Shi et al. 2017; Silvas et al. 2018). Adding to this complexity, studies on APOBEC3 editing have reached dissimilar conclusions as to the optimal secondary structure of the 5' T<u>C</u> target. For instance, a large number of A3A and A3G substrates were predicted to form a loop structure in innate immune cells and HEK293T cells (Sharma et al. 2015; Sharma et al. 2016). In a further mutagenesis study on three A3A substrates, *SDHB*, *APP*, and *TMEM109*, and on one A3G substrate *PRPSAP2*, Sharma, Baysal (2017) found that both A3A and A3G enzymes highly preferred to edit the respective 5'

TC and 5’ CC targets that resided within a 4 nt-loop in a stem-loop structure, with C located at the 3’ end of the loop followed by a +1G located at the 5’ end of the stem. In this structural context, changing the substrate at the -1 and +1 nucleotides greatly reduced A3A and A3G editing activities (Sharma, Baysal 2017). Nonetheless, a limitation of the study is that only four APOBEC3 editing substrates were examined and all were in the context of 5’ N(C/T)CG.

Another study (Holtz, Sadler, Mansky 2013) showed that A3G could also efficiently edit 5’ ACCA, 5’ CCCC, and 5’ TCCT, but not 5’ GCCG, in ssDNA oligonucleotides when these targets were in an open (unstructured) configuration. Furthermore, A3G poorly edits 5’ ACCA and 5’ CCCC targets when they are located in short loops (<7 nt and <6 nt, respectively). Summarily, A3G prefers to edit 5’ CC targets within a loop region but only in the context of 5’ NCCG.

A third study (McDaniel et al. 2020) examined the *in vitro* editing activities of all seven APOBEC3 enzymes on ssDNA oligonucleotides embedding 5’ NTCN motifs, where the 5’ TC targets were located within loop, stem, or open structures. McDaniel et al. (2020) found that the editing activities of all APOBEC3 enzymes, except A3F and A3H, were the highest when 5’ TC in the context of 5’ GTCG was located within a loop region. However, all seven APOBEC3 enzymes also had the lowest editing activities at 5’ GTCG in comparison to other 5’ NTCN motifs, and the editing activities of all APOBEC3 enzymes were the highest when 5’ ATCA, 5’ GTCA, 5’ CTCA, and 5’ CTCT were located in an open structure. In general, APOBEC3-edited 5’ (C/T)C targets in the context of 5’ N(C/T)CG prefer a loop region, but 5’ TC targets in the context of 5’ NTC(A/T) prefer an open structure.

6.3.4 Host immune responses exert selective pressures that shape the genomic composition of tissue-specific coronaviruses

The above observations allow for the formation of the hypothesis that APOBEC3 and ZAP exert selective pressure on coronavirus genomes. To test this, a variety of mammalian hosts and tissues should be considered because there may be both species-specific and tissue-specific differences in ZAP and APOBEC3 productions. If the short-lived mice should produce less ZAP and APOBEC3 than long-living mammals, then the mouse hepatitis virus (MHV) should also experience little selection targeting CpG by ZAP and C to U editing by APOBEC3 in comparison to other mammalian-specific coronaviruses. Moreover, coronaviruses regularly infect organ tissues exposed to the external environments such as the respiratory and digestive systems (Leung et al. 2003; Nicholls et al. 2003). We expect that if a coronavirus regularly infects host tissues that are abundant in ZAP, then its genome should display CpG deficiency in CG-rich motifs such as CN_XGNCG, to elude ZAP-mediated immune response. If, in addition, the regularly infected tissue is abundant in APOBEC3, then the viral genome should trend prevalently towards increased C to U mutations in the context of APOBEC3-preferred motifs. Conversely, if a species-specific coronavirus regularly infects host tissues that are deficient in either APOBEC3 or ZAP, there will be either no strong CpG deficiency or elevated U and decreased C contents, as these selective pressures will be weak.

Our investigation considered a comprehensive number of publicly available genomes for seven coronaviruses (the Betacoronaviruses: SARS-CoV-2, SARS-CoV, MERS, Bovine CoV, mouse hepatitis virus [MHV], and porcine hepatitis E virus [HEV], and the Alphacoronavirus canine respiratory coronavirus [CRCoV]) as well as transcriptomics studies of tissue-level *ZAP* and *APOBEC3* mRNA expressions in the five host species (human, cattle, dog, mice, and pig). We found that all surveyed coronaviruses, except MHV, regularly infect host tissues with high *ZAP* and *APOBEC3* mRNA expressions. Expectedly, all surveyed coronavirus genomes except MHV

are strongly CpG deficient. In addition, deficiency of CpG was detected in the context of ZAP-preferred motifs in SARS-CoV-2. Furthermore, a temporal and geographical analysis for single nucleotide polymorphisms (SNPs) in local SARS-CoV-2 regions showed that the occurrence of C to U mutations was strikingly more prevalent than other SNPs. The preferred motif and structural contexts of 5' UC to 5' UU mutations were consistent with those favorably edited by APOBEC3 enzymes, but 5' CC to 5' CU mutations were weakly explained by A3G editing preference. The genome compositions of viruses are subjected to adaptation when the virus regularly infects tissues expressing ZAP and APOBEC3 in abundance, but not when a virus infects tissues that lowly express these AVPs.

6.4 Materials and Methods

6.4.1 Retrieving and processing the *APOBEC3* and *ZC3HAV1* genes and their tissue level mRNA expressions in five mammalian species

The NCBI Nucleotide Database was queried for "*APOBEC3*" and "*ZC3HAV1*" as gene names, and "*Homo sapiens*", "*Bos taurus*", "*Canis lupus familiaris*", "*Mus musculus*", and "*Sus scrofa"* as species, and protein coding sequences of *APOBEC3* and *ZC3HAV1 (ZAP)* isoforms were extracted in FASTA format along with their biomaRt Ensembl Accession IDs.

To compare mRNA expressions of *APOBEC3* and *ZC3HAV1* among tissues, we retrieved publicly available RNA Sequencing and Microarray studies that each sampled total RNA in at least ten mammalian tissues (see Supplemental File S1). The five mammalian species that have extensive tissue-specific mRNA expressions are *Homo sapiens* (human), *Bos Taurus* (cattle), *Canis lupus familiaris* (dog), *Mus musculus* (mice), and *Sus scrofa* (pig). For *Homo sapiens*, tissue-

level mRNA expressions were retrieved in averaged FPKM values from all 171 RNA-Seq datasets in BioProject PRJEB4337 (Fagerberg et al. 2014), 48 RNA-Seq datasets in BioProject PRJEB2445, 20 RNA-Seq datasets in BioProject PRJNA280600 (Duff et al. 2015), and in median TPM values from all RNA-Seq datasets available in the GTEx Portal (Lonsdale et al. 2013). For *Mus musculus*, tissue-level mRNA expressions were retrieved in averaged FPKM values from all 741 RNA-Seq datasets in BioProject PRJNA66167 (mouse ENCODE consortium) (Yue et al. 2014) and in average TPM values from all 79 RNA-Seq datasets in BioProject PRJNA516470 (Naqvi et al. 2019). For *Sus scrofa*, tissue-level mRNA expressions were retrieved in averaged FPKM values from TISSUE 2.0 integrated datasets (Palasca et al. 2018). For *Canis lupus familiaris*, tissue-level mRNA expressions were retrieved in averaged fluorescence intensity units (FIU) from all 39 microarray datasets in BioProject PRJNA124245 (Briggs et al. 2011), and in averaged TPM values from all 75 RNA-Seq datasets in BioProject PRJNA516470 (Naqvi et al. 2019). Lastly, for *Bos taurus*, tissue-level mRNA expressions were retrieved in averaged FPKM values from 42 RNA-Seq datasets in the Bovine Genome Database (Shamimuzzaman et al. 2019). All selected studies have considered total RNA at the tissue level in healthy individuals, but they do not report cell type-specific mRNA expressions within most tissues (e.g., lung, liver, small intestine).

Given that mRNA expressions were extracted were from multiple independent sources (some reporting FPKM and others TPM) and thus not directly comparable between studies, we calculated the relative mRNA expression levels of APOBEC3 and ZAP isoforms among tissues in each independent source. Specifically, we calculated the proportion of mRNA expression (PME) as:

$$PME = \frac{\text{mRNA expression value in a specific tissue}}{\text{summed mRNA expression values in all tissues}} \quad (1)$$

To show that PME determines the relative mRNA expressions of a gene among tissues, we calculated the PME values for all 13 human genes that were determined to have the highest mRNA expressions in the lungs (marked as Tissue enriched) by The Human Protein Atlas database (Uhlén et al. 2015) (https://www.proteinatlas.org/humanproteome/tissue/lungs). They are *SFTPC*, *SFTPA2*, *SFTPA1*, *SCGB1A1*, *SFTPB*, *AGER*, *SCGB3A2*, *SFTA2*, *CACNA2D2*, *LAMP3*, *SFTPD*, *HTR3C*, *RTKN2*. If PME works as intended, then the PME values of these 13 genes should be high in the lung tissue in comparison to 52 other tissues reported in the GTEx database (Lonsdale et al. 2013). As expected, we found that 12 out of these 13 genes have the highest PME values in the lungs, but *CACNA2D2* has the highest PME value in the cerebellum and second highest PME values in the lungs (see Supplemental File S1). This is not unexpected because, based on the BRAIN ATLAS (Uhlén et al. 2015), the mRNA expression of *CACNA2D2* is also enhanced in the cerebellum.

PME values were calculated from averaged TPM values in 24 human tissues using all RNA-Seq datasets available in the GTEx Portal (Lonsdale et al. 2013), from averaged FPKM values in 26 cattle tissues using the Bovine Genome Database (Shamimuzzaman et al. 2019), from averaged FPKM values in 33 pig tissues using TISSUE 2.0 integrated datasets (Palasca et al. 2018), from averaged FPKM values in 17 mice tissues using all 741 RNA-Seq datasets in mouse ENCODE consortium (Yue et al. 2014), from averaged FPKM values in 12 mice tissues using 79 RNA-Seq datasets in BioProject PRJNA516470 (Naqvi et al. 2019), and from averaged fluorescence intensity units in 10 dog tissues using all 39 microarray datasets in BioProject PRJNA124245 (Briggs et al. 2011). For each AVP isoform, tissue-specific PMEs were designated as high if they are greater than averaged PME and low if they are less than averaged PME (see Supplemental File S1).

6.4.2 Processing and quantifying transcriptomic data from chimeric "human lung-only" mice to obtain AVP mRNA expressions in control vs. SARS-CoV-2-infected human lung epithelial cells

Transcriptomic data associated with a study exploring SARS-CoV-2 infection in chimeric human lung-only mice (LoM) (GSE155286) were retrieved from NCBI's Sequence Read Archive (SRA) database and a summary of the data collected is detailed in Table 6.1.

The data were first partitioned into gzipped forward and reverse read fastq files using fastq-dump from the NCBI SRA toolkit (version 2.10.8). The resulting fastq.gz files were trimmed for Illumina TruSeq3 adapters and reads averaging a phred quality score < 20 were discarded using trimmomatic version 0.39 (Bolger, Lohse, Usadel 2014). All surviving pairs from preprocessing were carried forward to quantification using kallisto (version 0.46.1) (Bray et al. 2016).

Table 6.1. Summary of the RNA-seq dataset used to quantify genes of interest across SARS-CoV-2 infection states. Column 2 describes the infection states of the human lung epithelial cells, with "Control" = samples collected prior to SARS-CoV-2 infection, and "# days after" = samples collected at # days after SARS-CoV-2 infection.

Series	Infection state	Sample	Experiment	Runs
GSE155286	Control	GSM4698496	SRX8839384	SRR12339593
				SRR12339594
		GSM4698497	SRX8839385	SRR12339595
				SRR12339596
	2 days after	GSM4698487	SRX8839375	SRR12339575
				SRR12339576
		GSM4698488	SRX8839376	SRR12339577
				SRR12339578
	6 days after	GSM4698490	SRX8839378	SRR12339581
				SRR12339582
		GSM4698491	SRX8839379	SRR12339583
				SRR12339584
	14 days after	GSM4698493	SRX8839381	SRR12339587
				SRR12339588
		GSM4698494	SRX8839382	SRR12339589
				SRR12339590

An index file for the human transcriptome was generated from the Ensembl FASTA reference file "Homo_sapiens.GRCh38.cdna.all.fa" containing all human cDNAs with Ensembl transcript IDs (Cunningham et al. 2019) using kallisto's index function. The resulting index was used to quantify transcript abundances using kallisto for each experiment detailed in Table 6.1, and 1000 bootstrap samples were computed for each experiment to act as a proxy for technical replicates during subsequent analysis using sleuth (version 0.30) (Pimentel et al. 2016).

The kallisto outputs, including bootstrapped values from the previous step, were processed using the sleuth R package. Ensembl transcript IDs were associated with their Ensembl Gene ID

and gene name using the biomaRt R package and a sleuth object was prepared using the Ensembl gene ID for aggregation. The sleuth object was then fitted with two models: a full model (alternative) that assumes transcript abundance varies based on the time after SARS-CoV-2 infection, and a reduced model (null) assuming that transcript abundance does not vary between samples. The two models were compared with the likelihood-ratio test, and the resulting transcript level p-values were then aggregated based on their associated Ensembl gene ID using Lancaster's method, which assigns weights based on transcript abundance. A Benjamini-Hochberg false discovery rate correction (Benjamini, Hochberg 1995) was then applied to the weighted p-values to account for multiple comparisons (Yi et al. 2018). AVP genes were then differentially assessed at the level of their corresponding transcripts and comparisons were drawn from heatmaps using natural log transformed TPM values with a 0.5 offset generated by sleuth. Only transcripts of interest with an Ensembl Biotype of "Protein coding" that demonstrated variations in expression levels were considered in subsequent analyses. All significantly differentially expressed genes between control and infected samples, with false discovery rate $q < 0.05$, are listed in Supplemental File S1.

6.4.3 Determining the regular habitats of coronaviruses infecting five mammalian species

Host tissues that are infected by SARS-CoV-2, SARS-CoV, and MERS in humans, Bovine CoV in cattle, CRCoV in dogs, MHV in mice, and HEV in pigs were identified through an exhaustive large-scale manual search for experimental evidence-based primary source studies published up until June 5, 2020. Only studies that showed results from clinical course, autopsy, and experimental infections were considered, but cross-host studies were excluded. In total, tissue infections were determined from 25 SARS-CoV studies, 11 SARS-CoV-2 studies, eight MERS

studies, 15 mouse hepatitis virus (MHV) studies, nine porcine hepatitis E virus (HEV) studies, 18 canine respiratory coronavirus (CRCoV) studies, and ten bovine coronavirus (Bovine CoV) studies (see references in Supplemental File S1). Next, the regular tissue habitats of viruses were determined based on commonness of viral detection in host tissues when all studies were considered. For example, among the 25 SARS-CoV-2 studies collected, some tissue infections (e.g., lungs and intestines) are recorded in many studies while other tissue infections are rarely recorded (e.g., stomach). To score the commonness of SARS-CoV-2 infection in a tissue, in the lungs for instance, we calculated commonness of detection (COD) as:

$$COD = \frac{\text{number of times lungs infection is recorded in all studies considered}}{\text{total number of recorded infections in all tissues in all studies considered}} \quad (2)$$

Note that the COD measurement should not be used to make specific comparisons to rank most to least regularly infected tissues, because manually curated study size biases COD measurements. However, COD does tell us which tissues were commonly infected by a virus. For example, among the 25 SARS-CoV-2 studies collected, viral detection was reported in the lungs in nine studies, the intestines in eight studies, the liver in four studies, the heart in three studies, the kidney in three studies, and the stomach in one study (Supplemental File S1). The COD values for the lungs and the intestines are therefore the highest. Hence, the lungs and the intestines are surely regular habitats of SARS-CoV-2. Similarly, we determined the regular habitat for SARS-CoV (human lungs), MERS (human lungs), MHV (mice brain), HEV (pig liver), CRCoV (dog intestines and lungs), and Bovine CoV (cattle intestines and lungs). These regular habitats have COD values higher than twice that of any other tissue, except dog lungs for CRCoV, whose COD value was at least 1.5 times that of other tissues.

6.4.4 Retrieving and processing the genomes of coronaviruses infecting five mammalian species

The genome, Accession ID, and Sample Collection Date of 28475 SARS-CoV-2 strains were retrieved from the China National Center for Bioinformation (CNCB) (https://bigd.big.ac.cn/ncov/variation/statistics?lang=en, last accessed May 16, 2020), among which 2666 strains were selected because they were annotated as having complete genome sequences and high sequencing quality. Additionally, the complete genomic sequences of 403 MERS strains, 134 SARS-CoV strains, 20 Bovine CoV strains, two CRCoV strains, 26 MHV strains, and ten HEV strains were downloaded from the National Center for Biotechnology Information (NCBI) Nucleotide Database (https://www.ncbi.nlm.nih.gov/) (see Supplemental File S2).

We computed the nucleotide and di-nucleotide frequencies in each viral genome. Among strains, some have long poly-A tails that are missing in others. Some also have a longer 5' untranslated region (5' UTR) than others. To make a fair comparison between strains, genomes were first aligned with MAFFT version 7 (Katoh, Standley 2013), with the slow but accurate G-INS-1 option for 134 SARS-CoV, 20 Bovine CoV, two CRCoV, 26 MHV, and ten HEV strains, and with the fast FFT-NS-2 option for large alignments for 2666 SARS-CoV-2 and 403 MERS strains. Next, using DAMBE version 7 (Xia 2018), the 5' UTR sequences were trimmed away until the first fully conserved nucleotide position, and the 3' UTR sequences were trimmed out up to the last fully conserved nucleotide position. Then, gaps were removed from each trimmed genome, and the global nucleotide and dinucleotide frequencies were computed in DAMBE under "Seq. Analysis|Nucelotide & di-nuc Frequency" (see Supplemental File S2). Additionally,

nucleotide and di-nucleotide frequencies were similarly computed for whole, untrimmed, genomes (see Supplemental File S3). Finally, the conventional index of CpG deficiency (I_{CpG}) (Cardon et al. 1994; Karlin, Mrázek, Campbell 1997) was calculated, using the formula below:

$$I_{CpG} = \frac{P_{CG}}{P_C P_G} \quad (3)$$

Where P_{CG} is the proportion of CG dinucleotides when all dinucleotide frequencies were considered, and P_C and P_G are proportions of C and G nucleotides, respectively. The index is expected to be proximal to 1 when CpG is not deficient or in excess, smaller than 1 if CpG is deficient and greater than 1 if CpG is in excess.

6.4.5 Determining the temporal and geographical patterns of SNPs in SARS-CoV-2 genomes

Among the 2666 SARS-CoV-2 genomes retrieved from CNCB (database last accessed on May 16, 2020), we randomly selected one genome at each unique collection date, inclusively between December 31, 2019 (Wuhan-Hu-1, first isolate) and May 6, 2020 (mink/NED/NB04), among those that have complete records of local region annotations and nucleotide sequences in NCBI (see Supplemental File S4). A total of 99 strains were retrieved across 127 days since SARS-CoV-2 (including strain Wuhan-Hu-1, MN908947) was first sequenced. For each of these 99 strains, the nucleotide sequence of 12 out of 13 viral regions (5' UTR, ORF1ab, S, ORF3, E, M, ORF6, ORF7a, ORF8, N, ORF10, and 3' UTR) were extracted from DAMBE in FASTA format, MAFFT aligned with the slow but accurate G-INS-1 option, and local nucleotide and dinucleotide frequencies were computed for each region (see Supplemental File S5). ORF7b was omitted from the analysis because it was not annotated in 30 out of 99 strains, including the reference genome Wuhan-Hu-1 (MN908947).

To determine the nucleotide mutation patterns over time at each viral region, each aligned sequence was grouped into one of six time ranges, and the time range within each group was determined as the number of days passed since the reference strain was collected (Wuhan-Hu-1, 2019-12-31). Note that the number of days between time intervals is unequal, because strains were grouped based on roughly equal sample size and not by equal number of days. Then, sequences within each group were pair-wise assessed for single nucleotide polymorphisms (SNPs) using DAMBE's "Seq. Analysis|Nucleotide substitution pattern" with reference genome = Wuhan-Hu-1 (MN908947) and Default genetic distance = F84, and the sum and identity of SNPs within each group was calculated (see Supplemental File S4).

To control for any confounding effects imposed by mutations that could arise in specific geographic areas, we repeated the above analysis for all high quality and complete genomes in a country-specific manner. Only three countries have sequenced large numbers of strains with unique collection dates, leading us to consider 80 strains from the United States, 39 strains from Australia, and 34 strains from China (see Supplemental File S4). Note that because sample collection dates vary from one country to another, the time intervals will differ among geographical locations. In addition, within a geographical location, the sample sizes and time intervals may differ slightly among viral regions because not all strains have complete annotations for every viral region. For example, all 34 strains from China have an annotated E region, but two out of the 34 strains are missing an annotation for ORF8. Nucleotide mutations in these strains were traced relative to the reference genome being the oldest available strain in each country: MN908947 in China (2019-12-31), MN985325 in the US (2020-01-19), and MT450920 in Australia (2020-01-25). The statistical significance of prevalence of C to U mutations relative to all other mutations was established using the non-parametric Wilcoxon rank-sum test with continuity correction.

6.4.6 Sequence context and structural analyses of C to U mutations in the SARS-CoV-2 genome

The count, location, and identity of all non-synonymous SNPs were determined for each MAFFT aligned protein coding region (e.g., ORF1ab, see Supplemental File S6) using DAMBE's "Seq. analysis|Codon substitution pattern, reference = Wuhan-Hu-1, MN908947". Next, the count, identity, and location of all SNPs at each viral region were determined using DAMBE's "Seq. analysis|Site-specific Nuc. Freq.". This output was then compared with the output that contains only non-synonymous substitutions to obtain the count, identity, and location of all synonymous substitutions. Similarly, the count, identity, and location of all SNPs in the two non-protein coding regions (5' UTR and 3' UTR) were determined using DAMBE's "Seq. analysis|Site-specific Nuc. Freq.".

The above outputs were further processed to retain SNPs with unique locations in each viral region. These outputs were used to determine the identities of the flanking nucleotides for all site-specific C to U mutations to generate the 5' N<u>C</u>, 5' NN<u>C</u>, and 5' NN<u>C</u>N motif contexts (underlined are the C to U mutation sites, and N is any nucleotide). Next, the total numbers of 5' NC, 5' NNC, and 5' NNCN motifs in the Wuhan-Hu-1 genome were determined using DAMBE's "Sequences|Extract motif context". Finally, these values were used to calculate the odds-ratio for each motif: the observed proportion of motifs with C to U mutations (e.g., number of 5' A<u>C</u> with C to U mutations divided by total number of 5' AC dinucleotides in Whuan-Hu-1 genome = 36/2023) divided by the expected proportion of C to U mutation (total number of C to U mutations divided by total number of C in Wuhan-Hu-1 genome = 98/5492). For example, the odds-ratio of 5' A<u>C</u> is (36/2023)/(98/5492) = 0.997.

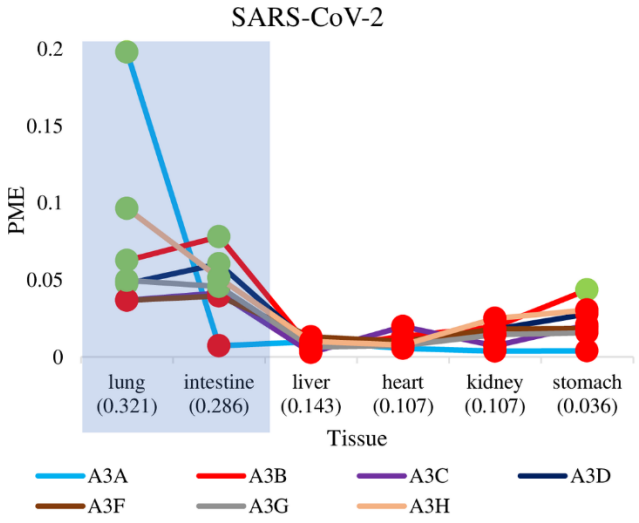
Next, for putatively edited C containing substrates on the Wuhan-Hu-1 genome, a 5 nt motif NNCNN was extended by 8 nt on either side to obtain a 21 nt sequence. To obtain the folding energy of the 21 nt sequence and obtain the secondary structure of the 5'NNCN motif, we used Minimum Folding Energy (MFE, - kcal/mol) via the Vienna RNA Folding Library (Hofacker 2003), with the following options: no lonely pairs, Temperature = 37°C (Supplemental File S6).

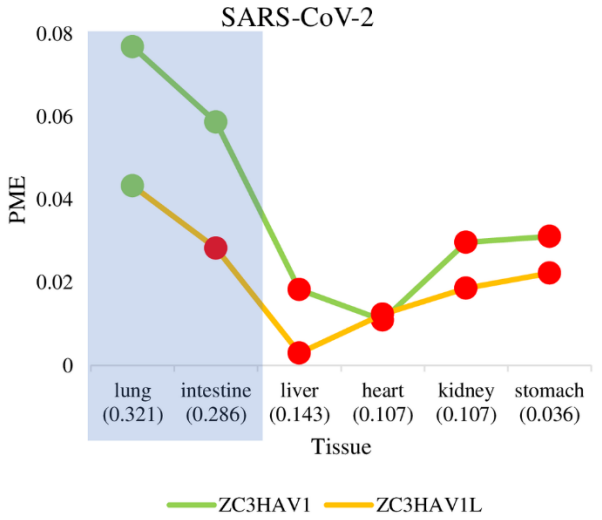
6.5 Results

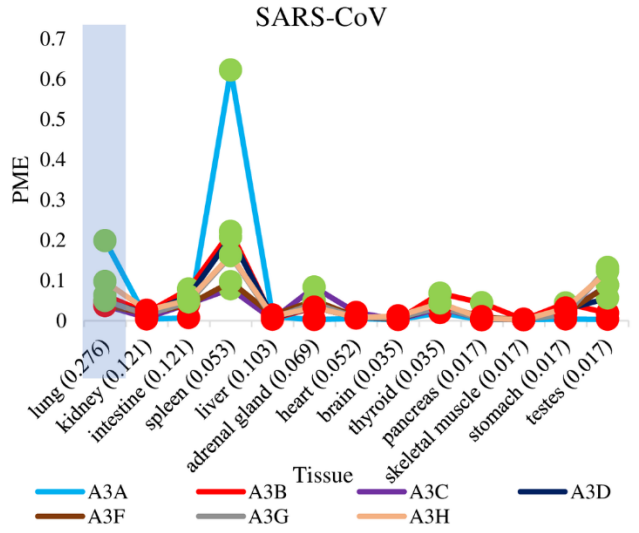
6.5.1 All surveyed coronaviruses except MHV regularly infect host tissues that highly express both ZAP and APOBEC3

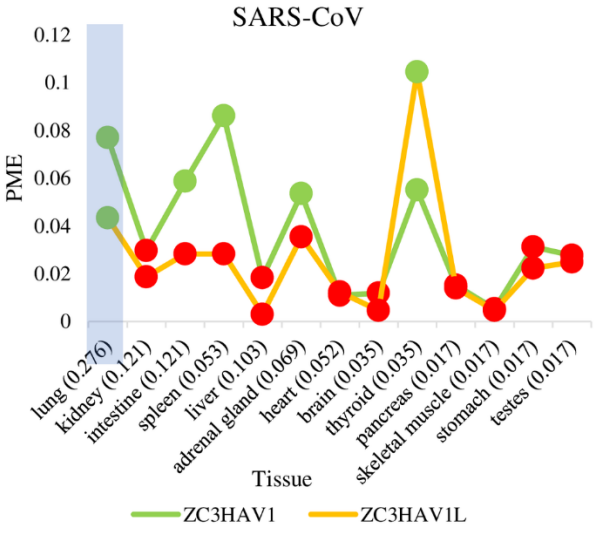
We determined which human tissues are regularly infected by coronaviruses and whether these tissues express ZAP and APOBEC3 in abundance. Figure S6.1 shows the tissue-specific mRNA expressions for AVP isoforms in humans and the number of tissue infection records for SARS-CoV-2, SARS-CoV, and MERS. For each susceptible tissue, Figure 6.1 shows the relative mRNA expressions of AVPs determined as high (in green) or low (in red). Furthermore, the regular habitats of each coronavirus were determined based on the highest COD. The lungs and the intestines are regular habitats of SARS-CoV-2 and both tissues contain high PMEs for many APOBEC3 isoforms (Figure 6.1: A3A, A3B, A3D, A3G, A3H in the lungs, and A3B, A3D, A3G, and A3H in the intestines) and for ZC3HAV1. Similarly, the regular habitats of SARS-CoV (lungs) and MERS (lungs) also contain high PMEs for some APOBEC3 and ZAP isoforms (Figure 6.1). Therefore, all three surveyed human coronaviruses can regularly infect host tissues where both *ZAP* and *APOBEC3* mRNAs are expressed in abundance and they display no strong preference for tissues deficient in either *ZAP* or *APOBEC3* transcripts.

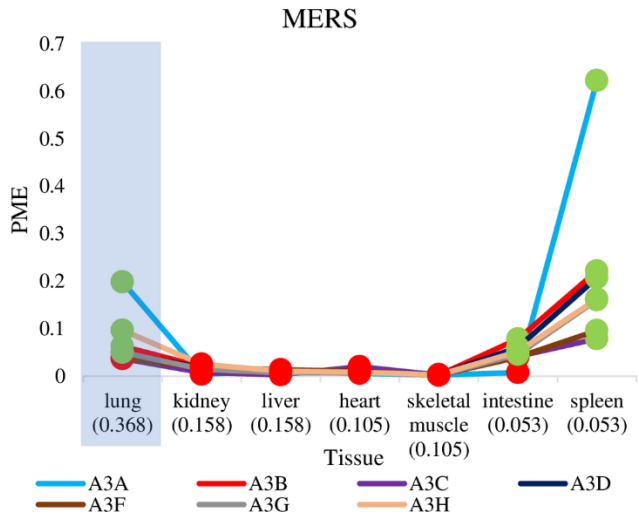
In an approach similar to Koning et al. (2009) and Refsland et al. (2010), we acquired the baseline levels of tissue-specific *APOBEC3* from total RNA in many tissues. Figure 6.1 is consistent with the findings of Koning et al. (2009) and Refsland et al. (2010), showing that *APOBEC3* mRNAs are abundant in the lung relative to other non-lymphoid tissues. Tissues such as the brain, liver, heart, skeletal muscle, and kidney are all deficient in both *ZAP* and *APOBEC3* mRNAs. The stomach, pancreas and testes abundantly express a subset of APOBEC3 enzymes but are ZAP deficient. In contrast, tissues of lymphoid organs including the spleen, adrenal gland, and thyroid express both AVPs in abundance.


SARS-CoV-2
PME
0.2
0.15
0.1
0.05
0
lung (0.321)
intestine (0.286)
liver (0.143)
heart (0.107)
kidney (0.107)
stomach (0.036)
Tissue
A3A
A3B
A3C
A3D
A3F
A3G
A3H


SARS-CoV-2
PME
0.08
0.06
0.04
0.02
0
lung (0.321)
intestine (0.286)
liver (0.143)
heart (0.107)
kidney (0.107)
stomach (0.036)
Tissue
ZC3HAV1
ZC3HAV1L


SARS-CoV
PME
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
lung (0.276)
kidney (0.121)
intestine (0.121)
spleen (0.053)
liver (0.103)
adrenal gland (0.069)
heart (0.052)
brain (0.035)
thyroid (0.035)
pancreas (0.017)
skeletal muscle (0.017)
stomach (0.017)
testes (0.017)
Tissue
A3A
A3B
A3C
A3D
A3F
A3G
A3H


SARS-CoV
PME
0.12
0.1
0.08
0.06
0.04
0.02
0
lung (0.276)
kidney (0.121)
intestine (0.121)
spleen (0.053)
liver (0.103)
adrenal gland (0.069)
heart (0.052)
brain (0.035)
thyroid (0.035)
pancreas (0.017)
skeletal muscle (0.017)
stomach (0.017)
testes (0.017)
Tissue
ZC3HAV1
ZC3HAV1L


MERS
PME
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
lung (0.368)
kidney (0.158)
liver (0.158)
heart (0.105)
skeletal muscle (0.105)
intestine (0.053)
spleen (0.053)
Tissue
A3A
A3B
A3C
A3D
A3F
A3G
A3H

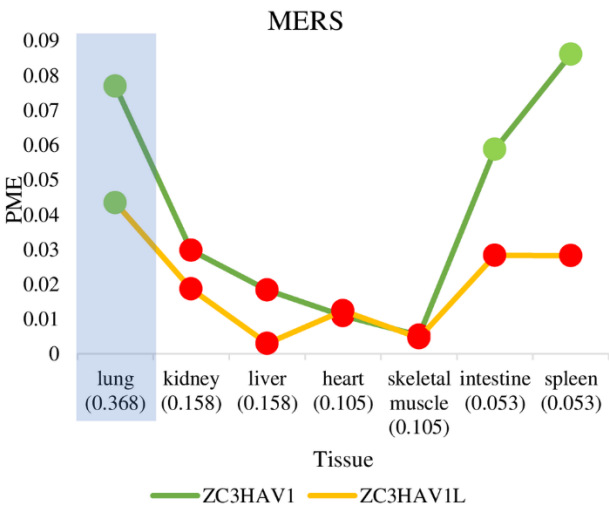

MERS
PME
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
lung (0.368)
kidney (0.158)
liver (0.158)
heart (0.105)
skeletal muscle (0.105)
intestine (0.053)
spleen (0.053)
Tissue
ZC3HAV1
ZC3HAV1L

Figure 6.1. Human tissues that are regularly infected by SARS-CoV-2, SARS-CoV, and MERS also have high mRNA expressions of *APOBEC3* and *ZAP* AVPs. The lines show the relative mRNA expressions in PME, for each *APOBEC3* and *ZAP* isoform, among tissues having records of SARS-CoV-2, SARS-CoV, and MERS infections. Dots highlighted in green and red are PME values that are greater and lower than the averaged PME values, respectively. These PME values were calculated based on averaged mRNA FPKMs retrieved from the GTEx Portal (Lonsdale et al. 2013). For each tissue, the commonness of viral detection (COD) score is appended in brackets next to tissue name. Shaded in light blue-gray are tissues that were determined to be regularly infected by the coronavirus (based on highest COD scores).

Retrieving averaged mRNA expression levels of *ZAP* and *APOBEC3* in four other mammalian species (cattle, dog, pig, mice) and their tissue-specific records of coronavirus infection (Figure 6.2, S6.2) reveals the tissues most susceptible to infection for these species (by highest CODs, tissues shaded in blue-gray) and the relative mRNA expressions (PMEs) for AVP isoforms in these tissues. Like human coronaviruses, these mammalian coronaviruses also regularly infect tissues exhibiting both high *APOBEC3* and *ZAP* mRNA expressions. Examples include HEV infecting pig liver, CRCoV infecting dog intestines and lungs, and Bovine CoV infecting cattle intestines. Conversely, while MHV regularly infects the brain in mice (Figure 6.2), PMEs for both *APOBEC3* and *ZAP* are low in this tissue.

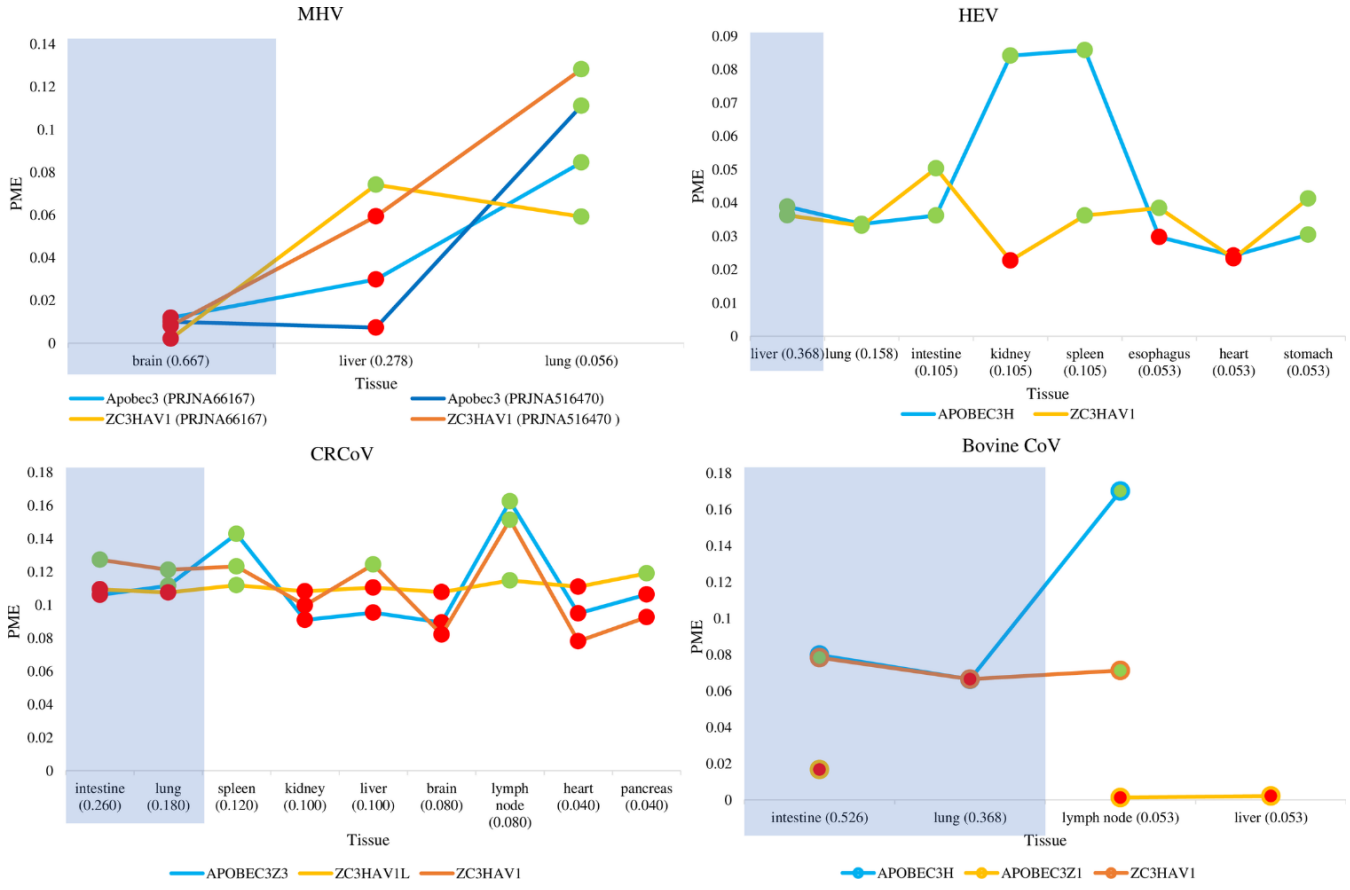


Figure 6.2. Pig, dog, and cattle tissues that are regularly infected by their respective coronaviruses (HEV, CRCoV, Bovine CoV) have high AVP mRNA expressions, but the mice brain that is regularly infected by MHV does not have high AVP mRNA expressions. The lines show the relative mRNA expressions (PME), for each APOBEC3 and ZAP isoform, among tissues where the virus is detected. Dots highlighted in green and red are PME values that are greater and lower than the averaged PME values, respectively. These PME values were calculated based on averaged mRNA expressions retrieved from the Bovine Genome Database (Shamimuzzaman et al. 2019), BioProject PRJNA124245 (Briggs et al. 2011), TISSUE 2.0 integrated datasets (Palasca et al. 2018), mouse ENCODE consortium (Yue et al. 2014) and BioProject PRJNA516470 (Naqvi et al. 2019). For each tissue, the commonness of viral detection (COD) score is appended in brackets next to tissue name. Shaded in light blue-gray are tissues that are regularly infected by the virus.

Taken together, Figures 6.1 and 6.2 show that lungs and intestines are regularly infected by five out of seven surveyed coronaviruses and exhibit high abundances of AVPs in all five mammals, except for lungs in cattle. This suggests that tissue-specific APOBEC3 and ZAP expressions may be correlated. Based on 24 human tissues, PMEs of APOBEC3 and ZAP are significantly positively correlated (e.g., for fitted regression line between 24 A3H and ZC3HAV1 values: coefficient of determination $R^2 = 0.43$, $P < 0.001$). Similarly, we found significant positive correlations between the PMEs of both AVPs in 17 mice tissues (APOBEC3 vs ZC3HAV1: $R^2 = 0.49$, $P = 0.0017$) and ten dog tissues (APOBEC3Z3 vs ZC3HAV1: $R^2 = 0.56$, $P = 0.021$). In contrast, there is no significant correlation between PMEs of both AVPs in 26 cattle tissues (APOBEC3H vs ZC3HAV1: $R^2 = 0.22$, $P = 0.34$) or 33 pig tissues (APOBEC3H vs ZC3HAV1: $R^2 = 0.11$, $P = 0.065$).

6.5.2 AVPs are expressed in lung epithelial cells, and in particular, the mRNA expressions of *ZAP*, *A3G*, and *ADAR* are upregulated in response to SARS-CoV-2 infection

Differential transcriptomic analysis of uninfected and infected LoM lung epithelial cell isolates revealed that among ADAR, AID, ZAP, APOBEC1, and APOBEC3 paralogues, transcripts that were found to be significantly differentially expressed between uninfected and infected human lung endothelial cells encoded A3G, ADAR, and ZAP. In all cases, the time after infection (on the time scale considered) was less of a contributing factor to expression levels than the intrinsic presence of infection (Figure 6.3, S6.3). This is evidenced by the consistent clustering of the uninfected control samples contrasted with the greater variance in the transcript-specific clustering of TPMs within infection time points. These results generally support the notion that

ADAR, A3G, and ZAP transcripts are either upregulated during SARS-CoV-2 infection relative to uninfected lung epithelial cells or remain at similar levels.

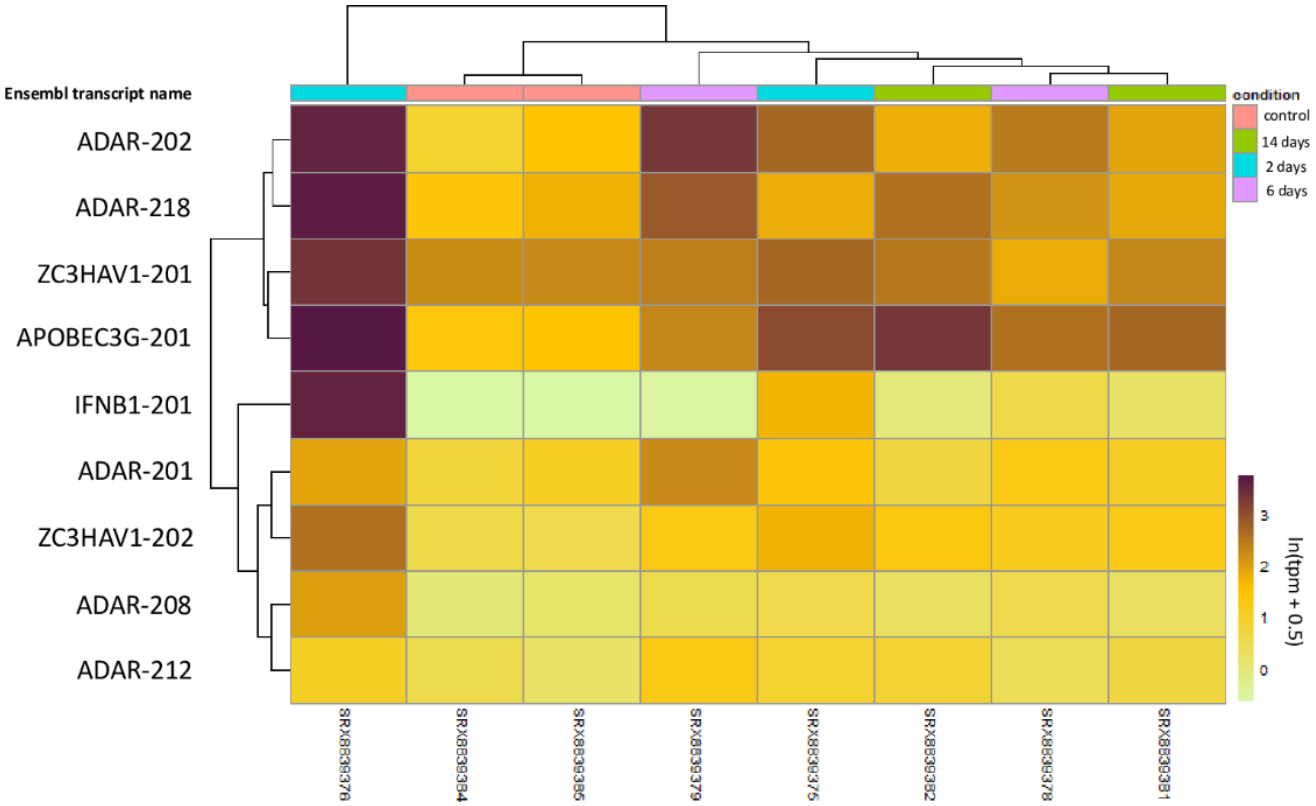


Figure 6.3. Differential expression of statistically significant transcripts of interest from LoM lung epithelial cell samples at varying time points after infection with SARS-CoV-2. Entries depict ln(tpm + 0.5) transformed fold-changes in kallisto-derived TPM values, from light green (lower) to dark purple (higher). Columns represent each experimental sample with its associated condition (time after infection) shown at the top. Each row represents a particular Ensembl transcript name, which is indicated to the left. Columns and rows are hierarchically clustered by similarity.

In particular, the results we observed for A3G are consistent with those observed during influenza A infection of A549 lung epithelial cells by Pauli et al. (2009) insofar as A3G was upregulated to the exclusion of other APOBEC3 paralogues and a corresponding significant

upregulation of IFNβ-encoding transcripts was generally observed in tandem, with the strongest coupling occurring in the samples with the greatest TPM fold-change (Figure 6.3). Apart from A3G, the sample from the SRX8839376 experiment demonstrated especially high upregulation of all transcripts of interest, followed by SRX8839375 and SRX8839379. All of these are from earlier infection time points (two and six days following infection), suggesting that sharper expression profile changes tend to happen earlier in infection. In contrast, the data make it clear that the cellular response to SARS-CoV-2 infection varies quite substantially. While the control samples have the most closely related expression profiles, the infected samples varied far more widely in their expression profiles and did not cluster strongly in a time-dependent manner. This emphasizes that immunological differences between individuals likely play an important role in combatting infection.

6.5.3 Coronaviruses targeting tissues with high AVP expressions exhibit decreased CpG and increased U content

Upon comparing the CpG and U contents between coronaviruses, we found those that regularly infected AVP-rich tissues tend to exhibit diminished CpG content in tandem with elevated U content. Conversely, MHV neither targeted AVP-rich tissues, nor did its genome indicate directional mutation with respect to CpG or U content. In both trimmed genomes (Figure 6.4a) and whole genomes (Figure S6.4a), MHV had the highest I_{CpG} (about 0.6 or higher) while SARS-CoV-2 had the lowest I_{CpG} (below 0.43 in all but two strains). As for all other coronaviruses surveyed, they also exhibited low $I_{CpG} < 0.5$ except for MERS being slightly higher. It should be noted that among the seven coronaviruses, I_{CpG} values also showed the greatest variation among MHV genomes but are much more constrained among the other six genomes (Figure S6.5a).

Nonetheless, in all seven coronaviruses, median I_{CpG} is the most deficient among I_{XpY} calculated (where X and Y are A, C, G, or U) and no other dinucleotides display strong deficiency or surplus (Figure S6.6).

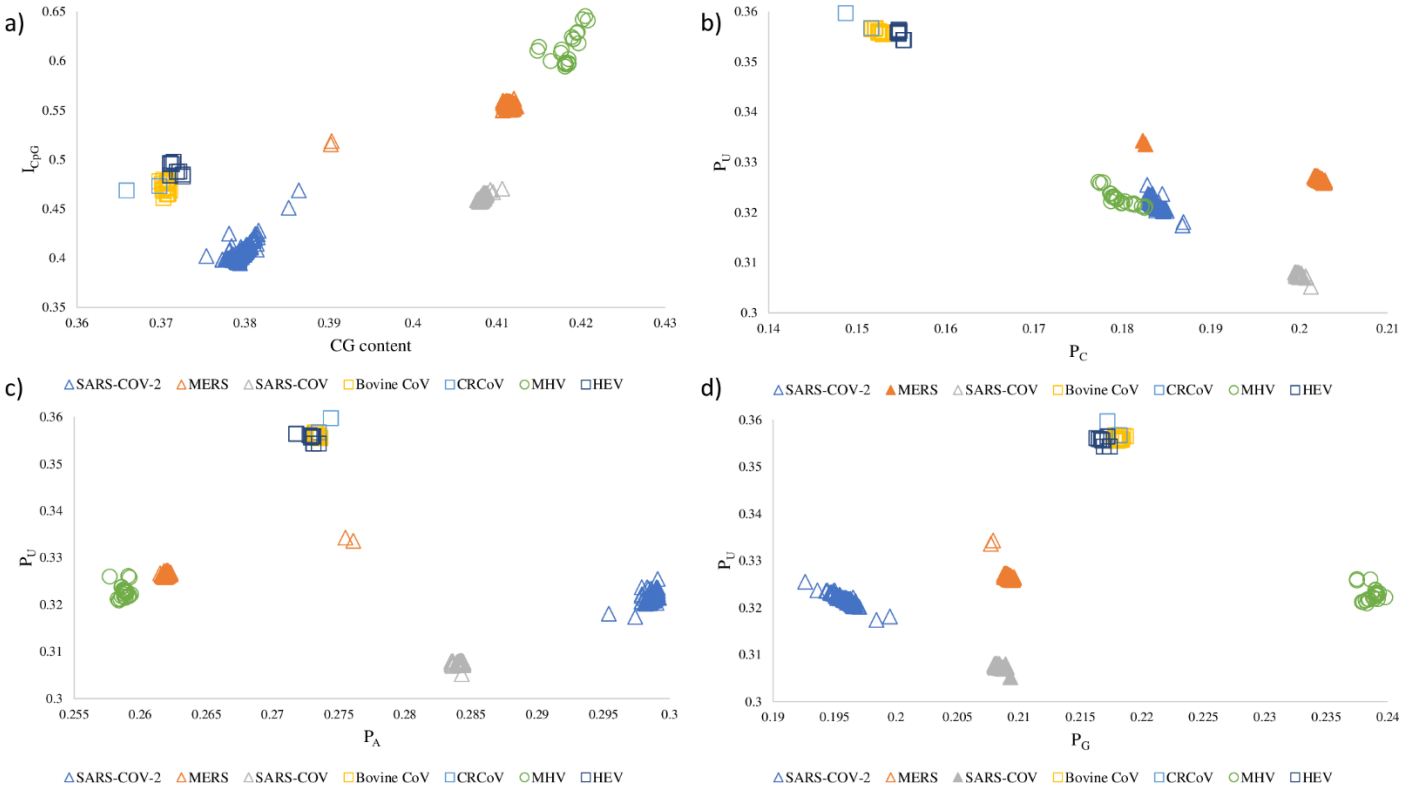


Figure 6.4. I_{CpG} and nucleotide compositions of seven mammalian-specific coronaviruses. All genomes were MAFFT aligned and all non-conserved sequence ends were trimmed (see Materials and Methods). Panel a) shows that SARS-CoV-2 has the smallest I_{CpG} in comparison to other coronaviruses from their natural hosts. Panels b), c) and d) compare the proportions of U (P_U) to those of A (P_A), C (P_C), and G (P_G), respectively. Each panel includes complete and high-quality sequence data of 2666 SARS-CoV-2 genomes, 403 MERS genomes, 134 SARS-CoV genomes, 20 Bovine CoV genomes, 2 CRCoV genomes, 26 MHV genomes, and 10 HEV genomes.

Figure 6.4 panels b, c, and d show that the proportion of U nucleotides (P_U) is inverse to the proportion of C nucleotides (P_C), but P_U does not correlate with P_A or P_G. Bovine CoV, CRCoV, and HEV all have very high P_U and conversely very low P_C. In comparison, MHV does not regularly infect tissues highly expressing APOBEC3 and has relatively reduced P_U and increased P_C (Figure 6.4b). Similar to I_{CpG}, P_U was least constrained in MHV relative to any other coronavirus (Figure S6.5b). Among human coronaviruses, genomic P_U is low in SARS-CoV-2 and MERS and especially in SARS-CoV (Figure 6.4b). These patterns persisted when I_{CpG} and P_U were re-analyzed using whole, untrimmed, genomes (Figure S6.4b).

6.5.4 Evidence of C to U directional mutation in SARS-CoV-2 viral regions

Above results demonstrate that viral genomes exhibit pronounced shifts towards CpG deficiency and elevated U content when the virus regularly infects tissues with high expression of both AVPs. However, human viruses share similar or lower global P_U relative to MHV, which predominantly infects AVP-deficient mice tissues (Figure 6.4). Hence, we examined whether there has been a history of P_U elevation in local SARS-CoV-2 regions over the span of the first four months since the virus was first isolated. We observed that most SNPs are C to U mutations (Figure 6.5a), and these mutations are prevalent at the 5' UTR and ORF1ab regions but infrequent at other viral regions (Figure 6.5b).

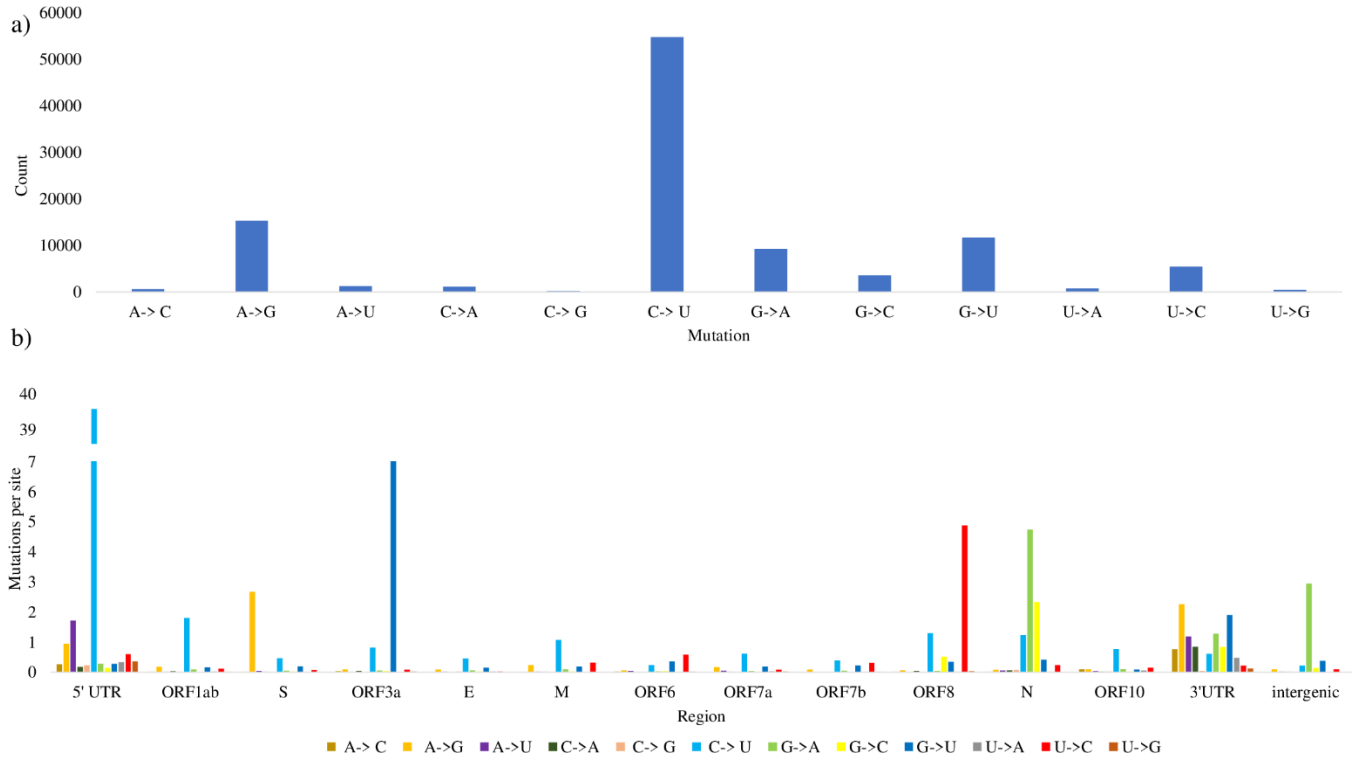


Figure 6.5. SNPs in 28474 SARS-CoV-2 (complete and incomplete) strains (samples collected up to 5-6-2020), with reference to strain Wuhan-Hu-1 (MN908947, 12-31-2019). Panel a) shows the frequency of each type of mutation in all isolates relative to the reference strain. Panel b) indicates the number of region-specific mutations normalized by region length (Mutations per site = count/sequence length) across all 13 viral regions and at the intergenic spaces. Indels and ambiguous point mutations were omitted from the analysis.

We next assessed temporal SNP patterns in each of 12 SARS-CoV-2 regions (excluding ORF7b) from a sample of 99 SARS-CoV-2 strains (see Materials and Methods). We observed a striking number of C to U mutations in aligned sequences between the reference and sampled strains (Figure 6.6), and the total number of C to U mutations trends upward over time in the 5' UTR and ORF1ab regions, but other regions did not exhibit any clear C to U mutation patterns.

Other notable substitution patterns were observed in the S region and ORF3a regions, namely: A to G mutations and G to U mutations, respectively.

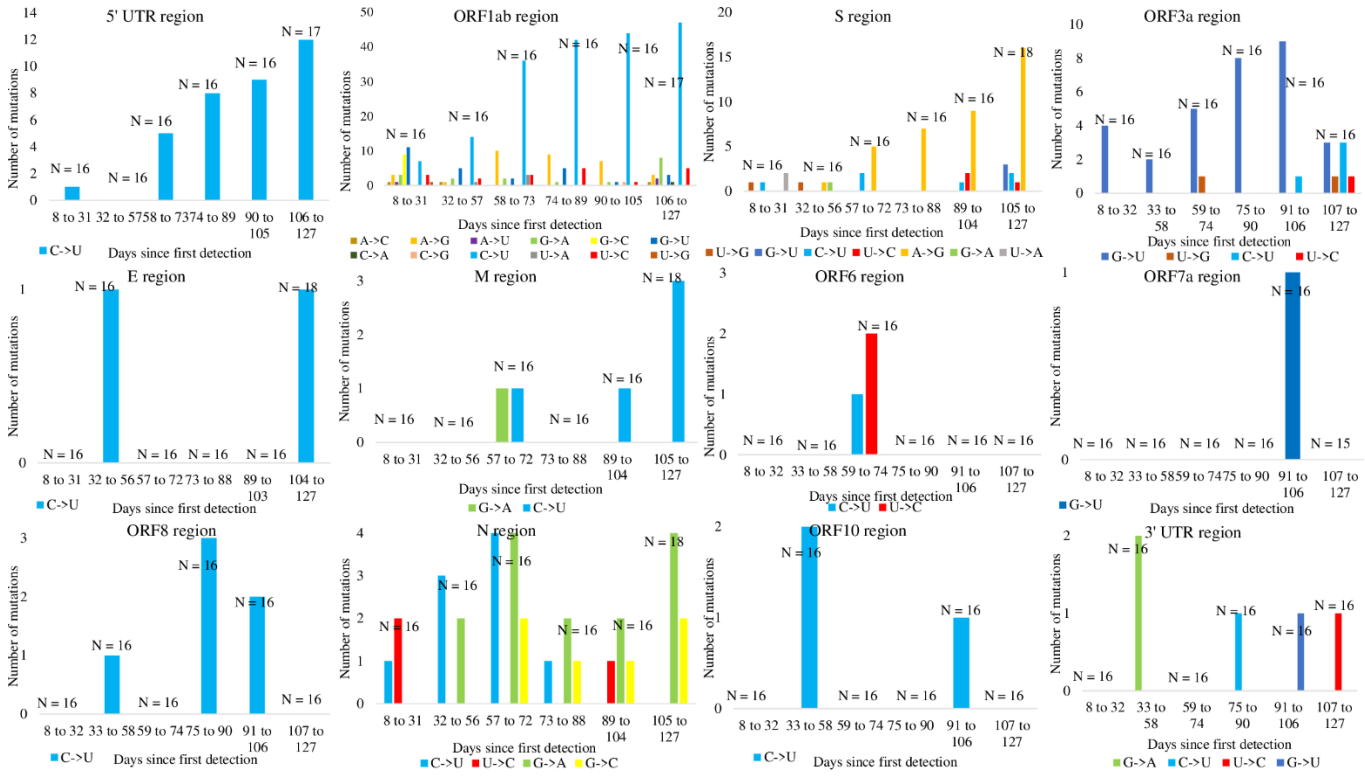


Figure 6.6. Temporal SNP patterns in 12 SARS-CoV-2 regions. Counts of C to U mutations are most prevalent and increase over time in the 5' UTR region and ORF1ab region. A to G mutations and G to U mutations are favoured in the S and ORF3a regions, respectively. In the eight other SARS-CoV-2 viral regions, SNPs are infrequent and display no obvious preference. A total of 99 complete and high-quality SARS-CoV-2 genomes with complete NCBI annotations were selected. These genomes were picked because they were each collected on a unique date, from the earliest sequenced strain Wuhan-Hu-1 (MN908947, 12-31-2019) to strain mink/NED/NB04 (MT457401, 5-6-2020), and each strain was grouped into one of six time ranges with roughly equal sample size. N denotes the number of strains per time range.

To control for potential geographical bias, such as a potential widespread C to U hypermutation before SARS-CoV-2 was transmitted outside of China, we show that the prevalence of C to U mutations in 5' UTR and ORF1ab regions persists when considering the temporal and geographical SNP patterns of SARS-CoV-2 strains isolated from three different countries: United States, Australia, and China (Figure 6.7, Figures S6.7 – S6.9). In all three countries, aligned sequences revealed the same pattern of C to U mutations we previously observed. Likewise, C to U mutations trended upwards over time in the 5' UTR and ORF1ab (Figure 6.7), but this bias was absent in other regions (Figures S6.7 – S6.9). The number of C to U mutations is significantly greater than for any other mutations in the ORF1ab region regardless of geographical constraint (all ORF1ab panels in Figures 6.5 and 6.6, Wilcoxon rank sum test with continuity correction: $P < 0.01$). Indeed, only C to U mutations were observed in the 5' UTR region of strains collected from the United States and Australia (Figure 6.7) and no two countries shared the same SNP patterns in other viral regions (Figures S6.7 – S6.9).

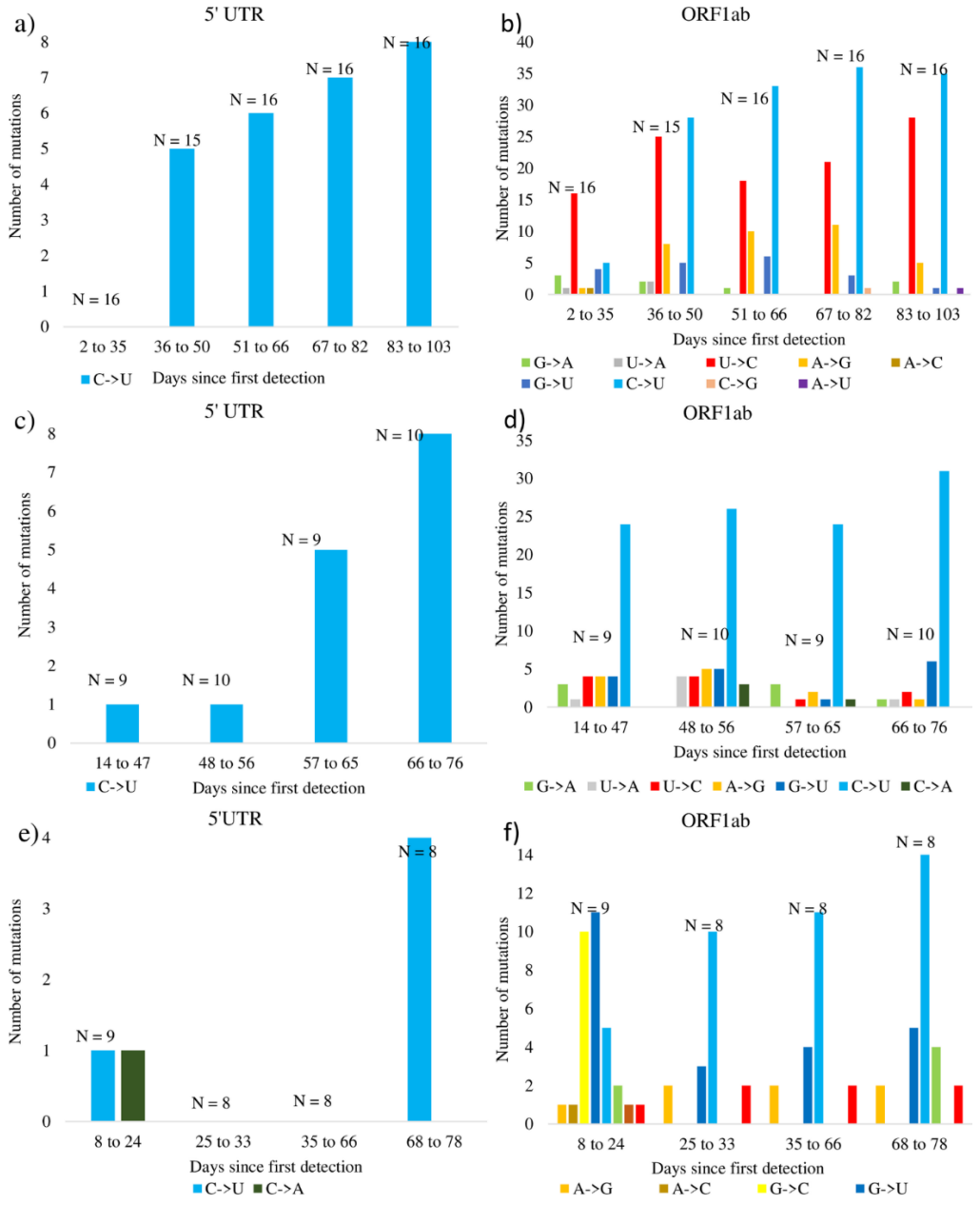


Figure 6.7. Geographical and temporal SNP patterns for samples of complete and high-quality SARS-CoV-2 strains. Panels respectively show SNPs within the 5' UTR and ORF1ab regions relative to the first isolate collected within the country. Panels a) and b) show SNPs from 79 strains relative to accession MN985325 collected in the United States. Panels c) and d) show SNPs from 38 strains relative to accession MT450920 collected in Australia. Panels e) and f) show SNPs from 33 strains relative to accession MN908947 collected in China.

We next investigated the sequence context for the prevalently observed C to U mutations. A total of 477 SNPs were observed when the reference genome (Wuhan-Hu-1) was compared to a sample of 98 SARS-CoV-2 strains. Over half of these SNPs were C to U mutations (262/477, Table S6.1), of which 144, 82, and 36 C to U mutations were synonymous, non-synonymous, and non-coding substitutions, respectively. These 262 C to U mutations were found at 98 unique nucleotide sites with respect to the reference genome of Wuhan-Hu-1 (Figure S6.10), with 92 unique sites located within protein coding regions (55 sites have synonymous substitutions and 37 sites have non-synonymous substitutions) and six unique sites that are not located within protein coding regions. Furthermore, the locations of unique sites subjected to C to U mutations were roughly evenly distributed across the Wuhan-Hu-1 genome (Figure S6.10a); these SNPs were not densely packed at any specific sequence region (Figure S6.10b).

6.5.5 C to U mutations in SARS-CoV-2 occur at known APOBEC3 recognition sites

Many of the aforementioned 98 unique C to U mutations sites occur at 5' C<u>C</u> and 5' U<u>C</u> (with C to U mutation sites underlined) dinucleotides embedded in motifs that facilitate APOBEC3 binding and editing in HIV-1, MLV, and SIV (Table 6.2). For each dinucleotide and trinucleotide motif, odds-ratios were calculated using the observed proportion of C to U mutations divided by the expected proportion of C to U mutations (see Materials and Methods). Among dinucleotides, only 5'C<u>C</u> and 5' U<u>C</u> have odds-ratios > 1 (observed > expected) with 1.136 and 1.111, respectively. In viruses such as HIV, MLV, and SIV, most studies are consistent in demonstrating that A3G prefers to edit 5' C<u>C</u> whereas the other six APOBEC3 enzymes prefer to edit 5' T<u>C</u> (Table 6.2).

Table 6.2. The preferred motif contexts of C to U mutations in the SARS-CoV-2 genome are the same as those previously identified in HIV, MLV, and SIV that were subjected to editing by APOBEC3 enzymes. Specifically, the preferred dinucleotides are 5' C<u>C</u> and 5' U<u>C</u>, and the preferred trinucleotides are 5' AC<u>C</u>, 5' CC<u>C</u>, and 5' UU<u>C</u> (by highest odds-ratio > 1 in bold). Underlined are sites subjected to C to U mutations. In red are non-APOBEC3 deaminases that were reported to have C to U/T editing ability in non-viral sequences.

Motifs	Hotspots subjected to editing in HIV, MLV, and SIV*	Motifs with C to U mutations in Wuhan-Hu-1	Total motifs in Wuhan-Hu-1	Odds-ratio
A<u>C</u>	A3D[5]	36	2023	0.997
AA<u>C</u>		9	615	0.82
CA<u>C</u>		9	459	1.099
GA<u>C</u>		3	340	0.494
UA<u>C</u>	AID[7]	15	609	1.38
C<u>C</u>	A3G[4,8]	18	888	**1.136**
AC<u>C</u>		11	376	**1.639**
CC<u>C</u>	A3G[1,2,3,6,7,14]	3	116	**1.449**
GC<u>C</u>		0	187	0
UC<u>C</u>	A3A[11], A3G[2]	4	209	1.073
G<u>C</u>		16	1168	0.768
AG<u>C</u>		5	301	0.931
CG<u>C</u>		2	97	1.155
GG<u>C</u>		2	223	0.503
UG<u>C</u>		7	547	0.717
U<u>C</u>	A3A[9], A3B[6,12], A3D[5,8], A3F[2,6,7,8,15], A3H[8,13,16], A1[17]	28	1413	**1.111**
AU<u>C</u>		3	339	0.496
CU<u>C</u>		4	287	0.781
GU<u>C</u>		5	269	1.042
UU<u>C</u>	A3C[10,14], A3F[3,14], A3H[18]	16	518	**1.731**

[1] (Yu et al. 2004); [2] (Liddament et al. 2004); [3] (Wiegand et al. 2004); [4] (Harris et al. 2003); [5] (Dang et al. 2006); [6] (Bishop et al. 2004); [7] (Kohli et al. 2009); [8] (Hultquist et al. 2011); [9] (Suspène et al. 2013); [10] (Adolph et al. 2017); [11] (Aguiar et al. 2008); [12] (Doehle, Schäfer, Cullen 2005); [13] (Harari et al. 2009); [14] (Langlois et al. 2005); [15] (Ebrahimi, Alinejad-Rokny, Davenport 2014); [16] (Ooms et al. 2013); [17] (Saraconi et al. 2014); [18] (Mitra et al. 2015).

* All preferred motifs subjected to RNA editing by APOBEC3 are based on HIV-1 mutagenesis studies except for [4] (Harris et al. 2003) which studied MLV. [2] (Liddament et al. 2004), [6] (Bishop et al. 2004), and [14] (Langlois et al. 2005) additionally studied MLV and [5] (Dang et al. 2006) additionally studied SIV.

Consensus motif for AID enzyme editing was determined from a mutagenesis study of *rpoB* gene in *Escherichia coli*: [7] (Kohli et al. 2009). Consensus motif for APOBEC1 enzyme editing was determined from mutagenesis study of chicken B-cell line DT40: [17] (Saraconi et al. 2014).

In the context of 5' NCC, the two most preferred trinucleotide motifs are 5' ACC (odds-ratio = 1.639) and 5' CCC (odds-ratio = 1.449). This observation is consistent with multiple studies showing that 5' CCC is preferred (Bishop et al. 2004; Wiegand et al. 2004; Yu et al. 2004; Kohli et al. 2009), although others found that 5' RCC may also be preferred (Liddament et al. 2004) by A3G editing in HIV-1 and MLV. When 5' NUC is considered, the preferred trinucleotide in SARS-CoV-2 is 5' UUC (odds-ratio = 1.731). This observation is also corroborated by multiple studies indicating that 5' TTC is preferred by all APOBEC3 enzymes except A3G (Feng et al. 2014; Chen, MacCarthy 2017; Martinez et al. 2019).

We further considered C to U mutations in the context of 5' NCCN and 5' NTCN. All studies summarized in Table 6.3 conclude that both the -2 and +1 positions flanking 5' NC influence the efficacy of APOBEC3 editing. Comparing between reported APOBEC3 enzyme activities within studies, activity levels were classified as preferred (++), less preferred (+), inefficient (-), and avoided (--) among motifs examined. A3D was excluded from Table 6.3 because its consensus target could not be specified beyond 5' (T/A)(T/A)C(G/T) (Takata et al. 2017); an A3D-preferred motif has not been established (McDaniel et al. 2020) because the catalytic properties of this enzyme are not fully characterized (Adolph, Love, Chelico 2018). Despite the lack of a strongly preferred consensus sequence among many APOBEC3 enzymes, most studies are consistent in reporting 5' CCC(A/T) as preferred targets of A3G, and 5' TTC(A/T) are among, if not the most, preferred motifs by all 5' TC editing APOBEC3 enzymes except A3B (Table 6.3).

Table 6.3. Number of unique C sites, in the context of 5’ N(U/C)CN motifs, that were subjected to mutation when the genome of Wuhan-Hu-1 was compared to 98 later sampled SARS-CoV-2 strains. Underlined are the mutated C sites. “Motifs in genome” indicates the number of observed motifs in the genome of Wuhan-Hu-1. Column 5 shows the consensus 5’ N(U/C)CN motifs as reported by surveyed studies. Columns 6 to 10 show the relative levels of APOBEC3 editing activities at select 5’ N(U/C)CN motifs tested by studies. Columns 11 to 13 show the structural configurations of 5’ (U/C)C dinucleotide sites in Wuhan-Hu-1 where mutations had occurred in later strains. The last column shows the preferred structural configurations of 5’ (U/C)C edited by APOBEC3 as reported by surveyed studies.

Motifs	Edited motifs	Motifs in genome	Odds-ratio	Consensus motif reports[a]	A3A[b] (Silvas)/(McDaniel)	A3B (McDaniel)	A3C (McDaniel)	A3F (Armitage)/(Wan)/((McDaniel)	A3H (McDaniel)/(Mitra)	UC (stem)	UC (loop)	UC (open)	A3 preferred structure
AUCA	1	137	0.368		--/++	-	-	+/-/-	+/na		1		open: ABCDGH (McDaniel)
AUCC	1	51	0.987		na/na	na	na	-/na/na	na/na			1	
AUCG	0	20	0.000		--/na	na	na	-/na/na	na/na				loop: A (Sharma)
AUCU	1	130	0.387		-/na	na	na	-/na/na	na/na		1		
CUCA	3	131	1.153	B , C , F (Bishop)	na/++	-	--	-/--/++	++/na			3	
CUCC	0	30	0.000		na/-	--	-	-/na/-	-/na				open: ABCFH (McDaniel)
CUCG	0	32	0.000		na/-	-	-	--/na/--	-/na				
CUCU	1	93	0.541		na/+	+	+	-/na/--	++/na			1	
GUCA	3	98	1.542	B, C (Bishop)	+/+	++	+	--/-/-	-/na		1	1	
GUCC	1	47	1.071		na/na	na	na	na/na/na	na/na	1			
GUCG	0	21	0.000		-/--	--	--	na/na/+	--/na				Loop: ABG (McDaniel); loop: A (Sharma); open: FH (Sharma)
GUCU	1	103	0.489		--/na	na	na	--/na/na	na/na	1			
UUCA	7	182	1.937	F (Liddament)	++/na	na	na	++/++/na	na/++		1	6	
UUCC	0	80	0.000		na/na	na	na	+/+/na	na/na				
UUCG	3	39	3.874		+/na	na	na	-/++/na	na/-	1	2		loop: A (Sharma)
UUCU	6	216	1.399	F (Liddament, Wiegand)	+/++	--	++	+/-/++	+/++			5	open: ABCFH (McDaniel)

Motifs	Edited motifs	Motifs in genome	Odds-ratio	Consensus motif reports	A3G (Yu, Ebrahimi, Armitage)	CC (stem)	CC (loop)	CC (open)	A3G preferred structure
ACCA	4	151	1.302		-/na/+			3	
ACCC	1	55	0.894		na/na/-	1			
ACCG	1	29	1.695		-/na/+			1	
ACCU	5	139	1.769		-/na/-	1	2	1	
CCCA	0	39	0.000	G (Bishop, Harris, Liddament, Wiegand)	++/++/++				
CCCC	1	14	3.512	G (Liddament)	na/-/+			1	
CCCG	0	13	0.000		++/na/+				loop: G (Sharma)
CCCU	2	49	2.007	G (Liddament)	+/na/+		1		
GCCA	0	78	0.000		--/--/--				
GCCC	0	17	0.000		na/na/--				
GCCG	0	17	0.000		na/na/na				
GCCU	0	75	0.000		na/na/na				
UCCA	2	86	1.143	G (Liddament)	na/na/+	1		1	
UCCC	0	29	0.000	G (Liddament)	na/na/-				
UCCG	0	15	0.000		na/na/-				
UCCU	2	79	1.245	G (Harris, Liddament)	na/na/-		1		

[a] – The seven APOBEC3 members were abbreviated in the table to show only the last letter of the enzyme (e.g., A3A = A). [b] – The relative APOBEC3 activity levels at surveyed motifs were designated with + and – symbols: "++" = preferred, "+" = less preferred, "– " = inefficient, "– – " = avoided. The "na" indicates that data is unavailable. Activity levels reported by different studies are separated by a "/" symbol, and the order of activity data corresponds to the order of cited references shown in the table header. In Brackets are the abbreviated citation showing the last name of only the first author; See PMID: 33351847 for full citation in table.

In SARS-CoV-2, the two tetranucleotides embedding the 5' UC editing target with the highest number of unique C to U mutations were 5' UUCA and 5' UUCU (odds-ratios 1.937 and 1.399, respectively), followed by 5' GUCA, 5' CUCA, and 5'UUCG (odds-ratios 1.542, 1.153, 3.874, respectively), and all except 5'UUCG are preferred APOBEC3 editing motifs (Table 6.3). However, the A3G-preferred 5' CC motifs 5' CCC(A/U) (e,g. (Armitage et al. 2008)) were not found and 5' ACC(A/U) were instead abundant in SARS-CoV-2 (Table 6.3). Nevertheless, 5' GCCN were avoided and no such motifs were observed in SARS-CoV-2.

Additionally, McDaniel et al. (2020) showed that the 5' UC targets in 5' NUC(A/U) motifs highly preferred an open structure configuration. Similarly, in SARS-CoV-2, the 5' UC within 5' NUC(A/U) motifs highly preferred open structure configurations (Table 6.3: 16 out of 23, see secondary structure details in Supplemental File S6). Of 5' (C/U)C targets in 5' N(C/U)CG that were reported to prefer the loop region by Sharma, Baysal (2017), three 5' NUCG motifs were observed in SARS-CoV-2 (all being 5' UUCG), two were found in the loop region and one was found in the stem region; in contrast, only one 5' NCCG (5' ACCG) motif was observed and it had an open structure (Table 6.3).

6.5.6 CpG deficiency is maintained in specific viral regions and I_{CpG} does not differ notably among SARS-CoV-2 genomes collected in the span of four months

Lastly, we performed a temporal analysis to determine whether there are differences in I_{CpG} within and between viral regions among the 99 SARS-CoV-2 strains. Within viral regions, there were no notable differences in I_{CpG} between strains sampled at different time intervals (Figure 6.8). However, there were notable differences in I_{CpG} between viral regions. In particular, the ORF1ab, S, and ORF6 regions had the lowest I_{CpG} values < 1, whereas the 5' UTR, E, and ORF10 regions

had the highest I_{CpG} values > 1. Thus, CpG content varies substantially across the SARS-CoV-2 genome (di Gioacchino et al. 2020; Digard et al. 2020).

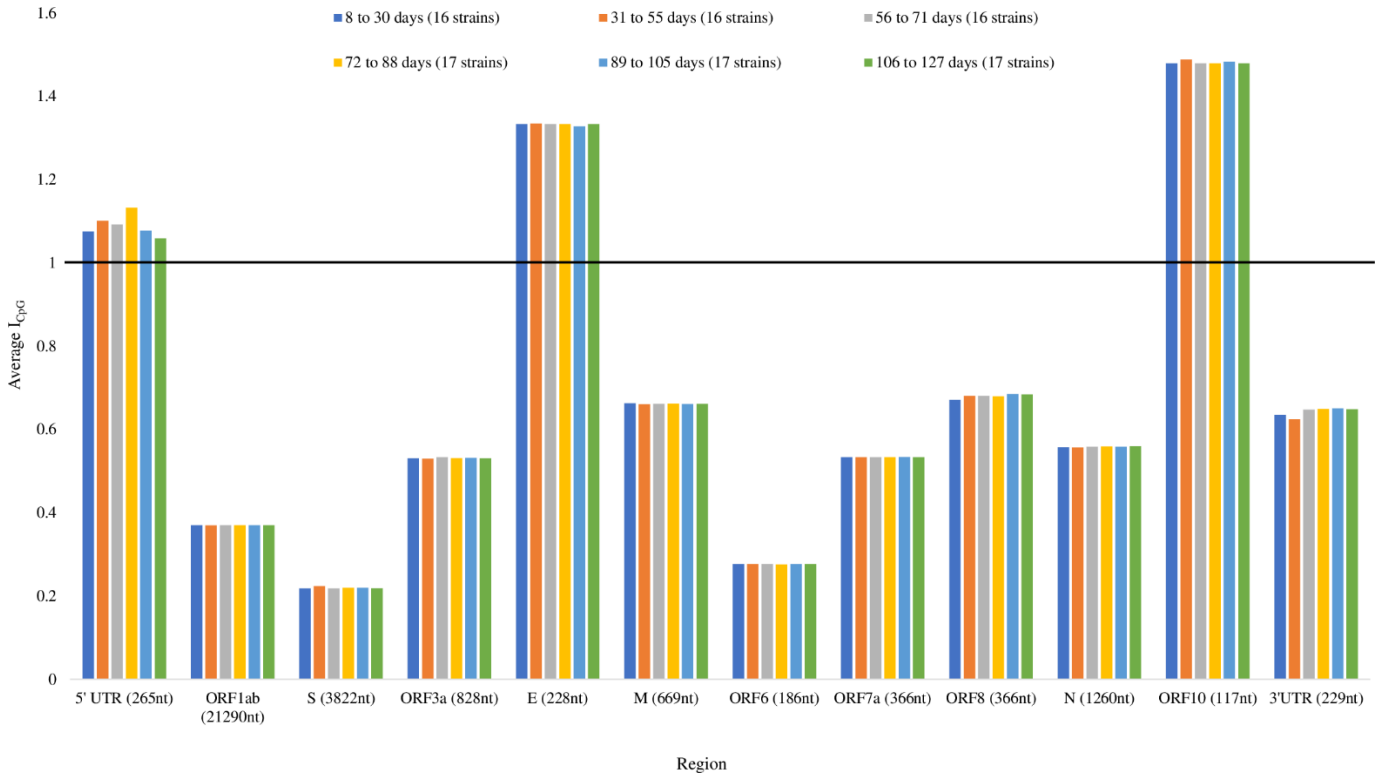


Figure 6.8. A temporal analysis for local I_{CpG} in a sample of 99 complete and high-quality SARS-CoV-2 strains with complete NCBI annotations. Each retrieved strain was collected on a unique day, regardless of geographical location, since the first isolated strain Wuhan-Hu-1 (MN908947, 12-31-2019) to strain mink/NED/NB04 (MT457401, 5-6-2020). Each strain was grouped into one of six time ranges with approximately equal sample size. I_{CpG} did not change substantially over the 127 days since first detection, but I_{CpG} values were not uniform across viral regions. I_{CpG} values were lowest in ORF1ab, S, and ORF6 regions, and the highest in the 5' UTR, E, and ORF10 regions. The horizontal black line highlights $I_{CpG} = 1$.

Next, we examined the CpG content in the SARS-CoV-2 genome in the context of CN_XGNCG motifs that were preferably recognized by ZAP in mice (Luo et al. 2020). When the reference Wuhan-Hu-1 was compared to the other 98 strains sequenced in the following four months, there were only 11 unique sites where either C or G, in the context of CpG, had been mutated, and only two out of the 11 mutations occurred in the context of CN_XGNCG. In addition, other C or G mutations, in the context of CpG, do not particularly prefer CG-rich sequences (Supplemental File S6). Note that we determined the number of mutations that have occurred at unique sites when referenced to the genome of Wuhan-Hu-1, because while a mutation at a given site may be carried by multiple later strains, the creation of such mutations could be derived from singular events. Nevertheless, there was a deficit in the total number of observed CG dinucleotides (Table 6.4: Obs/Exp ratio = 0.408) and CN_XGNCG motifs (Table 6.4: Obs/Exp ratio ranges from 0.309 to 0.619) at the genome of Wuhan-Hu-1.

Table 6.4. Total number of unique C and G mutations and those that occurred in the context of CpG motifs in SARS-CoV-2. Number of unique mutations were determined on the Wuhan-Hu-1 genome when it is compared to 98 later strains, in the context of CN_XGNCG (underlined are either C or G nucleotides that were mutated). N_X indicates the spacer sequence of length x = 4 nt to 8 nt. The "Observed number" indicates the total number of nucleotide and motifs observed in the genome of Wuhan-Hu-1, the "Expected number" is calculated based on the total nucleotide frequencies and the length of Wuhan-Hu-1 genome, and "Obs/Exp ratio" calculates Observed number/Expected number. The specific sequence contexts of all 11 unique C or G mutations that occurred at CpG are shown in Supplemental File S6.

Nucleotides and Motifs	Number of unique mutations at C or G	Observed number	Expected number	Obs/Exp ratio
C	98	5492		
G	66	5863		
CG	11	439	1076.802	0.408
GNCG	3	85	211.104	0.403
CN_4GNCG	0	12	38.777	0.309
CN_5GNCG	0	24	38.780	0.619
CN_6GNCG	1	14	38.783	0.361
CN_7GNCG	1	17	38.785	0.438
CN_8GNCG	0	22	38.788	0.567

6.6 Discussion

SARS-CoV-2 poses a serious global health emergency. Since its outset in Wuhan City, Hubei province of China in December 2019, the viral outbreak has resulted in over 20 million confirmed cases around the world (https://www.who.int/emergencies/diseases/novel-coronavirus-2019, last accessed August 12, 2020). The pandemic has prompted an immediate global effort to sequence the genome of SARS-CoV-2, and by May 2020 over 28000 strains have been publicly

deposited over the course of just four months. With a wealth of sequence data, we performed a comprehensive comparative genome study on SARS-CoV-2 and six other coronaviruses across five mammalian species, with the aim to understand how coronaviruses evolve in response to tissue-specific host immune systems.

We tested whether APOBEC3 and ZAP immune responses act as selective pressures to shape the genome of an infecting coronavirus. We note that ZAP is highly expressed in human lungs (Figure 6.1) and we observed that its expression is further upregulated in SARS-CoV-2 infected lung epithelial cells relative to the control (Figure 6.3). Our observations are compatible with the notion that cytoplasmic ZAP can bind to CpG dinucleotides to facilitate the degradation of viral transcripts. This idea, in conjunction with our observations, is corroborated by a recent study that found ZAP targets CpG to restrict SARS-CoV-2 replication in human lung cells (Nchioua et al. 2020). In contrast to ZAP, APOBEC3 enzymes are mostly expressed in immune cells such as $CD4^+$ T cells residing in tissues (Refsland et al. 2010). SARS-CoV-2 infection triggers T cell response in infected patients (Le Bert et al. 2020), and the ability of $CD4^+$ T cells to recognize a virus would then allow APOBEC3 enzymes to be packaged into the virions and cause RNA-editing (Harris, Dudley 2015).

We predicted that viral genomes should be driven towards reduced CpG dinucleotides to elude ZAP-mediated cellular antiviral defense, and increased U residues because of RNA editing by APOBEC3 proteins. In line with our expectations, we found compelling hallmarks of CpG deficiency as well as elevated U in tandem with lowered C contents in the genomes of SARS-CoV-2, SARS-COV, MERS, Bovine CoV, CRCoV, and HEV that regularly infected mammalian tissues expressing both AVPs in abundance (Figure 6.4). Unsurprisingly, these sequence trends

were absent from MHV genomes (Figure 6.4) as this virus regularly infects mice tissues that lowly express AVPs (Figure 6.2). Corroborating this observation, both I_{CpG} and P_U values showed the greatest variation among MHV strains (Figure S6.5), suggesting that MHV genomes are not constrained by either AVP. These results suggest that when a virus regularly infects host tissues that are abundant in ZAP and APOBEC3, these AVPs shape the molecular evolution of viral genomes in two ways: CpG deficiency allows the virus to evade ZAP-mediated antiviral defense, and elevated U content due to APOBEC3 editing activity.

Among three human coronaviruses, SARS-CoV-2 genomes exhibit the most CpG deficiency (Figure 6.4a). Many recent studies point to Bat CoV RaTG13 as the most closely related known relative of SARS-CoV-2 when the whole genome is considered (Andersen et al. 2020; Lai et al. 2020; Shang et al. 2020; Tang et al. 2020), and to the bat Rhinolophus affinis as a potential intermediate host or reservoir for SARS-CoV-2 (Liu et al. 2020). Indeed, the I_{CpG} values in SARS-CoV-2 are comparable to that of Bat CoV RaTG13 infecting Rhinolophus affinis but lower than that of many other coronaviruses surveyed (Xia 2020).

Nevertheless, global CpG is not more deficient in SARS-CoV-2 than many other highly pathogenic coronaviruses (di Gioacchino et al. 2020). Despite this, CpG deficiency largely fluctuates in local coding regions (di Gioacchino et al. 2020; Digard et al. 2020; Kim et al. 2020); the S, ORF1ab, and ORF6 regions have the most severe CpG deficiencies (Figure 6.8, $I_{CpG} < 0.4$), whereas the 5' UTR, E, and ORF10 regions have CpG surplus with no signs of CpG deficiency (Figure 6.8, $I_{CpG} > 1$). This may be surprising since one would expect that maintaining high CpG, regardless of its location, should have a detrimental effect on the virus. However, ZAP-mediated RNA degradation is cumulative (Takata et al. 2017). When CpG dinucleotides are added to

individual viral segment 1 or 2 in HIV-1, the inhibitory effect of ZAP is weak, but when the same CpG dinucleotides are added to both segments 1 and 2, the ZAP inhibition effect is strong. This implies that longer RNA sequences (ORF1ab and S) are more likely to be targeted by ZAP.

Moreover, a study on early SARS-CoV-2 genome evolution (di Gioacchino et al. 2020) suggests that the CG-rich N region is biased towards mutations lowering CpG content, whereas the CpG levels remain consistently low in the S region. Nonetheless, we found no notable change in I_{CpG} between 99 SARS-CoV-2 strains sampled in four months since its first detection (Figure 6.8), and occurrences of unique mutations at the CG dinucleotides in the context of the CG-rich CN_XGNCG motifs known to be preferred for ZAP targeting in mice (Luo et al. 2020) were rare (Table 6.4). This suggests that the evolutionary adaptation to CpG deficiency had not been a rapid process. Despite this, the first isolated SARS-CoV-2 genome is deficient in both CG dinucleotides and CN_XGNCG motifs (Table 6.4). Hence, SARS-CoV-2 may have preadapted to a low-CpG human environment as its closest RaTG13 counterpart is likewise CpG deficient (Nchioua et al. 2020). Altogether, our results are consistent with recent studies suggesting that, similar to the HIV-1 genome (Takata et al. 2017), the SARS-CoV-2 genome appears to be CpG deficient to evade ZAP recognition (Nchioua et al. 2020).

While APOBEC3 enzymes are highly expressed in immune cells (Koning et al. 2009; Refsland et al. 2010), they are also detected in mammary and lung epithelial cells (Pauli et al. 2009; Okeoma et al. 2010). An analysis of total RNA at the tissue level (Figure 6.1) showed that APOBEC3 enzymes are highly expressed in healthy lung tissues in comparison to other non-lymphoid tissues; this is consistent with results reported by Koning et al. (2009) and Refsland et al. (2010). Indeed, expressions of APOBEC3 enzymes are not confined to immune organs but are

dependent on tissue lymphocyte contents. To further localize antiviral protein expression, we examined the transcriptomic data of human lung epithelial cells (LoM) in the presence and absence of SARS-CoV-2 infection (Figure 6.3). Our findings were consistent with Pauli et al. (2009) insofar as we observed selective upregulation of A3G to the exclusion of other APOBEC3 paralogues in the lung epithelial cells during viral infection. This suggests that tissue-residing immune cells are predominantly contributing to the APOBEC3 levels observed in tissue total RNA (Figure 6.1), particularly with respect to the high abundance of A3A in lung tissue. It remains likely that tissue-residing immune cells are primarily responsible for variable APOBEC3 expressions at the tissue level.

A survey of complete SARS-CoV-2 genomes did not indicate drastically increased U and decreased C contents (Figure 6.4b). Nonetheless, over the span of four months since the virus was first isolated, there has been a history of P_U elevation and strong bias for C to U mutations relative to other substitutions. These C to U mutations are mostly located in the 5' UTR and ORF1ab regions (Figures 6.5 – 6.7), accounting for over half of all SNPs in SARS-CoV-2. That we observed the same prevalence of C to U mutations in the 5' UTR and ORF1ab regions in strains collected from three different countries (Figure 6.7) suggests that geographic patterns of sampling were not a confounding factor. Indeed, these results suggest that SARS-CoV-2 is consistently biased towards C to U mutations.

Consensus motifs embedding C to U mutations that are acted on by APOBEC3 enzymes have been experimentally verified in HIV-1. To support the hypothesis that C to U mutations in SARS-CoV-2 are catalyzed by APOBEC3 enzymes, we determined that the preferred C to U mutation hotspots in SARS-CoV-2 are the same as those in HIV-1. As summarized in Table 6.2,

most studies have shown that 5' C<u>C</u> and 5' CC<u>C</u> (with C to U mutation sites underlined) are the preferred consensus motifs subjected to RNA editing by A3G in HIV-1 and MLV. As for other APOBEC3 enzymes, 5' T<u>C</u> and 5' TT<u>C</u> are the preferred consensus motifs that are subjected to RNA editing in HIV-1, MLV, and SIV. Similarly, C to U mutations are prevalent in the aforementioned sequence contexts in SARS-CoV-2, suggesting that APOBEC3 enzymes may indeed edit the SARS-CoV-2 genome. Furthermore, among 5' N(C/U)<u>C</u>N mutations in SARS-CoV-2, the APOBEC3-preferred 5' UU<u>C</u>(A/U) were the most commonly observed (Table 6.3) and the embedded 5' U<u>C</u> targets preferred an open structure, akin to what was shown by McDaniel et al. (2020). However, the two 5' CC<u>C</u>(A/U) motifs that are preferably edited by A3G (Table 6.3) were not found and 5' NC<u>C</u>G motifs did not prefer the loop region as shown by Sharma, Baysal (2017). As APOBEC3 enzymes can be efficiently co-packaged into the same viral particle (Desimmie et al. 2016), these results suggest that while all A3A, A3B, A3C, A3F, and A3H could contribute to the prevalence of 5' UC to 5' UU mutations, the effect of A3G alone weakly explains 5' CC to 5' CU mutations in SARS-CoV-2. This is consistent with the observation made by Pauli et al. (2009) regarding A3G exhibiting no antiviral efficacy during influenza A infection despite its upregulation in that context.

It is not excluded that other deaminase enzymes may contribute to RNA editing in SARS-CoV-2. The preferred consensus motif that is subjected to editing by AID is 5' UA<u>C</u> in the *rpoB* gene in *E. coli* (Kohli et al. 2009); this may explain why C to U mutations were also preferred at 5' UA<u>C</u> in SARS-CoV-2 (Table 6.2). In addition, 5' T<u>C</u> is preferentially edited by APOBEC1 in chicken B-cells (Saraconi et al. 2014). Nonetheless, it is unknown whether these enzymes possess the ability to target viral genomes (Kohli et al. 2009; Saraconi et al. 2014). Another noteworthy observation is that A to G mutation was preferred in the S region and the numbers of A to G

mutations in this specific region were increasing over time (Figure 6.6). This mutation may be caused by the ADAR enzyme (Di Giorgio et al. 2020; Jiang 2020), which edits A into I and subsequently into G, in viruses that infect the lungs such as Influenza virus A and Measles virus (Suspène et al. 2011; Ward et al. 2011). Although ADAR was known for targeting double-stranded RNAs and not single-stranded RNAs (Zhao et al. 2004; O'Connell, Mannion, Keegan 2015; Eisenberg, Levanon 2018; Simmonds 2020), the secondary structure of viral genomes often contains local regions of base-pairings as possible substrate for ADAR. A survey of APOBEC1, AID, and ADAR expressions in 27 human tissues (Figure S6.11) shows that APOBEC1 and AID, but not ADAR, are most expressed in the small intestines. None are highly expressed in the lungs, but *ADAR* mRNA expression was upregulated in SARS-CoV-2 infected lung epithelial cells in comparison to the control (Figure 6.3, S6.3). Therefore, in addition to APOBEC3, other host deaminase systems such as A1, AID, and ADAR may act to edit the genome of SARS-CoV-2; however, both AID and A1 deaminases are DNA mutators that are not known to target viruses (Kohli et al. 2009; Saraconi et al. 2014).

The current study focuses on the evolutionary pressure that host immune systems exert onto viral genomes. Our aim is to prompt motivations for vaccine designs in the development of attenuated RNA viruses. Previous experimental works have shown that increasing CpG dinucleotides in CpG-deficient viral genomes drastically decrease viral replication and virulence (Burns et al. 2009; Tulloch et al. 2014; Antzin-Anduetza et al. 2017; Fros et al. 2017; Wasson et al. 2017; Trus et al. 2020), and in recent years several studies have proposed vaccine development strategies involving increased CpG to attenuate RNA viruses (Burns et al. 2009; Tulloch et al. 2014; Ficarelli et al. 2020; Trus et al. 2020). Increasing CpG content may provide a good starting point for strategies to attenuate SARS-CoV-2. On the other hand, because C to U deamination

cannot be proof-read by viral exonuclease Nsp14-ExoN (Eckerle et al. 2010; Smith et al. 2013; Victorovich et al. 2020), host innate deaminases may drive up the rate of evolution in viral genomes (Sadler et al. 2010; Di Giorgio et al. 2020). The possibility of APOBEC3 editing activity and its potential influence on the pathogenesis and drug resistance of viruses such as SARS-CoV-2 in the long term requires further investigation and scrutiny.

6.7 Data availability

Supplemental figures for Chapter 6 are available in Appendix F. Supplemental files are available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244025: Supplemental File S1 is the dataset containing reference compilation of virus regular habitats, tissue total RNA AVP mRNA expressions, and LoM AVP mRNA expressions. Supplemental File S2 is the dataset containing nucleotide and di-nucleotide frequencies in trimmed viral genomes. Supplemental File S3 is the dataset containing nucleotide and di-nucleotide frequencies in whole, un-trimmed, viral genomes. Supplemental File S4 is the dataset contains the global, local, temporal, and geographical SNP patterns in SARS-CoV-2 genomes. Supplemental File S5 is the dataset that contains the local CpG dinucleotide frequencies in a sample of 99 SARS-CoV-2 genomes. Supplemental File S6 is the dataset that contains all sequence and structural context analyses for C to U mutations.

6.8 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to X.X. [RGPIN/2018-03878], and NSERC Doctoral Scholarship to Y.W. [PGSD/2019-535291]. We thank Alibek Kruglikov and Heba Farookhi for discussion.

6.9 References

Adolph, MB, A Ara, Y Feng, CJ Wittkopp, M Emerman, JS Fraser, L Chelico. 2017. Cytidine deaminase efficiency of the lentiviral viral restriction factor APOBEC3C correlates with dimerization. Nucleic Acids Res 45:3378-3394.

Adolph, MB, RP Love, L Chelico. 2018. Biochemical Basis of APOBEC3 Deoxycytidine Deaminase Activity on Diverse DNA Substrates. ACS Infect Dis 4:224-238.

Aguiar, RS, N Lovsin, A Tanuri, BM Peterlin. 2008. Vpr.A3A chimera inhibits HIV replication. J Biol Chem 283:2518-2525.

Andersen, KG, A Rambaut, WI Lipkin, EC Holmes, RF Garry. 2020. The proximal origin of SARS-CoV-2. Nature Medicine 26:450-452.

Antzin-Anduetza, I, C Mahiet, LA Granger, C Odendall, CM Swanson. 2017. Increasing the CpG dinucleotide abundance in the HIV-1 genomic RNA inhibits viral replication. Retrovirology 14:017-0374.

Armitage, AE, A Katzourakis, T de Oliveira, JJ Welch, R Belshaw, KN Bishop, B Kramer, AJ McMichael, A Rambaut, AK Iversen. 2008. Conserved footprints of APOBEC3G on Hypermutated human immunodeficiency virus type 1 and human endogenous retrovirus HERV-K(HML2) sequences. J Virol 82:8743-8761.

Benjamini, Y, Y Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57:289-300.

Bishop, KN, RK Holmes, AM Sheehy, NO Davidson, SJ Cho, MH Malim. 2004. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr Biol 14:1392-1396.

Bolger, AM, M Lohse, B Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bray, NL, H Pimentel, P Melsted, L Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525-527.

Briggs, J, M Paoloni, QR Chen, X Wen, J Khan, C Khanna. 2011. A compendium of canine normal tissue gene expression. PLoS One 6:31.

Burns, CC, R Campagnoli, J Shaw, A Vincent, J Jorba, O Kew. 2009. Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. Journal of Virology 83:9957.

Cardon, LR, C Burge, DA Clayton, S Karlin. 1994. Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci U S A 91:3799-3803.

Chen, J, T MacCarthy. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comput Biol 13:e1005471-e1005471.

Chiu, Y-L, WC Greene. 2008. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. Annual Review of Immunology 26:317-353.

Cullen, BR. 2006. Role and Mechanism of Action of the APOBEC3 Family of Antiretroviral Resistance Factors. Journal of Virology 80:1067.

Cunningham, F, P Achuthan, W Akanni, et al. 2019. Ensembl 2019. Nucleic acids research 47:D745-D751.

Dang, Y, X Wang, WJ Esselman, YH Zheng. 2006. Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. J Virol 80:10522-10533.

Desimmie, BA, RC Burdick, T Izumi, et al. 2016. APOBEC3 proteins can copackage and comutate HIV-1 genomes. Nucleic acids research 44:7848-7865.

di Gioacchino, A, P Šulc, AV Komarova, BD Greenbaum, R Monasson, S Cocco. 2020. The heterogeneous landscape and early evolution of pathogen-associated CpG and UpA dinucleotides in SARS-CoV-2. bioRxiv : the preprint server for biology:2020.2005.2006.074039.

Di Giorgio, S, F Martignano, MG Torcia, G Mattiuz, SG Conticello. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances:eabb5813.

Digard, P, HM Lee, C Sharp, F Grey, E Gaunt. 2020. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. Virus evolution 6.

Doehle, BP, A Schäfer, BR Cullen. 2005. Human APOBEC3B is a potent inhibitor of HIV-1 infectivity and is resistant to HIV-1 Vif. Virology 339:281-288.

Duff, MO, S Olson, X Wei, SC Garrett, A Osman, M Bolisetty, A Plocik, SE Celniker, BR Graveley. 2015. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature 521:376-379.

Ebrahimi, D, H Alinejad-Rokny, MP Davenport. 2014. Insights into the motif preference of APOBEC3 enzymes. PLoS One 9:e87679-e87679.

Eckerle, LD, MM Becker, RA Halpin, et al. 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog 6:1000896.

Eisenberg, E, EY Levanon. 2018. A-to-I RNA editing — immune protector and transcriptome diversifier. Nature Reviews Genetics 19:473-490.

Fagerberg, L, BM Hallström, P Oksvold, et al. 2014. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. Molecular & Cellular Proteomics 13:397.

Feng, Y, TT Baig, RP Love, L Chelico. 2014. Suppression of APOBEC3-mediated restriction of HIV-1 by Vif. Frontiers in microbiology 5:450-450.

Ficarelli, M, I Antzin-Anduetza, R Hugh-White, AE Firth, H Sertkaya, H Wilson, SJD Neil, R Schulz, CM Swanson. 2020. CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. J Virol 94:01337-01319.

Fros, JJ, I Dietrich, K Alshaikhahmed, TC Passchier, DJ Evans, P Simmonds. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. Elife 29:29112.

Gillick, K, D Pollpeter, P Phalora, EY Kim, SM Wolinsky, MH Malim. 2013. Suppression of HIV-1 infection by APOBEC3 proteins in primary human CD4(+) T cells is associated with inhibition of processive reverse transcription as well as excessive cytidine deamination. J Virol 87:1508-1517.

Greenbaum, BD, AJ Levine, G Bhanot, R Rabadan. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 4:1000079.

Guo, X, J Ma, J Sun, G Gao. 2007. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. Proceedings of the National Academy of Sciences 104:151.

Harari, A, M Ooms, LC Mulder, V Simon. 2009. Polymorphisms and splice variants influence the antiretroviral activity of human APOBEC3H. J Virol 83:295-303.

Harris, RS, KN Bishop, AM Sheehy, HM Craig, SK Petersen-Mahrt, IN Watt, MS Neuberger, MH Malim. 2003. DNA deamination mediates innate immunity to retroviral infection. Cell 113:803-809.

Harris, RS, JP Dudley. 2015. APOBECs and virus restriction. Virology 480:131-145.

Hayward, JA, M Tachedjian, J Cui, AZ Cheng, A Johnson, ML Baker, RS Harris, L-F Wang, G Tachedjian. 2018. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. Molecular Biology and Evolution 35:1626-1637.

Hofacker, IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res 31:3429-3431.

Holtz, CM, HA Sadler, LM Mansky. 2013. APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. Nucleic Acids Res 41:6139-6148.

Hultquist, JF, JA Lengyel, EW Refsland, RS LaRue, L Lackey, WL Brown, RS Harris. 2011. Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. Journal of Virology 85:11220-11234.

Jiang, W. 2020. Mutation Profile of Over 4,500 SARS-CoV-2 Isolations Reveals Prevalent Cytosine-to-Uridine Deamination on Viral RNAs. Preprints.

Karlin, S, J Mrázek, AM Campbell. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899-3913.

Katoh, K, DM Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780.

Kim, D, J-Y Lee, J-S Yang, JW Kim, VN Kim, H Chang. 2020. The Architecture of SARS-CoV-2 Transcriptome. Cell 181:914-921.e910.

Kohli, RM, SR Abrams, KS Gajula, RW Maul, PJ Gearhart, JT Stivers. 2009. A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. J Biol Chem 284:22898-22904.

Koning, FA, EN Newman, EY Kim, KJ Kunstman, SM Wolinsky, MH Malim. 2009. Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. J Virol 83:9474-9485.

Lai, C-C, T-P Shih, W-C Ko, H-J Tang, P-R Hsueh. 2020. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. International Journal of Antimicrobial Agents 55:105924.

Lamers, MM, J Beumer, J van der Vaart, et al. 2020. SARS-CoV-2 productively infects human gut enterocytes. Science 369:50-54.

Langlois, MA, RC Beale, SG Conticello, MS Neuberger. 2005. Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. Nucleic Acids Res 33:1913-1923.

Le Bert, N, AT Tan, K Kunasegaran, et al. 2020. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature 584:457-462.

Leung, WK, KF To, PK Chan, HL Chan, AK Wu, N Lee, KY Yuen, JJ Sung. 2003. Enteric involvement of severe acute respiratory syndrome-associated coronavirus infection. Gastroenterology 125:1011-1017.

Liddament, MT, WL Brown, AJ Schumacher, RS Harris. 2004. APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. Curr Biol 14:1385-1391.

Liu, P, JZ Jiang, XF Wan, et al. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog 16.
Lonsdale, JJ ThomasM Salvatore, et al. 2013. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45:580-585.
Luo, X, X Wang, Y Gao, J Zhu, S Liu, G Gao, P Gao. 2020. Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. Cell Rep 30:46-52.
Martinez, T, M Shapiro, S Bhaduri-McIntosh, T MacCarthy. 2019. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus evolution 5:vey040-vey040.
McDaniel, YZ, D Wang, RP Love, MB Adolph, N Mohammadzadeh, L Chelico, LM Mansky. 2020. Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. Nucleic acids research 48:1353-1371.
Meagher, JL, M Takata, D Gonçalves-Carneiro, et al. 2019. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. Proc Natl Acad Sci U S A 116:24303-24309.
Milewska, A, E Kindler, P Vkovski, S Zeglen, M Ochman, V Thiel, Z Rajfur, K Pyrc. 2018. APOBEC3-mediated restriction of RNA virus replication. Scientific Reports 8:5960.
Mitra, M, D Singer, Y Mano, J Hritz, G Nam, RJ Gorelick, I-JL Byeon, AM Gronenborn, Y Iwatani, JG Levin. 2015. Sequence and structural determinants of human APOBEC3H deaminase and anti-HIV-1 activities. Retrovirology 12:3.
Nabel, CS, JW Lee, LC Wang, RM Kohli. 2013. Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. Proceedings of the National Academy of Sciences:201306345.
Naqvi, S, AK Godfrey, JF Hughes, ML Goodheart, RN Mitchell, DC Page. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. Science 365.
Nchioua, R, D Kmiec, JA Müller, et al. 2020. SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans. mBio 11:01930-01920.
Nicholls, JM, LL Poon, KC Lee, et al. 2003. Lung pathology of fatal severe acute respiratory syndrome. Lancet 361:1773-1778.
O'Connell, MA, NM Mannion, LP Keegan. 2015. The Epitranscriptome and Innate Immunity. PLoS genetics 11:e1005687-e1005687.
Odon, V, JJ Fros, N Goonawardane, I Dietrich, A Ibrahim, K Alshaikhahmed, D Nguyen, P Simmonds. 2019. The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. Nucleic acids research 47:8061-8083.
Okeoma, CM, AL Huegel, J Lingappa, MD Feldman, SR Ross. 2010. APOBEC3 proteins expressed in mammary epithelial cells are packaged into retroviruses and can restrict transmission of milk-borne virions. Cell host microbe 8:534-543.
Ooms, M, B Brayton, M Letko, SM Maio, CD Pilcher, FM Hecht, JD Barbour, V Simon. 2013. HIV-1 Vif adaptation to human APOBEC3H haplotypes. Cell host microbe 14:411-421.
Palasca, O, A Santos, C Stolte, J Gorodkin, LJ Jensen. 2018. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database 2018.

Pauli, EK, M Schmolke, H Hofmann, C Ehrhardt, E Flory, C Münk, S Ludwig. 2009. High level expression of the anti-retroviral protein APOBEC3G is induced by influenza A virus but does not confer antiviral activity. Retrovirology 6:1742-4690.

Pimentel, H, P Sturmfels, N Bray, P Melsted, L Pachter. 2016. The Lair: a resource for exploratory analysis of published RNA-Seq data. BMC Bioinformatics 17:490.

Refsland, EW, MD Stenglein, K Shindo, JS Albin, WL Brown, RS Harris. 2010. Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. Nucleic Acids Res 38:4274-4284.

Rodriguez-Frias, F, M Buti, D Tabernero, M Homs. 2013. Quasispecies structure, cornerstone of hepatitis B virus infection: mass sequencing approach. World journal of gastroenterology 19:6995-7023.

Sadler, HA, MD Stenglein, RS Harris, LM Mansky. 2010. APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. J Virol 84:7396-7404.

Saraconi, G, F Severi, C Sala, G Mattiuz, SG Conticello. 2014. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. Genome Biol 15:014-0417.

Shamimuzzaman, M, JJ Le Tourneau, DR Unni, CM Diesh, DA Triant, AT Walsh, A Tayal, GC Conant, DE Hagen, CG Elsik. 2019. Bovine Genome Database: new annotation tools for a new reference genome. Nucleic acids research 48:D676-D681.

Shang, J, G Ye, K Shi, Y Wan, C Luo, H Aihara, Q Geng, A Auerbach, F Li. 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 581:221-224.

Sharma, S, BE Baysal. 2017. Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. PeerJ 6.

Sharma, S, SK Patnaik, RT Taggart, BE Baysal. 2016. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Sci Rep 6.

Sharma, S, SK Patnaik, R Thomas Taggart, ED Kannisto, SM Enriquez, P Gollnick, BE Baysal. 2015. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nature Communications 6:6881.

Sharma, S, J Wang, E Alqassim, S Portwood, E Cortes Gomez, O Maguire, PH Basse, ES Wang, BH Segal, BE Baysal. 2019. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome Biol 20:019-1651.

Sheehy, AM, NC Gaddis, JD Choi, MH Malim. 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418:646-650.

Shi, K, MA Carpenter, S Banerjee, et al. 2017. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. Nat Struct Mol Biol 24:131-139.

Silvas, TV, S Hou, W Myint, E Nalivaika, M Somasundaran, BA Kelch, H Matsuo, N Kurt Yilmaz, CA Schiffer. 2018. Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions. Sci Rep 8:018-25881.

Simmonds, P. 2020. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. bioRxiv : the preprint server for biology:2020.2005.2001.072330.

Smith, EC, H Blanc, MC Surdel, M Vignuzzi, MR Denison. 2013. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog 9:15.

Suspène, R, MM Aynaud, JP Vartanian, S Wain-Hobson. 2013. Efficient deamination of 5-methylcytidine and 5-substituted cytidine residues in DNA by human APOBEC3A cytidine deaminase. PLoS One 8.

Suspène, R, V Petit, D Puyraimond-Zemmour, MM Aynaud, M Henry, D Guétard, C Rusniok, S Wain-Hobson, JP Vartanian. 2011. Double-stranded RNA adenosine deaminase ADAR-1-induced hypermutated genomes among inactivated seasonal influenza and live attenuated measles virus vaccines. J Virol 85:2458-2462.

Takata, MA, D Gonçalves-Carneiro, TM Zang, SJ Soll, A York, D Blanco-Melo, PD Bieniasz. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124-127.

Tang, X, C Wu, X Li, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review.

Trus, I, D Udenze, N Berube, C Wheler, M-J Martel, V Gerdts, U Karniychuk. 2020. CpG-Recoding in Zika Virus Genome Causes Host-Age-Dependent Attenuation of Infection With Protection Against Lethal Heterologous Challenge in Mice. Frontiers in immunology 10:3077-3077.

Tulloch, F, NJ Atkinson, DJ Evans, MD Ryan, P Simmonds. 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. Elife 3:e04531-e04531.

Uhlén, M, L Fagerberg, BM Hallström, et al. 2015. Proteomics. Tissue-based map of the human proteome. Science 347:1260419.

Victorovich, KV, G Rajanish, KT Aleksandrovna, KS Krishna, SA Nicolaevich, PV Vitoldovich. 2020. Translation-associated mutational U-pressure in the first ORF of SARS-CoV-2 and other coronaviruses. bioRxiv : the preprint server for biology:2020.2005.2005.078238.

Wan, L, T Nagata, M Katahira. 2018. Influence of the DNA sequence/length and pH on deaminase activity, as well as the roles of the amino acid residues around the catalytic center of APOBEC3F. Phys Chem Chem Phys 20:3109-3117.

Wang, SM, CT Wang. 2009. APOBEC3G cytidine deaminase association with coronavirus nucleocapsid protein. Virology 388:112-120.

Ward, SV, CX George, MJ Welch, LY Liou, B Hahm, H Lewicki, JC de la Torre, CE Samuel, MB Oldstone. 2011. RNA editing enzyme adenosine deaminase is a restriction factor for controlling measles virus replication that also is required for embryogenesis. Proc Natl Acad Sci U S A 108:331-336.

Wasson, MK, J Borkakoti, A Kumar, B Biswas, P Vivekanandan. 2017. The CpG dinucleotide content of the HIV-1 envelope gene may predict disease progression. Scientific Reports 7:8162-8162.

Wiegand, HL, BP Doehle, HP Bogerd, BR Cullen. 2004. A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. Embo J 23:2451-2458.

Xia, X. 2018. 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.

Yao, XH, ZC He, TY Li, et al. Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient: Cell Res. 2020 Jun;30(6):541-543. doi: 10.1038/s41422-020-0318-5. Epub 2020 Apr 28.

Yi, L, H Pimentel, NL Bray, L Pachter. 2018. Gene-level differential analysis at transcript-level resolution. Genome Biology 19:53.

Yu, Q, R König, S Pillai, K Chiles, M Kearney, S Palmer, D Richman, JM Coffin, NR Landau. 2004. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. Nat Struct Mol Biol 11:435-442.

Yue, FY ChengA Breschi, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355-364.

Zhao, Z, H Li, X Wu, Y Zhong, K Zhang, Y-P Zhang, E Boerwinkle, Y-X Fu. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evolutionary Biology 4:21.

Zhu, Y, G Chen, F Lv, X Wang, X Ji, Y Xu, J Sun, L Wu, Y-T Zheng, G Gao. 2011. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. Proceedings of the National Academy of Sciences 108:15834.

CHAPTER SEVEN

Conclusions

7.1 Coevolution between gene features and translation machinery in bacteria

Rapid bacterial replication depends on efficient translation. To improve translation efficiency, protein-coding genes coevolve with the translation machinery. I studied how coevolution acts on 1) translation initiation, between the Shine-Dalgarno (SD) sequence and the anti-Shine-Dalgarno (anti-SD) sequence, and on 2) translation elongation, between codon usage and tRNA abundance.

A prominent feature of bacterial translation initiation is the SD sequence found in the 5' untranslated region of mRNAs, which pairs with the anti-SD sequence located at the 3' terminus of mature 16S rRNA. A good SD signal that facilitates ribosome recruitment requires optimal SD/anti-SD pairing potential and location. Nonetheless, much about this interaction remains unclear. For example, knowing what constitutes the full extent of the anti-SD sequence is crucial to characterize optimal SD/anti-SD pairing, but the 3' end of 16S rRNA is often mis-characterized in bacteria. We undertook three integrative research investigations to better understand the SD-mediated translation initiation in bacteria. First, using our custom software ARSDA, we devised novel RNA-Seq-based approaches to reliably characterize the 3' end of mature 16S rRNAs with single base specificity. Our results were consistent with experimentally characterized 16S rRNAs and our method rectified 16S rRNA annotation errors in many surveyed species. Second, we discuss a new model ($D_{toStart}$) that explains SD/anti-SD pairing in its biological context to better define optimal complementarity and locational constraints in a species-specific manner. Third, we

applied our model in a case study to reveal evolutionary differences between Cyanobacteria and chloroplast translation initiation. We found that cyanophages well-mimic Cyanobacteria in SD usage because both have been under the same selection pressure for SD-mediated initiation. However, Chloroplasts do not retain optimal SD sequences because the need for SD-mediated initiation has been reduced in plastids as a result of host-symbiont coevolution.

To understand how translation elongation is optimized in bacteria, we studied the degree codon usage is shaped by tRNA availability. The degree to which codon usage depends on tRNA abundance in bacterial species is unclear because tRNA abundance is often inadequately approximated by tRNA copy numbers. To better understand the relationship between these two players, we established a new RNA-Seq-based approach to quantify tRNA levels. Our new estimator of tRNA availability is robust with results consistent with experimentally quantified tRNA abundances. Our method allowed us to 1) elucidate species-specific translationally optimal codons, 2) improve the predictive power of the commonly used tRNA adaptation index, tAI, and 3) explain that tRNA availability is optimized to codon usage in fast-growing species but not in slow-growing species.

7.2 Coronaviruses carry distinct genomic signatures in adaptation to their habitats

How do coronaviruses such as SARS-CoV-2 adapt to their hosts and what triggers host-switching? As obligate parasites, coronaviruses evolve to carry genomic signatures in adaptation to their host-specific environments. Answering these questions from a microbial genome evolution perspective will contribute to the global research efforts to battle against the current COVID-19 pandemic.

How do coronaviruses evolve towards host-specificity during cell entry? Coronaviruses encode a Spike protein that binds to host Angiotensin-converting enzyme 2 (ACE2) receptor to mediate cell entry. A high binding potential between these two players is a key determinant of viral infectivity. While SARS-CoV-2 can transmit efficiently in humans, it is less clear which other mammals are at risk of being infected. We performed large-scale comparative genome analyses to determine which coronaviruses carry SARS-like Spike proteins and which mammalian species carry human-like ACE2 receptors. This investigation aims to improve our ability to predict and control the current pandemic, to manage and protect wildlife and domesticated animals, and to elucidate the origin and evolution of human SARS viruses.

How do coronaviruses evade or adapt to host immune responses once they have entered host cells? Coronaviruses regularly infect host tissues that express antiviral proteins (AVPs) in abundance. Two AVPs that affect viral genomes are the zinc finger antiviral protein (ZAP) and the Apolipoprotein B Editing Complex 3 (APOBEC3). ZAP binds to CpG dinucleotides to facilitate the degradation of viral transcripts, while APOBEC3 prevalently deaminates C into U residues which could generate notable viral sequence variations. It follows that the genome of an infecting virus should trend towards reduced CpG content to evade ZAP and increased C to U mutations due to APOBEC3 editing. We examined changes to the genomic compositions of tissue-specific coronaviruses such as SARS-CoV-2 over their infection time course in distinct cellular environments with differential AVP expressions. We showed that coronaviruses that regularly infect tissues with abundant AVPs have CpG-deficient and U-rich genomes; whereas those that do not infect tissues with abundant AVPs do not share these sequence hallmarks. Understanding the evolutionary pressures exerted by host immune systems onto viral genomes may improve our

ability to control the spread of infection, and we expect our findings to be of interest to researchers seeking novel designs for an attenuated SARS-CoV-2 vaccine.

7.3 Future work on the coevolution between phage-bacteria

The research investigations discussed in this thesis are aimed at advancing our knowledge of microbial molecular biology and evolutionary genomics from two perspectives: 1) how gene features and translation machinery coevolve in bacteria, and 2) distinct genomic signatures that coronaviruses carry in adaptation to their hosts. As a future direction, these concepts apply to phage genome research to understand phage-bacteria adaptation.

Phage as an anti-bacterial agent must be efficient in killing bacteria, and consequently needs to replicate efficiently. As obligate parasites, phages rely on the host translation machinery for protein production (Bull et al. 2004). Their dependence on the host translation machinery entails that phage protein-coding genes adapt and evolve to acquire sequence features characteristic of host protein-coding genes, especially highly expressed ones (Xia 2019). One example of phage mimicry is at the Shine-Dalgarno sequence, as discussed in Chapter 2.

As proof of concept, Figure 7.1 illustrates the distinct SD/anti-SD adaptation between two phages, Phi 29 (encoding 60 genes) and S-SSM6a (encoding 305 genes), infecting *Bacillus subtilis* and *Microcystis aeruginosa*, respectively. Indeed, the distribution of SD/anti-SD pairing locations as measured by $D_{toStart}$ in phages is effectively confined within the same range as their host bacteria. This implies that, if we are to develop an effective phage against bacteria, then we should take into consideration the distinct SD location in $D_{toStart}$ in the host bacteria. Such theoretical predictions are only recently available, predicated on accurate methods to characterize the 3' end of mature

16S rRNA (constituting the anti-SD reference) in the host bacteria to measure $D_{toStart}$ (Wei, Silke, Xia 2017; Silke, Wei, Xia 2018). Nonetheless, while determining phage adaptation at translation initiation in host bacterial systems is expected to facilitate the design of phages with attenuated or optimized fitness and in our understanding on the differences in adaptation between lysogenic and lytic phages, such predictions require experimental validation.

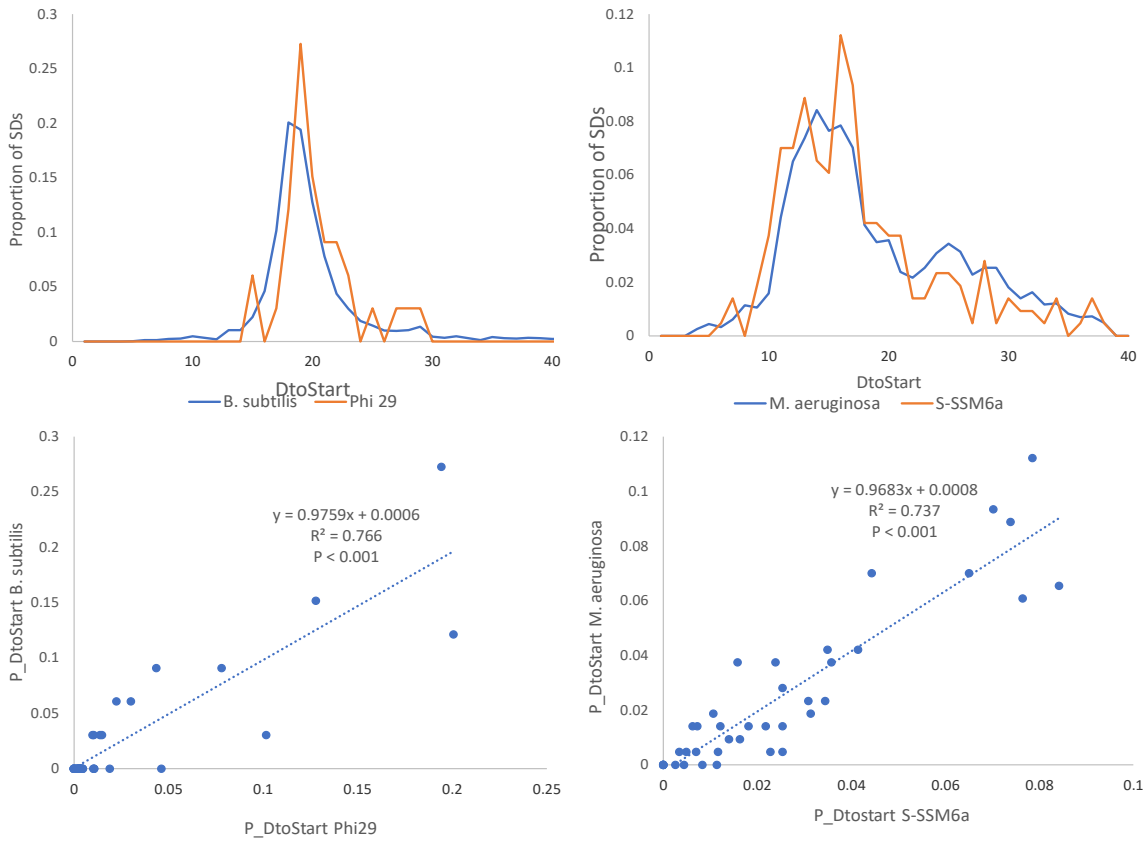


Figure 7.1. The SD/anti-SD pairing constraints as measured by $D_{toStart}$ differ between bacterial species such as *B. subtilis* and *M. aeruginosa*, but these distinct $D_{toStart}$ constraints are well-mimicked by their respective phages (Phi 29 and S-SSM6a). Analyses were conducted using all non-pseudo bacterial and phage genes.

Additionally, codon usage in phage genes should evolve towards mimicking that of host highly expressed genes. Between lytic and lysogenic phages, lysogenic phages are expected to have better codon usage adaptation with their host for two reasons: 1) a prophage and its lysogen share the same mutation rate as the host DNA, and 2) lysogenic phages have increased chance of recombining with host gene segments (Chithambaram, Prabhakaran, Xia 2014). Indeed, a simple measurement of codon usage similarities by the correlation between host and phage RSCUs (R_RSCU) (Chithambaram, Prabhakaran, Xia 2014) shows that lysogenic coliphages better mimic *E. coli* genes in terms of codon usage (Figure 7.2). In addition, one may reason that, if a bacteriophage encodes many tRNA genes in its own genome, especially when these tRNAs are rare in the host, then the phage codon usage will be less dependent on the host tRNA pool (Chithambaram, Prabhakaran, Xia 2014). While presence of phage-encoded tRNA genes can be predicted using computational approaches such as the tRNAscan-SE Search Server (Schattner, Brooks, Lowe 2005), the degree phage-encoded tRNAs affect host tRNA pool can be determined with approaches described in Chapter 3 upon availability of RNA-Sequencing in phage-infected bacteria.

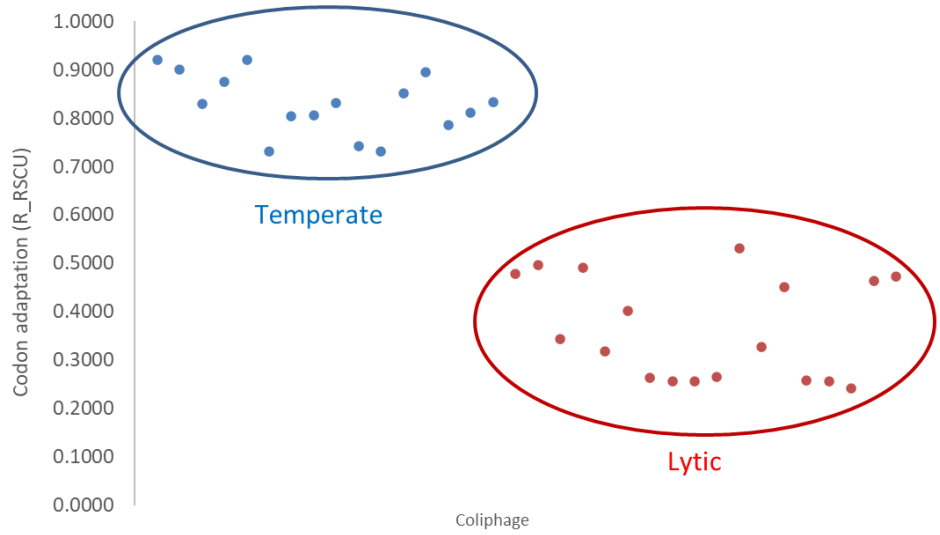


Figure 7.2 Distinct differences in coliphage codon usage adaptation to host *E. coli* genes, as measured by the correlation between RSCU of all phage genes and RSCU of all host genes (R_RSCU). An analysis on 34 coliphages show that temperate (lysogenic) phages can be clearly distinguished from lytic phages with higher degrees of codon usage adaptation to host genes. Data retrieved from Chithambaram, Prabhakaran, Xia (2014).

Above are some examples and concepts in testing phage adaptation by determining the levels that translation initiation and elongation are optimized, using approaches discussed in Chapters 2 and 3, respectively. Together with approaches to measure how gene features affect ribosome dynamics, as described in Chapter 4, will facilitate the development of an integrative computational model that can predict phage fitness and evolution in different host systems. As a general conceptual framework, such a computational model will 1) consider features that optimize translation initiation and elongation efficiencies in phage and bacteria, 2) pinpoint how these features affect ribosome dynamics, and 3) determine the level of phage adaptation to host systems. Additionally, such a model can be further expanded to examine phage host-range and how phage

evolve against anti-phage mechanisms in bacteria, with computational approaches discussed in Chapters 5 and 6, respectively. Together, these investigations will ultimately contribute to genetic engineering to make phage therapy more lethal in the arms race against the evolution of resistance.

7.4. References

Bull, JJ, MR Badgett, R Springman, IJ Molineux. 2004. Genome properties and the limits of adaptation in bacteriophages. Evolution 58:692-701.

Chithambaram, S, R Prabhakaran, X Xia. 2014. Differential codon adaptation between dsDNA and ssDNA phages in Escherichia coli. Mol Biol Evol 31:1606-1617.

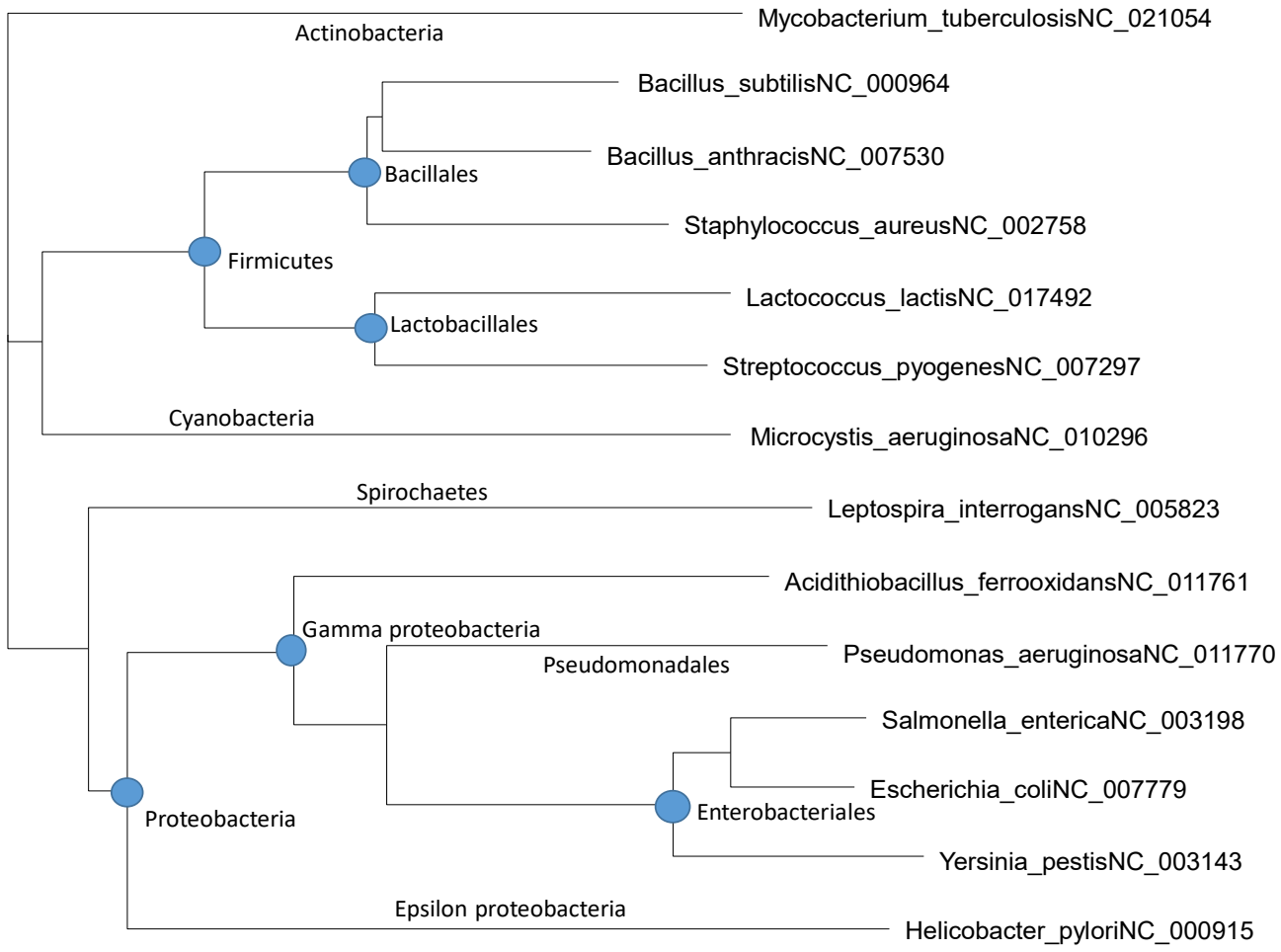
Schattner, P, AN Brooks, TM Lowe. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic acids research 33:W686-W689.

Silke, JR, Y Wei, X Xia. 2018. RNA-Seq-Based Analysis Reveals Heterogeneity in Mature 16S rRNA 3' Termini and Extended Anti-Shine-Dalgarno Motifs in Bacterial Species. G3 8:3973-3979.

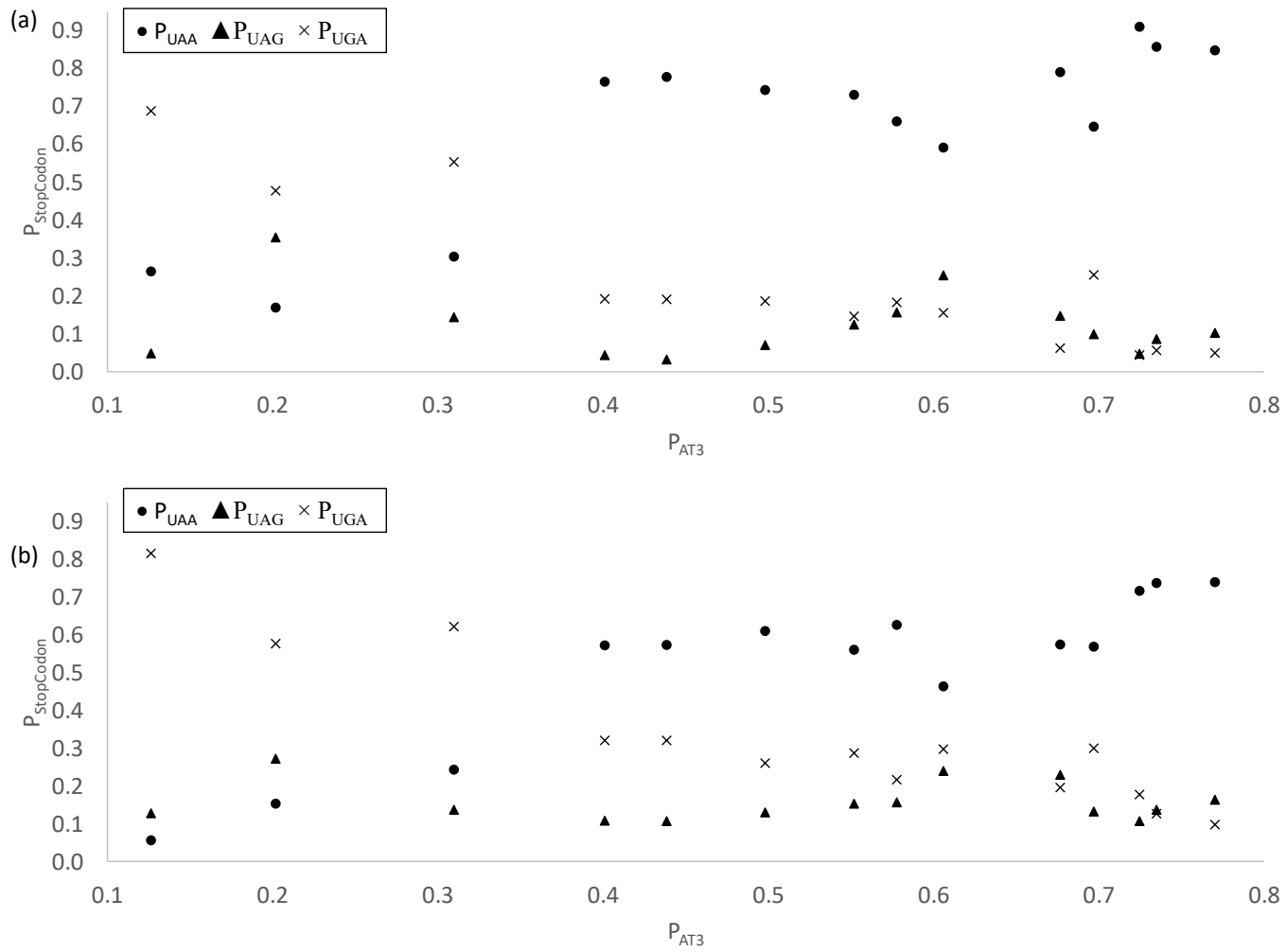
Wei, Y, JR Silke, X Xia. 2017. Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in Escherichia coli and Bacillus subtilis using RNA-Seq data. Sci Rep 7:017-17918.

Xia, X. 2019. Optimizing Phage Translation Initiation. OBM Genetics 3(4).

Appendix A: Supplemental Figures for Chapter 1

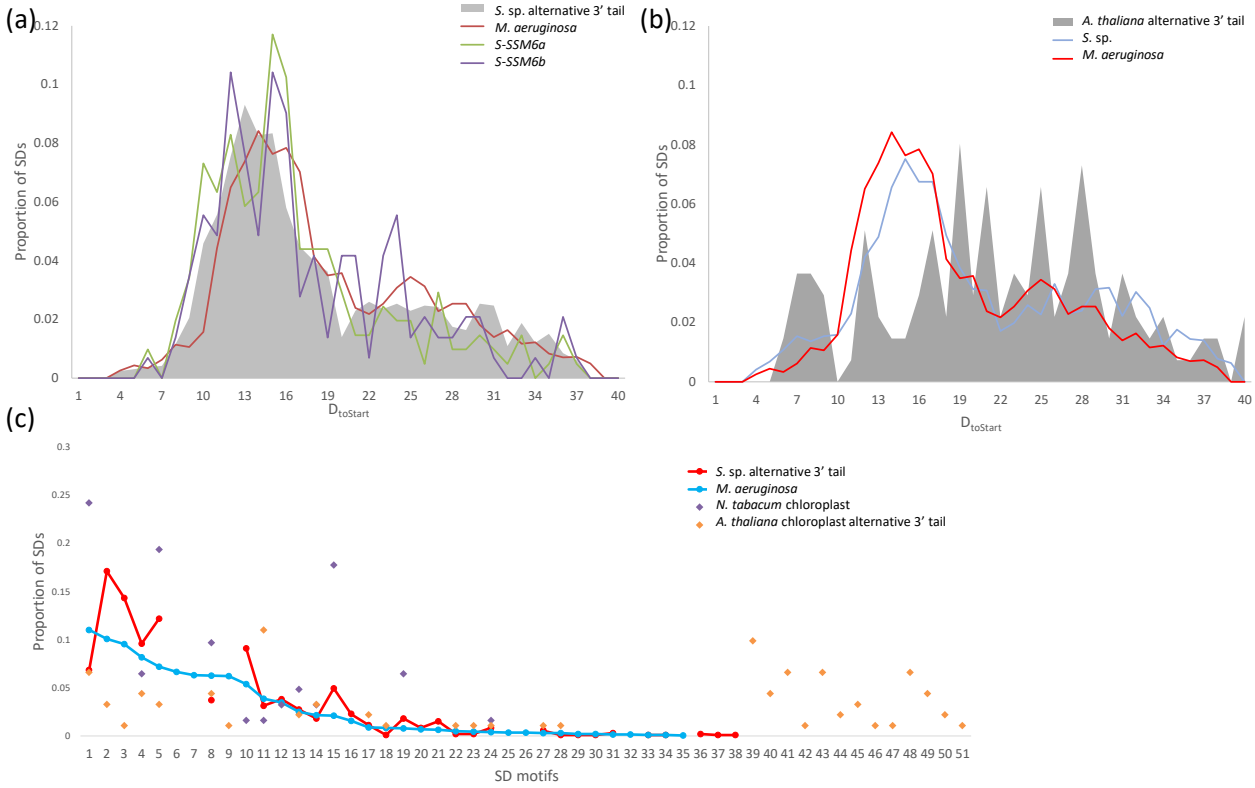


Supplemental Figure S1.1. PHYML phylogenetic tree built with MAFFT G-INS-i aligned small subunit ribosomal RNA sequences (ssu rRNA), with leaves denoted by species name and GenBank accession for genomes from which the ssu rRNA sequences are extracted. Only the first annotated ssu rRNA sequence is used. Adapted from Wei et al. 2016.



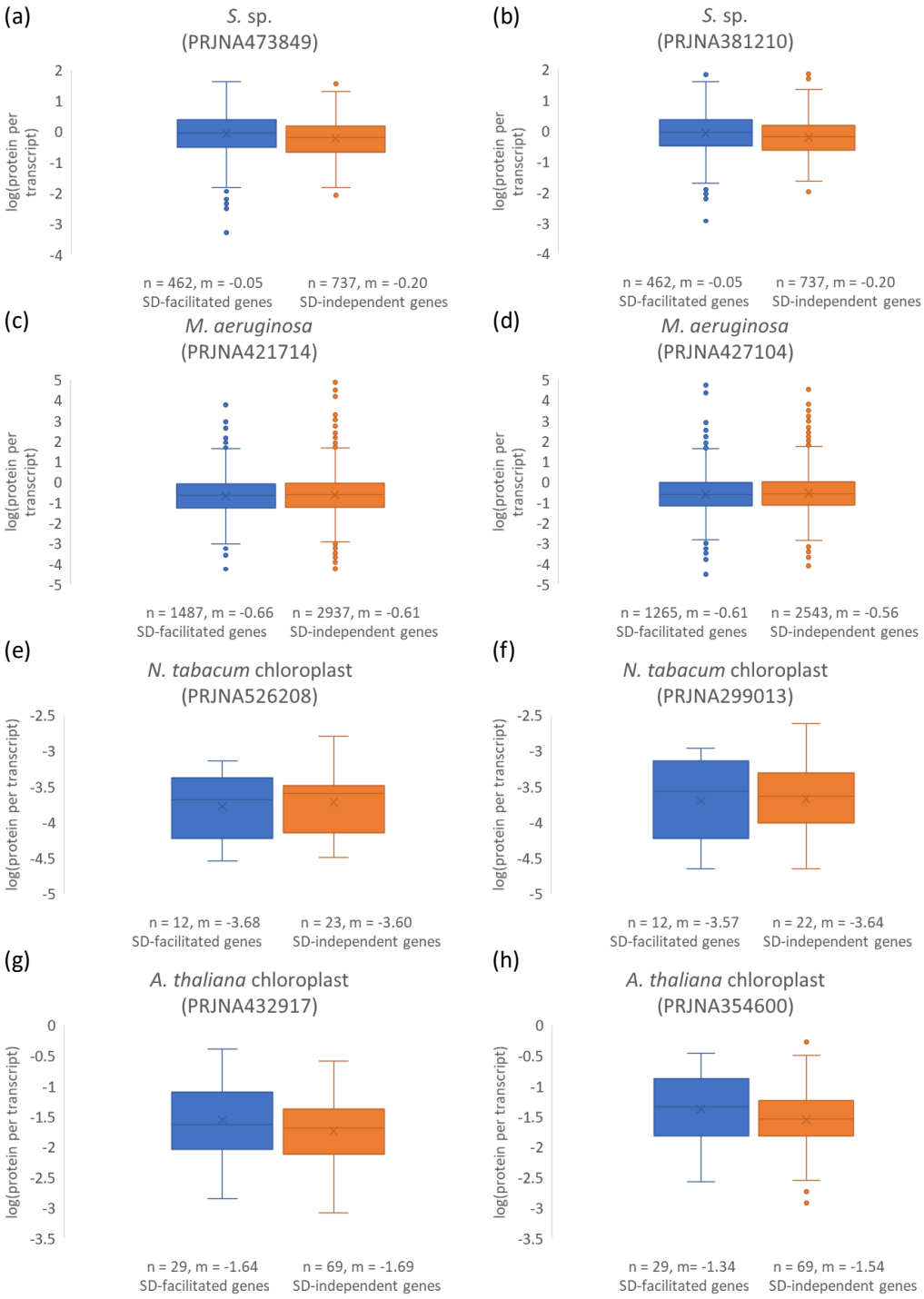
Supplemental Figure S1.2. In 14 bacterial species, UAA usage increases, UGA usage decreases, and UAG usage changes little with P_{AT3}, in a) HEGs and b) LEGs. Adapted from Wei et al. 2016.

Appendix B: Supplemental Figures for Chapter 2



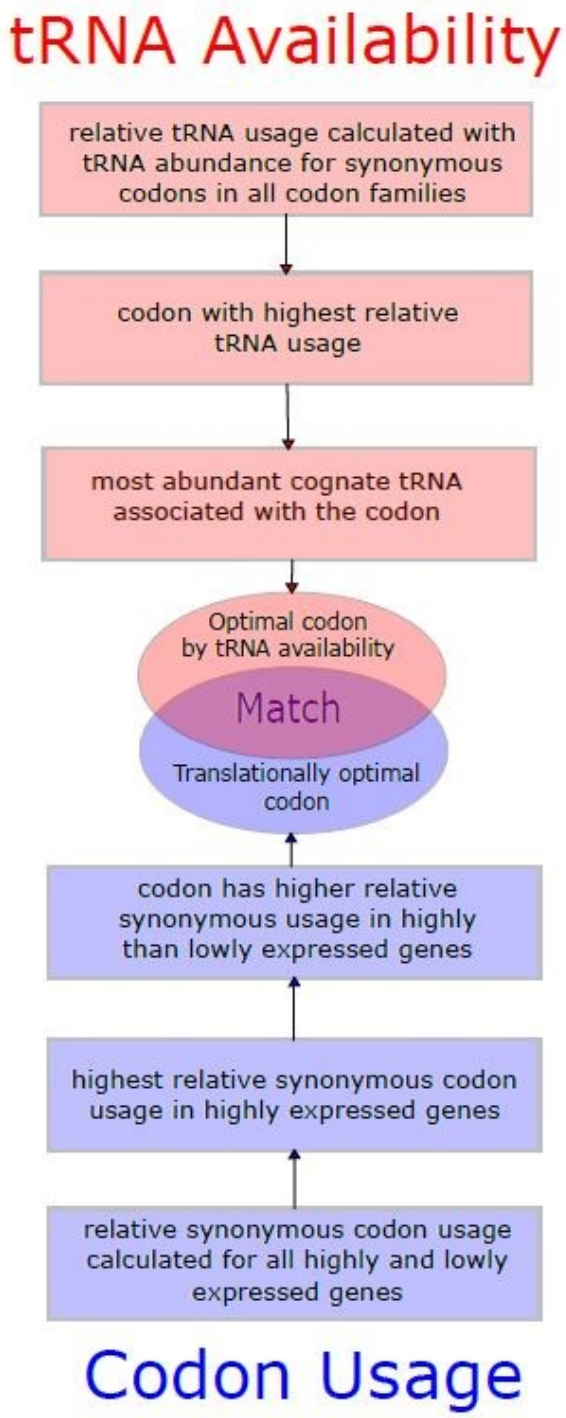
Supplemental Figure S2.1. Comparisons in $D_{toStart}$ constraints and SD motif usage among species when the alternative 3' tail sequences were used for *S.* sp. and *A. thaliana* chloroplast SD determination. (a) $D_{toStart}$ constraints resemble among Cyanobacteria and Cyanophages, but (b) differ between Cyanobacteria and *A. thaliana* chloroplasts. (c) SD motif usage at optimal $D_{toStart}$ is comparable between the two Cyanobacteria (*S.* sp. and *M. aeruginosa*) but differ between Cyanobacteria and chloroplasts. In (a) and (c) the alternative 16S rRNA 3' tail sequence 5'-GAUCACCUCCU-3' in *S.* sp. was used to determine putative SD sequences in *S.* sp. and Cyanophages S-SSM6a and S-SSM6b. In (b) the alternative 16S rRNA 3' tail sequence 5'-GAUCACCUCCUUUUCAG-3' was used to determine putative SD sequences form all CDSs in

A. thaliana chloroplasts. For all species, all non-pseudo coding DNA-Sequences are considered. All data points presented have non-zero SD proportions.

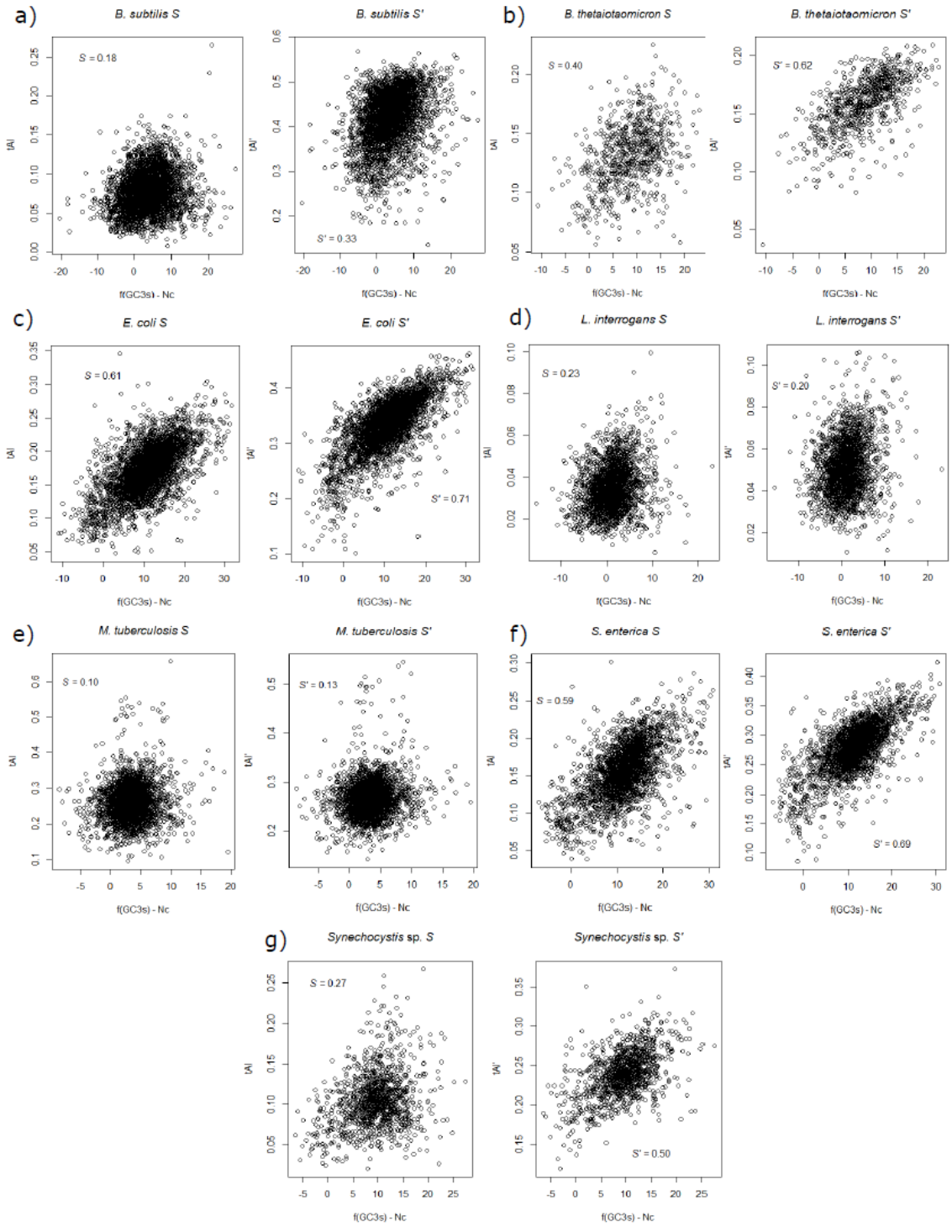


Supplemental Figure S2.2. Contrasts in protein per transcript between SD-facilitated genes (SD/anti-SD pair ≥ 4nt and SD within optimal $D_{toStart}$) and all others (SD-independent genes). All genes are non-pseudo with protein abundance > 0. mRNA transcript expressions (in average FPKM) were calculated using all wild type and control RNA-Seq experiments within the same study, with BioProject accession listed below species name. For each gene set, the number of genes (n) and median (m) are denoted under the boxplots.

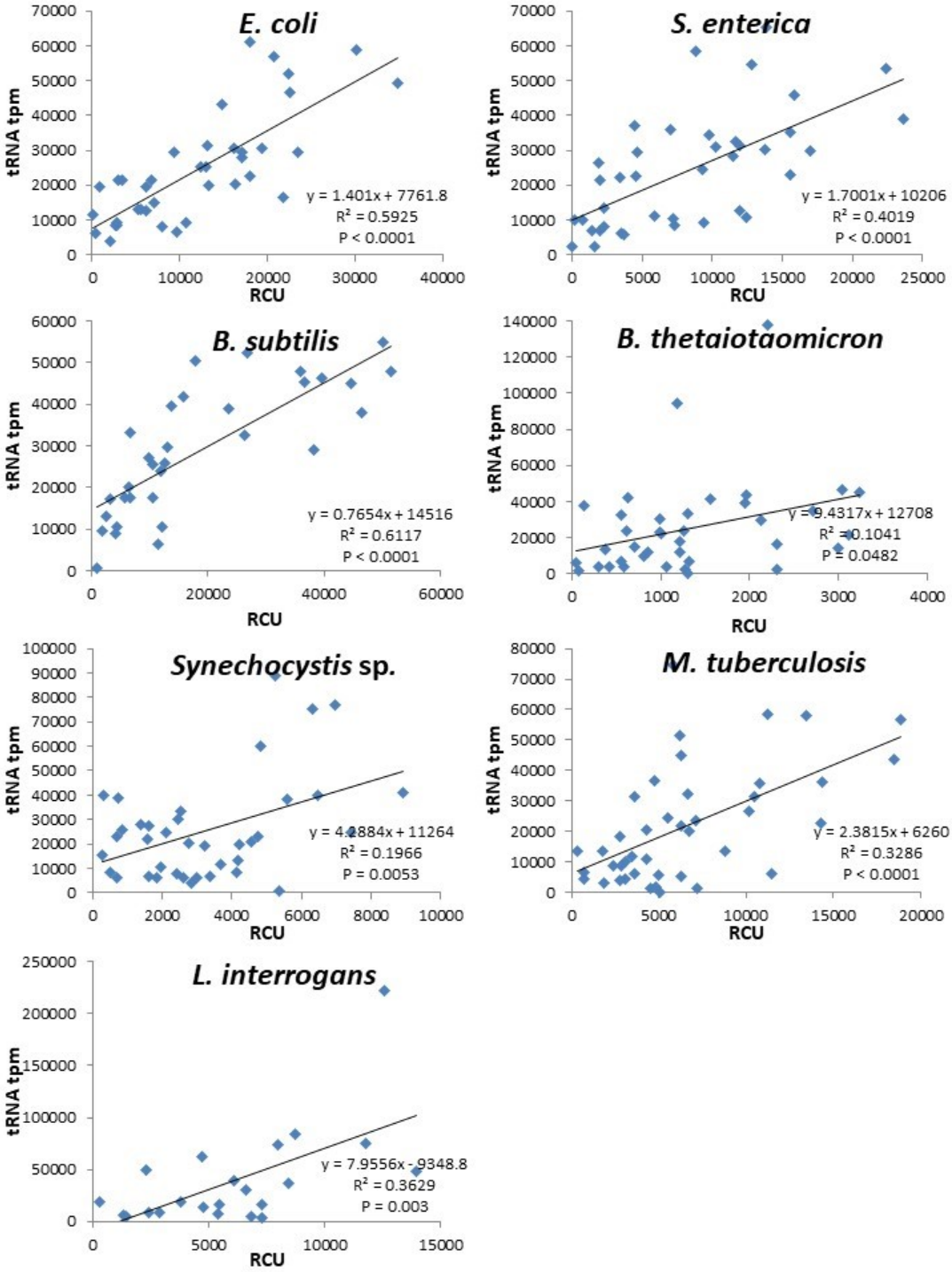
Appendix C: Supplemental Table and Figures for Chapter 3



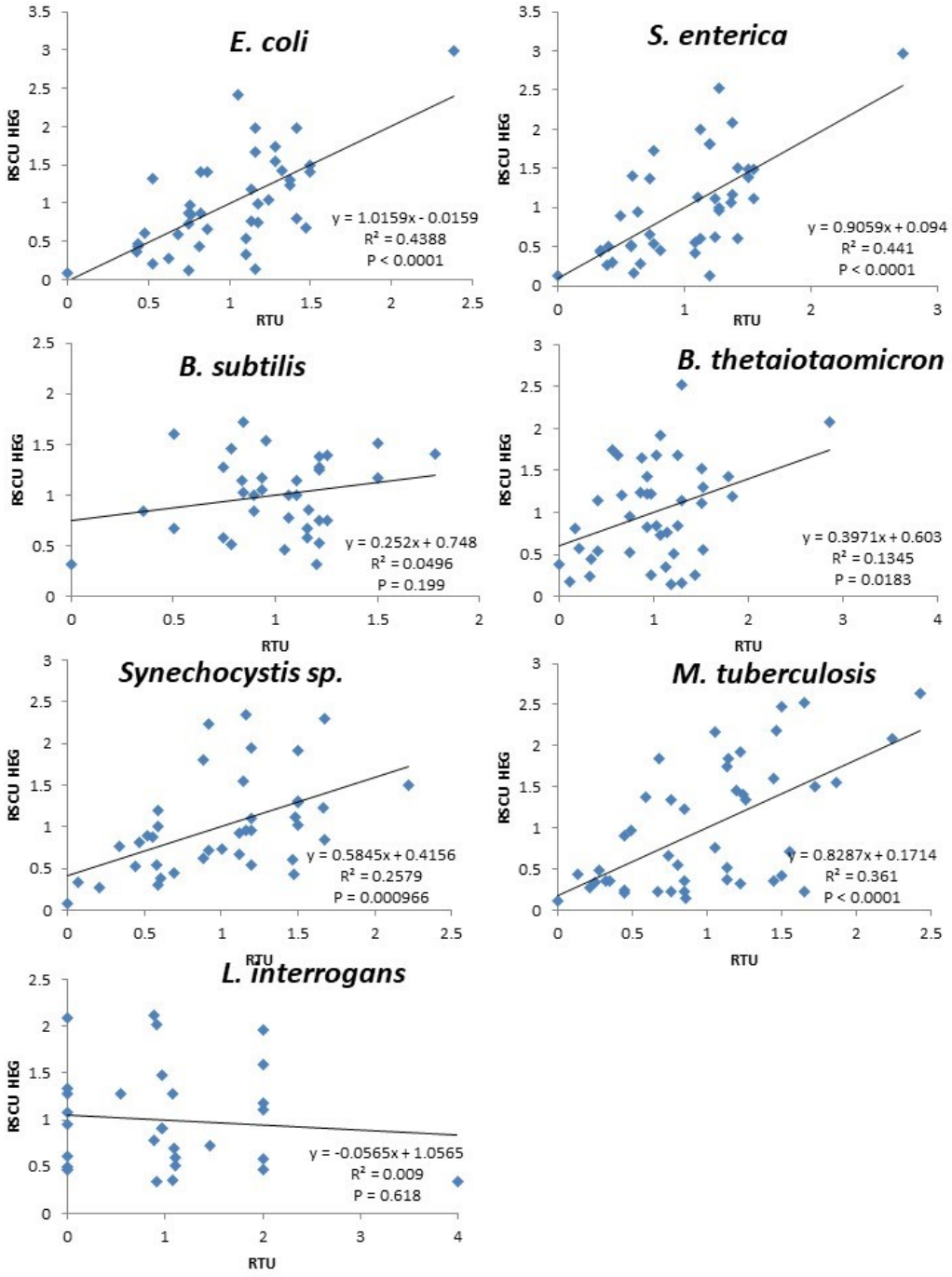
Supplemental Figure S3.1. A summary of methods in designating optimal codon through both tRNA availability and codon usage.



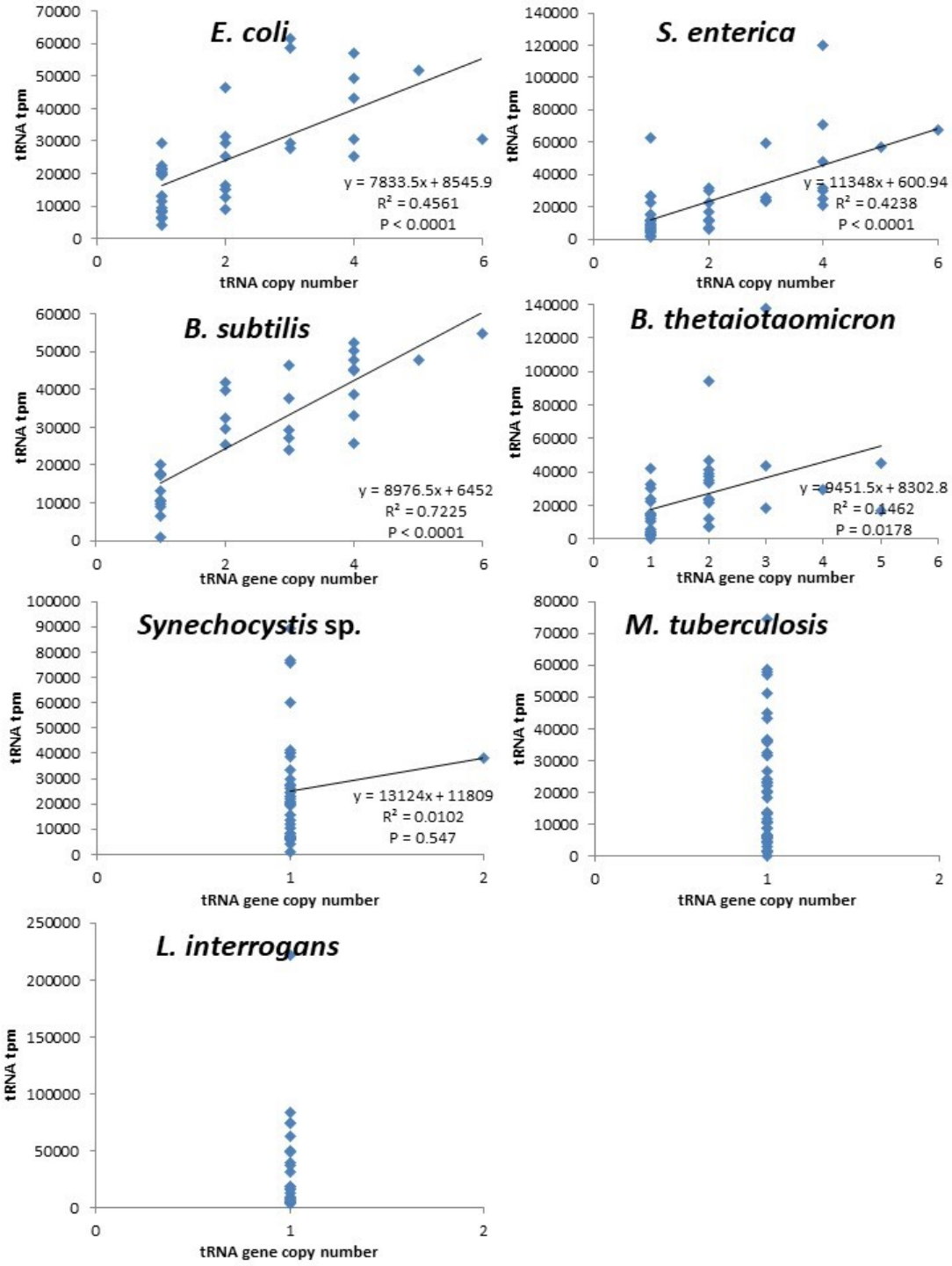
Supplemental Figure S3.2. The relationshipe between corrected Nc and tAI/tAI' values for all seven species. The *S* and *S'* correlation coefficients are shown on each plot. Within each species, tAI and tAI' differ signiificantly from one another ($p < 0.0001$).



Supplemental Figure S3.3. Relationship between readable codon usage (RCU) and tRNA tpm in HEGs determined by protein per transcript, in three fast-growing species (*E.* coli, *S.* enterica, and *B. subtilis*) and four slow-growing species (*B. thetaiotaomicron*, *S.* sp., *M. tuberculosis*, and *L. interrogans*).



Supplemental Figure S3.4. Relationship between relative tRNA tpm usage (TPU) and RSCU in HEGs determined by protein per transcript in three fast-growing species (*E.* coli, *S.* enterica, and *B. subtilis*) and four slow-growing species (*B. thetaiotaomicron*, *S.* sp., *M. tuberculosis*, and *L. interrogans*). Synonymous codons with RTU = 1 are results of lack of difference in total readable (cognate and wobble pairing) tRNA tpm and are omitted.



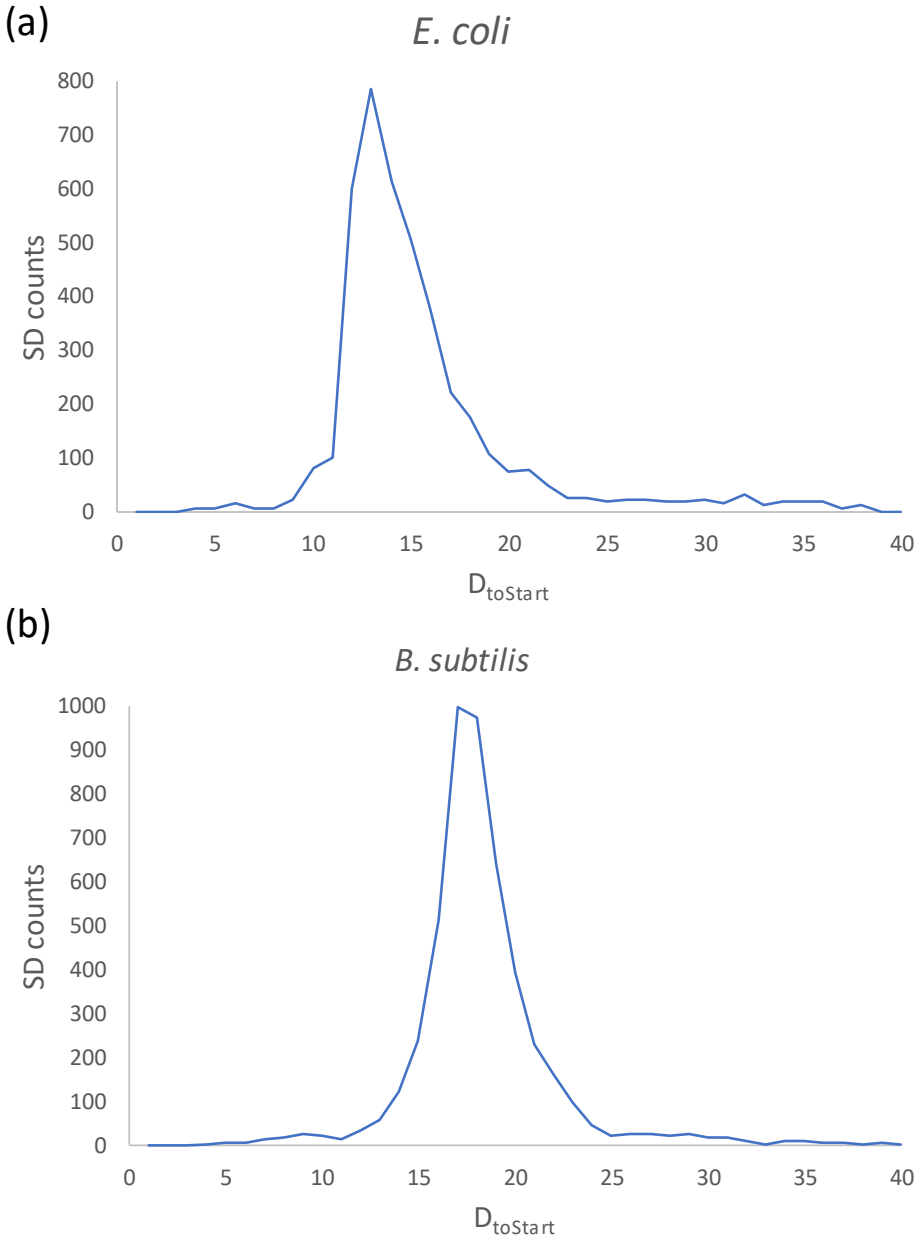
Supplemental Figure S3.5. Relationship between relative tRNA gene copy number and tRNA tpm values in three fast-growing species (*E.* coli, *S.* enterica, and *B. subtilis*) and four slow-growing species (*B. thetaiotaomicron*, *S.* sp., *M. tuberculosis*, and *L. interrogans*).

Supplemental Table S1. Number of genes determined to be translationally highly expressed (HEG) and lowly expressed (LEG) based on protein per transcript. The selected HEGs and LEGs are all genes found in the top and bottom 30% of genes by protein per transcript, respectively, in both SRX datasets. The numbers of ribosomal protein genes contained in top and bottom 30% ppm/tpm gene sets are also shown for each dataset.

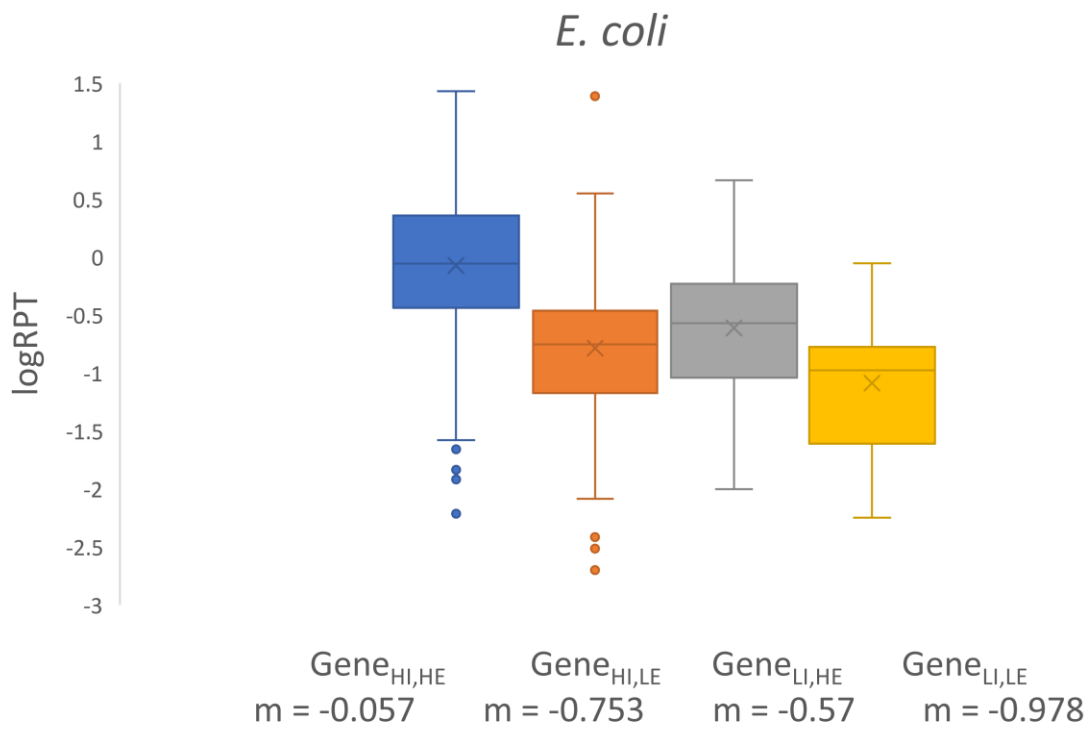
		Number of genes		**Ribosomal protein genes**			**Matching genes between the both gene sets**	
Species	SRX ID	Top 30% ppm/tpm	Bottom 30% ppm/tpm	Number of genes	In top 30% ppm/tpm	In bottom 30% ppm/tpm	Number of HEGs	Number of LEGs
E. coli	SRX515174	1110	1110	57	50	0	976	962
	SRX669653	1119	1119		55	0		
B. subtilis*	SRX515181	1056	1056	57	42	0	1056	1056
	SRX2804667	N/A	N/A		23	0		
B. thetaiotaomicron	SRX020805	186	186	55	17	5	114	150
	SRX860738	230	230		24	1		
L. interrogans	SRX2448246	735	735	54	39	1	505	543
	SRX405952	672	672		31	1		
M. tuberculosis	SRX1372108	995	995	58	48	4	866	901
	SRX4374910	996	996		48	3		
S. enterica	SRX1638989	866	866	59	45	1	649	646
	SRX1258668	784	784		52	0		
Synechocystis sp.	SRX347145	400	400	55	43	1	306	308
	SRX4145044	360	360		42	1		

*- *B. subtilis* SRX2804667 was not considered due to Illumina MiSeq protocol removing large RNA transcripts (e.g. protein-coding gene transcripts).

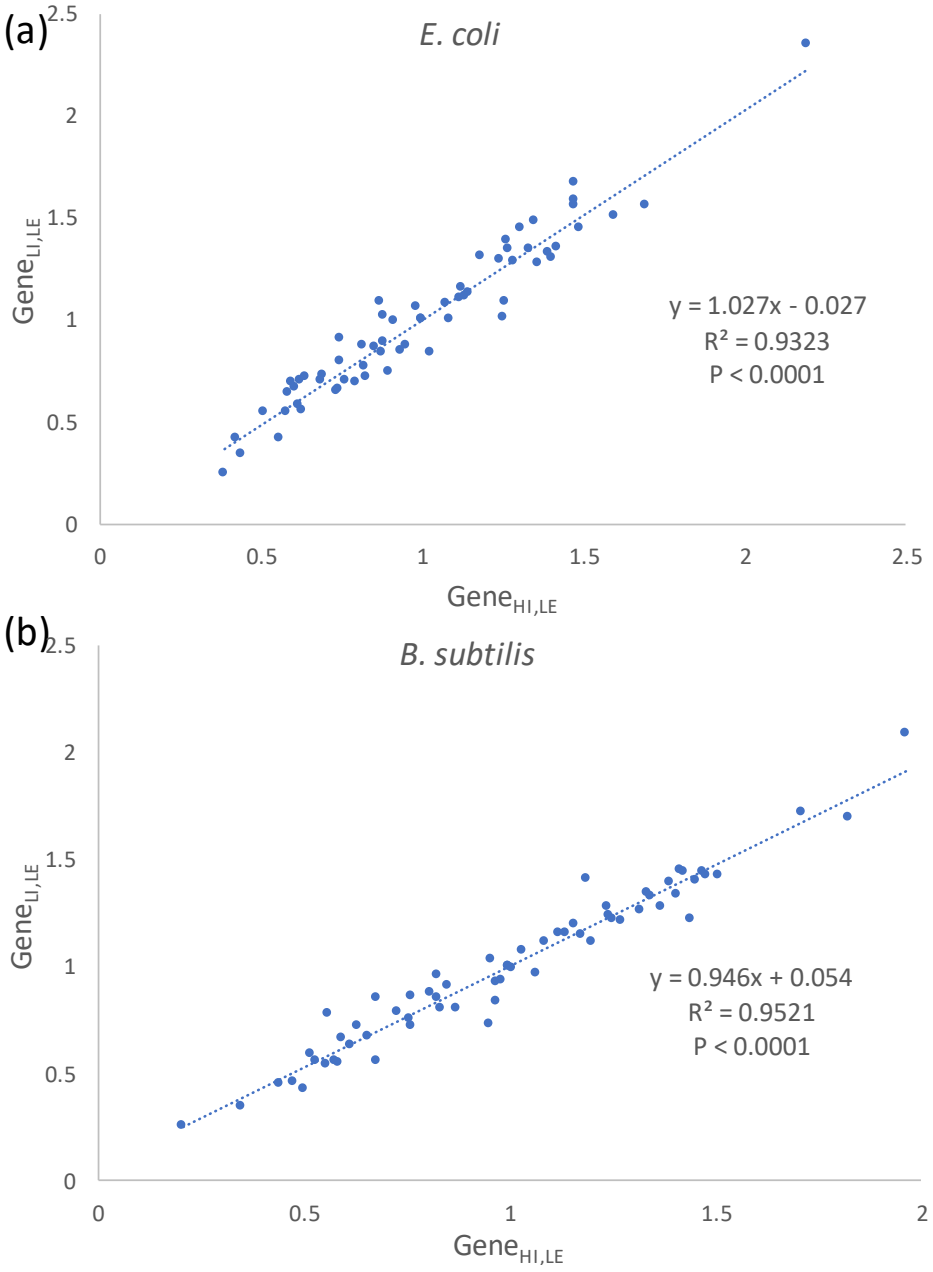
Appendix D: Supplemental Table and Figures for Chapter 4



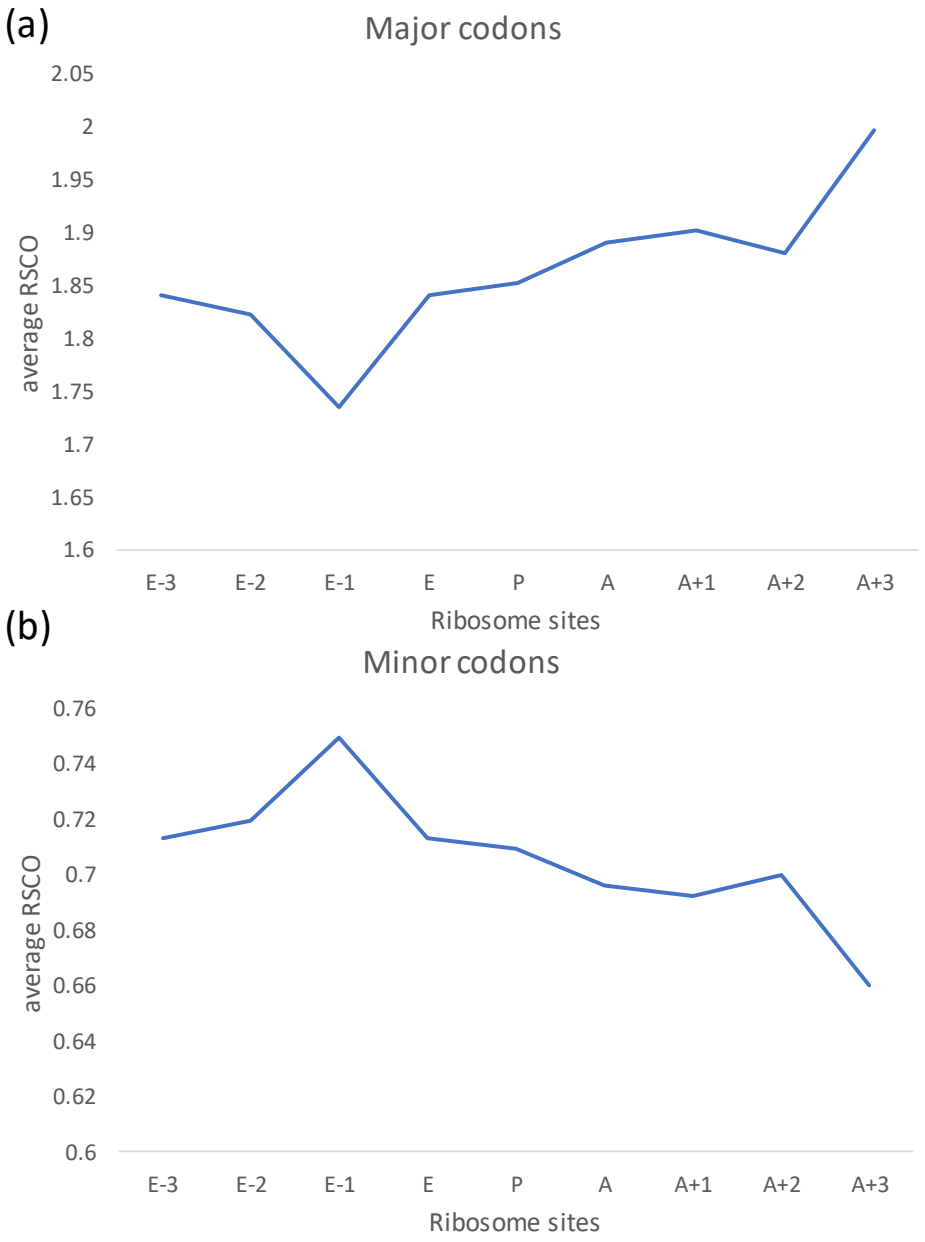
Supplemental Figure S4.1. SD $D_{toStart}$ is strongly constrained within a narrow range (sites 10 – 21 in *E. coli* and 15 – 21 in *B. subtilis*). All 4nt to 12nt putative SD sequences pairing to anti-SD sequence 5'-GAUCACCUCCUUA-3' in *E. coli* and 5'-GAUCACCUCCUUUCU-3' in *B. subtilis* were determined from all non-pseudo genes.



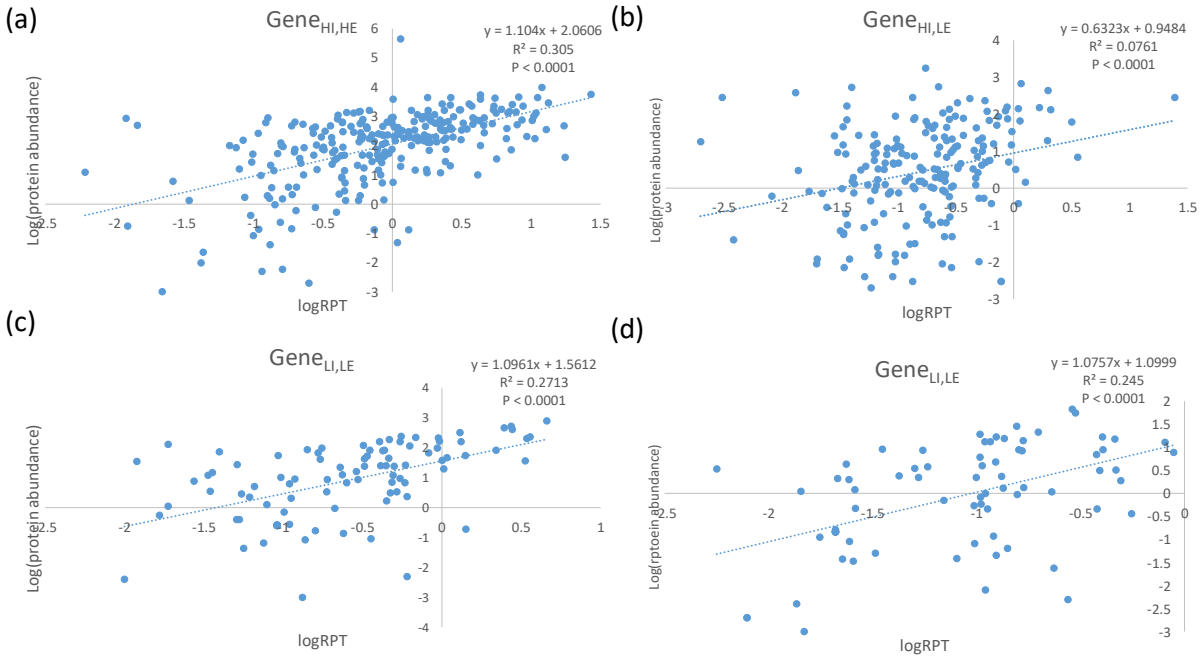
Supplemental Figure S4.2. Ribosomes per transcript (logRPT) in genes differing in translation initiation or elongation efficiencies in *E. coli*. logRPT was calculated as log(average ribosome FPKM/RNA FPKM) from Ribo-Seq datasets SRX4615480 and SRX4615482 and RNA-Seq dataset SRX515174. FPKM values were calculated in ARSDA. m denotes the median.



Supplemental Figure S4.3. Relative synonymous codon usage (RSCU) for 61 sense codons and 3 stop codons well correlate between genes in $Gene_{HI,LE}$ and genes in $Gene_{LI,LE}$.

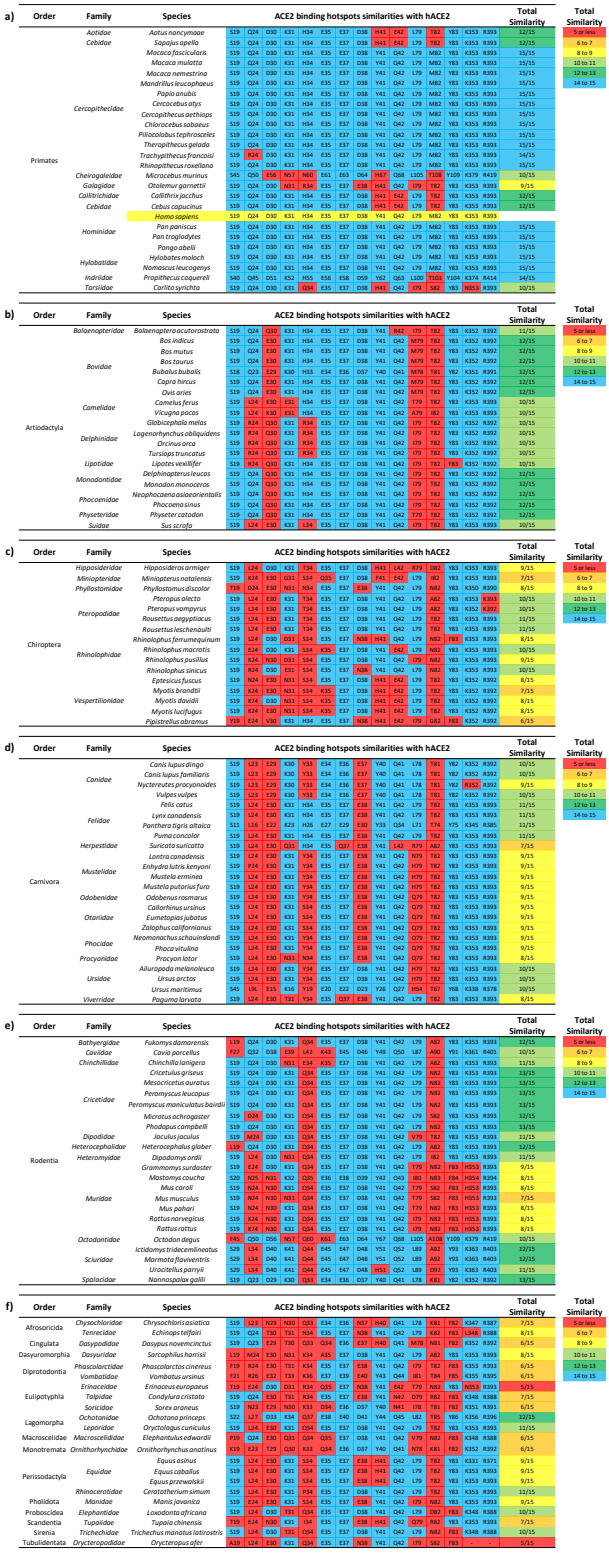


Supplemental Figure S4.4. The averaged site-specific ribosome occupancies over the entire CDSs (all from non-pseudo genes) for 15 major codons and 44 minor codons in *E. coli*, with Ribo-Seq dataset SRX4615480. There is no clear evidence that average $RSCO_A$ significantly differ from RSCO at other sites in both codon groups.

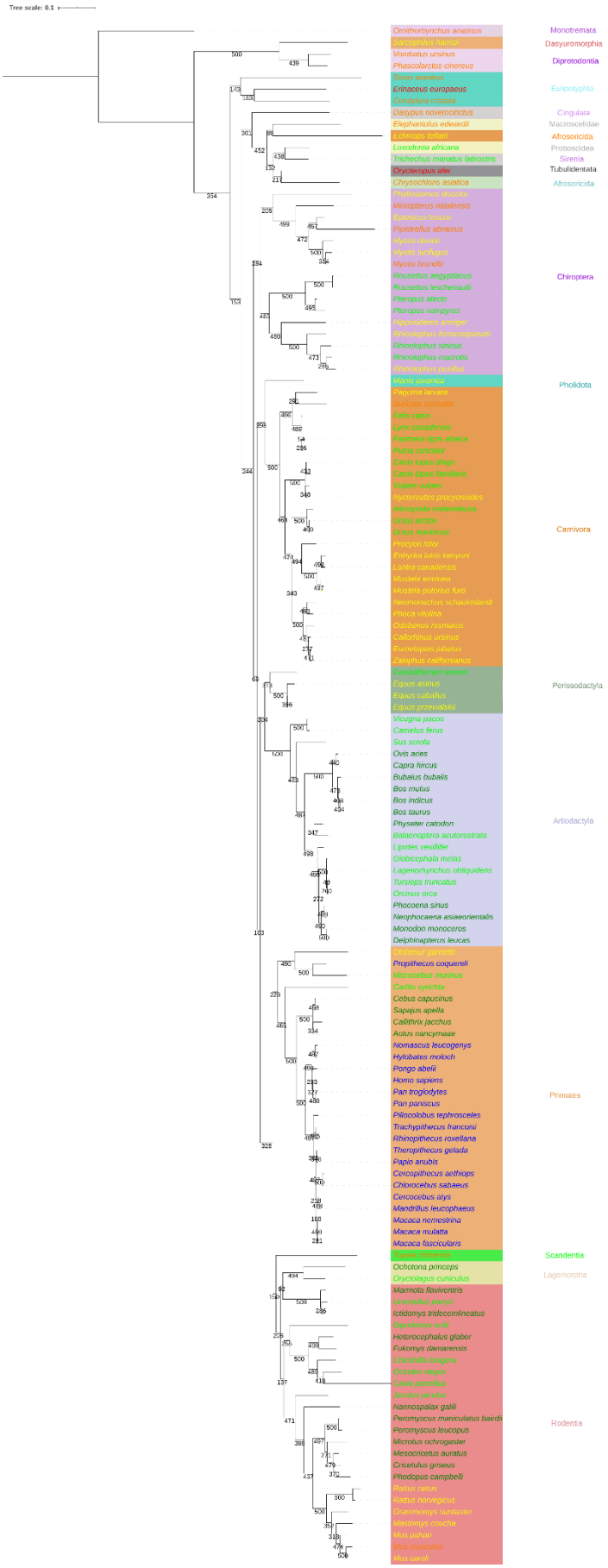


Supplemental Figure S4.5. Relationship between *E. coli* log protein abundance (Integrated protein abundance from PaxDB) and logRPT (calculated from Ribo-Seq datasets SRX4615480 and SRX4615482, and from RNA-Seq dataset SRX515174), in four sets of genes separated by translation initiation and elongation efficiencies. All gene sets contain non-pseudo and non-hypothetical genes with protein abundance > 0 (See Figure 4.2 for number of genes in each set).

Appendix E: Supplemental Figures for Chapter 5



Supplemental Figure S5.1. Comparisons of *ACE2* genes in 132 mammalian species at 15 key human ACE2 sites involved in SARS-CoV-2 S protein binding. These species belong to the orders a) Primates, b) Artiodactyla, c) Chiroptera, d) Canivora, and e) Rodentia, and in panel f) displays species belonging to 14 other orders. All *ACE2* genes are aligned using MAFFT with option G-INS-i. Highlighted blue and red are matching and mis-matching amino acid residues, respectively, between mammalian *ACE2* and human *ACE2*. Total similarity designates the total number of matching amino acid residues with respect to human *ACE2*. The total similarity score for mammalian *ACE2* in blue highlights perfect site similarities (14 to 15 matching sites), in green highlights high similarity (12 to 13 matching sites), in light green highlights medium-high similarity (10 to 11 matching sites), in yellow highlights medium similarity (8 to 9 matching sites), in orange highlights medium-low similarity (6 to 7 matching sites), and in red highlights low similarity (5 or less matching sites).


Tree scale: 0.1
Ornithorhynchus anatinus
Sarcophilus harrisii
Vombatus ursinus
Phascolarctos cinereus
Sorex araneus
Erinaceus europaeus
Condylura cristata
Dasypus novemcinctus
Elephantulus edwardii
Echinops telfairi
Loxodonta africana
Trichechus manatus latirostris
Orycteropus afer
Chrysochloris asiatica
Pteropus alecto
Pteropus vampyrus
Rhinolophus sinicus
Rhinolophus macrotis
Manis javanica
Panthera tigris altaica
Puma concolor
Canis lupus dingo
Canis lupus familiaris
Vulpes vulpes
Ursus arctos
Ursus maritimus
Procyon lotor
Lontra canadensis
Mustela putorius furo
Neomonachus schauinslandi
Phoca vitulina
Odobenus rosmarus
Callorhinus ursinus
Eumetopias jubatus
Ceratotherium simum
Equus asinus
Equus caballus
Equus przewalskii
Vicugna pacos
Camelus ferus
Sus scrofa
Ovis aries
Capra hircus
Bubalus bubalis
Bos mutus
Bos indicus
Bos taurus
Physeter catodon
Balaenoptera acutorostrata
Lipotes vexillifer
Globicephala melas
Lagenorhynchus obliquidens
Tursiops truncatus
Orcinus orca
Phocoena sinus
Neophocaena asiaeorientalis
Monodon monoceros
Delphinapterus leucas
Propithecus coquereli
Macaca nemestrina
Macaca mulatta
Macaca fascicularis
Ochotona princeps
Oryctolagus cuniculus
Marmota flaviventris
Ictidomys tridecemlineatus
Dipodomys ordii
Heterocephalus glaber
Fukomys damarensis
Chinchilla lanigera
Octodon degus
Cavia porcellus
Jaculus jaculus
Nannospalax galili
Peromyscus maniculatus bairdii
Peromyscus leucopus
Microtus ochrogaster
Mesocricetus auratus
Cricetulus griseus
Rattus norvegicus
Mus pahari
Mus musculus
Monotremata
Dasyuromorphia
Diprotodontia
Eulipotyphla
Cingulata
Macroscelidae
Afrosoricida
Proboscidea
Sirenia
Tubulidentata
Chiroptera
Pholidota
Carnivora
Perissodactyla
Artiodactyla
Primates
Scandentia
Lagomorpha
Rodentia

Supplemental Figure S5.2. Phylogenetic reconstruction based on 132 MAFFT G-INS-i aligned mammalian *ACE2* amino acid sequences. The tree is built with the maximum-likelihood-based PHYML approach, with best substitution model = JTT + G + I + F, Bootstrap = 500, and Outgroup = *Ornithorhynchus anatinus*. Mammalian species names are colored based on total similarity at *ACE2* against human *ACE2*: in blue highlights perfect site similarity (14 to 15 matching sites), in green highlights high similarity (12 to 13 matching sites), in light green highlights medium-high similarity (10 to 11 matching sites), in yellow highlights medium similarity (8 to 9 matching sites), in orange highlights medium-low similarity (6 to 7 matching sites), and in red highlights low similarity (5 or less matching sites).

Order	Family	Species	ACE2 binding hotspots similiarties with hACE2									Total Similarity
Afrosoricida	*Chrysochloridae*	*Chrysochloris asiatica*	N30	H40	K81	F82	S83	K347	D348	D349	F350	R351	4
	Tenrecidae	*Echinops telfairi*	T31	Y41	K82	F83	S84	L348	N349	D350	F351	R352	4
Artiodactyla	*Balaenopteridae*	*Balaenoptera acutorostrata*	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
	Bovidae	*Bos indicus*	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
		Bos mutus	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
		Bos taurus	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
		Bubalus bubalis	K30	Y40	T81	Y82	S83	K351	G352	D353	F354	R355	8
		Capra hircus	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
		Ovis aries	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
	Camelidae	*Camelus ferus*	E31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	8
		Vicugna pacos	E31	Y41	I82	Y83	P84	K353	G354	D355	F356	R357	8
	Delphinidae	*Globicephala melas*	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
		Lagenorhynchus obliquidens	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
		Orcinus orca	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
		Tursiops truncatus	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
	Lipotidae	*Lipotes vexillifer*	K31	Y41	T82	F83	P84	K352	G353	D354	F355	R356	8
	Monodontidae	*Delphinapterus leucas*	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
		Monodon monoceros	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
	Phocoenidae	*Neophocaena asiaeorientalis*	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
		Phocoena sinus	K31	Y41	T82	Y83	S84	K352	G353	D354	F355	R356	8
	Physeteridae	*Physeter catodon*	K31	Y41	T82	Y83	P84	K352	G353	D354	F355	R356	9
	Suidae	*Sus scrofa*	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
Carnivora	*Canidae*	*Canis lupus dingo*	K30	Y40	T81	Y82	P83	K352	G353	D354	F355	R356	9
		Canis lupus familiaris	K30	Y40	T81	Y82	P83	K352	G353	D354	F355	R356	9
		Nyctereutes procyonoides	K30	Y40	T81	Y82	P83	R352	G353	D354	F355	R356	8
		Vulpes vulpes	K30	Y40	T81	Y82	P83	K352	G353	D354	F355	R356	9
	Felidae	*Felis catus*	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
		Lynx canadensis	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
		Panthera tigris altaica	K23	Y33	T74	Y75	P76	K345	G346	D347	F348	R349	9
		Puma concolor	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
	Herpestidae	*Suricata suricatta*	Q31	Y41	A82	Y83	P84	K353	G354	D355	F356	R357	8
	Mustelidae	*Lontra canadensis*	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
		Enhydra lutris kenyoni	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
		Mustela erminea	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
		Mustela putorius furo	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
	Odobenidae	*Odobenus rosmarus*	K31	Y41	T82	Y83	P84	K353	H354	D355	F356	R357	8
	Otariidae	*Callorhinus ursinus*	K31	Y41	T82	Y83	P84	K353	H354	D355	F356	R357	8
		Eumetopias jubatus	K31	Y41	T82	Y83	P84	K353	H354	D355	F356	R357	8
		Zalophus californianus	K31	Y41	T82	Y83	P84	K353	H354	D355	F356	R357	8
	Phocidae	*Neomonachus schauinslandi*	K31	Y41	T82	Y83	P84	K353	H354	D355	F356	R357	8
		Phoca vitulina	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
	Procyonidae	*Procyon lotor*	N31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	8
	Ursidae	*Ailuropoda melanoleuca*	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
		Ursus arctos	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
		Ursus maritimus	K16	Y26	T67	Y68	P69	K338	G339	D340	F341	R342	9
	Viverridae	*Paguma larvata*	T31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	8
Chiroptera	*Hipposideridae*	*Hipposideros armiger*	K31	H41	D82	Y83	R84	K353	G354	D355	F356	R357	7
	Miniopteridae	*Miniopterus natalensis*	G31	F41	I82	Y83	P84	K353	G354	D355	F356	R357	7
	Phyllostomidae	*Phyllostomus discolor*	N31	Y41	N82	Y83	P84	K350	K351	D352	F353	R354	7
	Pteropodidae	*Pteropus alecto*	K31	Y41	A82	Y83	Q84	K353	G354	D355	F356	R357	8
		Pteropus vampyrus	K31	Y41	A82	Y83	Q84	K352	G353	D354	F355	R356	8
		Rousettus aegyptiacus	K31	Y41	T82	Y83	Q84	K353	G354	D355	F356	R357	8
		Rousettus leschenaulti	K31	Y41	T82	Y83	Q84	K353	G354	D355	F356	R357	8
	Rhinolophidae	*Rhinolophus ferrumequinum*	D31	H41	N82	F83	S84	K353	G354	D355	F356	R357	5
		Rhinolophus macrotis	K31	Y41	N82	Y83	P84	K353	G354	D355	F356	R357	9
		Rhinolophus pusillus	D31	Y41	N82	Y83	P84	K353	G354	D355	F356	R357	8
		Rhinolophus sinicus	E31	Y41	N82	Y83	P84	K353	G354	D355	F356	R357	8
	Vespertilionidae	*Eptesicus fuscus*	N31	H41	T82	Y83	P84	K352	N353	D354	F355	R356	6
		Myotis brandtii	N31	H41	T82	Y83	P84	K352	G353	D354	F355	R356	7
		Myotis davidii	N31	H41	T82	Y83	P84	K352	G353	D354	F355	R356	7
		Myotis lucifugus	N31	H41	T82	Y83	P84	K352	G353	D354	F355	R356	7
		Pipistrellus abramus	K31	H41	G82	F83	P84	K352	N353	D354	F355	R356	6
Cingulata	*Dasypodidae*	*Dasypus novemcinctus*	T30	H40	N81	F82	S83	K352	G353	D354	F355	R356	5
Dasyuromorphia	*Dasyuridae*	*Sarcophilus harrisii*	N31	Y41	A82	Y83	P84	K353	G354	D355	F356	R357	8
Diprotodontia	*Phascolarctidae*	*Phascolarctos cinereus*	T31	Y41	T82	F83	P84	K353	G354	D355	F356	R357	7
	Vombatidae	*Vombatus ursinus*	T33	Y43	T84	F85	P86	K355	G356	D357	F358	R359	7
Eulipotyphla	*Erinaceidae*	*Erinaceus europaeus*	D31	Y41	N82	Y83	P84	N353	G354	D355	F356	R357	7
	Talpidae	*Condylura cristata*	T31	Y41	R82	F83	Q84	K348	G349	D350	F351	R352	6
	Soricidae	*Sorex araneus*	N30	Y40	T81	F82	P83	K351	N352	D353	Y354	R355	5
Lagomorpha	*Ochotonidae*	*Ochotona princeps*	K34	Y44	T85	Y86	P87	K356	D357	D358	F359	R360	8
	Leporidae	*Oryctolagus cuniculus*	K31	Y41	T82	Y83	P84	K353	R354	D355	F356	R357	8
Macroscelidae	*Macroscelididae*	*Elephantulus edwardii*	Q31	Y41	N82	F83	P84	K348	G349	D350	F351	R352	7
Monotremata	*Ornithorhynchidae*	*Ornithorhynchus anatinus*	Q30	Y40	K81	F82	N83	K352	N353	D354	F355	R356	5
Perissodactyla	*Equidae*	*Equus asinus*	K31	H41	T82	Y83	P84	K331	G332	D333	F334	R335	8
		Equus caballus	K31	H41	T82	Y83	P84	K353	G354	D355	F356	R357	8
		Equus przewalskii	K31	H41	T82	Y83	P84	K353	G354	D355	F356	R357	8
	Rhinocerotidae	*Ceratotherium simum*	K31	Y41	T82	Y83	P84	K353	G354	D355	F356	R357	9
Pholidota	*Manidae*	*Manis javanica*	K31	Y41	N82	Y83	Q84	K353	H354	D355	F356	R357	7
Primates	*Aotidae*	*Aotus nancymaae*	K31	H41	T82	Y83	P84	K353	Q354	D355	F356	R357	6
	Cebidae	*Sapajus apella*	K31	H41	T82	Y83	P84	K353	Q354	D355	F356	R357	7
	Cercopithecidae	*Macaca fascicularis*	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Cercocebus atys	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Cercopithecus aethiops	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Chlorocebus sabaeus	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Macaca mulatta	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Macaca nemestrina	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Mandrillus leucophaeus	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Papio anubis	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Piliocolobus tephrosceles	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Theropithecus gelada	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Rhinopithecus roxellana	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Trachypithecus francoisi	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
	Cheirogaleidae	*Microcebus murinus*	N57	H67	T108	Y109	P110	K379	G380	D381	F382	R383	7
	Galagidae	*Otolemur garnettii*	N31	H41	T82	Y83	P84	K353	D354	D355	F356	R357	6
	Callitrichidae	*Callithrix jacchus*	K31	H41	T82	Y83	P84	K353	Q354	D355	F356	R357	7
	Cebidae	*Cebus capucinus*	K31	H41	T82	Y83	P84	K353	Q354	D355	F356	R357	7
	Hominidae	*Homo sapiens*	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	
		Pan paniscus	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Pan troglodytes	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Pongo abelii	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
	Hylobatidae	*Hylobates moloch*	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
		Nomascus leucogenys	K31	Y41	M82	Y83	P84	K353	G354	D355	F356	R357	10
	Indriidae	*Propithecus coquereli*	K52	Y62	T103	Y104	P105	K374	G375	D376	F377	R378	9
	Tarsiidae	*Carlito syrichta*	K31	H41	S82	Y83	P84	N353	S354	D355	F356	R357	6
Proboscidea	*Elephantidae*	*Loxodonta africana*	T31	Y41	D82	F83	P84	K348	G349	D350	F351	R352	7
Rodentia	*Bathyergidae*	*Fukomys damarensis*	K31	Y41	A82	Y83	P84	K353	N354	D355	F356	R357	8
	Caviidae	*Cavia porcellus*	E39	Y49	A90	Y91	P92	K361	N362	D363	F364	R365	7
	Chinchillidae	*Chinchilla lanigera*	N31	Y41	A82	Y83	P84	K353	D354	D355	F356	R357	7
	Cricetidae	*Cricetulus griseus*	K31	Y41	N82	Y83	S84	K353	G354	D355	F356	R357	8
		Microtus ochrogaster	K31	Y41	S82	Y83	S84	K353	D354	D355	F356	R357	7
		Mesocricetus auratus	K31	Y41	N82	Y83	S84	K353	G354	D355	F356	R357	8
		Peromyscus leucopus	K31	Y41	N82	Y83	S84	K353	G354	D355	F356	R357	8
		Peromyscus maniculatus bairdii	K31	Y41	N82	Y83	S84	K353	G354	D355	F356	R357	8
		Phodopus campbelli	K31	Y41	N82	Y83	P84	K353	E354	D355	F356	R357	8
	Dipodidae	*Jaculus jaculus*	K31	Y41	T82	Y83	P84	K353	N354	D355	F356	R357	8
	Heterocephalidae	*Heterocephalus glaber*	K31	Y41	A82	Y83	S84	K353	D354	D355	F356	R357	7
	Heteromyidae	*Dipodomys ordii*	N31	Y41	I82	Y83	S84	K353	G354	D355	F356	R357	7
	Muridae	*Grammomys surdaster*	K31	Y41	N82	F83	S84	H353	G354	D355	F356	R357	6
		Mastomys coucha	K32	Y42	N83	F84	S85	H354	G355	D356	F357	R358	6
		Mus caroli	K31	Y41	S82	F83	S84	H353	G354	D355	F356	R357	6
		Mus musculus	N31	Y41	S82	F83	S84	H353	G354	D355	F356	R357	5
		Mus pahari	K31	Y41	N82	F83	S84	H353	G354	D355	F356	R357	6
		Rattus norvegicus	K31	Y41	N82	F83	S84	H353	G354	D355	F356	R357	6
		Rattus rattus	K31	Y41	N82	F83	S84	H353	G354	D355	F356	R357	6
	Octodontidae	*Octodon degus*	N57	Y67	A108	Y109	P110	K379	N380	D381	F382	R383	7
	Sciuridae	*Ictidomys tridecemlineatus*	K41	Y51	A92	Y93	P94	K363	G364	D365	F366	R367	9
		Marmota flaviventris	K41	Y51	A92	Y93	P94	K363	G364	D365	F366	R367	9
		Urocitellus parryii	K41	H51	D92	Y93	P94	K363	G364	D365	F366	R367	8
Scandentia	*Spalacidae*	*Nannospalax galili*	K30	Y40	K81	Y82	S83	K352	G353	D354	F355	R356	8
Sirenia	*Tupaiidae*	*Tupaia chinensis*	K31	H41	R82	Y83	P84	K353	N354	D355	F356	R357	7
Tubulidentata	*Trichechidae*	*Trichechus manatus latirostris*	T31	Y41	N82	F83	P84	K348	G349	D350	F351	R352	7
	Orycteropodidae	*Orycteropus afer*	K31	Y41	S82	F83	S84	G346	K3547	G348	D349	F350	2

Total similarity
0 to 1
2 to 3
4 to 5
6 to 7
8 to 9
10

Supplemental Figure S5.3. Comparisons of *ACE2* genes in 132 mammalian species at 10 key human ACE2 sites involved in SARS-CoV S protein binding. These species belong to the orders Primates, Artiodactyla, Chiroptera, Canivora, Rodentia, and 14 other orders. All *ACE2* genes are aligned using MAFFT with option G-INS-i. Highlighted blue and red are matching and mis-matching amino acid residues, respectively, between mammalian *ACE2* and human *ACE2*. Total similarity designates the total number of matching amino acid residues with respect to human *ACE2*. The total similarity score for mammalian *ACE2* in blue highlights perfect site similarities (10 matching sites), in green highlights high similarity (8 to 9 matching sites), in light green highlights medium-high similarity (6 to 7 matching sites), in yellow highlights medium similarity (4 to 5 matching sites), in orange highlights medium-low similarity (2 to 3 matching sites), and in red highlights low similarity (1 or less matching sites).

Appendix F: Supplemental Table and Figures for Chapter 6

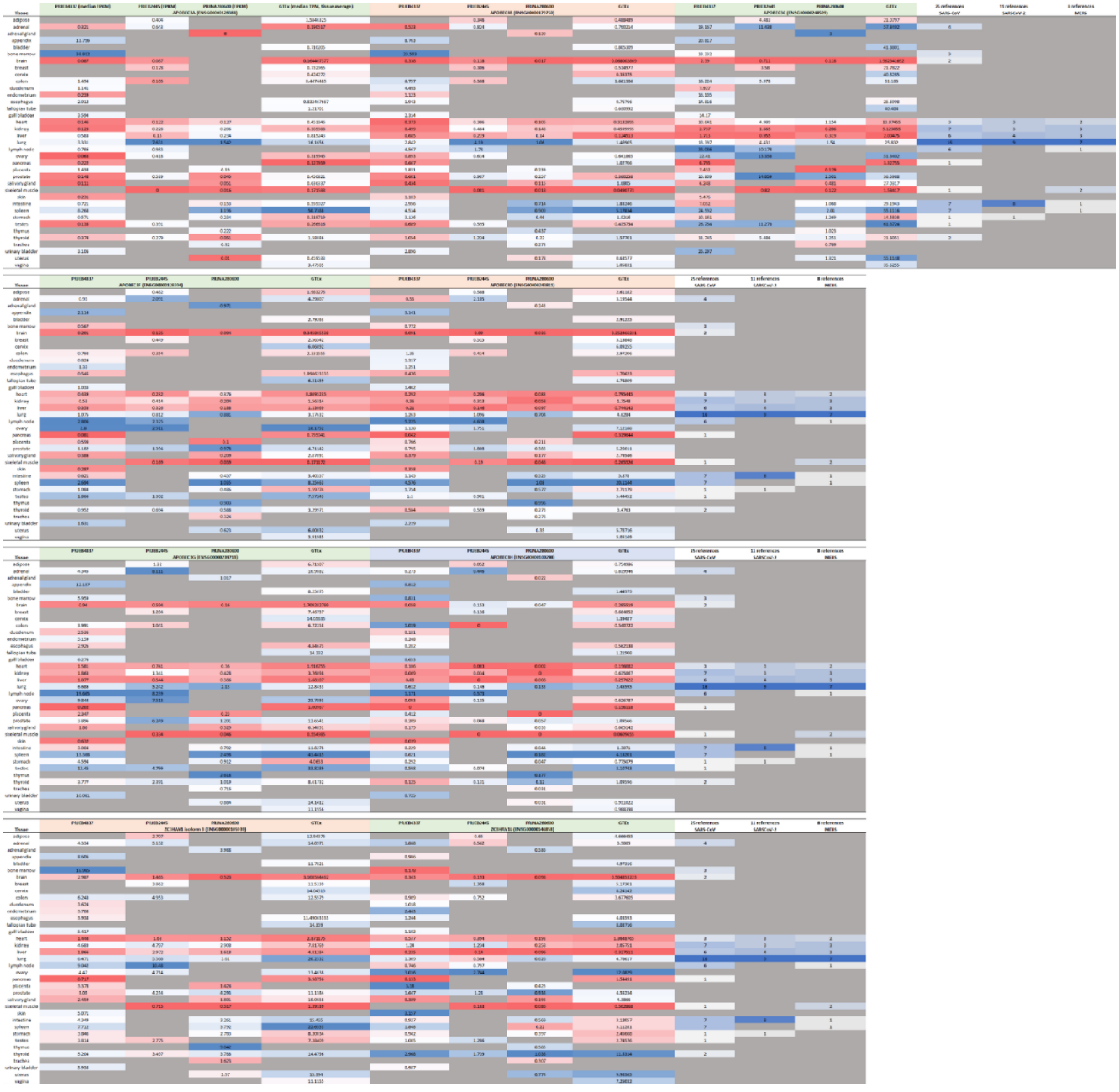


Figure S6.1. Averaged human tissue-specific mRNA expressions (in FPKM and TPM) from four independent studies (PRJEB4337, PRJEB2445, PRJNA280600, and GTEx) and regular tissue habitats of SARS-CoV, SARS-CoV-2 and MERS (in the last three columns). The color spectrum from blue (higher) to white (median) to red (lower) indicates the comparative tissue-specific mRNA expressions of seven APOBEC3 isoforms (A3A, A3B, A3C, and A3D, A3F, A3G, A3H) and two ZAP isoforms (ZC3HAV1 and ZC3HAV1L) within each independent study. Similarly, for tissue habitats of the viruses, the color spectrum (from more blue to white) indicates the prevalence of tissue infection observed from independent studies (observed most commonly to least commonly across different studies, respectively). Light grey indicates tissues with no expression data or for which we encountered no peer-reviewed reports of infection.

Tissue	Apobec3 (ENSMUSG00000009585)		ZC3HAV1 (ENSMUSG00000029826)		15 references
	PRJNA66167 (average FPKM)	PRJNA516470 (average TPM expression)	PRJNA66167 (average FPKM)	PRJNA516470 (average TPM expression)	Murine CoV (MHV)
adipose	6.7485	9.313692333	5.035	16.690011	
adrenal	4.618	6.271135167	6.119	21.8915105	
bladder	11.881		6.091		
brain	1.545	2.2344235	0.198	1.427979833	12
duodenum	2.999		2.968		
heart	3.65	3.0484375	1.771	3.128852833	
intestine	5.780666667	12.956496	5.929333333	25.97670933	
kidney	5.427		2.372		
liver	3.8595	1.607701333	6.8	10.33109233	5
lung	10.945	24.6179815	5.424	22.307956	1
mammary gland	13.893		8.221		
ovary	6.049		7.301		
placenta	8.61		7.664		
skeletal muscle		2.098281667		2.2478465	
spleen	21.781	121.3567817	10.019	55.866762	
stomach	4.625		2.269		
testis	5.106	1.192741	0.604	0.931320333	
thymus	11.723		12.819		
thyroid		36.741307		13.164933	

Tissue	APOBEC3Z3 (ENSCAFG00000001361)		ZC3HAV1L (ENSCAFG00000025152)		ZC3HAV1 (ENSCAFG00000004102)		18 references
	PRJNA516470 (average tissue TPM expression)	PRJNA124245 (average microarray counts)	PRJNA516470	PRJNA124245	PRJNA516470	PRJNA124245	Canine CCoV
adipose	0.862516167		6.2554245		11.90297883		
adrenal	0.404936		6.148822333		21.44471392		
brain	0.189064583	3.63742	5.136939167	3.3429625	10.67451	3.4624775	4
intestine	1.4546535	4.3188475	4.94227625	3.3938725	14.72036083	5.3667	13
heart	0.0834275	3.85984	2.470184333	3.44175	2.979832	3.2913525	2
kidney		3.7027925		3.3562325		4.209065	5
liver	0.464394667	3.88141	2.152363333	3.4251675	10.48709683	5.2497175	5
lung	3.709102833	4.5385275	7.093286833	3.3347825	30.22677333	5.1139125	9
lymph node		6.616965		3.558905		6.3848875	4
muscle	0.055350167		1.086068333		1.499525667		
pancreas		4.325896667		3.691123333		3.904663333	2
pituitary	1.394461417		21.9540575		39.07547483		
skeletal muscle		3.792595		3.39584		3.15464	
skin	1.176864083		9.392302167		21.526278		
spleen	5.767154083	5.8154675	5.429668167	3.4722925	19.40309767	5.1977125	6
testis	0.062242625		0.387509125		1.55394975		
thyroid	0.574273125		12.6971745		14.39405667		

Tissue	Bovine Genome Database (average FPKM)			10 references
	APOBEC3H (ENSBTAG00000007755)	APOBEC3Z1 (ENSBTAG00000037800)	ZC3HAV1 (ENSBTAG00000021617)	Bovine CoV
adrenal gland	29.727776	1.086749	13.683403	
eye (average)	12.094164	59.979467		
cerebral cortex	12.748904		17.009167	
esophagus		0.313111		
gall bladder		0.093025	15.412585	
heart ventricle		0.07895	13.411163	
laryngeal cartilage		0.408039		
liver		0.142727		1
lung	11.709772		16.534698	7
lymph node	29.98797367	0.08924	17.742007	1
nasal mucosa		3.9424085		
oviductal ampulla	11.763007			
renal medulla		0.081365		
spleen	13.3909175	0.075877	24.736925	
testis	16.235922			
tongue		0.126493		
urethra	11.731789			
uterine endometrium		0.085553	15.722054	
uterus	12.785157			
myometrium			16.379176	
subcutaneous adipose tissue			19.783173	
longissimus thoracis			15.513075	
bone marrow			15.025097	
mammary gland			14.38763	
intestine (average of cecum, ileum, duodenum, jejunum)	14.063233	1.1451785	19.47974967	10
stomach (average of abomasum, rumen, omasum, reticulum)		0.770561	13.83319233	

Tissue	TISSUE 2.0 integrated datasets (average FPKM)		8 references
	APOBEC3H (ENSSSCP00000000095)	ZC3HAV1 (ENSSSCP00000017495)	Porcine HEV
adrenal	0.479	1.635	
bladder	0.992	1.384	
bone marrow	0.349	1.147	
brain	1.069	0.684	
colon	0.956	2.176	
duodenum	0.512	1.714	
esophagus	0.79	1.877	1
eye	0.314	1.315	
gall bladder	0.454	1.384	
heart	0.642	1.142	1
intestine	0.961	2.459	2
kidney	2.233	1.109	2
liver	1.032	1.771	7
lung	0.893	1.617	3
lymph node	2.311	1.969	
ovary	1.362	1.501	
pancreas	0.699	0.901	
placenta	0.28	1.131	
placenta	0.28	1.131	
rectum	0.365	1.666	
salivary gland	0.229	1.164	
skeletal muscle	0.158	0.909	
spinal cord	0.679	1.694	
spleen	2.277	1.768	2
stomach	0.808	2.017	1
testis	0.962	1.687	
throat	0.322	1.246	
thymus	1.371	1.73	
thyroid gland	0.322	1.246	

Figure S6.2. Averaged host tissue-specific mRNA expressions (in FPKM and TPM) and regular tissue habitats of Murine MHV (MHV), Bovine CoV, Canine CoV (CRCoV), and Porcine HEV (HEV). The color spectrum from blue (higher) to white (median) to red (lower) indicates the comparative tissue-specific mRNA expressions of APOBEC3 and ZAP (ZC3HAV1) isoforms within each independent study. Similarly, for tissue habitats of the viruses, the color spectrum (from more blue to white) indicates the prevalence of tissue infection observed from independent studies (observed most commonly to least commonly across different studies, respectively). Light grey indicates tissues with no expression data or for which we encountered no peer-reviewed reports of infection.

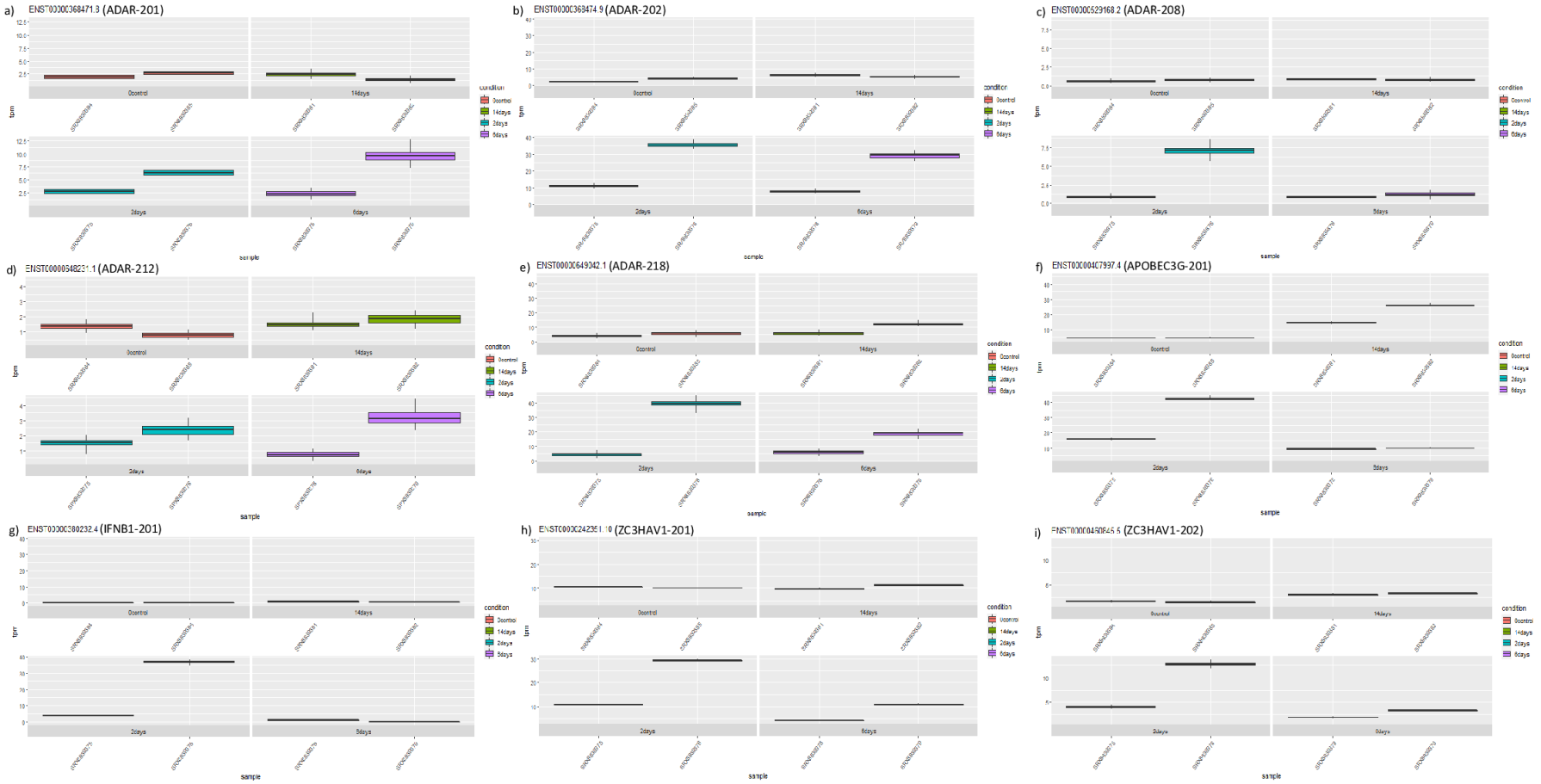


Figure S6.3. The variation within and between samples for all significantly variable transcripts of interest from Figure 5.3. The y axes show the number of transcripts per kilobase million (TPM) generated deterministically from estimated counts pseudo-aligned by kallisto, and the variation shown in each experiment is a proxy for technical replicates from 1000 bootstrap samples. The x axes group the experimental samples by condition.

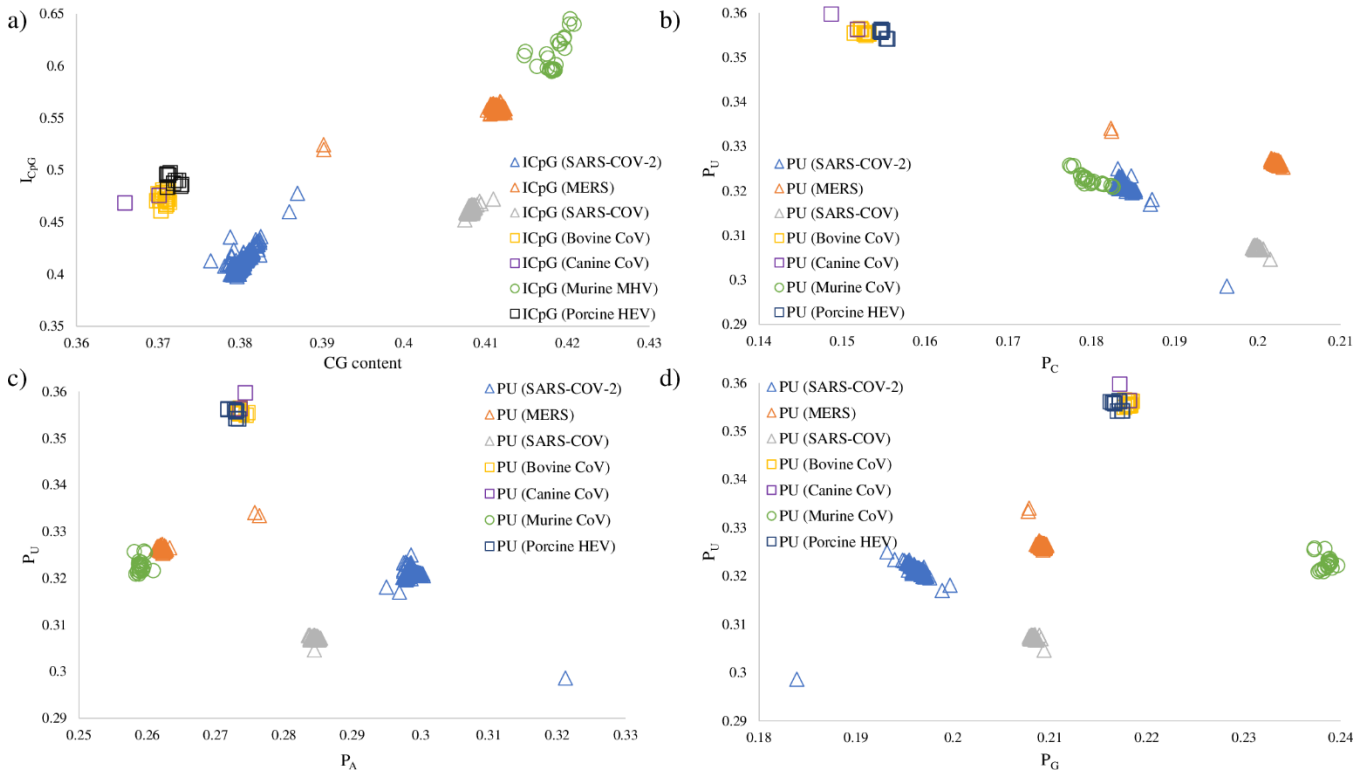


Figure S6.4. I_{CpG} and nucleotide proportions for seven coronaviruses with complete, untrimmed, genomes and host information. Panel a) shows that SARS-CoV-2 has the least I_{CpG} in comparison to other coronaviruses from their natural hosts. Panels b), c) and d) respectively show that the P_U negatively correlates with P_C but not with P_A or P_G; P_U is highest among Bovine CoV, Canine CoV (CRCoV), and Porcine HEV (HEV) but lowest among Murine MHV (MHV) and human coronaviruses. Each panel includes 2666 SARS-CoV-2 genomes, 403 MERS genomes, 134 SARS-CoV genomes, 20 Bovine CoV genomes, two CRCoV genomes, 26 MHV genomes, and ten HEV genomes.

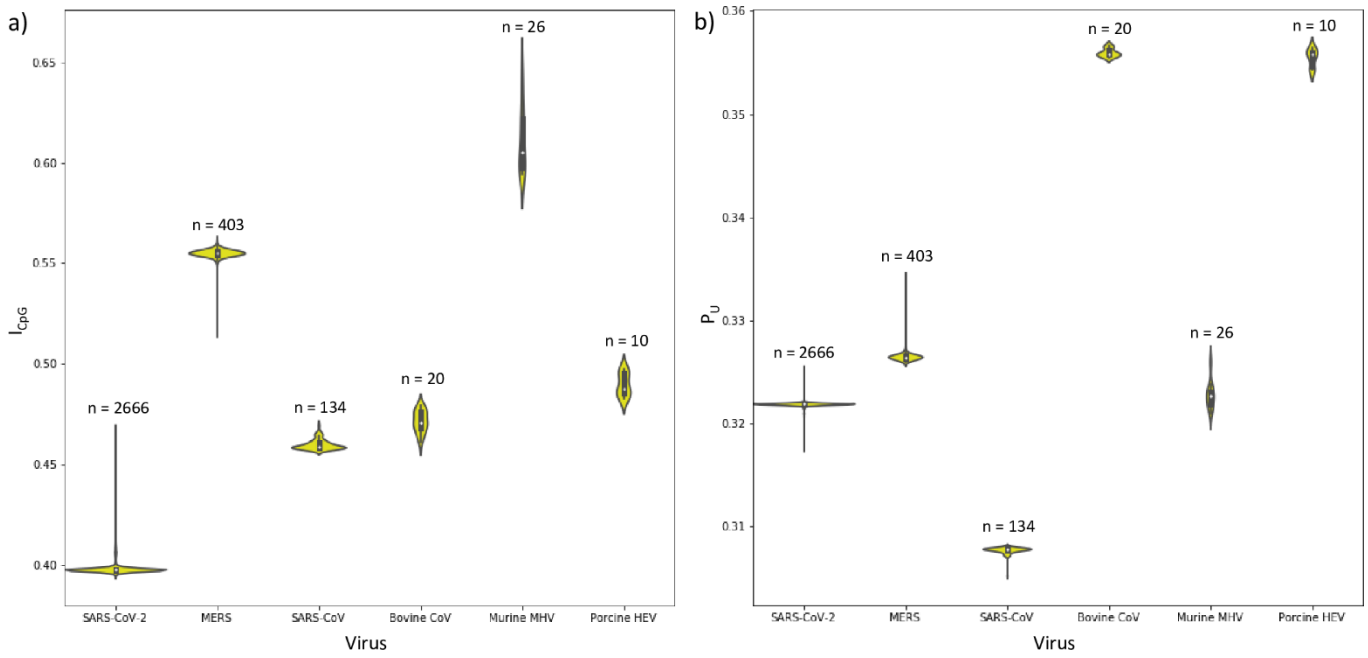


Figure S6.5. Variations of a) I_{CpG} and b) P_U among trimmed genomes in six coronaviruses. Canine CoV (CRCoV) was omitted because only two genomes had been identified. The sample size for each category is denoted by 'n'. Median I_{CpG} is represented by a white dot, black rectangles represent the interquartile range. The width of yellow regions corresponds with the frequency range of I_{CpG} and P_U values.

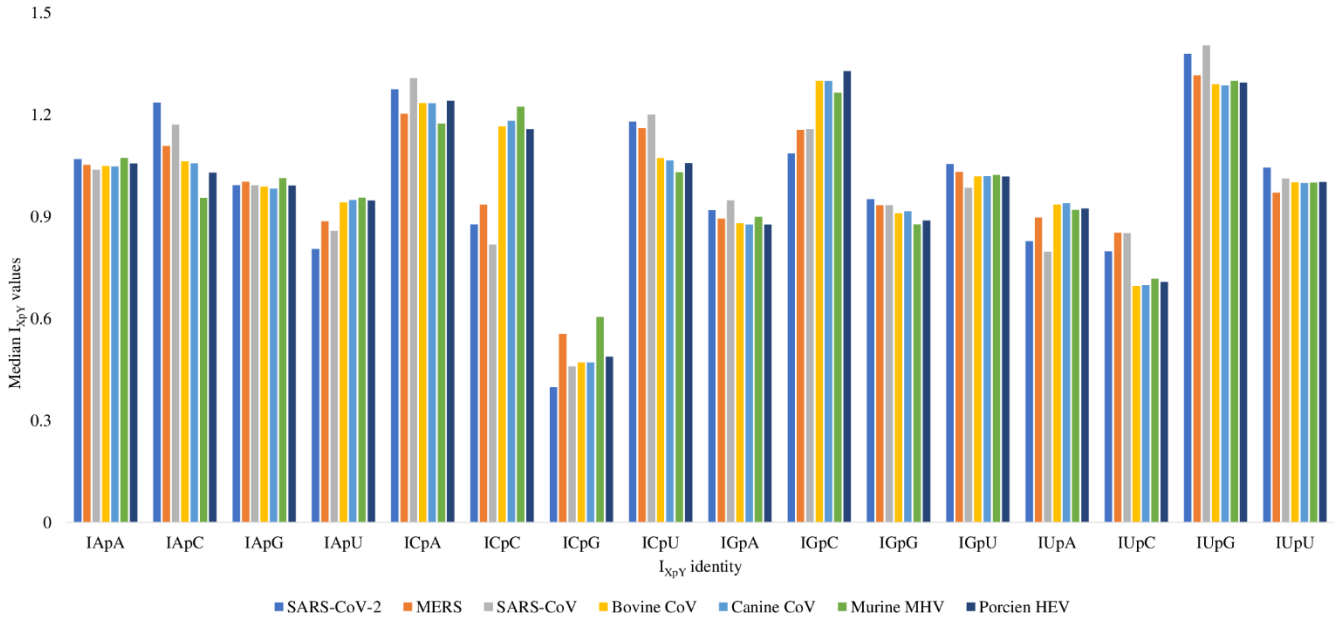


Figure S6.6. Median I_{XpY} shows I_{CpG} is most deficient among all 16 XY dinucleotide combinations (where X and Y are A, C, G or U) in all seven coronaviruses. Each bar value displays the median I_{XpY} calculated from 2666 SARS-CoV-2 genomes, 403 MERS genomes, 134 SARS-CoV genomes, 20 Bovine CoV genomes, two CRCoV genomes, 26 MHV genomes, and ten HEV. All genomes are complete, with ends trimmed after MAFFT alignment, and have host information.

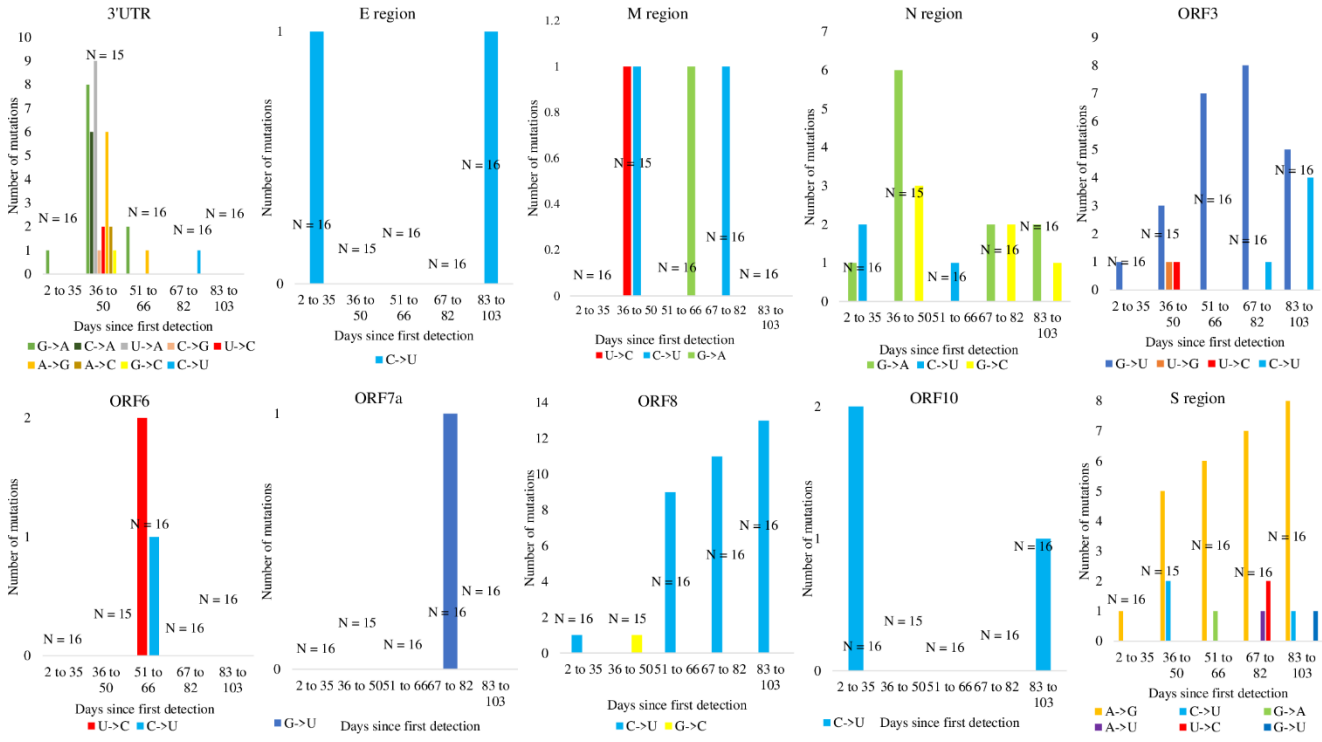


Figure S6.7. Geographical and temporal SNP patterns in 80 complete and high quality SARS-CoV-2 strains collected in the United States. Panels respectively show mutations within the E region, M region, N region, ORF3a, ORF6, ORF7a, ORF8, ORF10, S region, and 3' UTR, in pair-wise comparison between 79 strains and the oldest strain collected in the United States (accession MN985325, sampled 2020-01-19). N denotes the number of samples per time range.

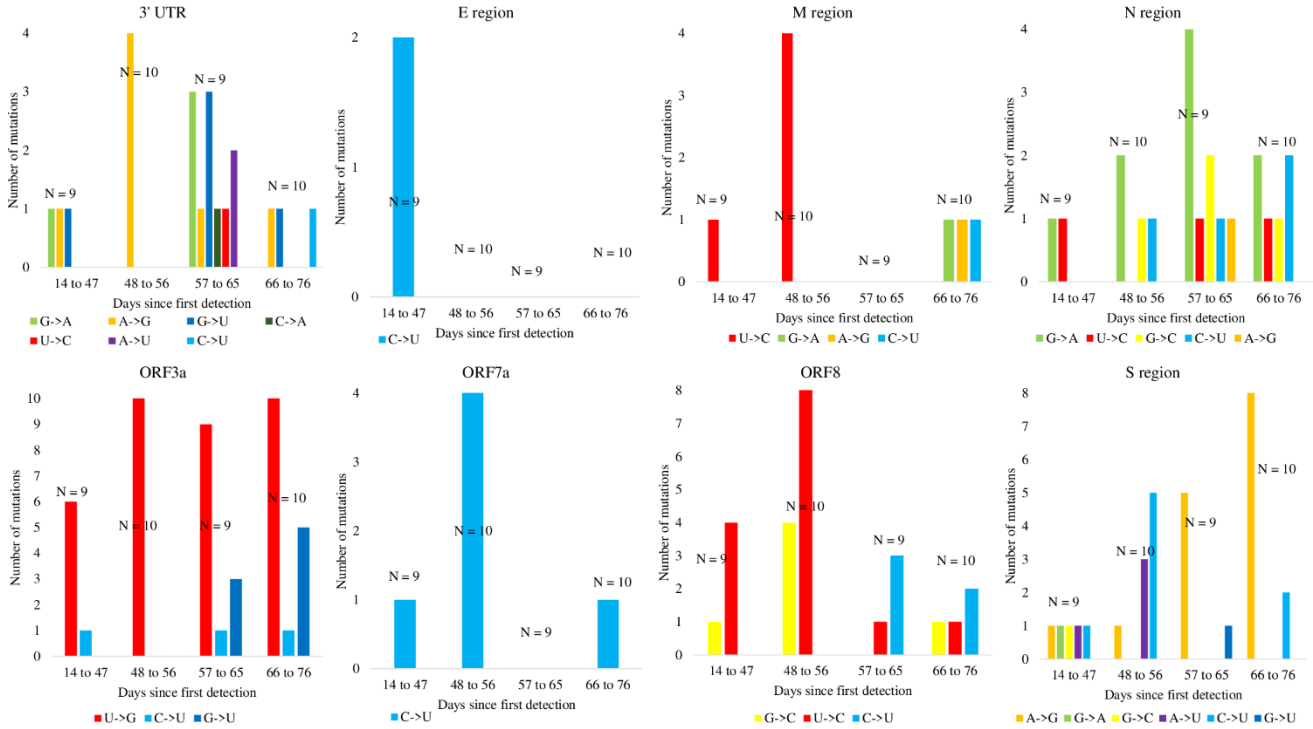


Figure S6.8. Geographical and temporal SNP patterns in 39 complete and high quality SARS-CoV-2 strains collected in Australia. Panels respectively show mutations within the N region, M region, E region, ORF3a, ORF7a, ORF8, S region, and 3' UTR, in pair-wise comparison between 38 strains and the oldest strain collected in Australia (accession MT450920, sampled 2020-01-25). ORF6 and ORF10 regions were omitted because there were no observed mutations in these regions. N denotes the number of samples per time range.

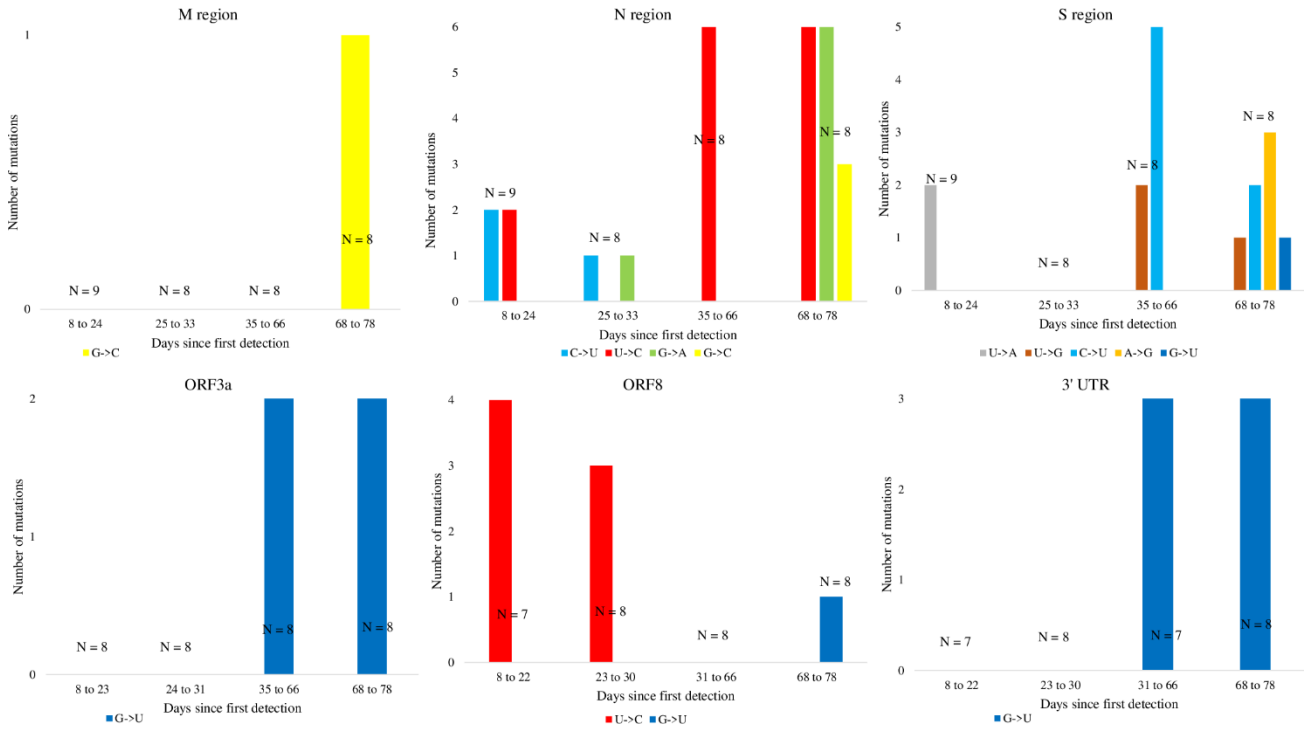


Figure S6.9. Geographical and temporal SNP patterns in 34 complete and high quality SARS-CoV-2 strains collected in China. Panels respectively show mutations within the M region, N region, ORF3a, ORF8, S region, and 3' UTR, in pair-wise comparison between 33 strains and the oldest strain collected in the China (accession MN908947, sampled 2019-12-31). E, ORF6, ORF7a, and ORF10 regions were omitted because there were no observed mutations in these regions. N denotes the number of samples per time range.

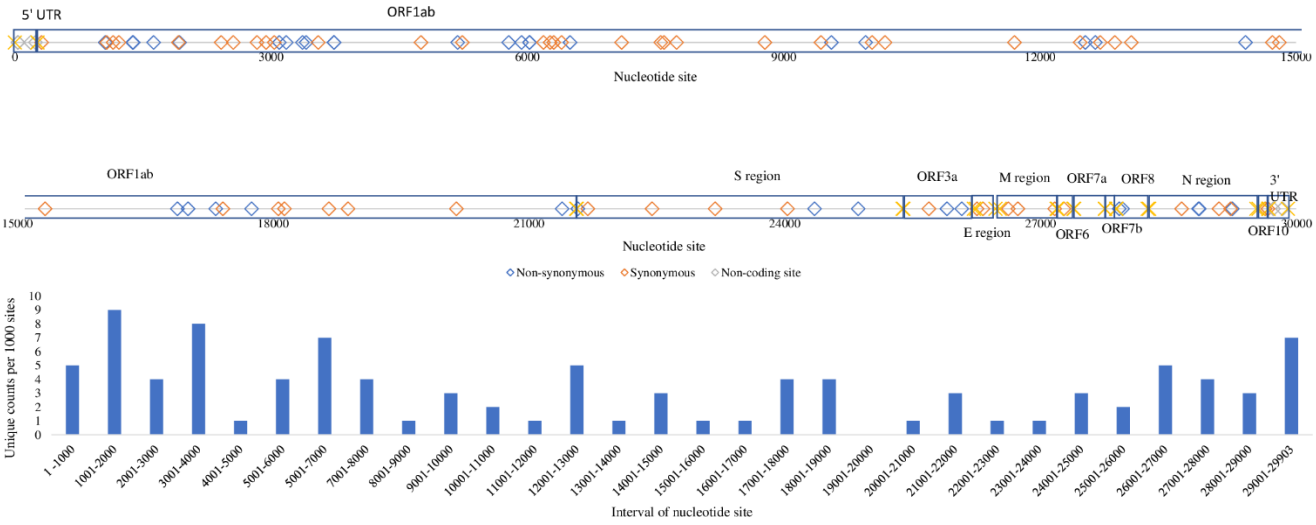


Figure S6.10. Site-specific C to U mutations when the reference SARS-CoV-2 strain Wuhan-Hu-1 (accession MN908947, sampled 2019-12-31) was compared to 98 SARS-CoV-2 genomes collected worldwide with unique collection dates. A) The locations of 98 unique sites having C to U mutations in the Wuhan-Hu-1 genome with annotated viral regions. B) The total count number of unique C to U mutations sites per 1000 nucleotide bases in the Wuhan-Hu-1 genome.

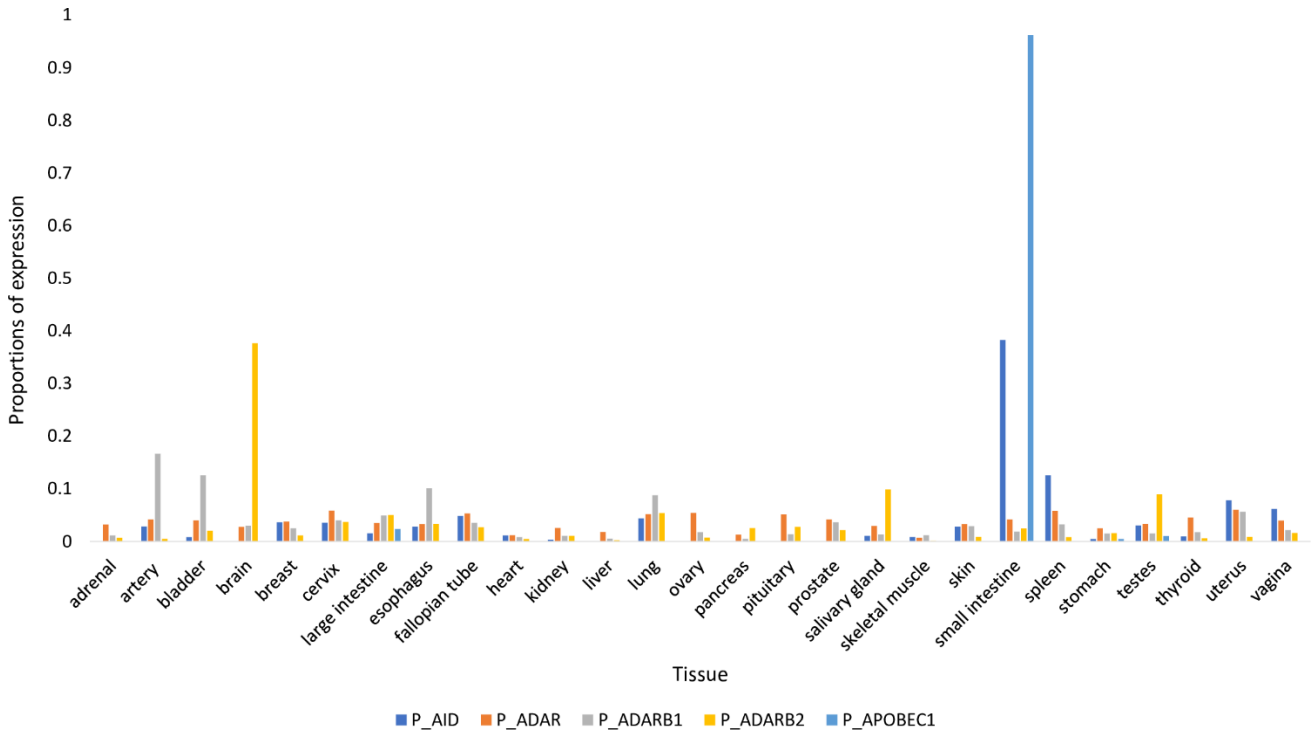


Figure S6.11. The relative mRNA expressions of *AID*, *ADAR*, and *APOBEC1* in 27 human tissues. "Proportions of expression" on the y axis is measured by tissue median TPM/sum tissue median TPM for each gene. Human tissue-specific mRNA expressions, in median TPM values, were retrieved from all RNA-Seq datasets available in the GTEx Portal.

Table S6.1. The summed number of C to U mutations and other mutations (non-C to U) that were observed at each viral region when the reference SARS-CoV-2 strain Wuhan-Hu-1 (accession MN908947) was compared to 98 SARS-CoV-2 genomes collected worldwide with unique collection dates.

Viral region	Sequence length	Synonymous (C to U)	Non-Synonymous (C to U)	Non-coding (C to U)	Synonymous (non C to U)	Non-synonymous (non C to U)	Non-coding (non-C to U)
5' UTR	265	-	-	35	-	-	0
ORF1	21291	124	66	-	26	75	-
S	3819	3	3	-	5	44	-
ORF3a	828	2	2	-	0	34	-
E	225	2	0	-	0	0	-
M	666	5	0	-	0	1	-
ORF6	186	1	0	-	2	0	-
ORF7a	363	0	0	-	0	1	-
ORF8	363	0	6	-	0	0	-
N	1257	4	5	-	3	20	-
ORF10	114	3	0	-	0	0	-
3' UTR	229	-	-	1	-	-	4
Total	29112	144	82	36	36	175	4