THE ROLES OF CODON USAGE IN TRANSLATION AND TRANSCRIPTION

APPROVED BY SUPERVISORY COMMITTEE

Yi Liu, Ph.D.

Benjamin Tu, Ph.D.

Philip Thomas, Ph.D.

Xuewu Zhang, Ph.D.

DEDICATION

To my parents Ruiji Zhao, Xihong Wang and my wife Ruoxi Zheng.

THE ROLES OF CODON USAGE IN TRANSLATION AND TRANSCRIPTION

by

FANGZHOU ZHAO

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2020

THE ROLES OF CODON USAGE IN TRANSLATION AND TRANSCRIPTION

Publication No. _____

Fangzhou Zhao, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2020

Supervising Professor: Yi Liu, Ph.D.

Codon usage biases are found in all eukaryotic and prokaryotic genomes that can regulate gene expression. Although codon usage has been previously shown to regulate translation elongation speed in fungal systems, its effect in animal systems is not clear. In our first study, using a *Drosophila* cell-free translation system to directly compare the velocity of translation elongation, we demonstrated that optimal codons speed up translation elongation while non-optimal codons slow it down. In addition, codon usage regulates ribosome movement and stalling on mRNA during translation. Finally, we showed that codon usage affects protein structure and function both *in vitro* and in *Drosophila* cells, potentially by regulating co-translational protein folding process. This study indicates that the effects of codon usage on

translation elongation speed is a conserved mechanism from fungi to animals and there is a codon usage "code" fine tuning translation elongation speed to achieve optimal co-translational folding.

In addition to the role of codon usage in translation, it has also been shown to play a role in regulating gene transcription of reporter genes and there is a global correlation between codon usage and mRNA levels. To further characterize the role of codon usage in gene transcription, we performed nuclear RNA-seq in *Neurospora* and found that there is a genome-wide strong correlation between gene codon usage bias and nuclear RNA levels, suggesting that codon usage has a global role in impacting transcription in a translation-independent manner. To uncover the underlying mechanism, we performed a genetic screening by performing RNA-seq in over 250 *Neurospor*a single gene knockout strains and identified 18 mutants with significantly reduced correlation between codon usage and mRNA levels. In addition, the deletion of these 18 genes resulted in mRNA level changes in a codon usage-dependent manner. Interestingly, most of these identified genes, such as set-2, are predicted to play a role in regulating transcription or chromatin structures. Together, these results further established the role of synonymous codon sequences in regulating gene transcription and identified potential factors contributing to gene transcription level in a codon dependent manner.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Zhao, F., Zhou, Z., & Liu, Y. Multiple chromatin regulatory factors influence the codon usage effect on transcription. (In preparation)

Zhao, F., Yu, C. H., & Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. *Nucleic acids research*, *45*(14), 8484-8492.

Yang, Q., Yu, C. H., **Zhao, F**., Dang, Y., Wu, C., Xie, P., ... & Liu, Y. (2019). eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic acids research*, *47*(17), 9243-9258.

Yu, C. H., Dang, Y., Zhou, Z., Wu, C., **Zhao, F**., Sachs, M. S., & Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular cell*, *59*(5), 744-754.

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CAI – Codon Adaptation Index

CBI – Codon Bias Index

RSCU – Relative Synonymous Codon Usage

tAI – tRNA Adaptation Index

mRNA – messenger RNA

tRNA – transfer RNA

RNA – Ribonucleic acid

MNase – Micrococcal Nuclease

SNP – Single Nucleotide Polymorphism

ORF – Open Reading Frame

FRQ – FREQUENCY gene

FRET – Fluorescence and Förster Resonance Energy Transfer

RNAP II – RNA polymerase II

H3K9me3 – Histone H3 Lysine 9 trimethylation

KO – Knockout

KD – Knockdown

Luc – Luciferase

5’ UTR – 5’ Untranslated Region

3’ UTR – 3’ Untranslated Region

TFA – Time of First Appearance

2P-seq – Poly(A)-Primed Sequencing

AUC – Area Under Curve

CDS – Coding Sequence

CTD – C-terminal Domain

S2p – Phosphorylated Ser-2

ChIP – Chromatin Immunoprecipitation

PCR – Polymerase Chain Reaction

RT-qPCR – Quantitative reverse transcription PCR

CHAPTER ONE

INTRODUCTION

1.1 Codon Usage Bias

Codons are the triplet nucleotide sequences decoding amino acids during mRNA translation process. Since there are 61 codons decoding 20 amino acids, each amino acid is decoded by 2 to 6 codons except for methionine and tryptophan. Codons decoding same amino acids are defined as synonymous codons. For a long time, these synonymous codons were thought to be redundant. However, these so-called synonymous codons are not used at same frequencies in each gene or genome and certain synonymous codons are preferred in different genomes. This phenomenon is called as codon usage bias, which has been found in all prokaryotic and eukaryotic genomes [1-4]. The existence of codon usage biases suggest that synonymous codons are not redundant and may have additional functions other than just decoding amino acids.

1.1.1 The Evolutionary Driving Force of Codon Usage Bias

Since the discovery of codon usage bias, its evolutionary driving force has been under investigation. Both mutational bias and natural selection were thought to shape codon usage profiles in all organisms [5-8]. Evidences supporting the effect of mutational bias were mostly from bioinformatic analyses focusing on the correlation between base composition of

coding regions and that of proximal non-coding regions [9]. On the other hand, natural selection was also believed to play an important role in shaping codon usage biases [10-13]. Early studies revealed that certain synonymous codons were used predominantly in highly expressed genes comparing to other synonymous codons [2, 14, 15]. Importantly, optimizing the codon usage of either endogenous or exogenous genes according to these highly expressed genes could increase the protein yield in the host organism, a method that has been widely used in heterologous gene expression to obtain high level of protein products [16]. In addition to the effect on protein expression level, codon usage was also shown to affect translation accuracy while amino acids decoded by frequently used codons are less prone to be mistranslated [17, 18]. Together, these studies suggested that gene codon usage has multiple roles in regulating protein production process.

1.1.2 Measurement of Codon Usage Bias

Codon usage bias could be measured either on gene-level or on codon-level. Many different indexes were developed to represent the extent of gene codon usage bias (Figure 1-1). Relative Synonymous Codon Usage (RSCU) is a commonly used codon-level index which describes how much of the codon frequency is deviated from random equal usage [2]. On the other hand, multiple gene-level indexes are also used, including CAI [19], CBI [12] and tAI [20, 21]. To calculate CAI (Codon Adaptation Index), a group of highly expressed genes or the entire genome are selected as reference. Then different codons are given different weights based on their frequencies in the reference set and CAI is calculated as the geometric mean of the weights of codons in target genes. tAI (tRNA adaptation index) is

similarly calculated except that the weights of different codons are obtained based on their cognate and near cognate tRNA copy numbers and codon-anticodon base pairing. CBI (Codon Bias Index) describes the extent of optimal (preferred) codon usage in each gene (Figure 1-1). Although these different indexes can be used to measure codon usage biases, they are usually tightly correlated with each other. In my studies, I mainly used CBI to describe codon usage bias due to its relatively large range.

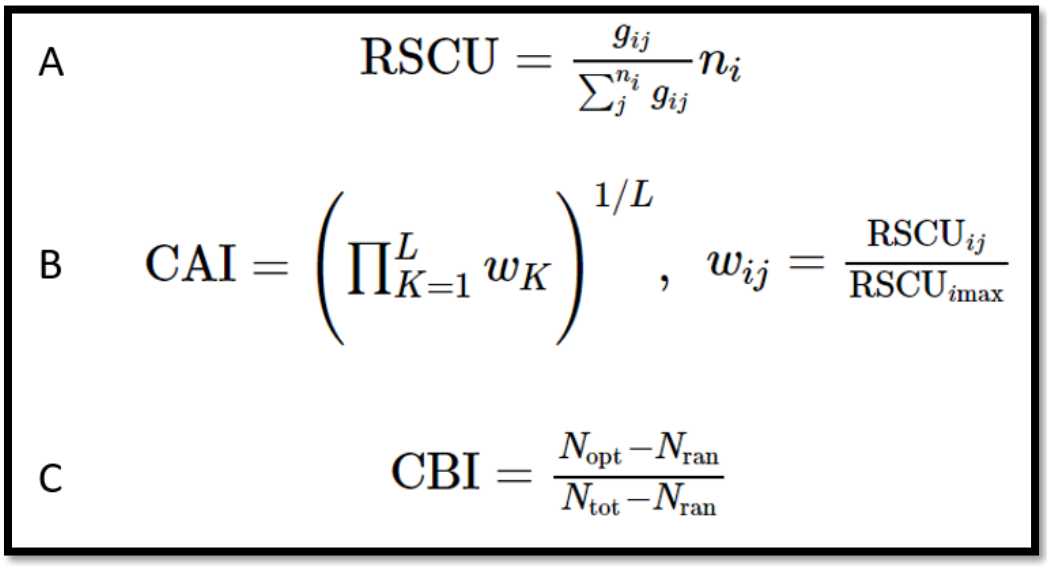

Figure 1-1 Different Indexes for Measuring Codon Usage Bias (Adapted from [22]).

(**A**) "g_{ij} is the observed number of the *i*th codon for the *j*th amino acid, which has n_i kinds of synonymous codons." (**B**) w_{ij} represents the weight of codon j in codon family i calculated as shown in the right. $\mathrm{RSCU}_{i\mathrm{max}}$ is the highest RSCU for codons in codon family i. L represents the total number of codons in a given gene. (**C**) "N_{opt} represents the total number of occurrences of the superior codon in the gene. N_{ran} represents the sum of the number of occurrences of the superior codon when all the synonymous codons are random in a certain

protein; N_{tot} represents the occurrence number of the amino acid corresponding to the superior codon in the gene."

1.1.3 Codon Usage Bias in Different Organisms

Although codon usage bias has been found in all known prokaryotes and eukaryotes, the extent and pattern of codon usage bias are different in different organisms. My studies were performed in *Neurospora crassa* and *Drosophila melanogaster*. Both organisms show relatively strong codon usage biases, meaning that optimal codons are much more frequently used than alternative synonymous codons at genome level. However, in higher eukaryotes like mouse and human, the levels of codon usage bias are much lower [11]. This could be due to stronger mutational pressure and weaker natural selection in vertebrates [8, 11]. On the other hand, codon usage was found to be adapted to tRNA pools in many organisms due to selection pressure [1, 23-25]. And higher eukaryotes like mouse and human may have different tRNA expression patterns in different tissues, which could potentially mask the level of codon usage bias on genome level. Although tRNA abundances are thought to be the main determinant of codon optimality, direct evidences are still lacking due to technical limitations on the quantification of functional tRNA pools. In addition, the nucleotide preferences are also different in different organisms. Organisms such as *Neurospora, Drosophila* and mammals prefer C and G ending codons, while yeast and *Arabidopsis* show opposite preferences for A and T ending codons. This phenomenon could be explained by the differences of tRNA pools and may be useful in distinguishing between the effect of codon usage and GC content.

1.2 The Roles of Codon Usage in Translation

Since codons are recognized and decoded during translation process, most effects of codon usage are thought to be dependent on mRNA translation. Translation process can be further divided into translation initiation, elongation and termination. Evidences showed that codon usage can affect all three steps of translation either directly or indirectly, while the effect on translation elongation plays the most important role. Translation initiation can be affected by codon usage through ribosome queueing, a process in which elongating ribosomes move slowly on non-optimal codons causing the upstream ribosomes queue to the start codon and interfere with the initiation process [26]. On the other hand, optimal codons may also feedback to translation initiation by facilitating ribosome movement, which helps the recycling of ribosomes and increases local ribosome pool [3]. Recently, our lab also showed that non-optimal codons may cause pre-mature translation termination in an eRF1 dependent manner [27]. Importantly, this pre-mature termination process relies on the pausing of ribosomes on non-optimal codons during translation elongation.

1.2.1 The Effect of Codon Usage on Translation Elongation Speed

The studies mentioned above led to the hypothesis that ribosomes move faster on optimal codons and slower on non-optimal codons. However, this hypothesis was under debate for long time with many conflicting experimental data [28-31]. Early studies mostly relied on protein overexpression and indirect measurement of ribosome movement, which

suffered from limited ability to accurately measure translation elongation speed. Recently, ribosome profiling technique was developed in which ribosome protected fragments are isolated for high-throughput sequencing to reflect genome-wide ribosome occupancy [32]. Since ribosomes are captured during active translation elongation, regions associated with fast-moving ribosomes have less chance to be protected and detected by sequencing while regions with slow-moving ribosomes or stalled ribosomes are more likely to have ribosome peaks. However, in the initial ribosome profiling studies, correlation between codon usage and ribosome occupancy were not observed [32-35]. To explain this result, a balance model in which codon usage matches their cognate tRNA expression levels to reach optimal translation efficiency has been proposed [35]. In this model, although the supply of tRNAs for optimal codons is higher, the demand for those tRNAs is also very high and vice versa for rare codons. Therefore, translation elongation speeds are not significantly different between optimal and rare codons if the supply matches the demand. However, this model failed to consider the competition between optimal and rare tRNAs in the local environment and that the chances of these tRNAs being incorporated at any given position are different. Later it is known that ribosome profiling results can be influenced by many factors including sequencing depth, experimental conditions, cloning biases and bioinformatic methods [36-40]. In addition, the calculation of correlation between codon usage and ribosome occupancy requires codon-level sensitivity, meaning that the codons at the A-sites of ribosomes need to be determined by the relative position of ribosome protected fragments. In addition, the determination of A-site codon also strongly relies on precise enzymatic cleavage of nuclease digestion at the edge of each ribosome [41]. Some nucleases like MNase (Micrococcal

Nuclease) is known to have cutting biases which may strongly influence the determination of A-site codons. Together, these results indicated that ribosome profiling technique may be limited in its codon level sensitivity to detect the relationship between codon usage bias and translation elongation speed.

To directly determine the effect of codon usage on translation elongation speed, we used *Neurospora* and yeast in vitro translation systems to monitor translation elongation of reporter genes [42]. This approach ensured the same amount of mRNA in the reaction and can synchronize the translation of reporter genes with different codon usage. We demonstrated clearly that optimal codons speed up translation elongation while non-optimal codons slow it down, which resolved the long-standing debate on the role of codon usage on elongation speed (Figure 1-2). Interestingly, multiple synonymous SNPs in human genes were found to affect protein structure and function [43, 44], yet no evidence existed whether the influence is related to altered translation elongation speed. Furthermore, although codon optimization of KRAS gene was shown to increase protein expression level and oncogenesis [45, 46], it is not clear whether codon usage or GC content underlies the regulation. Given these findings, it is important to show whether the effect of codon usage on translation elongation speed is conserved in animal systems, which we will discuss in detail in CHAPTER TWO.

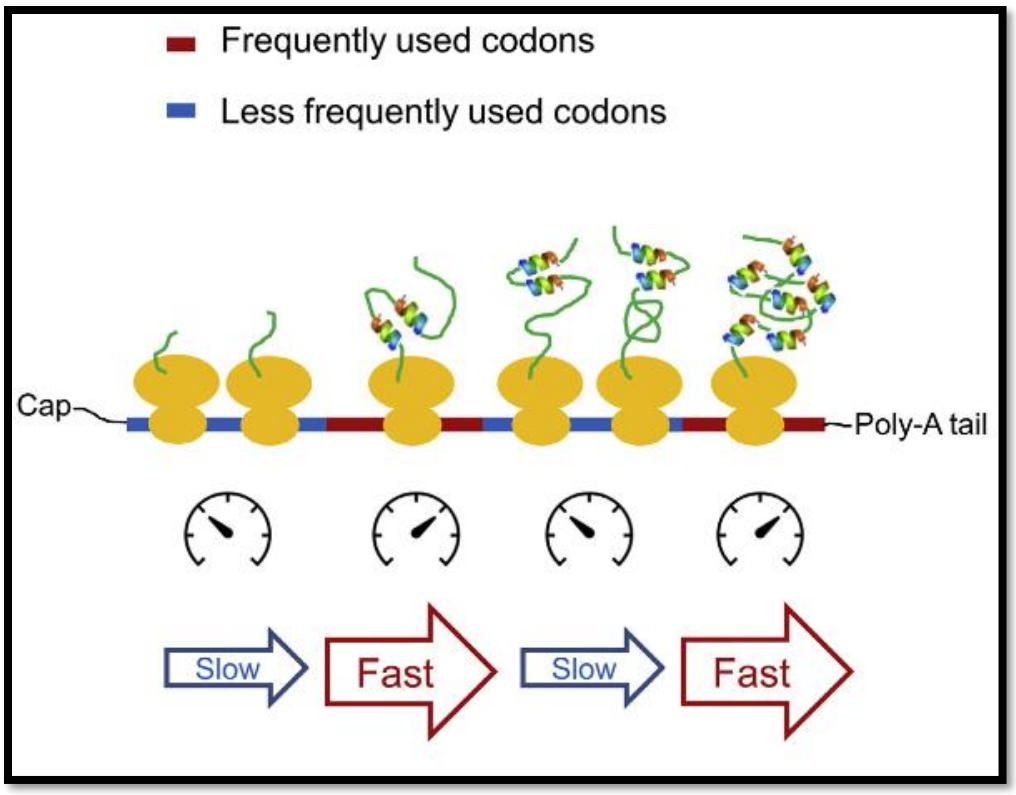

Figure 1-2 Codon Usage Affects Translation Elongation Speed. (Redirected from [42])

Frequently used codons are translated faster, and non-preferred codons are translated slower.

1.2.2 Physiological Significance of Non-optimal Codons in Translation Elongation

As mentioned previously, codon usage pattern is under natural selection for translation efficiency and accuracy. In this sense only optimal codons should be selected. Additionally, decrease of translation elongation speed caused by rare codons may lead to pre-mature translation termination or interfere with translation initiation. Despite these negative effects of rare codons, most genes still use rare codons at many locations and many genes even use rare codons extensively in their ORFs. This could be due to the balance of mutational pressure and natural selection, or due to the lack of selection at specific locations [47]. On the

other hand, it is also possible that non-optimal codons have important biological functions, which reflects alternative selection pressures other than translation efficiency and accuracy. Indeed, our lab showed previously that non-optimal codons are very important for the structure and function of the *Neurospora* core clock gene *frequency* (*frq*) [48]. The *frq* gene contains many non-optimal codons and its expression level oscillates periodically, driving the circadian rhythm in Neurospora. When we optimized only the N-terminal or middle part of FRQ gene, the circadian rhythm was abolished with alterations of the phosphorylation profile, structure and function of FRQ protein. We also report the similar observation for the *Drosophila period* gene in vivo [49]. Multiple studies in *Escherichia coli* also suggested that codon usage or translation elongation speed can affect protein structure and function of heterologous protein products [50-56]. Moreover, bioinformatic studies revealed surprising association between codon usage and the structural organization of proteins in different organisms [57-61]. To address the mechanism underlying the effect of codon usage on protein structure, it was reported that codon usage can affect co-translational folding process by controlling translation elongation speed, leading to distinct protein conformation [55]. In this study, co-translational folding process was monitored in real-time by FRET (Fluorescence and Förster Resonance Energy Transfer). Together, these studies indicated that there is a codon usage "code" within genetic codons which fine tunes translation elongation speed to allow for optimal co-translational folding.

Besides the important function of non-optimal codons in regulating co-translational folding, their negative effect on protein expression level is also beneficial in some specific

situations. Although higher protein expression levels or higher expression efficiencies are generally preferred due to their energy saving capacity, in some cases the protein levels of specific genes need to be kept low. It is reported recently that the poor codon usage of KRAS gene is very important to decrease its protein level and oncogenesis capacity [45, 46]. Similarly, two closely related Toll-like receptors TLR7 and TLR9 have dramatically different protein expression levels, which is important for the immune responses to nucleic acids [62-65]. Disruption of the balanced expression of the two proteins may lead to autoimmune diseases. Interestingly, the poor codon usage of TLR7 contributes largely to the limitation of its protein expression level [66]. In summary, although the genome expression pattern is tightly regulated by many cis- and trans-factors temporally and spatially, I hypothesize that codon usage largely determines the basal expression level for all the genes in a given genome.

1.3 Codon Usage and mRNA level

We mentioned previously that highly expressed genes are enriched in optimal codons and codon optimization serves as an important strategy to increase heterologous protein expression. However, it is often ignored how exactly protein expression level is affected by codon usage since it is not practically important so long as the protein product reaches satisfactory level. It was proposed for long time that codon usage regulate protein levels by affecting translation efficiency, which is usually defined as the protein amount produced per mRNA in a given time period. It should be noted that higher translation elongation speed by

optimal codons does not necessarily lead to higher translation efficiency, unless translation elongation is the rate limiting step of translation process. In fact, recent studies showed that translation initiation is the rate limiting step in most situations and it is mainly controlled by mRNA secondary structure near the start codon [67-69]. These results indicated that translation efficiency may not be the major explanation for how codon usage regulate gene expression level. Combining the information in 1.2 and 1.3, the impact of major mRNA sequence features on translation process and the consequences on gene expression and protein products can be summarized as in Figure 1-4.

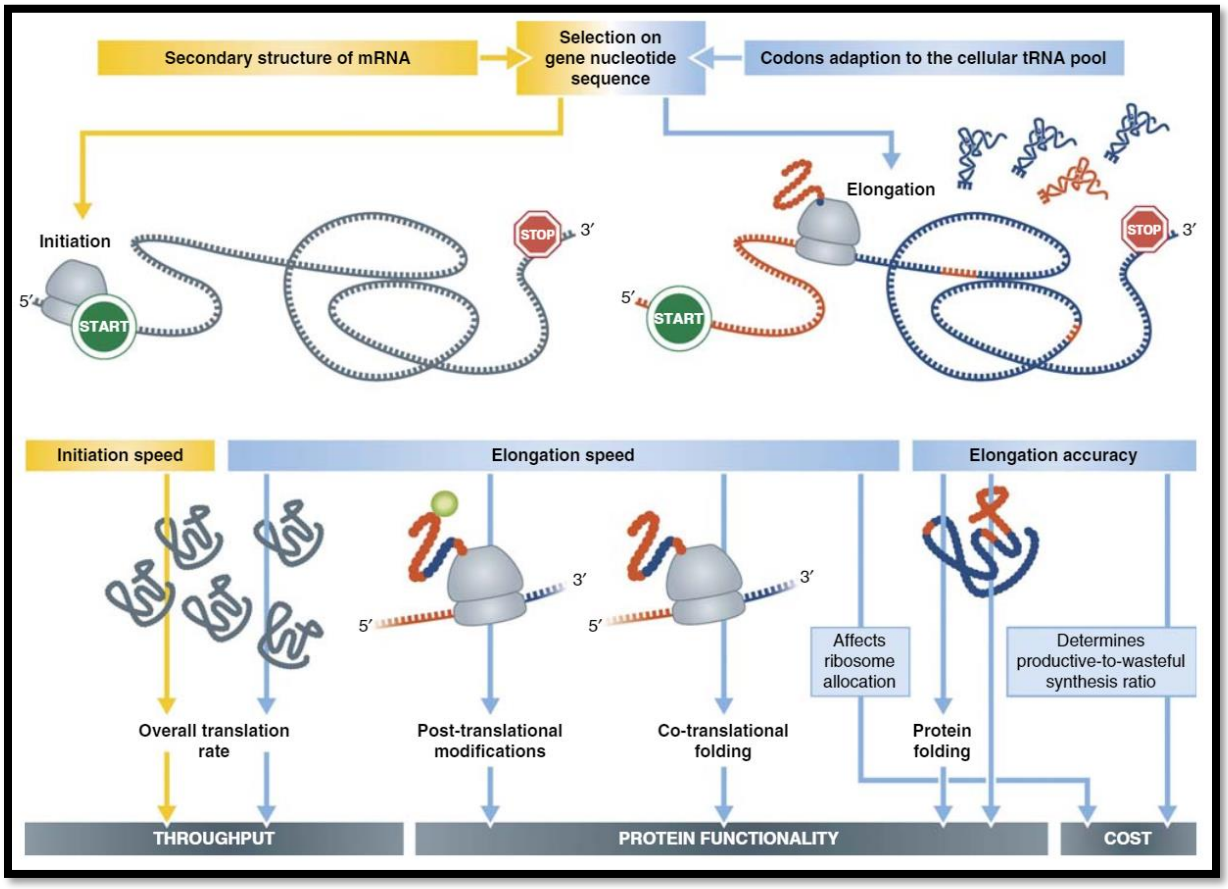

Figure 1-3 The Effect of mRNA Secondary Structure and Codon Usage on Translation and the Corresponding Consequences. (Redirected from [70])

Interestingly, multiple studies showed that codon optimization could also greatly boost mRNA level in many situations and there is also a genome wide correlation between codon usage and mRNA level for many organisms like yeast and Neurospora [71-76]. Steady state mRNA level is determined by both mRNA transcription and degradation. It is more reasonable to hypothesize that codon usage affects mRNA stability since mRNA degradation mostly happens post-transcriptionally and can be influenced by translation process which is the time when codons are recognized. Indeed, many recent studies demonstrated that codon usage affects mRNA stability by controlling ribosome movement kinetics in multiple organisms [71-74]. A DEAD-box containing helicase Dhh1 were shown to be involved in the recognition of slow-moving ribosomes and targeting corresponding mRNAs for degradation (Figure 1-5) [77]. On the other hand, higher translation elongation speed was shown to increase the binding of Pab1 with mRNA poly(A) tail and these well-protected mRNAs are harder to be degraded by CCR4-NOT complex (Figure 1-6) [78].

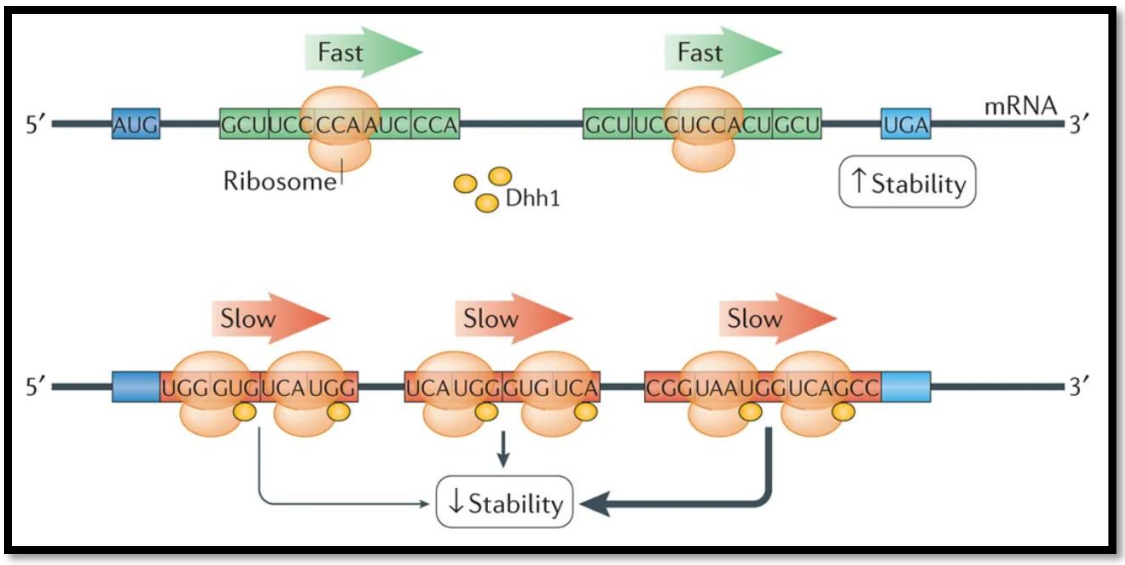

Figure 1-4 Non-optimal Codons Decrease mRNA Stability in a Dhh1-dependent Mechanism. (Redirected from [79])

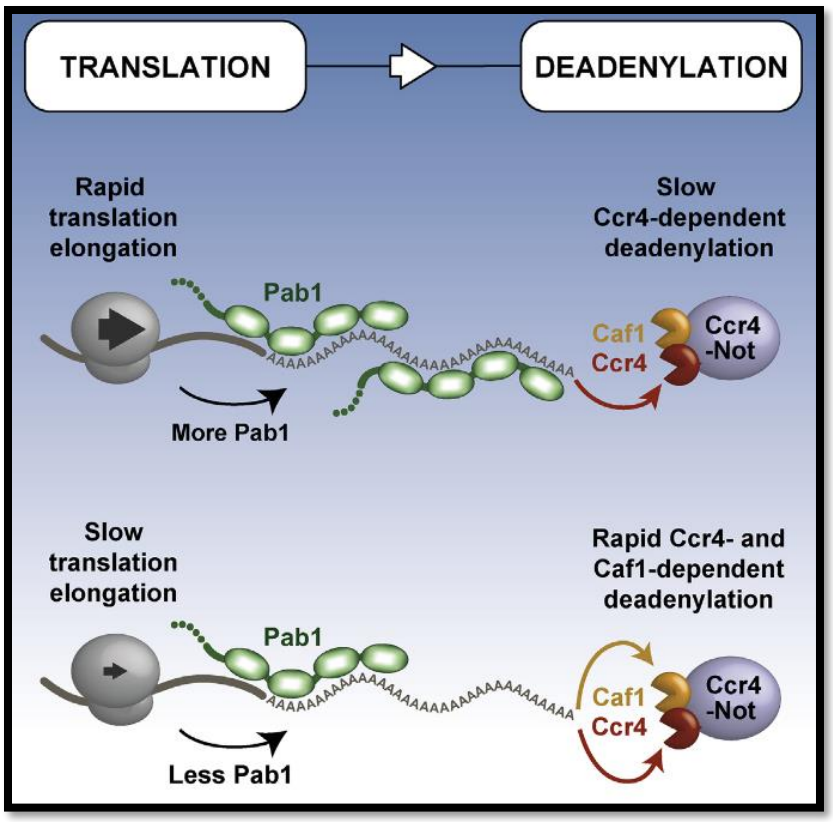

Figure 1-5 mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. (Redirected from [78])

However, in some studies although dramatic increase of mRNA levels was observed after codon optimization, the stability of these optimized mRNAs was not improved significantly [66, 75, 76]. More importantly, the elevation of mRNA levels in these situations was largely independent of translation, which was demonstrated by various experiments. In addition, elevated RNAP II enrichment on codon optimized reporter genes were frequently observed. These findings raised a counterintuitive possibility that codon usage can affect mRNA transcription in a translation independent manner. The biggest question is how codon

usage exerts its function without being recognized by tRNAs. We previously showed in Neurospora that wildtype luciferase gene can induce heterochromatin marked by high level of H3K9me3 when integrated into Neurospora genome [75]. And in H3K9 methyltransferase KO mutant, the effect of codon optimization on luciferase mRNA level disappeared. When we optimized the codon usage of KRAS gene in mammalian cells, we also observed elevated enrichment of active epigenetic markers [76]. These discoveries hinted that codons may not be recognized in the form of triplet nucleotides in these specific situations. Instead, codon usage may co-evolve with other sequence elements which are involved in the regulation of chromatin structure and mRNA transcription. Interestingly, we recently reported evidences of co-evolution between codon usage and transcription termination machinery, which may be responsible for the pre-mature termination at some rare codon positions [80]. In CHAPTER THREE we showed in Neurospora that codon usage affects genome wide transcription level and we identified multiple factors which may be responsible for the regulation. In summary, this study may reveal another important aspect of regulation by codon usage and connect different sequence elements for gene expression control in a broader view.

CHAPTER TWO

CODON USAGE REGULATES PROTEIN STRUCTURE AND FUNCTION BY AFFECTING TRANSLATION ELONGATION SPEED IN DROSOPHILA CELLS

2.1 Introduction

Due to the redundancy of triplet genetic codons, most amino acids can be encoded by two to six synonymous codons. Synonymous codons are not used with equal frequencies, a phenomenon called codon usage bias that is observed in all organisms [1-4]. Selection for efficient and accurate translation was proposed to be a major cause of codon usage bias [10, 18, 48, 81]. On the other hand, studies using protein expression in *Escherichia coli* suggested that translation rate and synonymous codon usage can affect protein folding and functions [50-56]. Strongly supporting this hypothesis, we previously demonstrated that the codon usage of the *Neurospora* circadian clock gene, *frequency* (*frq*), is critical for the structure and function of FREQUENCY (FRQ) protein *in vivo* [48]. More recently, we showed that the codon usage plays a similar role for the *Drosophila* period gene *in vivo* [49]. Importantly, these *in vivo* studies showed that codon usage influences protein structures in a location-specific manner. Consistent with this conclusion, bioinformatic analyses have revealed correlations between codon usage and protein structural motifs in different organisms [57-61]. Together, these studies led to the hypothesis that there is a codon usage 'code' within

genetic codons that regulates translation elongation speed to permit optimal co-translational protein folding.

Most of these proposed roles of codon usage are based on its effect on translation elongation speed. However, earlier studies addressing this issue based on protein overexpression in *E. coli* and indirect measurement of ribosome movement led to conflicting conclusions [28-31]. Moreover, ribosome profiling, a powerful method that uses deep sequencing of the ribosome-protected fragments (RPF), initially found no correlations between codon usage and levels of RPF in different organisms [31-35]. Ribosome profiling results are now known to be influenced by experimental conditions, sequencing depth, cloning/sequencing biases and the bioinformatic methods used [36-40]. Furthermore, ribosome profiling relies on precise enzymatic 5′ end cleavage of the RPF to allow accurate A site determination, which are often difficult due to digestion biases and different experimental conditions [41]. These results suggest that although ribosome profiling has codon-level resolution, it might not have codon-level sensitivity to detect the effect of codon usage on translation elongation speed.

It should be noted that although the implicated role of codon usage in regulating protein expression levels led to the hypothesis that codon usage impacts protein expression levels by affecting translation efficiency [14, 82-84], a role for codon usage in translation elongation speed does not necessarily affect translation efficiency. In fact, recent studies suggest that translation efficiency is mainly determined by the efficiency of translation initiation, a process that is mostly affected by RNA structure but not codon usage near the start codon

[67-69]. In the current study, we only focus on the role of codon usage in regulating translation elongation speed.

To determine the effect of codon usage on translation elongation speed, we previously used *Neurospora* and yeast *in vitro* translation systems to directly monitor the speed of protein translation elongation [42]. Our results demonstrate that codon usage plays an important role in regulating translation elongation speed on mRNA *in vitro* and *in vivo* in these fungal systems. However, the effect of codon usage on translation elongation rate in animal systems is still not known. Although single synonymous SNPs in human genes have been shown to associate with altered protein conformation and function [43, 44], there is no clear evidence showing that these SNPs regulate translation elongation. In addition, although codon manipulation has also been shown to affect KRAS expression and oncogenesis in mice, the role of codon usage in regulating gene expression in mammalian cells can also be explained by difference in GC contents in genes [45, 46, 66, 85]. More recently, we demonstrated that the codon usage of *Drosophila Period* gene is important for protein structure and function *in vivo* [49]. These results raise the possibility that codon usage is a conserved mechanism from fungi to animals that influences translation elongation speed.

Like mammals and *Neurospora*, the *Drosophila melanogaster* genome has a strong codon bias for G/C at the wobble positions [10, 48, 86-88]. Changing codon usage of *Drosophila* genes is known to alter protein expression levels [49, 89] but the mechanism of such an effect is not clear. In addition, codon optimality was also suggested to play a potential role in splicing regulation in *Drosophila* [90]. Previous ribosome profiling using *Drosophila* cells could not allow robust A site assignment due to enzyme cleavage

biases [41]. In this study, we used *Drosophila in vitro* translation system to determine the effect of codon usage on translation elongation. Our results demonstrate that codon usage affects translation elongation speed and local ribosome movement on mRNA. In addition, we show that codon usage influences protein structure and activity *in vitro* and in *Drosophila* cells. Together, our results suggest that the effect of codon usage on translation elongation is a conserved mechanism regulating protein structure and function from fungi to animals.

2.2 Materials and Methods

2.2.1 Cell Culture and Preparation of S2 Cell-free Translation Extract

The preparation of translation extract was modified from a protocol previously described [91]. *Drosophila* Schneider 2(S2) cells (kindly provided by Dr Jin Jiang) were cultured in Schneider's *Drosophila* Medium (Gibco) supplemented with 10% heat-inactivated fetal bovine serum (Sigma) and 1% penicillin-streptomycin (10 000 units penicillin and 10 mg streptomycin/ml, Sigma) at 27°C. After reaching confluence, S2 were harvested by centrifugation at 1000 × *g* for 4 min and washed by 1 × phosphate buffered saline for three times. Cell pellets were resuspended in 2× volumes of hypotonic buffer (10 mM HEPES-KOH pH 7.6, 10 mM potassium acetate, 0.5 mM magnesium acetate, 5 mM Dithiothreitol) and incubated on ice for 40 min to 1 h. Cells were then homogenized in a Dounce homogenizer by 20–30 strokes on ice and the final concentration of potassium acetate was adjusted to 50 mM. The cell extract was centrifuged at 16 000 × *g* for 10 min at

4°C. The supernatant was aliquoted, snap frozen in liquid nitrogen and stored at −80°C before use.

2.2.2 Plasmid Construction

OPT and dOPT luciferase constructs were designed based on the *Drosophila* codon usage table (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=7227). Except for the first 10 codons, all other codons are either the most preferred or the least preferred codons in the OPT or dOPT construct, respectively. Gene synthesis was carried out by Genescript and the resulting genes were cloned into pJI 204 vector containing a T7 promoter and 30-nt poly-A sequence [92]. The pJI 204 vector was modified to include a NheI site following the poly-A sequence to enable plasmid linearization for *in vitro* transcription and a 5 c-Myc tag at the N terminus to allow detection of nascent peptides. N-OPT, M-OPT and C-OPT constructs were created by replacing the N-terminal part (codon 11–223, count from the first ATG of luciferase gene), middle part (codon 224–423) and C-terminal part (codon 424–550) of dOPT sequence with the corresponding optimized sequence, respectively. Constructs for S2 cell expression were created by cloning different versions of luciferase sequences into pUAST vector using NotI/XbaI sites. The pUAST vector and a ubiquitin-Gal4 construct for co-transfection were kindly provided by Dr Jin Jiang. The sequence information of the basic OPT and dOPT luciferase genes are provided in APPENDIX A.

2.2.3 In Vitro Transcription Assay

To obtain mRNA for *in vitro* translation, the plasmids were linearized by NheI and then transcribed using HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB). 3′-0-Me-m7G (5′) ppp (5′) G anti-reverse cap structure analog (NEB) was added following the manufacturer's instruction. The mRNA products were subsequently purified by LiCl precipitation, quantified by Nanodrop (Thermo Scientific), aliquoted and stored at −80°C before use. Denaturing agarose gel electrophoresis was used to determine the quality of the mRNA products.

2.2.4 Cell-free Translation Assay

To monitor the kinetics of luciferase activity in real time, a reaction mixture (20 mM HEPES-KOH pH 7.6, 0.5 mM spermidine, 8 mM creatine phosphate, 0.2 mM GTP, 1 mM ATP, 20 μM complete amino acids [Promega], 100 mM potassium acetate, 1 mM magnesium acetate, 0.13U/μl creatine phosphate kinase, 50 μM luciferin, 0.2U/μl SUPERase•In RNase Inhibitor [Invitrogen], 3 μl total volume per reaction) was prepared and aliquoted into 96-well plate. A total of 1 μl of the mRNA template (~60–180 ng) and 8 μl cell-free translation extract were successively added to the reaction mixture. Afterward, the luminescence signal of the reaction was recorded every 30 s using FLUOstar Optima (BMG) at 26°C. The concentrations of potassium acetate and magnesium acetate in the reaction mixture were optimized for different batches of cell-free translation extracts.

To examine the protein products of the cell-free translation assays, the reactions were incubated in a 26°C water bath and stopped at the indicated time points by adding sodium

dodecyl sulphate (SDS) sample buffer, followed immediately by heating at 90°C. The samples were subsequently denatured and analyzed by western blot.

For the harringtonine chase experiment, 20 ng/ml harringtonine was added to *in vitro* translation assay after 6 min of reaction to inhibit translation initiation. Afterward, the reaction mixture was withdrawn at the indicated time points and subjected to western blot analysis.

2.2.5 Isolation of Ribosome Associated Nascent Chains

The method was modified from previously described [93]. *In vitro* translation reaction was terminated at 15 min by adding cycloheximide to a final concentration of 0.5 mg/ml. Translation products are carefully layered on top of sucrose cushion (0.5M sucrose, 25 mM HEPES, pH 7.5, 80 mM KOAc, 1 mM $Mg(OAc)_2$) and then centrifuged in a TL 100.3 rotor (Beckman Coulter) at 100 000 rpm for 5 min at 4°C. The ribosomal pellets are washed once with 25 mM HEPES (pH 7.5), 80 mM KOAc and 1 mM $Mg(OAc)_2$ and then resuspended in 1× SDS sample buffer for western blot analysis. The supernatant is concentrated and mixed with SDS sample buffer to the same volume as pellet samples for western blot analysis.

2.2.6 S2 Cell Transfection

S2 cell transfection was carried out following the standard calcium phosphate transfection protocol in six-well plates. A total of 1.5 μg luciferase construct and 1.5 μg ubiquitin-Gal4 construct were co-transfected for each transfection. Cells were harvested 48–

60 h afterward and lysed in 1× Passive Lysis Buffer (Promega) according to the manufacturer's instruction. Cell lysate was obtained by centrifugation at top speed for 1 min at 4°C and stored at −80°C.

To determine *Luc* mRNA stability, 10 μg/ml actinomycin D was added to stop transcription and cell samples were collected at indicated time points for RNA purification. Purified RNA samples were examined by northern blot analysis to determine *Luc* mRNA levels.

2.2.7 Limited Trypsin Digestion Assay

For trypsin digestion of *in vitro* translated protein products, 0.5 mg/ml cycloheximide was added to terminate reaction after 20 min of *in vitro* translation. A total of 5 μg/ml trypsin was then added to the reaction and samples were withdrawn at indicated time points for western blot analysis. For trypsin digestion of S2 cell extract, total cell extract was used. Total protein concentration of each sample was adjusted to 1 μg/μl. The trypsin digestion was carried out at 26°C by adding 300 ng/ml trypsin. Digestion products were withdrawn from the reaction at indicated time points and the denatured protein samples were analyzed by western blot.

2.3 Results

2.3.1 Optimal Codons Speed Up Translation Elongation while Non-Optimal Codons Slow It Down in Drosophila Cell-free System

To determine the role of codon usage on translation elongation speed, we established a cell-free translation system using *Drosophila* S2 cells to compare translation speeds of firefly luciferase (*Luc*) mRNAs with different codon preferences. All *Luc* mRNAs have identical 5′ and 3′ untranslated regions (UTR) and the same poly-A length. The first ten codons of the *Luc* open reading frame were also identical in these mRNAs to ensure the same translation initiation context. The use of the cell-free system allows the separation of transcription and translation processes so that translation can be synchronized. Because luciferase is known to be rapidly folded (within a few seconds) after translation [94], the time when luminescence signal is first detected after start of translation (time of first appearance (TFA)) should reflect the speed of translation process. Because different *Luc* mRNAs have the same translation initiation context, the difference in TFA values should mainly reflect the difference in translation elongation speed of different *Luc* mRNAs.

We first generated two *Luc* mRNAs by *in vitro* transcription: OPT (the entire *Luc* ORF codon optimized according to the *Drosophila* codon usage table) and dOPT (the least preferred codon is used for every amino acid). Afterward, same amount of OPT and dOPT mRNAs were separately translated in the S2 translation extracts supplemented with luciferin and luciferase activity was monitored in real-time. As shown in Figure 2-1A and B, the TFA of the OPT mRNA was ~4 min earlier than that of the dOPT mRNA. To confirm this result, we performed Western blot analysis using a luciferase antibody to detect the production of full-length luciferase protein at different time points (Figure 2-1C). As expected, the appearance of full-length luciferase for the OPT mRNA was also about 4 min earlier than that of the dOPT mRNA. It should be noted that the cell-free translation extract contained

endogenous mRNA from S2 cells and could compete with luciferase mRNA for translation. By determining the total level of mRNA, we estimated that the luciferase mRNA added in the reaction is about 1–2% of total mRNA.

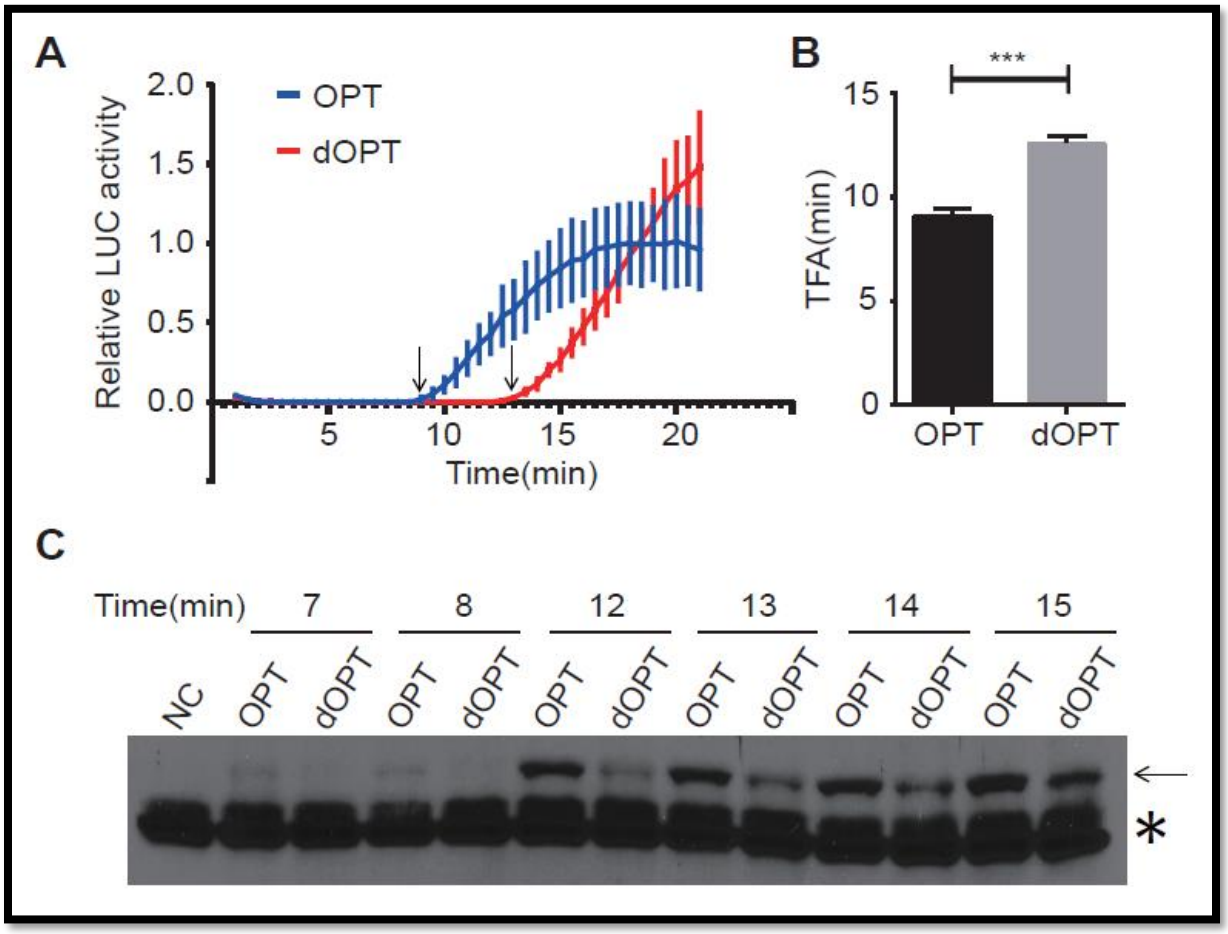

Figure 2-1 Translation Elongation Speed of OPT and dOPT mRNAs

(**A**) Real-time measurement of firefly luciferase activity using S2 cell-free translation system and the indicated *Luc* mRNA templates. TFA values are indicated by arrows. Luciferase signals were normalized to the signal level of the OPT mRNA at 21 min. Error bars represent standard deviations of three replicates. (**B**) Quantification of the TFA values of OPT and dOPT *Luc* mRNAs. ***$P < 0.001$. (**C**) Western blot analysis using luciferase antibody

showing the production of the full-length luciferase (indicated by the arrow) in the *in vitro* translation assays. The asterisk indicates the non-specific bands recognized by the luciferase antibody. NC: Negative Control.

To determine whether the effect observed is due to accumulative effect of codon usage along the ORF and not due to change of a specific mRNA structure, we generated three additional *Luc* mRNAs based on the dOPT mRNA. In the N-OPT, M-OPT and C-OPT mRNAs, the N-terminal region (codons 11–223), middle region (codons 224–423) and C-terminal region (codons 424–550) of the luciferase ORF was replaced by preferred codons, respectively. If the effect we observed above is due to change of mRNA structure at a specific location, one of these *Luc* mRNAs should have the same TFA as the OPT mRNA and the other two *Luc* mRNAs should be similar to the dOPT mRNA. Instead, we found that the TFA values of the N-OPT, M-OPT and C-OPT mRNAs were all significantly higher than that of the OPT mRNA but lower than that of the dOPT mRNA (Figure 2-2). These results suggest that codon usage regulates the speed of translation elongation. In addition, the effect of codon usage on elongation rate is due to cumulative effects of codon usage in the coding sequence rather than changes in specific RNA structure.

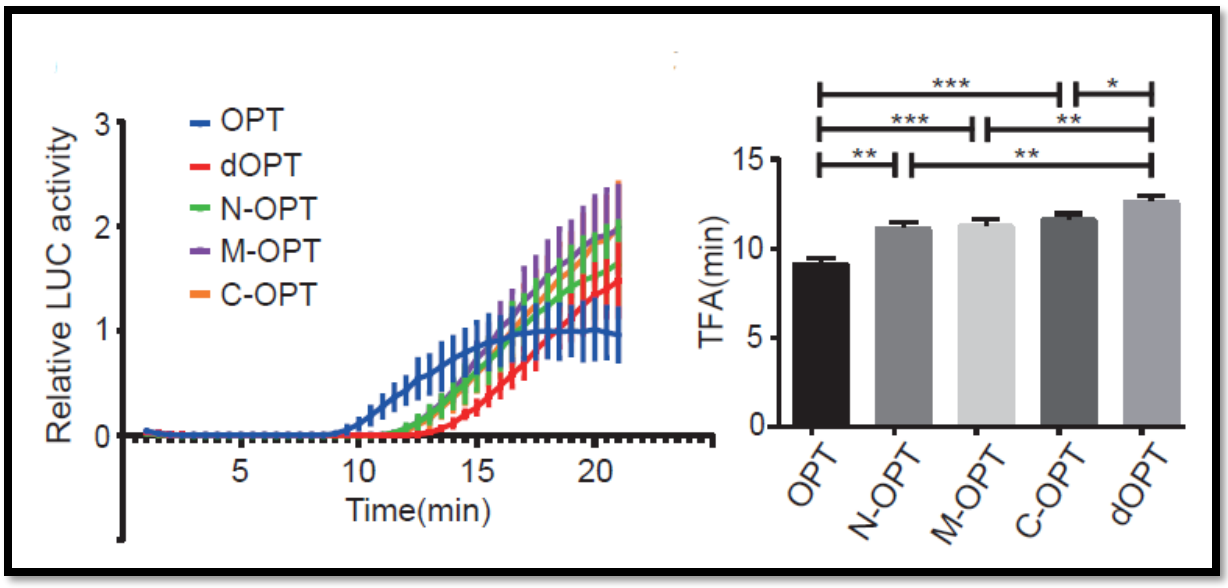

Figure 2-2 Translation Elongation Speed of Partially Optimized Luciferase mRNAs

Left: Real-time measurement of firefly luciferase activity using the S2 cell-free translation system and the indicated *Luc* mRNA templates. mRNA concentrations are adjusted to produce similar activity levels. Error bars represent standard deviations of three replicates. **Right:** Quantification of the TFA values of the indicated Luc mRNAs. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

2.3.2 mRNA Concentration Does Not Affect TFA in a Wide Dynamic Range

Although we used in vitro transcription to generate mRNAs for cell-free translation, we did not use same amounts of mRNAs for different versions of luciferase reporters. Most times we adjusted the mRNA concentrations so that different mRNAs in comparison generate similar luminescence levels at the plateau. This should eliminate the possibility that lower TFA is caused by higher basal luciferase activity. However, it is important to demonstrate that mRNA concentrations in the cell-free translation system do not affect TFA

in a practically reasonable range. Indeed, we tested mRNA concentrations from 1.5 ng/ul to 15 ng/ul in the cell-free translation system using OPT mRNA and did not observe any significant alteration of TFA (Figure 2-3). It should be noted that all the in vitro translation assays in this study used mRNA concentrations within the tested range. Importantly, this result also argues against the balance model mentioned previously (see 1.2.1) since the demand of tRNAs changed dramatically in these conditions while translation elongation speed was not affected.

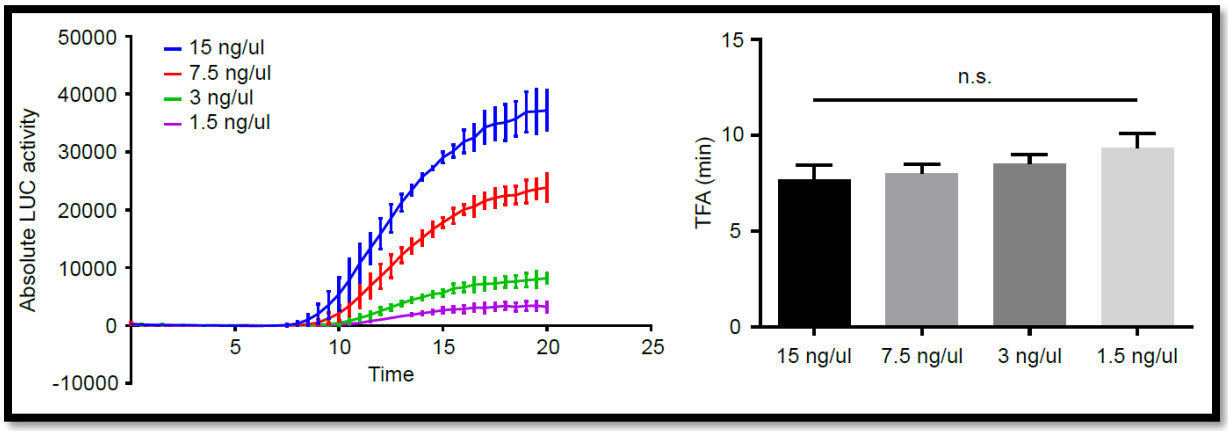

Figure 2-3 TFA of OPT Luciferase with Different mRNA Concentrations

Left: Real-time measurement of firefly luciferase activity using the S2 cell-free translation system and OPT *Luc* mRNA template with indicated concentrations. Error bars represent standard deviations of three replicates. **Right:** Quantification of the TFA values for different mRNA concentrations. No significant difference was observed for any pair of conditions.

2.3.3 Codon Usage Influences Local Ribosome Stalling on mRNA

To further determine the effect of codon usage on ribosome movement on mRNA, we generated *Luc* mRNAs in which the luciferase protein is tagged by five consecutive c-Myc

tags at the N-terminus so that nascent luciferase peptides can be detected by Western blot analysis using a c-Myc monoclonal antibody. After 15 min of translation in the S2 system, the production of luciferase protein from different mRNAs was examined by western blot analysis. Translation of the OPT and dOPT mRNAs resulted in dramatically different profiles of nascent luciferase peptides (Figure 2-4A). For the dOPT mRNA, most of the peptides seen are intermediate LUC species with a low level of the full-length protein. In contrast, translation of OPT mRNA resulted in mostly the full-length protein and much less intermediate LUC species. Remarkably, when N-OPT, M-OPT and C-OPT mRNAs, which each has a different region (codon 11–223, 224–423 or 424–550) of *Luc* codon-optimized in the dOPT background, are individually translated, we observed specific disappearance or reduction of the protein intermediate species within the expected molecular weight ranges. Furthermore, when the codons between 224–333 in the OPT luciferase gene were replaced by the least preferred codons in a series of 20 codon-windows, luciferase intermediates corresponding to the predicted molecular sizes appeared in the *in vitro* translated luciferase products (Figure 2-4B). Together, these results further demonstrate the role of codon usage in regulating elongation speed and show that codon usage regulates local ribosome movement on mRNA. Preferred codons speed up ribosome movement on mRNA and non-optimal codons cause ribosome stalling on mRNA.

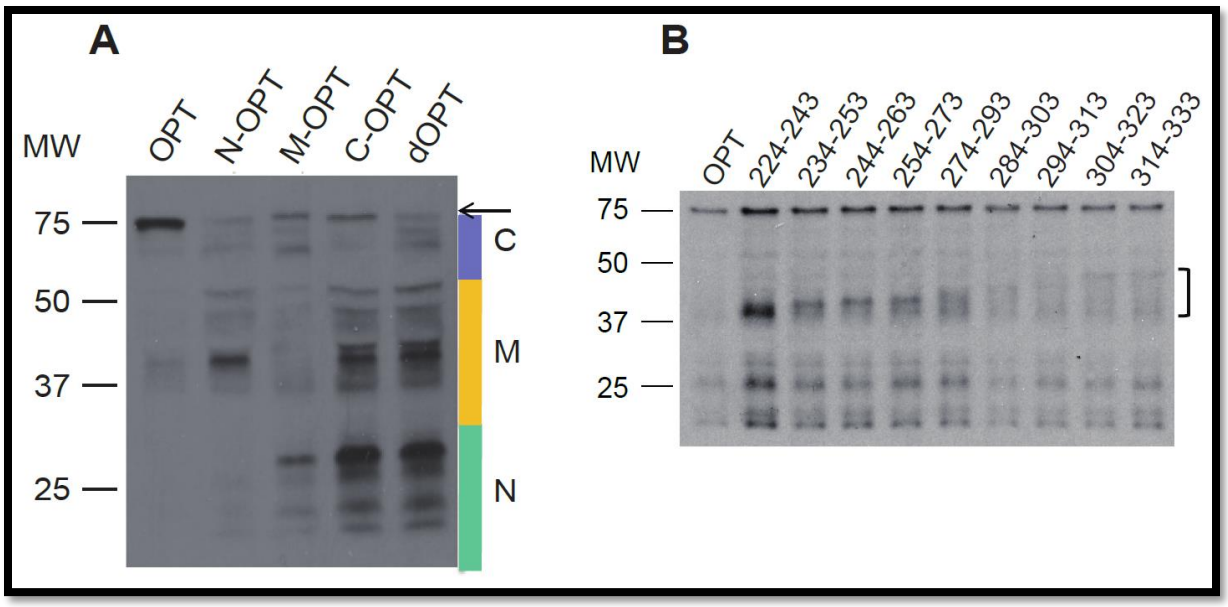

Figure 2-4 Non-optimal Codon Usage Results in Local Ribosome Stalling on mRNA

(**A**) Western blot analysis of denatured protein samples using c-Myc antibody showing the production of full-length and nascent luciferase peptides from the S2 cell-free extract translated *Luc* mRNAs. A five c-Myc tag was added to the N-terminal end of the *Luc* ORF for all constructs so that all nascent peptides could be detected by c-Myc antibody. The bars on the right represent approximate location of the N-terminal, middle part and C-terminal regions of LUC that were codon optimized, respectively. The arrow indicates the full-length LUC protein. The loadings for different samples were adjusted to produce similar full-length luciferase protein levels. (**B**) Western blot analysis showing the accumulation of nascent luciferase peptides at the locations of codon de-optimization of the Myc-OPT mRNA. The numbers on the top indicate the 20-amino-acid windows in which luciferase codons were de-optimized. The bracket on the right indicates the region including all scanning windows.

2.3.4 The Fate of Stalled Ribosomes on mRNA

To determine the fate of stalled ribosomes on mRNA, we added harringtonine, a translation initiation inhibitor, 6 min after the start of translation of the OPT mRNA and the translation of luciferase was monitored at different time points by western blot analysis. At 6 min of translation, only luciferase intermediates were observed (Figure 2-5). After the addition of harringtonine, the luciferase intermediates gradually disappeared or reduced, and the full-length luciferase appeared. This result indicated that at least part of the intermediate bands were caused by temporary ribosome pausing and these ribosomes were still capable of translating full-length products. However, some of the intermediate bands remained even after prolonged incubation time. It was not clear whether this was due to decreased potency of the extracts or pre-mature translation termination.

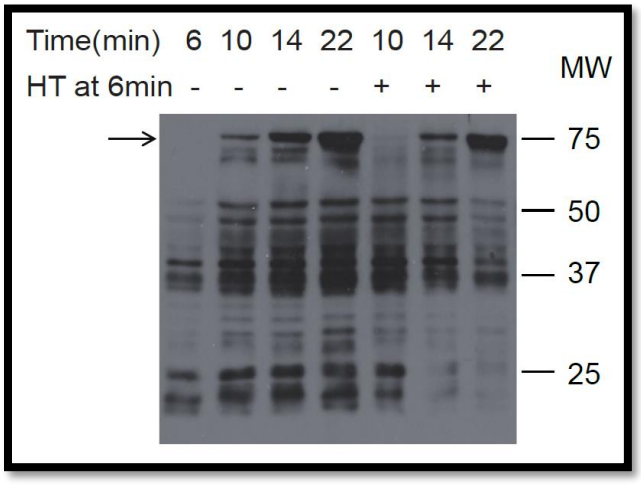

Figure 2-5 Part of Stalled Ribosomes Can Continue Translation.
Western blot analysis showing the profiles of luciferase nascent peptides at the indicated time points with/without harringtonine treatment. For samples treated by harringtonine, the drug was added after six min of *in vitro* translation reaction.

To address the above question, it is important to know whether the nascent peptides are associated with ribosomes. By isolation of ribosome-associate nascent chains, we found that in contrast to the full-length protein, the majority of low molecular weight luciferase species

were associated with ribosomes (Figure 2-6A), which indicated ribosome pausing and is consistent with the harringtonine treatment assay. On the other hand, some of the nascent peptides, especially when translating dOPT luciferase mRNA, were released into the supernatant, indicating that translation was pre-maturely terminated on these mRNAs. This result is consistent with our recent finding that rare codons may cause pre-mature translation termination in a eRF1 dependent manner [27]. Importantly, these intermediate species are not degradation products of the full-length protein since the *in vitro* translated luciferase protein species were very stable (Figure 2-6B).

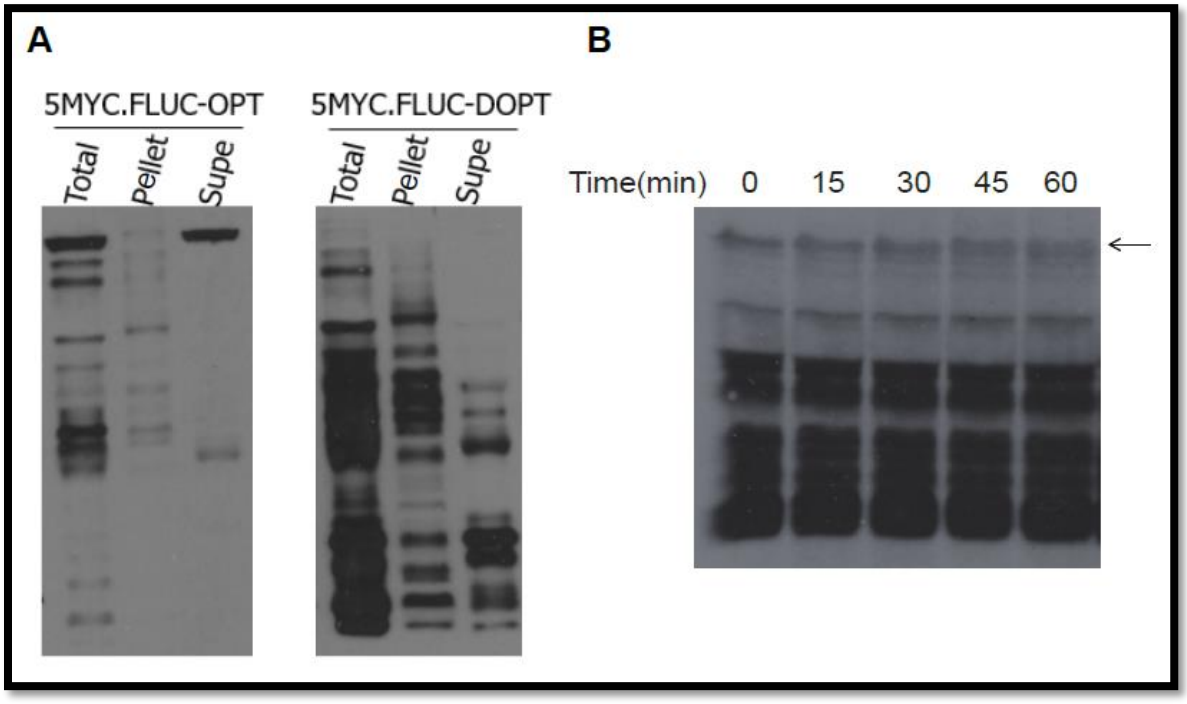

Figure 2-6 Part of Nascent Peptides were Released from Ribosomes.

(**A**) OPT or dOPT mRNAs were translated in vitro for 15 minutes, followed by sucrose precipitation. Total input protein, pellet and supernatant fractions are shown by western blot against c-myc tag. (**B**) Cycloheximide was added after 20 minutes of translation to stop the

reaction. The protein samples were then harvested at the indicated time points. The arrow indicates full-length protein products.

Interestingly, we also found that pre-mature translation termination may happen through frameshifting event for dOPT luciferase. When performing systematic mapping for intermediate bands, we found a slippery sequence in the N-terminus of dOPT luciferase which contains 7 consecutive adenosines. This slippery sequence associates with a strong intermediate band in the N-terminal region and when the AAAAAA codons were substituted by synonymous AAGAAG codons (DOPTΔlysine), the intensity of the intermediate band dramatically decreased (Figure 2-7). To show that the intermediate band is dependent on ribosome frameshifting, we mutated one (DOPTΔstop1) or two clusters (DOPTΔstop2) of the out-of-frame stop codons downstream of the slippery sequence and the intermediate band also shifted higher corresponding to the next out-of-frame stop codons. We know that the canonical -1 frameshifting requires not only the slippery sequence, but also a strong mRNA secondary structure downstream to slow down the ribosome [95]. However, we did not find any conserved mRNA secondary structure following the slippery sequence in dOPT mRNA. This result indicated that rare codons may have similar function with mRNA secondary structure in programmed frameshifting to slow down ribosome movement. Together, we identified two different fates of stalled ribosomes which is pausing and pre-mature termination, while pre-mature termination can happen through non-canonical frameshifting event specifically for dOPT luciferase mRNA.

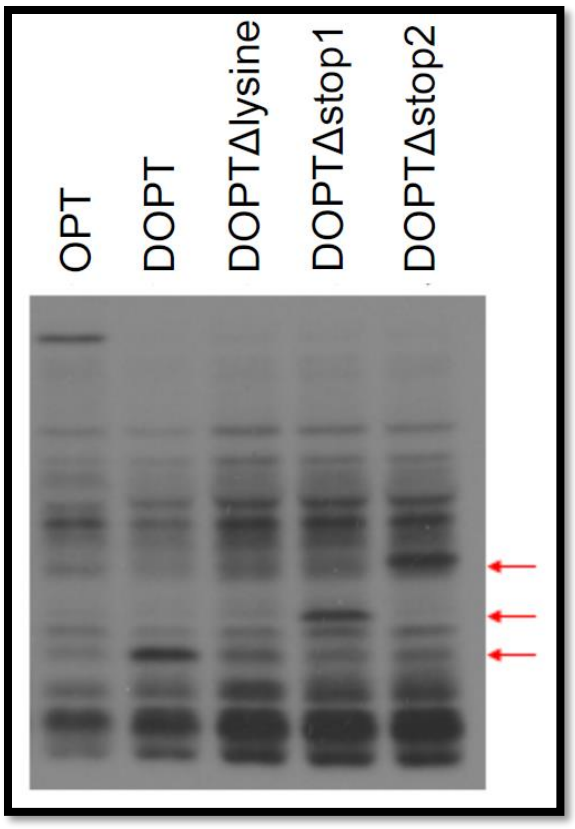

Figure 2-7 Frameshifting in the Translation of dOPT mRNA
Western blot analysis showing the profiles of luciferase nascent peptides translated from indicated mRNAs and blotted against the 5myc tag at the N-terminus. Arrows indicate the pre-maturely terminated nascent peptide caused by frameshifting. In DOPTΔlysine construct, 2 consecutive lysine codons AAAAAA in the N-terminus were mutated to AAGAAG. In the two DOPTΔStop constructs, only the out-of-frame stop codons were mutated.

2.3.5 A Specific 5-codon Sequence is Sufficient to Induce Ribosome Stalling Independent of Sequence Context

The scanning window assay described in Figure 2-4B was part of the attempt to investigate the minimal regions whose codon usage affects the intensity of one specific intermediate band in western blot. We thought that by identifying the minimal regions we may obtain more information on why ribosomes prefer to stall on these positions. It should be noted that codon usage may not be the major determinant of ribosome stalling positions since for most of the luciferase mRNAs used in this study, the intermediate products always appear at same positions regardless of their codon usage. We hypothesize that amino acid

sequence determines where the ribosomes stall, while codon usage determines the intensity of stalling. Although we finally narrowed down the minimal regions to the length of 5 codons, only a few intermediate bands showed significant intensity change after codon manipulation. Unfortunately, we did not have enough information to infer the features that may cause ribosome stalling. However, using one of the identified 5-codon sequences we were able to investigate the context dependency of ribosome stalling. Codons 366-370 were identified from the scanning window assay which influences the intensity of a specific intermediate band in western blot (Figure 2-8, lane 1-2). When we insert either OPT (OPT.insOPT) or DOPT (OPT.insDOPT) version of this 5-codon sequence after codon 185 of luciferase, only DOPT version of the sequence induce an intermediate band at the new position (Figure 2-8, lane 3-4). Together, these results indicated that specific DOPT sequences as short as 5 codons are sufficient to induce significant ribosome stalling independent of their sequence context.

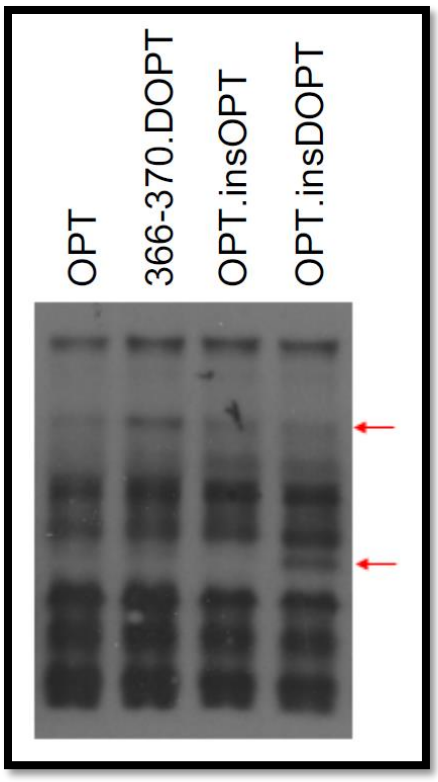

Figure 2-8 A 5-codon Non-optimal Sequence Sufficient to Induce Ribosome Stalling
Western blot analysis showing the profiles of luciferase nascent peptides translated from indicated mRNAs and blotted against the 5myc tag at the N-terminus. Arrows indicate the intermediate product associated with codon 366-370 at the original position (upper) or the inserted position (lower).

2.3.6 Codon Usage Affects Protein Structure In Vitro and in S2 cells

Since protein folds co-translationally, we predicted that the effect of codon usage on translation elongation speed should affect the time available for co-translational folding process, thus affecting protein structure. To test this hypothesis, we used limited trypsin digestion assay to probe the structural difference of the full-length luciferase proteins produced from OPT and dOPT *Luc* mRNAs by *in vitro* translation. As shown in Figure 2-9A, the OPT luciferase protein was found to be significantly more sensitive to trypsin digestion than the dOPT protein. To further confirm this conclusion *in vivo*, OPT and dOPT luciferase expressing constructs were individually transfected into S2 cells and we examined the trypsin sensitivity of the expressed luciferase in the S2 total cell extract. As

expected, the S2 cell-expressed full-length OPT luciferase was found to be more sensitive to trypsin digestion than the dOPT protein (Figure 2-9B). It is worth noting that no luciferase intermediates could be observed for the S2 expressed luciferase, likely due to their rapid degradation in cells. Importantly, the trypsin digestion of the full-length OPT luciferase resulted in luciferase fragments that could not be detected for the digestion of dOPT protein (Figure 2-9B), which serves as another evidence of their structural difference. Together, these *in vitro* and *in vivo* results indicate that codon usage affects protein structure.

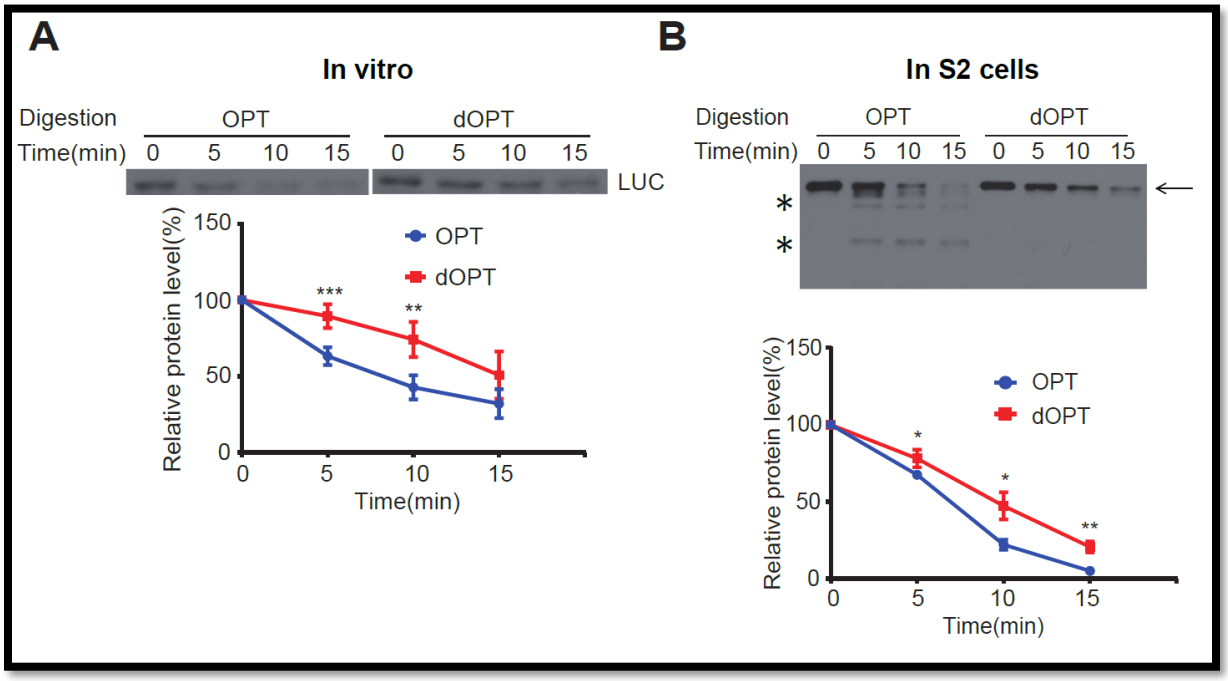

Figure 2-9 Trypsin Sensitivity of LUC Proteins Translated in vitro and in S2 cells.

(**A**) Limited trypsin digestion assay for the full-length OPT and dOPT luciferase proteins translated by the S2 cell-free system for 20 min. A representative western blot was shown in the upper panel. The lower panel shows the densitometric analysis results of LUC levels from independent experiments. $^{**}P < 0.01$, $^{***}P < 0.001$. $n = 5$ for OPT and $n = 4$ for dOPT

mRNAs. Area under curve (AUCs) for OPT and dOPT samples are 860.1 ± 42 and 1195 ± 62.64, respectively. (**B**) Limited trypsin digestion assay for luciferase protein produced in S2 cells after transfecting the OPT or dOPT expressing construct. A representative western blot was shown in the upper panel. The lower panel shows the densitometric analysis results of LUC levels from three independent experiments. *$P < 0.05$, **$P < 0.01$. AUCs for OPT and dOPT samples are 710.7 ± 13.73 and 927.8 ± 38.47, respectively.

2.3.7 Codon Usage Affects Protein Specific Activity in S2 Cells

The altered luciferase structure shown above suggests that codon usage may affect luciferase activity in cells. To confirm this, we transfected OPT and dOPT *Luc* constructs into S2 cells and measured luciferase activity and luciferase protein levels. As shown in Figure 2-10A and B, although the luciferase activity from cells expressing dOPT *Luc* was about 50% of the OPT *Luc*, the luciferase protein level from the OPT *Luc* was more than 20-fold higher than that of dOPT *Luc*. Thus, the specific activity of dOPT luciferase is about 10-fold of that of OPT luciferase (Figure 2-10C). Thus, codon usage of luciferase mRNA greatly affects luciferase activity in *Drosophila* cells.

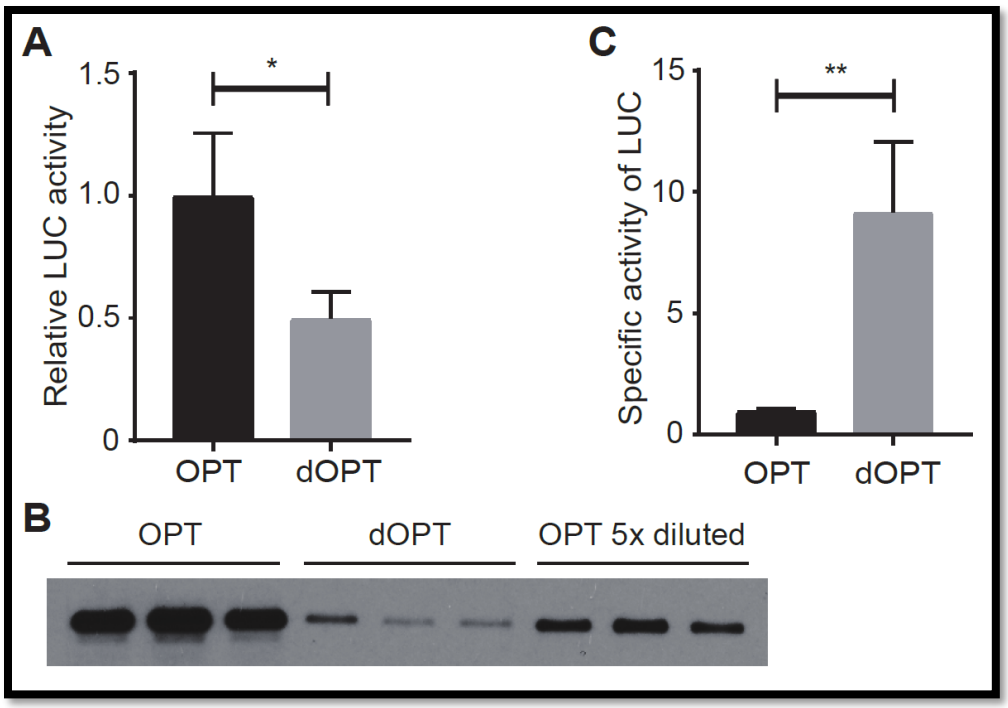

Figure 2-10 Specific Activity of LUC Proteins Translated in S2 cells.

(**A**) Relative luciferase activity of S2 cells transfected with OPT and dOPT constructs. Same amount of protein extracts was used for each sample. Data were obtained from three independent experiments. (**B**) Western blot analysis showing the levels of the full-length luciferase in the OPT or dOPT transfected S2 cells from three replicates. (**C**) Specific activity of OPT and dOPT luciferase proteins calculated by dividing luciferase activity level by corresponding protein level. $^{**}P < 0.01$. The signals from the OPT 5×diluted samples in (**B**) were used to determine the levels of the OPT LUC protein.

2.3.8 Codon Usage Affects mRNA Stability in Drosophila S2 cells

In addition to the difference in protein levels, the steady state level of OPT *Luc* mRNA was much higher than that of the dOPT mRNA (Figure 2-11A), which largely explains the

low level of luciferase protein in cells expressing dOPT construct. To determine whether the difference in mRNA levels is due to altered mRNA stability, we measured mRNA degradation rate after the addition of actinomycin D, a transcription inhibitor. As expected, the OPT *Luc* mRNA was significantly more stable than the dOPT mRNA after adding actinomycin D (Figure 2-11B and C). Therefore, similar to the reports in other organisms [71-73, 77, 96], codon usage also affects mRNA stability in Drosophila cells.

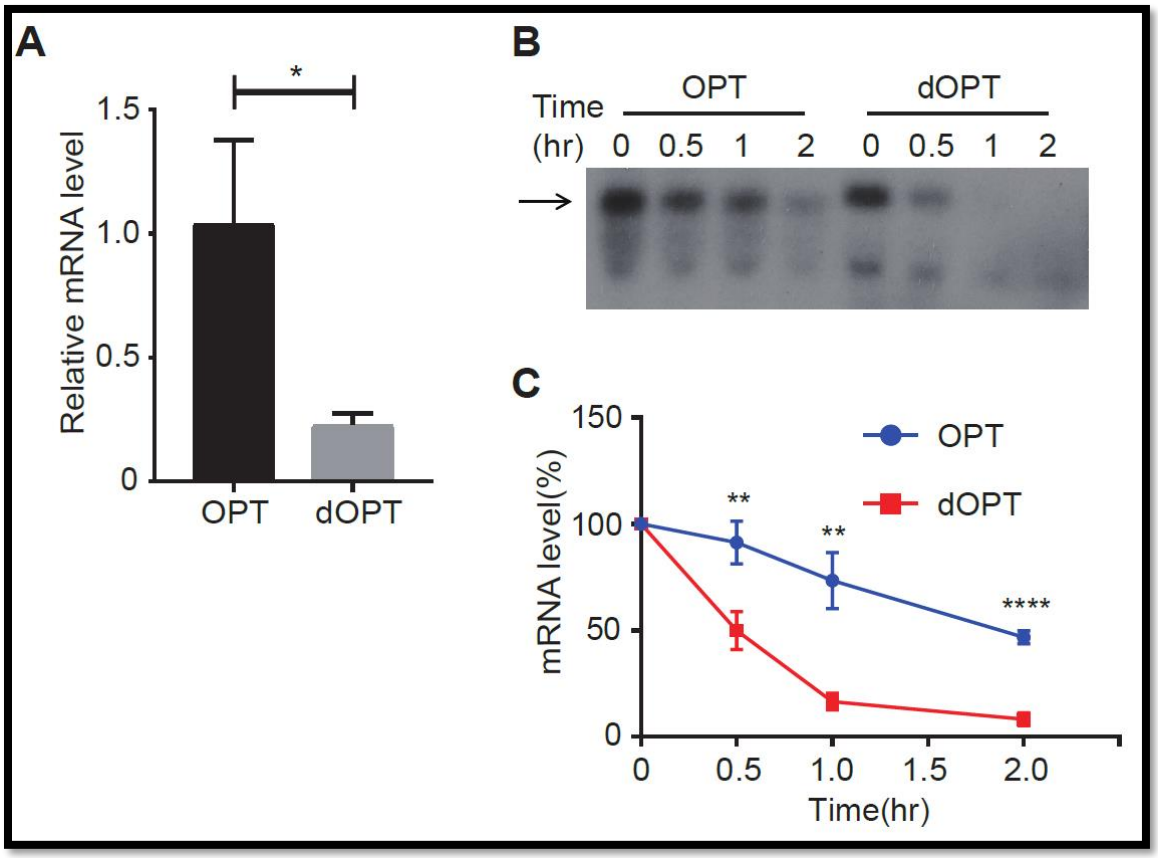

Figure 2-11 Codon Usage Influences mRNA Stability in S2 Cells.

(**A**) qRT-PCR analysis showing the relative levels of *Luc* mRNA from the OPT or dOPT transfected S2 cells. Error bars represent standard deviations from three independent replicates. *$P < 0.05$. (**B**) Representative northern blot analysis result showing the level

of *Luc* mRNAs at the indicated time points after the addition of actinomycin D. The arrow indicates the full-length luciferase mRNA. (**C**) Quantification of three independent results of the experiment in (B) was shown. ** $P < 0.01$. **** $P < 0.0001$. AUCs for OPT and dOPT samples are 149.1 ± 8.306 and 66.5 ± 4.128, respectively.

2.4 Discussion

Although codon usage has been previously shown to regulate co-translational protein folding in fungi by affecting translation elongation speed, its effect in animal systems is not known. By using a cell-free *Drosophila* S2 translation system, we demonstrated here that codon usage plays an important role in regulating the speed of translation elongation: optimal codons enhance the rate of elongation while non-optimal codons slows it down (Figure 2-1 and 2-2). In addition, as shown by the accumulation of nascent intermediate peptides during translation (Figure 2-4), we showed that ribosome stalling occurs at the sites of non-optimal codons, indicating that codon usage regulates local ribosome movement speed on mRNA. Furthermore, we showed *in vitro* and in S2 cells that different codon usage biases led to structural differences of proteins with the same amino acid sequence, resulting in proteins with different functional activity (Figure 2-9 and 2-10). Together with our previous study showing the importance of codon usage for the function of a circadian clock protein in *Drosophila* [49], these results demonstrate that the role of codon usage in regulating protein structure and function by affecting translation elongation speed is conserved in fungi and *Drosophila*.

Bioinformatic analyses have uncovered correlations between codon usage and protein structural motifs in different fungal organisms [57-61]. In addition, we recently showed that non-optimal codons are preferentially used in intrinsically disordered regions, and optimal codons are more frequently used in structured domains in yeast, *Neurospora*, *Caenorhabditis elegans* and *Drosophila* systems [60]. Importantly, the functional importance of such correlations *in vivo* was confirmed by structure-based codon manipulation in the *Neurospora* and *Drosophila* circadian clock genes [49, 60]. Together with our current study, these studies suggest the existence of a codon usage code within genetic codons that generates rhythms of translation elongation rate optimal for protein structure and function in eukaryotic organisms.

Recently, codon optimality has been shown to affect mRNA stability due to its role in affecting translation elongation in budding yeast [71, 77]. A similar role for codon usage was also proposed in zebrafish and *Xenopus* systems [73, 96]. Consistent with these studies, we showed that codon usage manipulation of the luciferase gene altered *Luc* mRNA stability in *Drosophila* cells (Figure 2-11). Therefore, codon usage may also affect mRNA stability due to its role in regulating translation elongation rate in animals.

CHAPTER THREE

MULTIPLE CHROMATIN REGULATORY FACTORS INFLUENCE THE CODON USAGE EFFECT ON TRANSCRIPTION

3.1 Introduction

Codon usage bias can be found in all organisms and it is defined as the unequal usage of synonymous codons between different genes or within a single gene [1, 3]. The evolutionary driving force of codon usage bias has been discussed intensively. It is widely accepted that mutational bias and natural selection both contribute to shaping codon usage pattern, while selection pressure for higher translation efficiency and accuracy plays a more important role [2, 4, 10, 18, 35, 70, 82]. As a result, codon optimization was thought to be able to increase translation efficiency [14, 83, 84] and frequently used in heterologous gene expression [16]. On the other hand, although some synonymous codons are not preferred on the genome level, they are necessary in some genes to produce functional protein products [48, 49, 58, 60] or keep protein expression level of oncogene at a low level to prevent tumorigenesis [46]. As the basis of most regulatory roles of codon usage, it is important to know how codon usage affects translation elongation process. Recently, many experimental evidences were reported regarding the effects of codon usage on translation elongation and co-translational folding [40, 42, 50, 51, 55, 57, 97]. Interestingly, we recently showed that non-preferred codons may cause pre-mature translation termination in an eRF1 dependent manner, which suggested that

translation termination is also regulated by codon usage [27]. While these discoveries indicated the complexity of translational regulation by codon usage, little is known about the effects of codon usage beyond translation.

Although most studies on codon usage were focused on translational regulation, more and more evidences suggested that translation efficiency is mainly determined by translation initiation efficiency, which is strongly dependent on mRNA secondary structure in the proximal region of the start codon [67-69]. These discoveries questioned on how codon usage regulates protein expression level. Interestingly, multiple reports showed recently in E.coli, yeast, zebrafish and human cells that codon usage could affect mRNA level by regulating mRNA stability [71-74]. Further studies in yeast revealed two different translation dependent mechanisms underlying the effects of codon usage on mRNA stability. The first mechanism is dependent of the DEAD-box protein Dhh1 which recognizes slow-moving ribosomes and targets the corresponding mRNAs for degradation [77]. More recently, another mechanism was proposed in which transcripts with different translation elongation speeds are protected differently by Poly(A)-binding protein Pab1, leading to distinct mRNA decay processes by Ccr4-Not complex [78]. Our lab recently also showed in both Neurospora crassa and mammalian cells that codon optimization of either endogenous or exogenous reporter genes could directly increase their mRNA levels [75, 76]. More importantly however, we found that the elevation of mRNA level was not likely caused by different RNA stability, while evidences of increased transcriptional level were obtained. Similar observations were also reported when researchers performed codon optimization on TLR7 gene in mammalian cells [66]. Collectively, these studies suggested that codon usage affects

mRNA level by regulating both mRNA transcription and mRNA stability. However, whether codon usage also affects genome-wide mRNA transcription and the underlying mechanisms remain unknown.

Using multiple high-throughput sequencing techniques, we showed in this study that on the genome level, mRNA transcription is associated with codon usage similarly as total mRNA level. Combining with the codon optimized reporter assays in our previous studies, it is reasonable to hypothesize that codon usage has a regulatory effect on mRNA transcription process genome wide. From a blind screen of ~250 Neurospora crassa knockout strains, we identified 18 candidate factors which potentially mediate the effect of codon usage on mRNA level. Further analyses on two specific strains Δset-2 and ΔFKH1 revealed that their regulation of mRNA level is largely on transcriptional level. Overall, we provided evidences suggesting that multiple factors and pathways may be involved in regulating genome wide mRNA transcription based on the codon usage of different genes.

3.2 Materials and Methods

3.2.1 Neurospora crassa Strains and Culturing

The strains used in this study are listed in APPENDIX B. All the strains were cultured in the same condition except for the strains used in total mRNA sequencing by Joint Genome Institute. Briefly for mRNA-seq strains, they were inoculated from the knockout library of Fungal Genetics Stock Center onto minimal slants (3% sucrose, 1x Vogel's salt solution [98] and 1.5% agar). After 8~21 days of growth, a suspension of the conidia was inoculated into

50ml liquid medium (2% sucrose and 1x Vogel's salt solution) in a 150mm petri dish. Depending on the growth rate of the strains, the mycelium was collected after 30~48 hours and separated into 3 replicates, each in 50ml liquid medium. The mycelium was grown on a shaker at 30 °C for 24 hours under constant light. The tissue samples were then collected for subsequent experiments. For other experiments, 2% glucose was used instead of sucrose for all liquid medium and the tissue samples were grown at room temperature. Specifically, for nuclear RNA-seq experiment, higher amount of mycelium was inoculated in 1L liquid medium to obtain more tissue samples for nuclei extraction.

3.2.2 Total mRNA Sequencing

Total RNA was extracted from tissue powder by TRIzol reagent and treated with DNase. Standard RNA sequencing libraries were then created and sequenced on NovaSeq by Joint Genome Institute.

3.2.3 Nuclear RNA Extraction

From each tissue sample prepared for nuclear RNA extraction, a small proportion was separated for total RNA extraction as a control. 3.5~4g of frozen tissue was ground into powder with 2g of glass beads in liquid nitrogen. 8ml of cold Buffer A (1M sorbitol, 7% ficoll, 20% glycerol, 5mM Mg(OAc)2, 3mM CaCl2, 50mM Tris.HCl pH 7.5, 3mM DTT added before use) was added and incubated on ice for 10min with stirring. The samples were then filtered through 2 layers of cheesecloth and the volume brought to 8ml by Buffer A. Subsequently, 16ml of Buffer B (10% glycerol, 5mM Mg(OAc)2, 25mM Tris.HCl pH 7.5)

was added slowing with continuous mixing. The samples were then layered onto 10ml of cold Buffer A/B (2.5:4) and centrifuge at 3000g at 4 °C for 7min to pellet cell debris. 1ml of the supernatant was kept as total fraction and the rest was layered onto 5ml of cold Buffer D (1M sucrose, 10% glycerol, 5mM Mg(OAc)2, 25mM Tris.HCl pH 7.5, 1mM DTT added before use). Nuclei were pelleted at 9400g at 4 °C for 15min. 1ml of the supernatant was kept as cytosolic fraction and a small amount of the pellet was collected in SDS buffer for western blot analysis. Nuclear RNA was extracted from the rest of the pellet by TRIzol Reagent.

3.2.4 Nuclear RNA Sequencing

Nuclear RNA and corresponding total RNA samples were treated with DNase before library construction. After quality control, total RNA samples were used in making standard polyA mRNA sequencing libraries for Illumina, while nuclear RNA samples did not go through polyA purification process to keep both mature and pre-mature RNAs. All the libraries were then sequenced on NextSeq 500.

3.2.5 RNA Sequencing Data Processing

For total RNA sequencing, data analyses were performed by Joint Genome Institute following a standard pipeline. In the analyses of correlation between mRNA level and codon usage bias, a consistent set of 4136 genes with moderate to high expression levels were used. When we selected candidate strains with significantly lower correlation coefficients comparing to wildtype strains, we set the false discovery rate to be 0.02. To analyze the codon usage bias of differentially expressed genes, we selected genes with FPKM over 1 and

fold change over 2 comparing to wildtype strains. Specifically, we excluded the genes without statistical significance, regardless of their fold change. For nuclear RNA sequencing, raw reads were first aligned to the Neurospora crassa genome using Bowtie2 [99] with the "--local" parameter in order to align both mature and pre-mature mRNA fragments. Raw counts for each gene were then obtained by HTseq [100] and used as the input for DEseq2 [101] for differential expression analyses. For total RNA sequencing controls, STAR [102] was used instead of Bowtie2 as a splicing-sensitive aligner. Due to relatively lower sequencing depth, differentially expressed genes were selected based on the adjust p values (less than 0.1) instead of using fold change cutoff.

3.2.6 2P-seq and ChIP-seq Data Processing

Processed data of 2P-seq and ChIP-seq were obtained from previous studies in our lab [75, 80]. The same set of 4136 genes selected from RNA-seq were used in both analyses. For 2P-seq data, we used raw counts to perform correlation analyses since each read represents a single transcript in 2P-seq. For ChIP-seq data, reads per million mapped reads (RPM) of each gene was normalized by the RPM of negative control (no antibody) and the ratio was used as the relative enrichment of RNA polymerase II.

3.2.7 Calculation of Codon Bias Index

The CBIs for all Neurospora crassa transcripts were calculated using CodonW [103]. For genes with multiple variants, only the CBI of one variant was used to represent the CBI of the gene.

3.2.8 Neurospora crassa Growth Rate Measurement

We used race tube to measure the growth rate of a selected set of stains. Each strain was inoculated at one end of the tube and grew toward the other end at room temperature under constant light. The length between the front line and the start line was measured every 24 hours and the average growth rate was represented as millimeter per day.

3.2.9 Clustering of Candidate Strains

We first generated a distance matrix between the 18 candidate strains by the following steps. For all the candidate strains, genes with average FPKM less than 1 were excluded from further analyses. Then for any pair of the strains, their differentially regulated genes were combined and the distance between them was calculated as the spearman correlation of Log_2(Fold change) of these genes between the two strains. Finally, the distance matrix was used for hierarchical clustering and plotting the heatmap.

3.2.10 ChIP-qPCR Experiment

Neurospora tissue samples grown in liquid medium were fixed by 1% formaldehyde for 15min and then stopped by 0.1M glycine. The tissue samples were harvested and ground to powders in liquid nitrogen. Tissue powders were then suspended in 8ml of ChIP lysis buffer per 1g of tissue powder for sonication using Bioruptor. The chromatin was fragmented to ~200bp and the supernatant was diluted to 1ug/ul total protein concentration. Aliquots of lysate containing 500ug of total protein were incubated with 1~2ul of RNAP II CTD

(Abcam; ab26721) or S2p (Abcam; ab5095) antibody overnight at 4 °C. 1/10 of the lysates were saved as input samples. Then 20ul of magnetic G protein coupled beads were added and incubated for 3~4 hours. The beads were then sequentially washed by lysis buffer, low-salt wash buffer, high-salt wash buffer, LNDET (0.25 M LiCl, 1% NP40, 1% deoxycholate, 1 mM EDTA) buffer and TE buffer. Next, the beads were eluted twice using elution buffer (1% SDS, 0.1M $NaHCO_3$). The input samples and eluted samples were de-crosslinked by 0.2M NaCl at 65 °C overnight and treated with proteinase K at 45 °C for 1 hour. Then DNA was extracted using Phenol: Chloroform: Isoamyl Alcohol (25:24:1, v/v) used in qPCR experiments. qPCR primers target either the 5' UTR or the 3' end of CDS of target genes. The immunoprecipitated DNA levels were first normalized by tubulin level and then the ratios of normalized DNA levels between Δset-2 mutant and wildtype strain were used as fold enrichment. The target genes were selected according to total and nuclear RNA-seq.

3.3 Results

3.3.1 Codon Usage Associates with Genome-Wide mRNA Transcription

Using total RNA sequencing, we first confirmed the correlation between mRNA level and codon usage bias seen in our previous study [75]. We selected a consistent set of ~4000 highly expressed genes to perform all the correlation analyses. And we observed a Pearson correlation coefficient of 0.58 between Codon Bias Index (CBI) and normalized total mRNA level (Figure 3-1A, left), which is consistent with our previous data. To investigate whether the effect of codon usage is mediated by mRNA transcription, we performed nuclear RNA

sequencing which could eliminate the potential effect of translation mediated mRNA decay [77]. From nuclear RNA sequencing we obtained a slightly lower correlation coefficient (0.49), which indicated that mRNA transcription contributes to a large proportion of the genome-wide correlation (Figure 3-1A, right). To verify the transcriptional regulation, we also analyzed the 2P-seq data and RNA polymerase II ChIP-seq data performed in our previous studies. In 2P-seq, poly(A) selected mRNAs are partially digested by RNase T1 and then the 3' end mRNA fragments associated with poly(A) tails are selectively reverse transcribed by a poly(T) primer and go through normal library construction process for high-throughput sequencing [104]. Our lab previously used this method to identify non-canonical transcription termination sites. However, in this study we only used 2P-seq to obtain mRNA expression level, which is represented by the raw counts of 3' end mRNA fragments. Surprisingly, in this analysis the nuclear RNA level even showed higher correlation coefficient with CBI, potentially due to different experimental conditions and genetic background of the strains used comparing to mRNA-seq (Figure 3-1B).

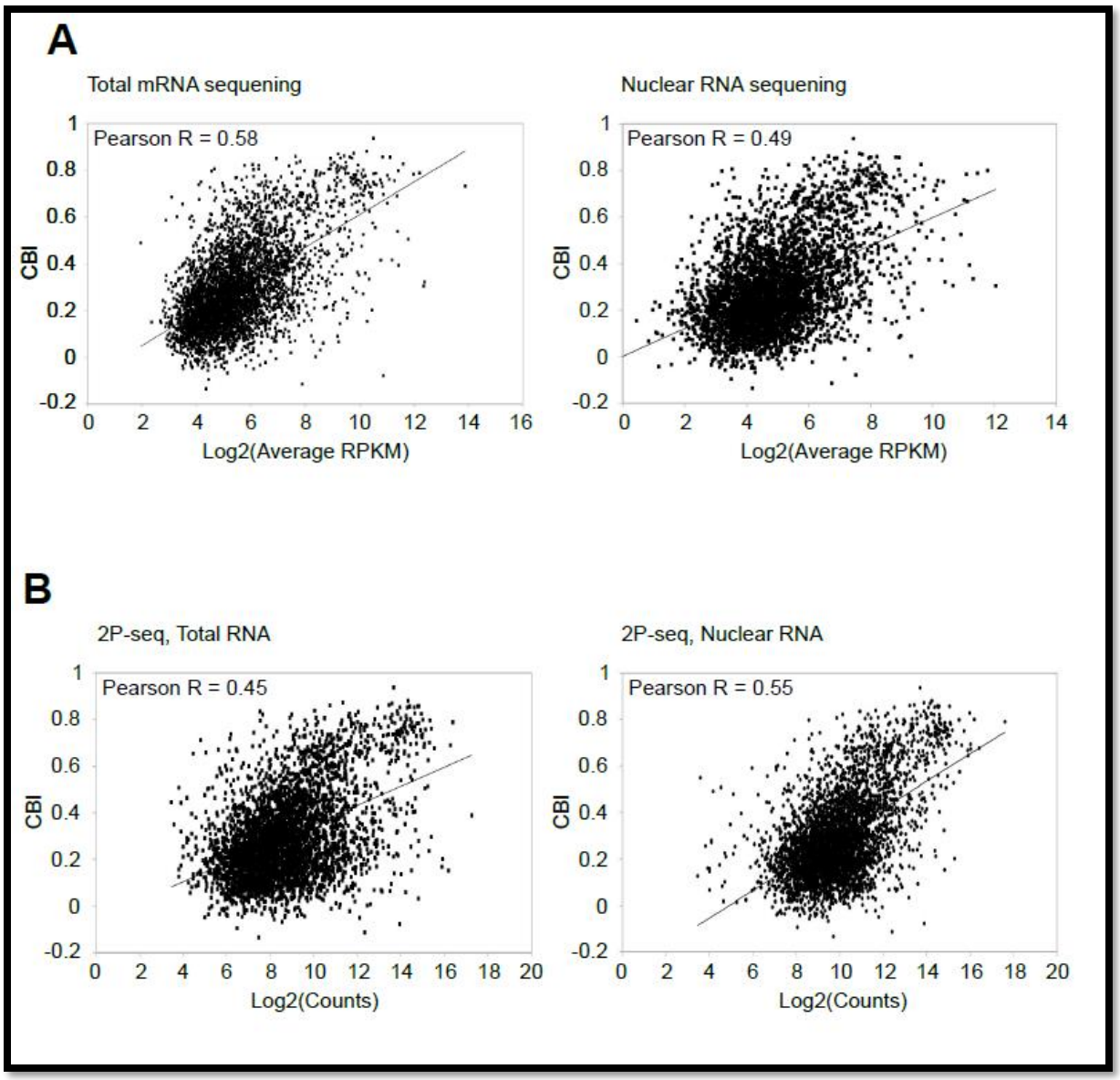

Figure 3-1 Genome-wide Association between Codon Usage Bias and Nuclear mRNA Level Measured by mRNA-seq or 2P-seq.

(**A**) Pearson correlation between CBI and total mRNA level (left) or nuclear mRNA level (right). Correlation coefficients were indicated at upper left corner. mRNA levels were measured by RNA sequencing of FGSC4200 wildtype strain and represented as

Log_2(Average RPKM/FPKM). RPKM/FPKM were averaged for each gene from 3 replicates. (**B**) Same analyses as in (A). mRNA levels were measured by 2P-seq of 87-3 wildtype strain and represented by Log_2(Raw Counts). Each panel showed a representative analysis from 2 replicates.

Additionally, in ChIP-seq experiments using two different RNA polymerase II antibodies targeting either non-phosphorylated C-terminal domain or ser-2 phosphorylated C-terminal domain, we could also observe moderate positive correlation between codon usage bias and RNAP II enrichment (Figure 3-2). Although RNAP II enrichment can better reflect mRNA transcription level, it is known that ChIP-seq experiment is not very quantitative due to antibody specificity and sensitivity, signal normalization method, etc. Collectively, we used 3 different high-throughput sequencing techniques to show the potential association between codon usage bias and mRNA transcription. Although we did not directly measure mRNA transcription level, the evidences we presented here strongly indicated genome-wide transcriptional regulation by codon usage considering the reporter assays previously shown by our lab [75]. More importantly, we showed clearly that the correlation between steady state mRNA level and CBI is largely independent of translation in Neurospora, which reflects additional roles of codon usage not relying on tRNA recognition. Therefore, the potential mechanisms underlying the correlation worth further investigation no matter whether genome-wide transcription is indeed affected by codon usage or not.

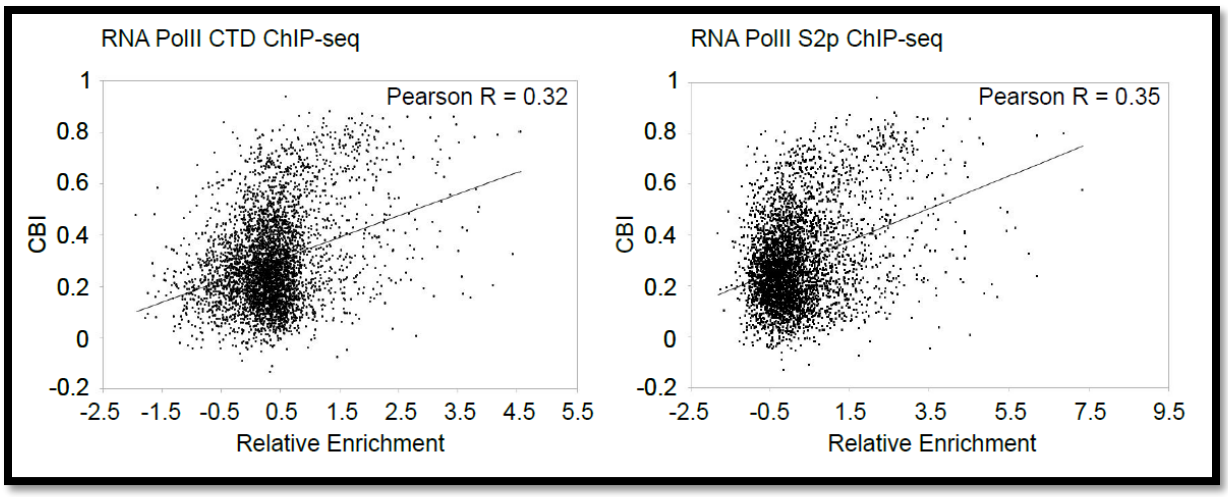

Figure 3-2 Genome-wide association between codon usage bias and RNAP II Enrichment measured by ChIP-seq.

Pearson correlation between CBI and RNA polymerase II enrichments. RNAP II occupancy was measured by ChIP-seq of 87-3 wildtype strain using antibodies targeting either non-phosphorylated C-terminal domain (left) or ser-2 phosphorylated C-terminal domain (right) of RNAP II. The relative enrichment was calculated as the ratio of normalized reads between experimental group and no-antibody control group.

3.3.2 Frequencies of Individual Optimal or Non-Optimal Codons Correlates Reversely with Total and Nuclear mRNA levels

Although CBI values can reflect synonymous codon usage bias on gene level, it does not show how individual codons associate with mRNA level in the genome. To address this question, we analyzed the correlation between single codon frequency in all genes and their mRNA levels for all the codons except the 3 stop codons. Surprisingly, we found that all the optimal codons positively correlated with mRNA levels, while all the rare codons negatively

correlated with mRNA levels on genome level (Figure 3-3A). Optimal and rare codons are defined by their usage frequencies in the genome. In addition, we obtained very similar correlation for all the codons when we used nuclear mRNA level to perform the same analysis (Figure 3-3B). Both the correlation coefficients and the rankings of all codons are very similar between total and nuclear samples (Figure 3-4), which is consistent with the previous analyses using CBI to measure codon usage bias. These results indicated that mRNA transcription could potentially be regulated on single codon level, although different codons may have different effects.

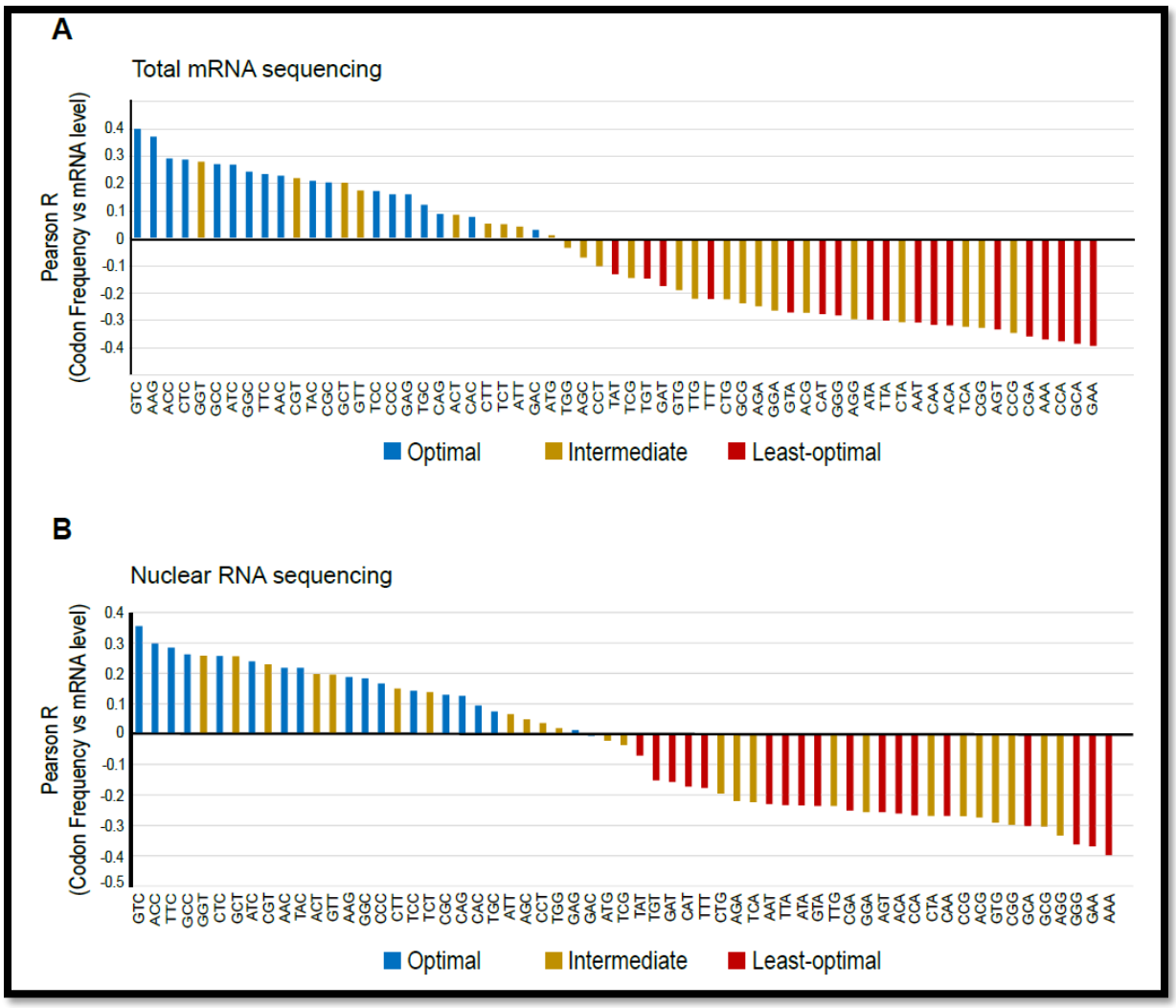

Figure 3-3 Codon Level Correlation with Total or Nuclear mRNA Level.

(**A**) Pearson correlation coefficients for all the 61 codons calculated between codon frequencies and mRNA levels. mRNA levels were measured by total RNA-seq of FGSC4200 wildtype strain and averaged over 3 replicates. Optimal codons (blue) are the ones most frequently used in Neurospora crassa genome within each synonymous codon family. Conversely, least-optimal codons (red) are least frequently used. And the rest are classified

as intermediate (yellow). (**B**) Same analysis as in (A). mRNA levels were measured by nuclear RNA-seq of FGSC4200 wildtype strain and averaged over 2 replicates.

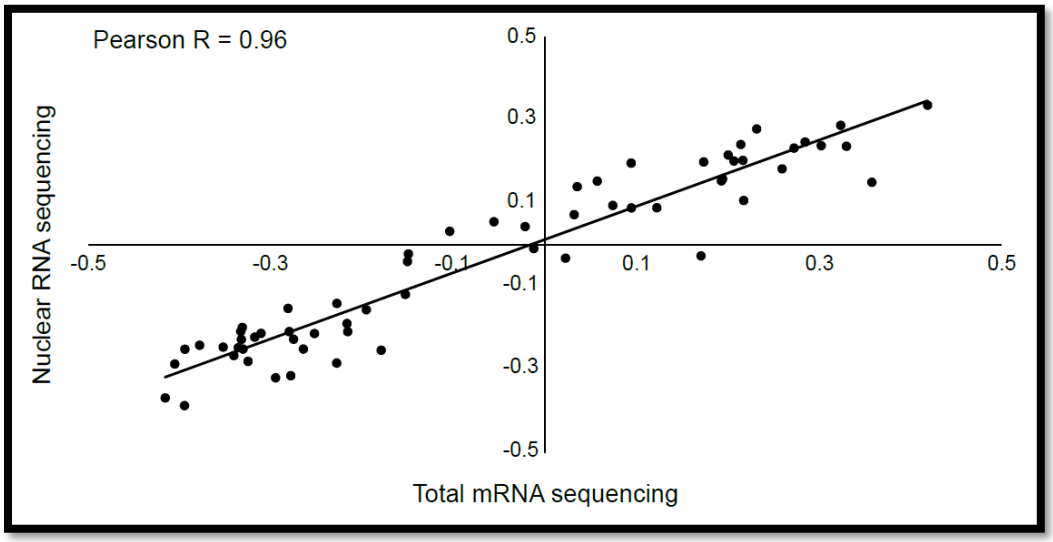

Figure 3-4 Individual Codons Correlate Similarly with Total or Nuclear RNA.

The correlation coefficients between individual codon frequencies and nuclear mRNA levels are plotted against the correlation coefficients between individual codon frequencies and total mRNA levels. The two sets of correlations are highly similar indicated by R value of 0.96.

3.3.3 Multiple Factors Mediate the Correlation between mRNA Level and Codon Usage Bias

To find out the factors mediating the correlation between mRNA level and CBI, we performed a systematic screen for ~250 Neurospora crassa knockout strains using mRNA sequencing. The knockout genes include most of the known transcription factors, chromatin remodeling factors, histone modifiers, histone variants, etc. The screening process is briefly shown in Figure 3-5. For each strain, we compared the correlation coefficient between

mRNA level and CBI to that of wildtype strain. And 18 strains with significantly lower correlation coefficients were selected as candidates (Figure 3-6A). The proteins knocked out in these strains can potentially mediate codon usage effect on mRNA level. Among the 18 candidate genes there were 10 transcription factors, 3 potential chromatin remodeling factors, 2 histone deacetylases, 2 methyltransferases and 1 RNA binding effector protein. Importantly, although some of the candidate strains showed lower growth rate, the decrease of correlation coefficients is not likely a secondary effect. Using an example batch of strains, we showed that the correlation coefficients are not associated with the strains' growth rate (Figure 3-6B). It should be noted that growth rate is only one simple measurement of the strains' physiological status and this experiment cannot exclude the possibility that the decrease of correlation coefficients is the secondary effect of other alterations in the mutant strains. Interestingly, we observed significantly higher correlation coefficients when we cultured wildtype strains at lower temperature, while poor medium or rich medium did not affect the correlation dramatically (data not shown). These data indicated that the correlation between codon usage bias and mRNA level is under very complex regulation, which is consistent with the fact that many candidate factors stood out from the screening.

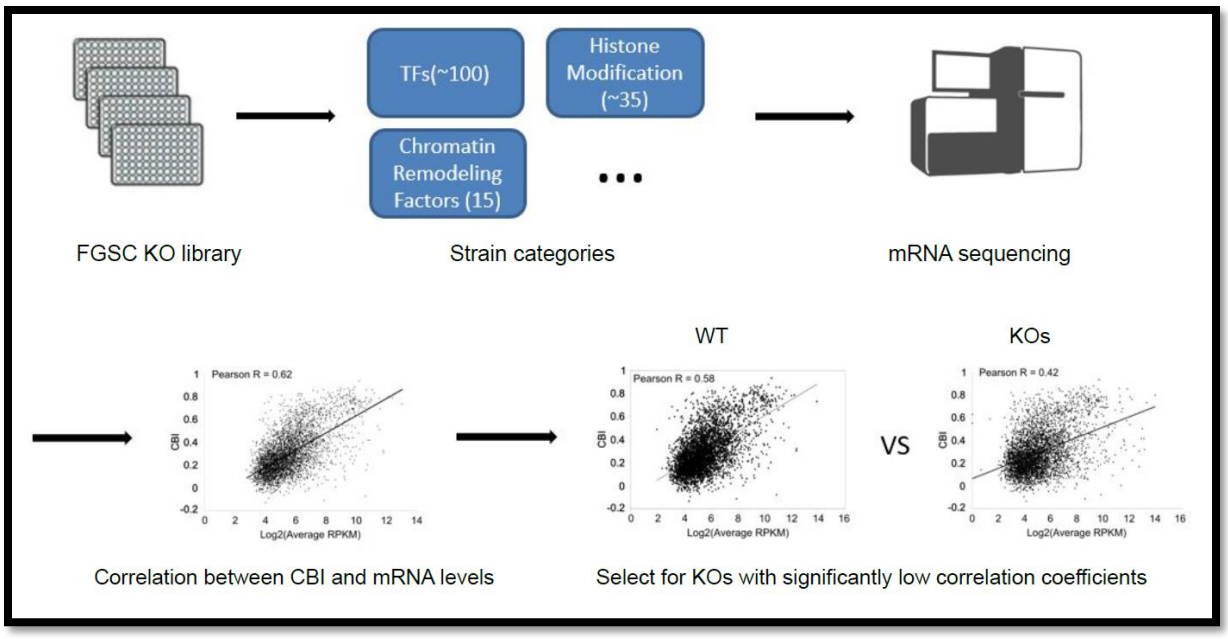

Figure 3-5 Schematic of Genetic Screening Process

The KO strains were inoculated from Neurospora KO libraries. Total RNA samples were prepared from these strains and went through library construction and high-throughput mRNA-seq. For each strain, the Pearson correlation between CBI and normalized mRNA levels was calculated and compared to the correlation of wildtype strain. After correction for multiple comparisons, strains with significantly lower correlation coefficients were selected as candidates.

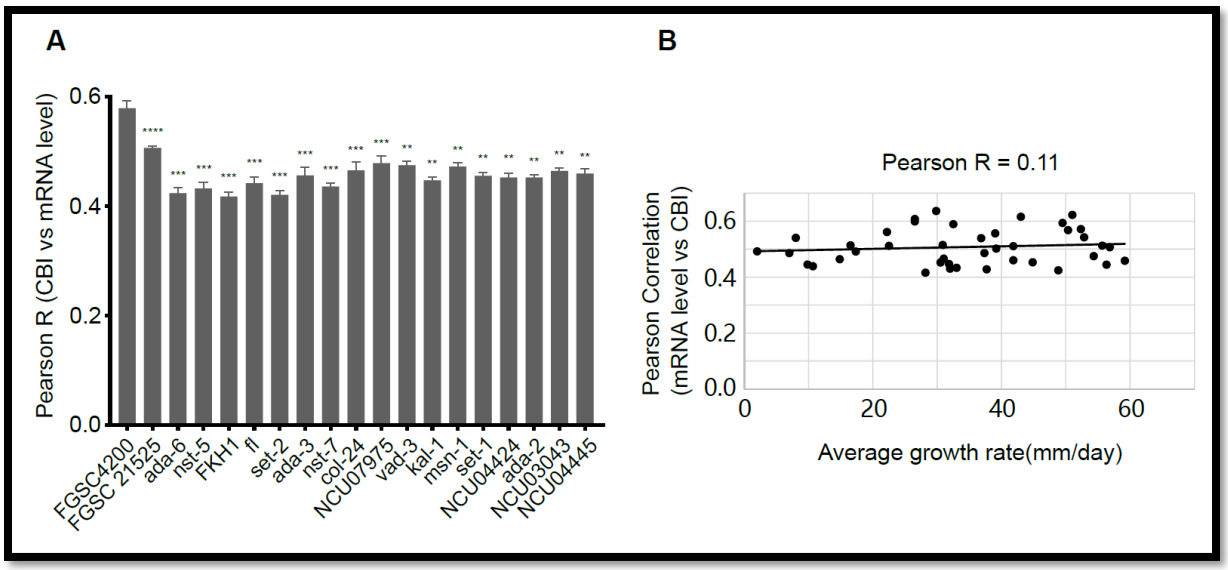

Figure 3-6 Pearson Correlation Coefficients (CBI vs mRNA level) of Candidate Strains are Significantly Lower than That of Wildtype Strain.

(A) The Pearson R values of candidate strains were compared to that of wildtype strain using bar plot. 3 replicates of each strain were used for statistical analysis. **P < 0.01, ***P < 0.001, ****P < 0.0001. (B) The Pearson correlation coefficients of a selected set of strains were plotted against their average growth rate.

We showed the scatter plot of five representative candidate strains with lowest correlation coefficients comparing to wildtype strain (Figure 3-7). Of these candidate factors, set-2 is an H3K36 specific methyltransferase; nst-5 is an NAD dependent deacetylase; FKH1 and ada-6 are transcription factors; NCU04424 is a potential chromatin remodeling factor. It's worth noting that ΔDhh1 and ΔDbp2 strains were also included in our screen. These two proteins were previously shown to have regulatory effects on codon-mRNA correlation in a translation dependent manner in yeast [77, 105]. However, we did not observe significant

decrease of correlation coefficient after knocking out either of the two genes (Figure 3-8). This is consistent with our previous findings that the correlation between mRNA level and CBI is largely translation independent in Neurospora crassa (Figure 3-1 and 3-2). Specifically, we are not arguing against the regulatory effects of codon usage on mRNA stability. In fact, we previously also observed that codon usage can influence mRNA stability in specific situations in Neurospora crassa, although the effect was not dramatic and universal for most genes examined. However, according to the deep sequencing results, this mechanism involving mRNA stability may not account for the genome-wide correlation between codon usage and mRNA levels, or the codon usage effects on mRNA levels may not be mediated by these two proteins in Neurospora.

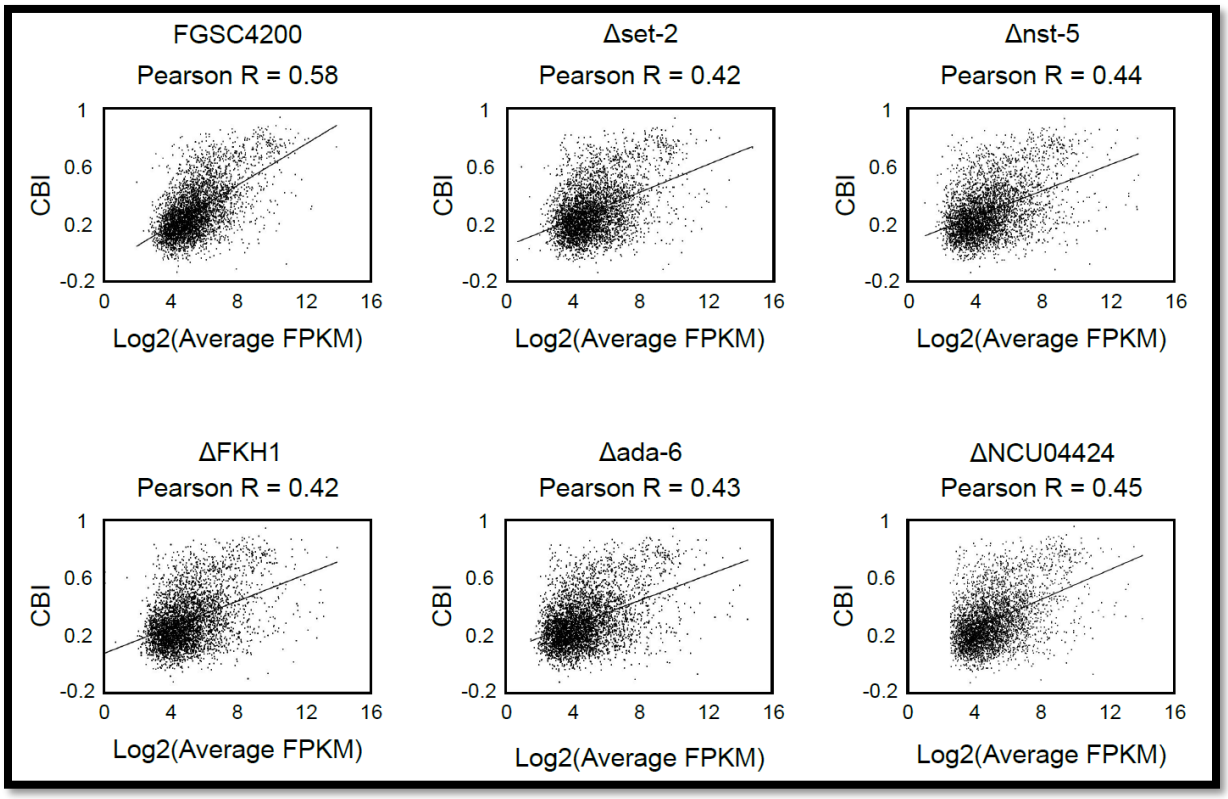

Figure 3-7 Scatter Plots of the Correlation Between CBI and Total mRNA Level.

Pearson correlations are shown as scatter plots for wildtype and 5 candidate strains with lowest Pearson R values. mRNA levels were averaged from 3 replicates of total mRNA-seq.

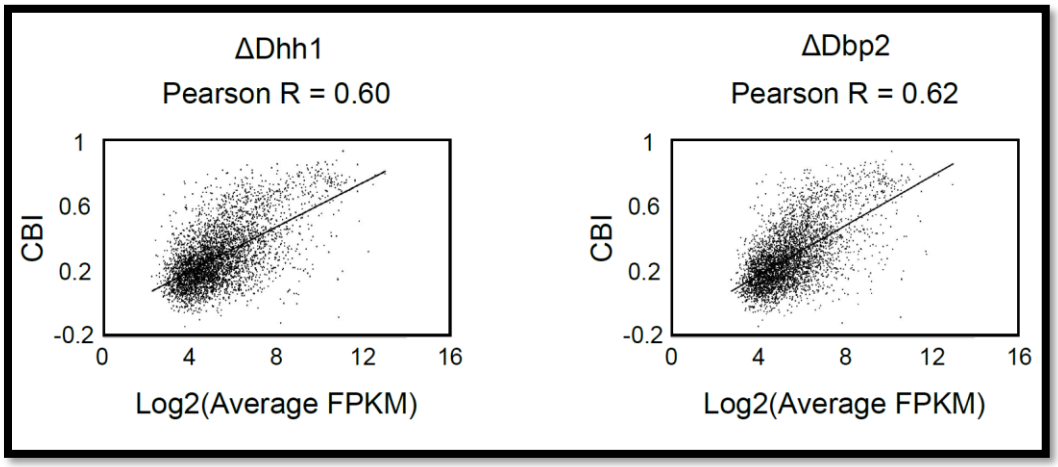

Figure 3-8 Dhh1 and Dbp2 Do Not Affect the Correlation between CBI and mRNA level.

Pearson correlations are shown as dot plots for ΔDhh1 and ΔDbp2. mRNA levels were averaged from 3 replicates of total mRNA-seq. These two strains serve as representative control strains with no significant decrease of correlation coefficients.

3.3.4 Divergent Effect of Candidate Proteins on Genes with Different Codon Usage Bias

After knocking out the target genes in candidate strains, we could observe large amounts of genes being up- or down-regulated. Using unchanged genes as controls, we found that in most candidate strains the CBIs of up-regulated genes were significantly lower than those of the controls, while the CBIs of down-regulated genes were significantly higher (Figure 3-9). It should be noted that many factors we examined are known to function as activators or

repressors and the decrease of correlation coefficients observed in Figure 3-7 might be caused by the coincidence that their target genes happen to have lower or higher CBIs. However, the bi-directional regulation we observed in Figure 3-9A suggested that this is not likely the case and the regulations could be dependent on the codon usage of target genes. Consistent with previous correlation analyses, differentially expressed genes in ΔDhh1 strain did not show significant CBI change (Figure 3-9B, upper). In addition, although the knockout of histone deacetylase-2 leads to dramatically decreased growth rate and large amounts of differentially regulated genes in Δhda-2 mutant strain, the differentially regulated genes did not show specific CBI preferences (Figure 3-9B, lower). Importantly, on single codon level we also observed decreased correlation coefficients between codon frequencies and mRNA levels for almost all the codons in a representative candidate strain Δset-2 (Figure 3-10). Interestingly, the rankings and overall pattern of the correlation coefficients remained almost the same in Δset-2 mutant, indicating the potential involvement of other regulatory factors and the robustness of the correlation between codon usage and mRNA level.

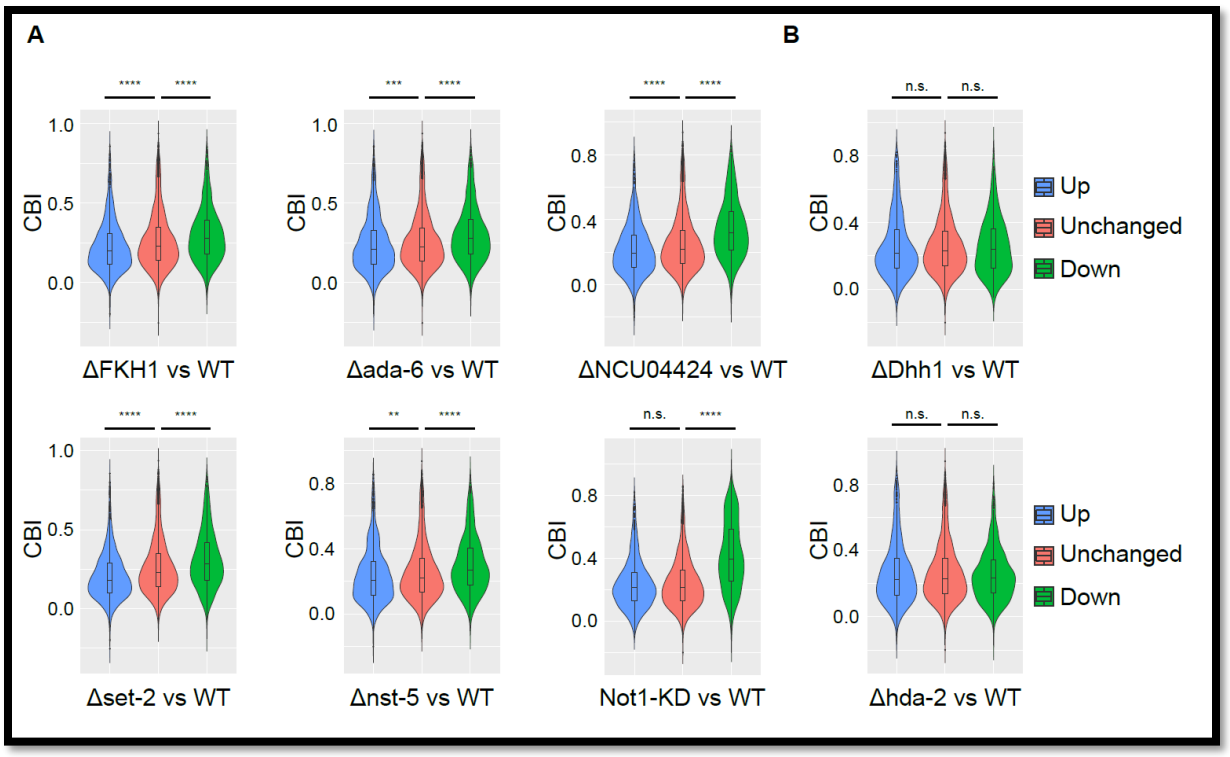

Figure 3-9 Bi-directional Regulation of Candidate Factors Associated with Codon Usage.

(**A**) Violin plots of CBI distribution for differentially regulated genes and unchanged genes in 6 representative candidate strains. **P <0.01, ***P<0.001, ****P<0.0001. (**B**) Violin plots of CBI distribution for differentially regulated genes and unchanged genes in two representative control strains. No significant CBI differences were found between differentially regulated genes and unchanged genes.

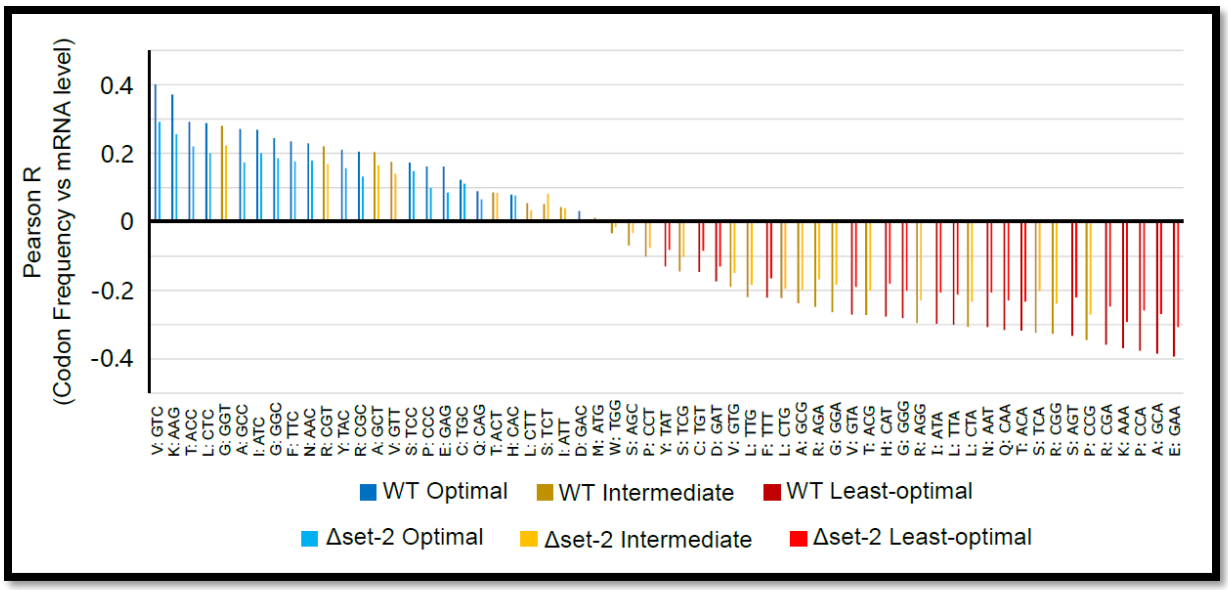

Figure 3-10 Decreased Correlation between Individual Codon Frequency and mRNA Level in Δset-2 Mutant Strain.

Pearson correlation coefficients for all 61 codons between codon frequencies and mRNA levels. Data obtained from total RNA-seq of FGSC4200 wildtype strain and Δset-2 mutant strain were plotted side by side. mRNA levels were averaged over 3 replicates. Definition of codon optimality is the same as mentioned in Figure 3-3.

Codon optimization has been widely used to increase protein expression level in many organisms. And we demonstrated previously in Neurospora crassa that many codon optimized reporters, including both endogenous and exogenous ones, showed elevated mRNA levels similar to the change of protein levels [75]. Given the divergent effects of our candidate genes on genome level, we sought to investigate whether the codon optimization effects in wildtype strain could be dampened in knockout strains. Indeed, when we expressed

codon optimized and wildtype luciferase genes in wildtype strain, we could observe dramatic difference on protein level (Figure 3-11A, black bar) and mRNA level (Figure 3-11B, black bar). However, when we expressed these two genes in set-2 mutant strain, the fold differences were significantly decreased on both protein level (Figure 3-11A) and mRNA level (Figure 3-11B).

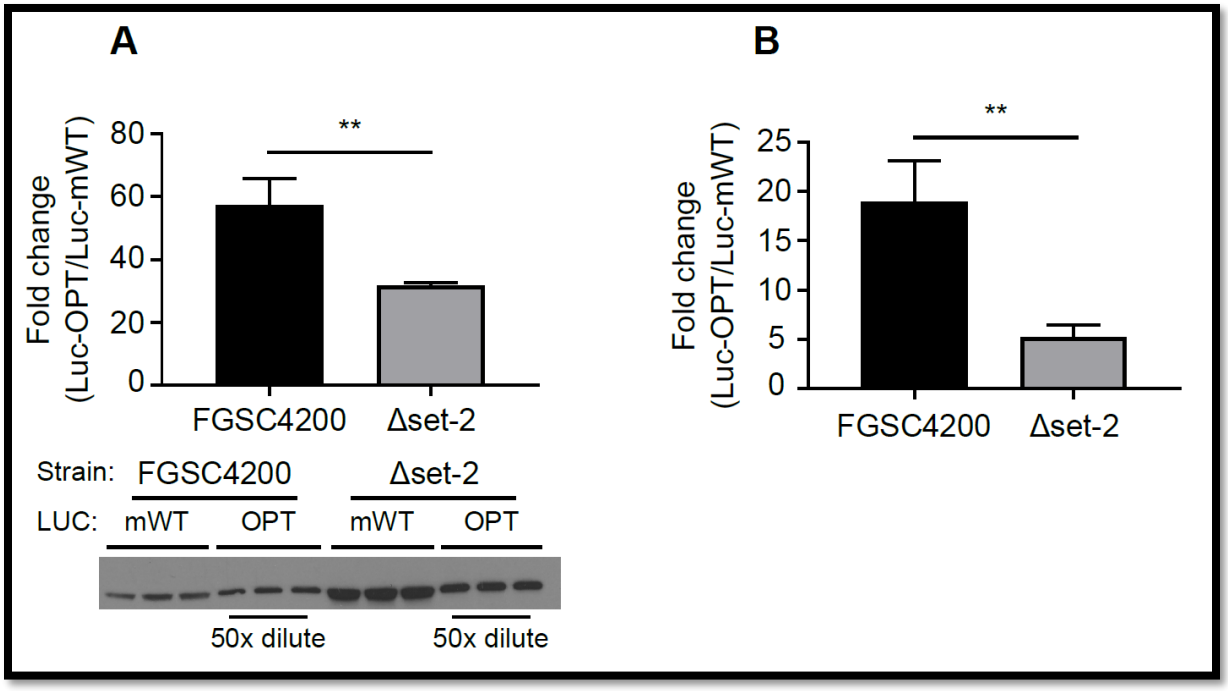

Figure 3-11 The Effect of Codon Optimization is Dampened in Δset-2 Mutant Strain.

(**A**) Western blot measurement of protein levels of codon-optimized (OPT) and wildtype luciferase (mWT) expressed in either FGSC4200 wildtype strain or Δset-2 mutant strain (lower). OPT protein was diluted by 50 times for quantification purpose. The bar plot shows the quantification of fold differences between codon-optimized and wildtype proteins in the two strains (upper). **P<0.01. Statistics were calculated from 3 replicates. (**B**) RT-qPCR quantification of mRNA fold differences between codon-optimized and wildtype luciferase

expressed in either FGSC4200 wildtype strain or Δset-2 mutant strain. **P<0.01. Statistics were calculated from 3 replicates.

3.3.5 Hierarchical Clustering of Candidate Strains

To further investigate the interrelationship between the candidate factors, we performed a hierarchical clustering for the 18 strains based on the correlation of their differentially expressed genes (Figure 3-12). The candidate strains could be clustered into 2 major groups, with Δset-2, ΔFKH1 and Δscp160 as outliers. It is interesting to note that both groups contain a mixture of transcription factors and epigenetic regulators, which indicates that their codon usage associated effects might be independent of their known functions. Furthermore, although Δset-2 and ΔFKH1 cannot be clustered with most other strains, they have the most severely affected correlations among all candidate strains. This indicates that multiple pathways and mechanisms might be involved in regulating mRNA level in a codon usage dependent manner.

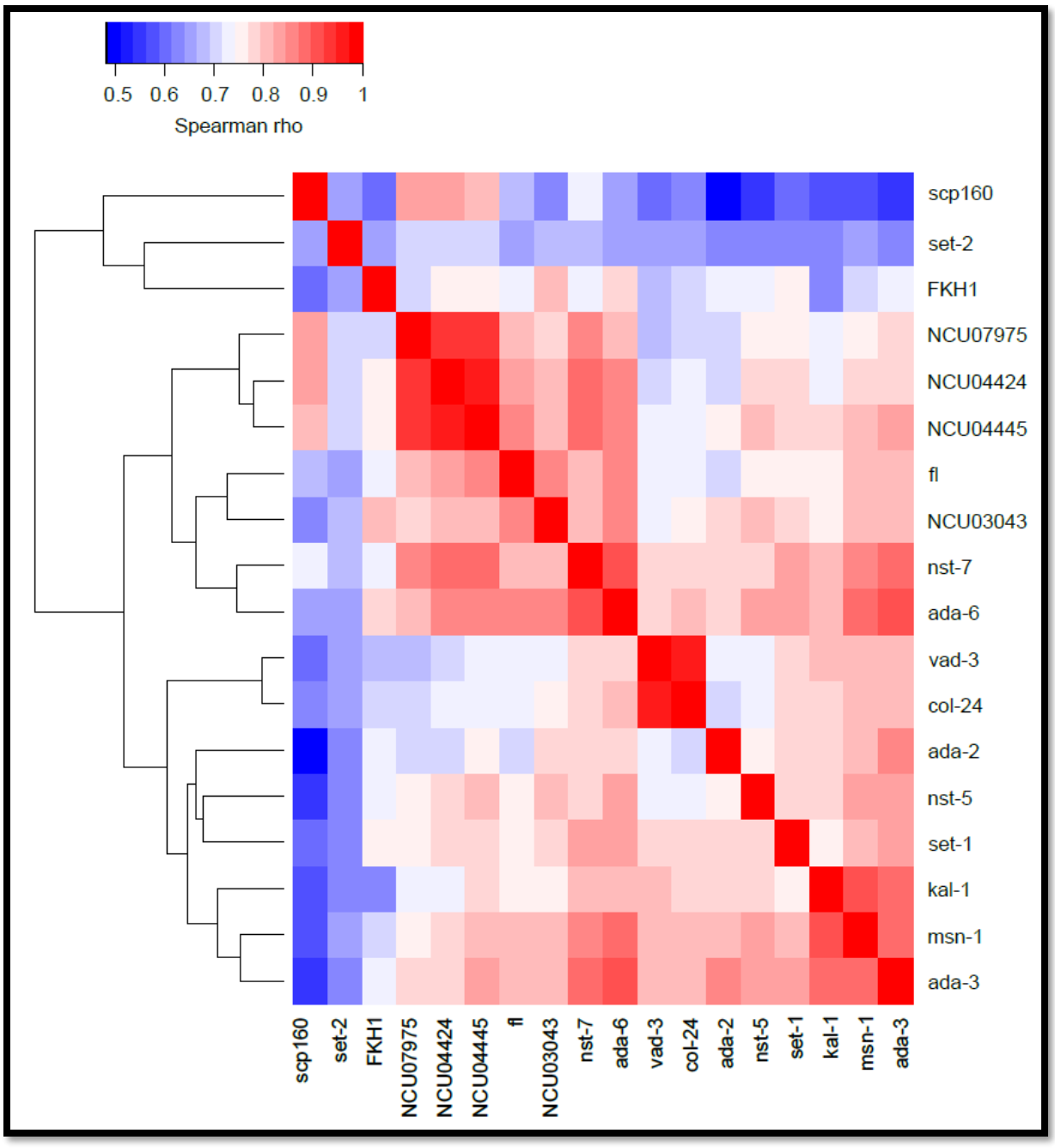

Figure 3-12 Hierarchical Clustering of Candidate Strains.

The heatmap shows the spearman correlation calculated using the fold change of differentially regulated genes in any pair of candidate strains. Then the strains were

hierarchically clustered using the spearman rho values as the distances between each pair of candidate strains.

3.3.6 Candidate Factors Regulate Genome Wide mRNA Transcription

Now that we know codon usage can affect genome-wide mRNA transcription and multiple factors are potentially involved in establishing this correlation, it is important to show that the effects of these factors are linked to mRNA transcription. To address this question, we performed nuclear RNA sequencings for 2 representative knockout strains Δset-2 and ΔFKH1. As expected, the correlation coefficients between nuclear RNA level and CBI were significantly decreased in both strains comparing to wildtype strain (Figure 3-13A). The overall correlation levels are lower than what we observed from total RNA-seq, potentially due to the influences of nucleic extraction process. However, the correlation differences between wildtype strain and mutant strains are very similar to what we obtained from total RNA-seq, which indicates that the effects of set-2 and FKH1 proteins are largely within the nucleus. On the other hand, although we also observed bi-directional regulation in Δset-2 strain, the CBI difference in down-regulated group was much more dramatic and we did not even observe any significant CBI difference in the up-regulated group of ΔFKH1 strain (Figure 3-13B). This observation could be due to the existence of additional regulation by the two factors in the cytosol, or simply different experimental conditions of nuclear RNA-seq. Importantly, we also performed total RNA sequencing using the same tissue samples to examine the overlap of differentially expressed genes between total RNAs and nuclear RNAs. As shown in Figure 3-13C, a large proportion of the genes either up- or down-

regulated in total RNA-seq also showed same up- or down-regulation in nuclear RNA-Seq, respectively. This finding strongly suggested that the regulation of candidate factors on mRNA level is largely on transcription level.

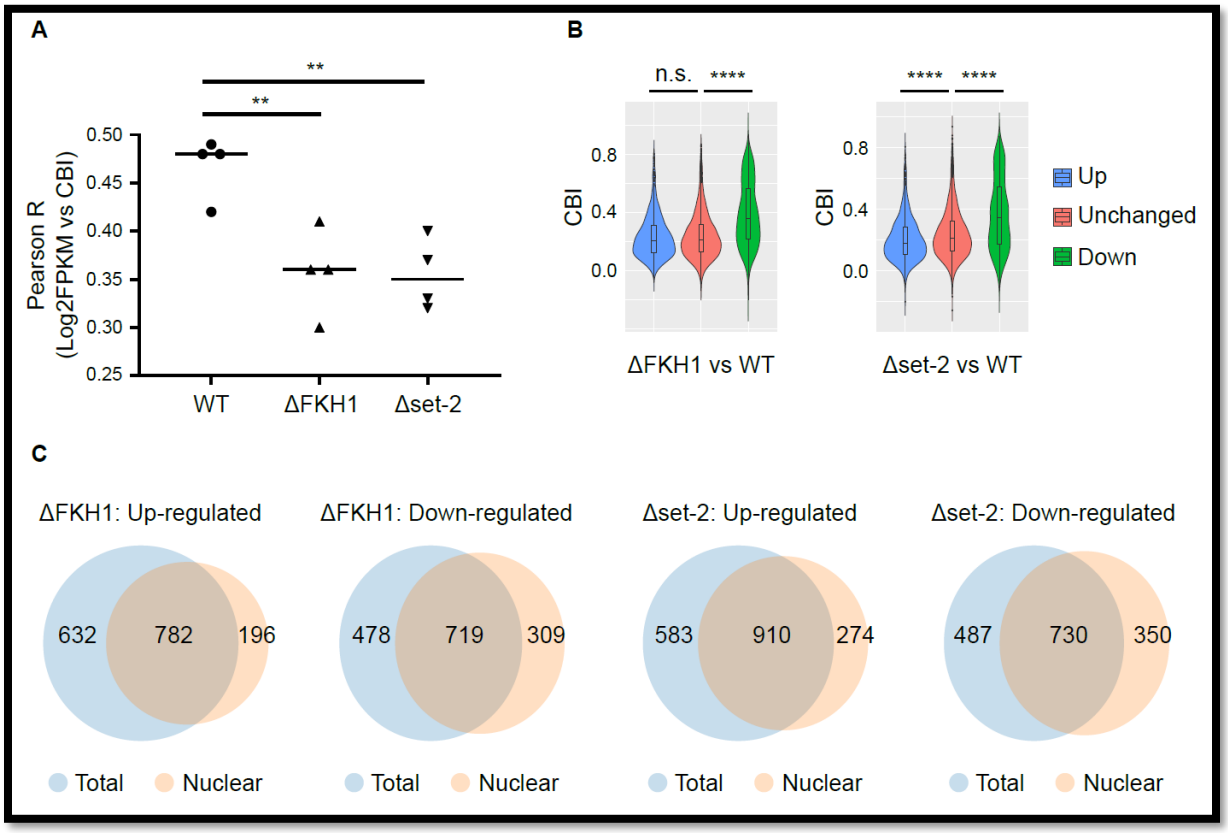

Figure 3-13 The effects of Δset-2 and ΔFKH1 are Largely Translation Independent.

(**A**) Pearson correlation coefficients between CBI and nuclear RNA level in FGSC4200 wildtype strain and two KO strains. 4 replicates for each strain were plotted and the medians are indicated by the horizontal lines. **P<0.01. (**B**) Violin plots of CBI distribution for differentially regulated genes and unchanged genes in Δset-2 and ΔFKH1 strains. Differentially regulated genes were obtained from 2 replicates of nuclear RNA-seq

experiments. (**C**) Overlaps of up- or down-regulated genes between total RNA-seq and nuclear RNA-seq were plotted for Δset-2 and ΔFKH1 mutant strains.

To further validate that the nuclear mRNA level change after knocking out set-2 is driven by altered mRNA transcription, we performed ChIP-qPCR experiment using 2 different RNAP II antibodies targeting either non-phosphorylated C-terminal domain (CTD) or Ser-2 phosphorylated CTD (S2p). Multiple genes either up- or down-regulated in both total and nuclear RNA-seq were selected as reporters for examination of RNAP II enrichment on them. Indeed, we saw elevated (Figure 3-14A) or decreased (Figure 3-14B) RNAP II enrichments on up-regulated or down-regulated genes, respectively. Although the fold enrichments on some genes did not show statistical significance due to variations between replicates, the overall trend was consistent for most of the genes examined. Collectively, these data indicated that the candidate factors found in our screen process are potentially able to regulate mRNA transcription in a codon usage dependent manner.

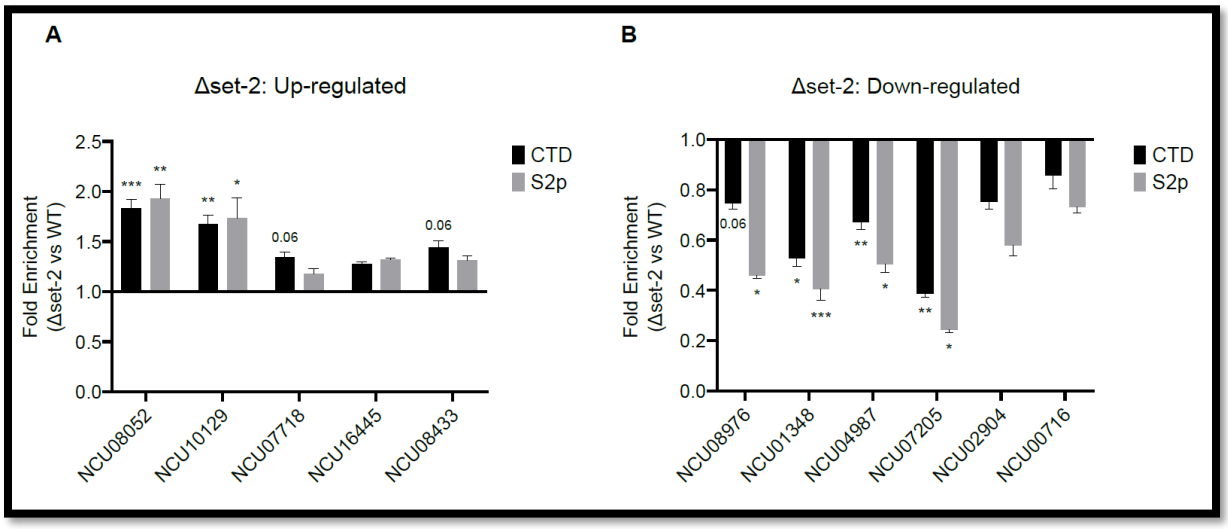

Figure 3-14 Differentially Regulated Genes in Δset-2 Mutant Show Consistent Change of RNAP II Enrichment Level.

(**A**) ChIP-qPCR experiment to measure the fold enrichment of RNAP II on 5 representative target genes between Δset-2 mutant and FGSC4200 wildtype strain. The 5 target genes all showed up-regulation in both total and nuclear RNA-seq experiment as shown in the figure title. qPCR primers targeted either 5' UTR or the 3' end of CDS. Fold enrichments were calculated as described in 3.2.10. The same antibodies as in Figure 3-2 were used and represented as CTD or S2p. (**B**) Similar experiment as A except that the target genes all showed down-regulation in total and nuclear RNA-seq.

3.4 Discussion

Since codons are recognized during translation elongation process, it was generally thought that codon usage does not affect mRNA transcription in the nucleus. Although recently multiple studies showed that steady-state mRNA level could be affected by codon usage [71-74], the underlying mechanisms remain controversial. We know that steady-state mRNA level is determined by mRNA transcription and degradation. But most of the current mechanistic models are focused on mRNA stability due to the coupling between translation and mRNA degradation [77, 78]. However, emerging evidences suggested that codon usage can also regulate mRNA level independent of translation [66, 75, 76]. While these studies showed elevated transcription level of reporter genes after codon optimization, we still do not know whether codon usage is a general regulatory factor on genome transcription level. To

address this issue, we used both RNA-seq and 2P-seq to measure nuclear mRNA levels and analyzed their correlation with codon usage bias. Although nuclear mRNA level is not a direct measurement of transcription level, it is largely determined by transcription and more importantly, it is independent of translation except for some rare, non-canonical nuclear translation events [106-108]. While RNA polymerase II ChIP-seq data can reflect transcription level more directly, it also suffers from limited quantification capacity. Combining these 3 different techniques and previous reporter assays [75, 76], we believe that the consistent correlations we observed indicated an important role for codon usage in regulating genome-wide transcription (Figure 3-1 and 3-2).

Next, we asked how codon usage regulates genome-wide transcription. Since we had little knowledge of the role of codon usage independent of translation, we decided to perform a blinded screen to seek for knockout strains with compromised correlations between mRNA level and codon usage bias (Figure 3-5 and 3-6). Our previous studies indicated that genes with different codon optimality could be associated with different level of histone modifications and chromatin structures [75, 76]. So, we specifically picked factors that interact with DNA sequence or affect chromatin structure to perform the screen. The lowest correlation coefficient we obtained from the screen is 0.42. Although this is not an extremely low correlation coefficient, it is reasonable since too dramatic alteration of the correlation may lead to lethality. And more importantly, the screen highlighted 18 factors from distinct protein families and pathways, which regulate different groups of genes (data not shown). This indicates that the genome-wide correlation between mRNA level and codon usage bias is controlled by multiple factors. For many strains we could observe large amounts of

differentially regulated genes, but the correlation coefficient between mRNA level and CBI remained unchanged. Specifically, ΔDhh1 and ΔDbp2 strains did not show expected correlation change, given their reported roles on codon dependent mRNA stability in yeast. This suggested that different organisms probably rely on different mechanisms to exert the effect of codon usage on mRNA level.

Since the screen process was only based on compromised correlation coefficients between mRNA and CBI, it is possible that some of the candidate factors happen to regulate a group of genes with distinct codon usage bias and the underlying mechanism is not related to codon usage. Further studies are necessary for all the candidates to validate their association with codon usage and the underlying mechanisms. However, the purpose of the screen in this study was to find out candidates for following investigations. We only performed some preliminary validations for several candidate strains using reporter genes (Figure 3-11). In Δset-2 mutant strain, we observed compromised codon effects on mRNA level marked by decreased fold change between codon optimized and wildtype transcripts. But it's worth noting that not all the reporter genes we examined showed expected changes in candidate strains (data not shown). In addition to the possibility we mentioned that the candidates are false positives, this could also be attributed to the fact that different candidate factors regulate different groups of genes. Moreover, a recent study indicated that gene promoters may also play an important role in determining codon effects on mRNA level [105]. Finally, we used nuclear RNA-seq and ChIP-qPCR to show that the regulation of set-2 and FKH1 is largely on transcriptional level (Figure 3-13 and 3-14), which links the genome-wide correlation and the mediating effects of candidate factors.

During further mechanistic studies, an important question will be how codons are recognized independent of translation. In fact, we have discussed on a similar question in our recent study which showed codon usage could affect pre-mature transcription termination [80]. We presented evidences of co-evolution between codon usage bias and transcription termination machinery. Similarly, we hypothesized that the correlation between mRNA transcription and CBI is based on the co-evolution of codon usage bias and the recognition sequences of the candidate factors identified in this study.

CHAPTER FOUR

CONCLUSIONS AND FUTURE DIRECTIONS

In CHAPTER TWO, we demonstrated in an animal system that codon usage directly regulates translation elongation speed. Optimal codons are translated faster, and rare codons are translated slower. The slow translation speed mediated by rare codons may have further impacts on ribosome movement and translation kinetics, including ribosome pausing, pre-mature translation termination and even frameshifting. However, it should also be noted that faster translation speed is not always better. First, higher translation speed does not necessarily lead to higher translation efficiency and protein expression since translation efficiency is mainly determined by translation initiation. Second, we showed that proteins translated by optimal codons are structurally and functionally different from same proteins translated by rare codons. Although optimized luciferase protein produced way more full-length protein product, the activity of optimized protein was much lower than de-optimized one. Combined with the reported effects of translation elongation speed on co-translational folding in other studies, we presented an important role of rare codons in fine tuning translation elongation speed to achieve optimal co-translational folding and proper protein structure and function.

Although codon usage mediated translation elongation speed is hypothesized to be the "code" within codons that regulates co-translational folding, it is not clear how exactly this

"code" is linked to protein structural motifs or domains. We do not know the exact structural alterations caused by codon manipulation since we only obtained evidences from limited trypsin digestion, which was merely an implication of structural difference. For proteins with dramatic functional alterations after codon manipulation, it will be interesting to directly compare the protein crystal structures translated endogenously to those translated from codon manipulated mRNAs. On the other hand, it should be noted that not all the rare codons in wildtype genes are important for their structures and functions. These codon usage differences may also result from mutational bias and the lack of selection pressure as mentioned previously. So, if we want to increase protein expression level while maximally preserving original protein structure and function, it is important to know which codon positions are truly necessary to be kept non-optimized.

In CHAPTER THREE, we presented evidences indicating that codon usage can affect genome-wide transcription process independent of translation. By systemic genetic screening, we identified multiple candidate proteins mediating the correlation between codon usage bias and mRNA level. Importantly, the effects of these candidate factors are also largely translation independent. Our study provides additional possibilities for how codon usage may influence mRNA level other than its known effects on translation dependent mRNA degradation. Together with the effects of codon usage on mRNA stability, our study provided a more reasonable explanation on how codon usage affects protein expression level.

Since we identified many candidate factors potentially mediating the correlation between codon usage bias and mRNA transcription, future studies will first be focusing on validating the codon dependent effects of these factors and investigating the mechanisms

underlying their regulations. It is critically important to examine how these factors may interact with DNA sequences and whether the recognition motifs co-evolve with codon usage patterns. In addition, although nuclear RNA-seq in this study indicated translation independent effects of codon usage, other experiments like GRO-seq are necessary to validate the involvement of transcription process. Genome-wide mRNA stability profiling is also beneficial to examine the proportions of codon usage effects on different pathways. On the other hand, since translation independent effects of codon usage were suggested in this study, it is worth investigating whether codon usage can also affect other nuclear processes including mRNA splicing, capping, polyadenylation, etc. Finally, the conservation of the proposed codon usage effects should be tested in higher eukaryotes.

APPENDIX A
Sequences of Luciferase Variants

OPT:

ATGGAGCAAAAGCTCATTTCTGAAGAGGACTTGAATGAAATGGAGCAAAAGCTC

ATTTCTGAAGAGGACTTGAATGAAATGGAGCAAAAGCTCATTTCTGAAGAGGAC

TTGAATGAAATGGAGCAAAAGCTCATTTCTGAAGAGGACTTGAATGAAATGGAG

CAAAAGCTCATTTCTGAAGAGGACTTGAATGAAATGGAGAGCTTGGGCGACATG

GAAGACGCCAAAAACATAAAGAAAGGCCCCGCCCCCTTCTACCCCCTGGAGGAT

GGCACCGCCGGCGAGCAGCTGCACAAGGCCATGAAGCGCTACGCCCTGGTGCCC

GGCACCATCGCCTTCACCGATGCCCACATCGAGGTGGATATCACCTACGCCGAG

TACTTCGAGATGAGCGTGCGCCTGGCCGAGGCCATGAAGCGCTACGGCCTGAAC

ACCAACCACCGCATCGTGGTGTGCAGCGAGAACAGCCTGCAGTTCTTCATGCCC

GTGCTGGGCGCCCTGTTCATCGGCGTGGCCGTGGCCCCCGCCAACGATATCTACA

ACGAGCGCGAGCTGCTGAACAGCATGGGCATCAGCCAGCCCACCGTGGTGTTCG

TGAGCAAGAAGGGCCTGCAGAAGATCCTGAACGTGCAGAAGAAGCTGCCCATC

ATCCAGAAGATCATCATCATGGATAGCAAGACCGATTACCAGGGCTTCCAGAGC

ATGTACACCTTCGTGACCAGCCACCTGCCCCCCGGCTTCAACGAGTACGATTTCG

TGCCCGAGAGCTTCGATCGCGATAAGACCATCGCCCTGATCATGAACAGCAGCG

GCAGCACCGGCCTGCCCAAGGGCGTGGCCCTGCCCCACCGCACCGCCTGCGTGC

GCTTCAGCCACGCCCGCGATCCCATCTTCGGCAACCAGATCATCCCCGATACCGC

CATCCTGAGCGTGGTGCCCTTCCACCACGGCTTCGGCATGTTCACCACCCTGGGC

TACCTGATCTGCGGCTTCCGCGTGGTGCTGATGTACCGCTTCGAGGAGGAGCTGT

TCCTGCGCAGCCTGCAGGATTACAAGATCCAGAGCGCCCTGCTGGTGCCCACCCT

GTTCAGCTTCTTCGCCAAGAGCACCCTGATCGATAAGTACGATCTGAGCAACCTG

CACGAGATCGCCAGCGGCGGCGCCCCCCTGAGCAAGGAGGTGGGCGAGGCCGT

GGCCAAGCGCTTCCACCTGCCCGGCATCCGCCAGGGCTACGGCCTGACCGAGAC

CACCAGCGCCATCCTGATCACCCCCGAGGGCGATGATAAGCCCGGCGCCGTGGG

CAAGGTGGTGCCCTTCTTCGAGGCCAAGGTGGTGGATCTGGATACCGGCAAGAC

CCTGGGCGTGAACCAGCGCGGCGAGCTGTGCGTGCGCGGCCCCATGATCATGAG

CGGCTACGTGAACAACCCCGAGGCCACCAACGCCCTGATCGATAAGGATGGCTG

GCTGCACAGCGGCGATATCGCCTACTGGGATGAGGATGAGCACTTCTTCATCGTG

GATCGCCTGAAGAGCCTGATCAAGTACAAGGGCTACCAGGTGGCCCCCGCCGAG

CTGGAGAGCATCCTGCTGCAGCACCCCAACATCTTCGATGCCGGCGTGGCCGGC

CTGCCCGATGATGATGCCGGCGAGCTGCCCGCCGCCGTGGTGGTGCTGGAGCAC

GGCAAGACCATGACCGAGAAGGAGATCGTGGATTACGTGGCCAGCCAGGTGAC

CACCGCCAAGAAGCTGCGCGGCGGCGTGGTGTTCGTGGATGAGGTGCCCAAGGG

CCTGACCGGCAAGCTGGATGCCCGCAAGATCCGCGAGATCCTGATCAAGGCCAA

GAAGGGCGGCAAGATCGCCGTGTAA

dOPT:

ATGGAAGACGCCAAAAACATAAAGAAAGGCCCTGCACCTTTTTATCCTTTAGAA

GACGGGACTGCAGGGGAACAATTACATAAAGCAATGAAAAGATATGCATTAGTA

CCTGGGACTATAGCATTTACTGACGCACATATAGAAGTAGACATAACTTATGCA

GAATATTTTGAAATGTCTGTAAGATTAGCAGAAGCAATGAAAAGATATGGGTTA

AATACTAATCATAGAATAGTAGTATGTTCTGAAAATTCTTTACAATTTTTTATGCC

TGTATTAGGGGCATTATTTATAGGGGTAGCAGTAGCACCTGCAAATGACATATAT

AATGAAAGAGAATTATTAAATTCTATGGGGATATCTCAACCTACTGTAGTATTTG

TATCTAAAAAAGGGTTACAAAAAATATTAAATGTACAAAAAAAATTACCTATAA

TACAAAAAATAATAATAATGGACTCTAAAACTGACTATCAAGGGTTTCAATCTAT

GTATACTTTTGTAACTTCTCATTTACCTCCTGGGTTTAATGAATATGACTTTGTAC

CTGAATCTTTTGACAGAGACAAAACTATAGCATTAATAATGAATTCTTCTGGGTC

TACTGGGTTACCTAAAGGGGTAGCATTACCTCATAGAACTGCATGTGTAAGATTT

TCTCATGCAAGAGACCCTATATTTGGGAATCAAATAATACCTGACACTGCAATAT

TATCTGTAGTACCTTTTCATCATGGGTTTGGGATGTTTACTACTTTAGGGTATTTA

ATATGTGGGTTTAGAGTAGTATTAATGTATAGATTTGAAGAAGAATTATTTTTAA

GATCTTTACAAGACTATAAAATACAATCTGCATTATTAGTACCTACTTTATTTTCT

TTTTTTGCAAAATCTACTTTAATAGACAAATATGACTTATCTAATTTACATGAAAT

AGCATCTGGGGGGGCACCTTTATCTAAAGAAGTAGGGGAAGCAGTAGCAAAAA

GATTTCATTTACCTGGGATAAGACAAGGGTATGGGTTAACTGAAACTACTTCTGC

AATATTAATAACTCCTGAAGGGGACGACAAACCTGGGGCAGTAGGGAAAGTAGT

ACCTTTTTTTGAAGCAAAAGTAGTAGACTTAGACACTGGGAAAACTTTAGGGGT

AAATCAAAGAGGGGAATTATGTGTAAGAGGGCCTATGATAATGTCTGGGTATGT

AAATAATCCTGAAGCAACTAATGCATTAATAGACAAAGACGGGTGGTTACATTC

TGGGGACATAGCATATTGGGACGAAGACGAACATTTTTTTATAGTAGACAGATT

AAAATCTTTAATAAAATATAAAGGGTATCAAGTAGCACCTGCAGAATTAGAATC

TATATTATTACAACATCCTAATATATTTGACGCAGGGGTAGCAGGGTTACCTGAC

GACGACGCAGGGGAATTACCTGCAGCAGTAGTAGTATTAGAACATGGGAAAACT

ATGACTGAAAAAGAAATAGTAGACTATGTAGCATCTCAAGTAACTACTGCAAAA

AAATTAAGAGGGGGGGTAGTATTTGTAGACGAAGTACCTAAAGGGTTAACTGGG

AAATTAGACGCAAGAAAAATAAGAGAAATATTAATAAAAGCAAAAAAAGGGGG

GAAAATAGCAGTATAA

APPENDIX B
Major Strains Used in CHAPTER THREE

Name	Genotype	Source
wildtype	wildtype mat a	FGSC 4200
ΔNCU03897	ΔNCU03897::HygR mat a	FGSC 21525
Δada-6	ΔNCU04866::HygR mat a	FGSC 11022
Δnst-5	ΔNCU00203::HygR mat a	FGSC 14806
ΔFKH1	ΔNCU00019::HygR mat a	FGSC 11437
Δfl	ΔNCU08726::HygR mat a	FGSC 11044
Δset-2	ΔNCU00269::HygR mat A	FGSC 15505
Δada-3	ΔNCU02896::HygR mat a	FGSC 11070
Δnst-7	ΔNCU07624::HygR mat a	FGSC 16002
Δcol-24	ΔNCU05383::HygR mat a	FGSC 11019
ΔNCU07975	ΔNCU07975::HygR mat A	FGSC 11337
Δvad-3	ΔNCU06407::HygR mat a	FGSC 11017
Δkal-1	ΔNCU03593::HygR mat a	FGSC 11129
Δmsn-1	ΔNCU02671::HygR mat a	FGSC 11345
Δset-1	ΔNCU01206::HygR mat A	FGSC 15827
ΔNCU04424	ΔNCU04424::HygR mat a	FGSC 11752
Δada-2	ΔNCU02017::HygR mat a	FGSC 11108
ΔNCU03043	ΔNCU03043::HygR mat a	FGSC 11224
ΔNCU04445	ΔNCU04445::HygR mat a	FGSC 11754
ΔDhh1	ΔNCU06149::HygR mat a	FGSC 13191
ΔDbp2	ΔNCU07839::HygR mat a	FGSC 15506
Δhda-2	ΔNCU02795::HygR mat A	FGSC 11158

BIBLIOGRAPHY

1. Ikemura, T., *Codon usage and tRNA content in unicellular and multicellular organisms.* Molecular biology and evolution, 1985. **2**(1): p. 13-34.

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

3. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nature Reviews Genetics, 2011. **12**(1): p. 32.

4. Comeron, J.M., *Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence.* Genetics, 2004. **167**(3): p. 1293-1304.

5. Duret, L., *Evolution of synonymous codon usage in metazoans.* Current opinion in genetics & development, 2002. **12**(6): p. 640-649.

6. Bulmer, M., *The selection-mutation-drift theory of synonymous codon usage.* Genetics, 1991. **129**(3): p. 897-907.

7. Akashi, H., R.M. Kliman, and A. Eyre-Walker, *Mutation pressure, natural selection, and the evolution of base composition in Drosophila,* in *Mutation and Evolution*. 1998, Springer. p. 49-60.

8. Sharp, P.M., et al., *Codon usage: mutational bias, translational selection, or both?* 1993, Portland Press Limited.

9. Bernardi, G., *Isochores and the evolutionary genomics of vertebrates.* Gene, 2000. **241**(1): p. 3-17.

10. Akashi, H., *Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy.* Genetics, 1994. **136**(3): p. 927-935.

11. Yang, Z. and R. Nielsen, *Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage.* Molecular biology and evolution, 2008. **25**(3): p. 568-579.

12. Bennetzen, J.L. and B.D. Hall, *Codon selection in yeast.* Journal of Biological Chemistry, 1982. **257**(6): p. 3026-3031.

13. Sharp, P. and W.-H. Li, *The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias.* Molecular biology and evolution, 1987. **4**(3): p. 222-230.

14. Duret, L. and D. Mouchiroud, *Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis.* Proceedings of the National Academy of Sciences, 1999. **96**(8): p. 4482-4487.

15. Castillo-Davis, C.I. and D.L. Hartl, *Genome evolution and developmental constraint in Caenorhabditis elegans.* Molecular biology and evolution, 2002. **19**(5): p. 728-735.

16. Gustafsson, C., S. Govindarajan, and J. Minshull, *Codon bias and heterologous protein expression.* Trends in biotechnology, 2004. **22**(7): p. 346-353.

17. Kurland, C., *Translational accuracy and the fitness of bacteria.* Annual review of genetics, 1992. **26**(1): p. 29-50.

18. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-352.

19. Sharp, P.M. and W.-H. Li, *The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications.* Nucleic acids research, 1987. **15**(3): p. 1281-1295.

20. Dos Reis, M., L. Wernisch, and R. Savva, *Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K‐12 genome.* Nucleic acids research, 2003. **31**(23): p. 6976-6985.

21. Sabi, R., R. Volvovitch Daniel, and T. Tuller, *stAIcalc: tRNA adaptation index calculator based on species-specific weights.* Bioinformatics, 2016. **33**(4): p. 589-591.

22. Gun, L., et al., *Comprehensive Analysis and Comparison on the Codon Usage Pattern of Whole Mycobacterium tuberculosis Coding Genome from Different Area.* BioMed research international, 2018. **2018**.

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

24. Bulmer, M., *Coevolution of codon usage and transfer RNA abundance.* Nature, 1987. **325**(6106): p. 728.

25. Moriyama, E.N. and J.R. Powell, *Codon usage bias and tRNA abundance in Drosophila.* Journal of molecular evolution, 1997. **45**(5): p. 514-523.

26. Chu, D., et al., *Translation elongation can control translation initiation on eukaryotic mRNAs.* The EMBO journal, 2014. **33**(1): p. 21-34.

27. Yang, Q., et al., *eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons.* Nucleic acids research, 2019. **47**(17): p. 9243-9258.

28. Sørensen, M.A., C. Kurland, and S. Pedersen, *Codon usage determines translation rate in Escherichia coli.* Journal of molecular biology, 1989. **207**(2): p. 365-377.

29. Bonekamp, F., et al., *Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli.* Journal of bacteriology, 1989. **171**(11): p. 5812-5816.

30. Chevance, F.F., S. Le Guyon, and K.T. Hughes, *The effects of codon context on in vivo translation speed.* PLoS genetics, 2014. **10**(6): p. e1004392.

31. Charneski, C.A. and L.D. Hurst, *Positively charged residues are the major determinants of ribosomal velocity.* PLoS biology, 2013. **11**(3): p. e1001508.

32. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* science, 2009. **324**(5924): p. 218-223.

33. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.

34. Li, G.-W., E. Oh, and J.S. Weissman, *The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria.* Nature, 2012. **484**(7395): p. 538.

35. Qian, W., et al., *Balanced codon usage optimizes eukaryotic translational efficiency.* PLoS genetics, 2012. **8**(3): p. e1002603.

36. Artieri, C.G. and H.B. Fraser, *Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation.* Genome research, 2014. **24**(12): p. 2011-2021.

37. Lareau, L.F., et al., *Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments.* Elife, 2014. **3**: p. e01257.

38. Gardin, J., et al., *Measurement of average decoding rates of the 61 sense codons in vivo.* Elife, 2014. **3**: p. e03735.

39. Nakahigashi, K., et al., *Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo.* BMC genomics, 2014. **15**(1): p. 1115.

40. Weinberg, D.E., et al., *Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation.* Cell reports, 2016. **14**(7): p. 1787-1799.

41. Dunn, J.G., et al., *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.* Elife, 2013. **2**: p. e01179.

42. Yu, C.-H., et al., *Codon usage influences the local rate of translation elongation to regulate co-translational protein folding.* Molecular cell, 2015. **59**(5): p. 744-754.

43. Kimchi-Sarfaty, C., et al., *A" silent" polymorphism in the MDR1 gene changes substrate specificity.* Science, 2007. **315**(5811): p. 525-528.

44. Liu, M.-L., T. Zang, and C.-L. Zhang, *Direct lineage reprogramming reveals disease-specific phenotypes of motor neurons from human ALS patients.* Cell reports, 2016. **14**(1): p. 115-128.

45. Lampson, B.L., et al., *Rare codons regulate KRas oncogenesis.* Current Biology, 2013. **23**(1): p. 70-75.

46. Pershing, N.L., et al., *Rare codons capacitate Kras-driven de novo tumorigenesis.* The Journal of clinical investigation, 2015. **125**(1): p. 222-233.

47. Smith, N.G. and A. Eyre-Walker, *Why are translationally sub-optimal synonymous codons used in Escherichia coli?* Journal of molecular evolution, 2001. **53**(3): p. 225-236.

48. Zhou, M., et al., *Non-optimal codon usage affects expression, structure and function of clock protein FRQ.* Nature, 2013. **495**(7439): p. 111.

49. Fu, J., et al., *Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD.* Genes & development, 2016. **30**(15): p. 1761-1775.

50. Komar, A.A., T. Lesnik, and C. Reiss, *Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation.* FEBS letters, 1999. **462**(3): p. 387-391.

51. Spencer, P.S., et al., *Silent substitutions predictably alter translation elongation rates and protein folding efficiencies.* Journal of molecular biology, 2012. **422**(3): p. 328-335.

52. Zhang, G., M. Hubalewska, and Z. Ignatova, *Transient ribosomal attenuation coordinates protein synthesis and co-translational folding.* Nature structural & molecular biology, 2009. **16**(3): p. 274.

53. Siller, E., et al., *Slowing bacterial translation speed enhances eukaryotic protein folding efficiency.* Journal of molecular biology, 2010. **396**(5): p. 1310-1318.

54. Sander, I.M., J.L. Chaney, and P.L. Clark, *Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design.* Journal of the American Chemical Society, 2014. **136**(3): p. 858-861.

55. Buhr, F., et al., *Synonymous codons direct cotranslational folding toward different protein conformations.* Molecular cell, 2016. **61**(3): p. 341-351.

56. Purvis, I.J., et al., *The efficiency of folding of some proteins is increased by controlled rates of translation in vivo: a hypothesis.* Journal of molecular biology, 1987. **193**(2): p. 413-417.

57. Pechmann, S., J.W. Chartron, and J. Frydman, *Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo.* Nature structural & molecular biology, 2014. **21**(12): p. 1100.

58. Pechmann, S. and J. Frydman, *Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding.* Nature structural & molecular biology, 2013. **20**(2): p. 237.

59. Zhou, T., M. Weems, and C.O. Wilke, *Translationally optimal codons associate with structurally sensitive sites in proteins.* Molecular biology and evolution, 2009. **26**(7): p. 1571-1580.

60. Zhou, M., et al., *Nonoptimal codon usage influences protein structure in intrinsically disordered regions.* Molecular microbiology, 2015. **97**(5): p. 974-987.

61. Warnecke, T. and L.D. Hurst, *GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution.* Molecular systems biology, 2010. **6**(1).

62. Fukui, R., et al., *Unc93B1 restricts systemic lethal inflammation by orchestrating Toll-like receptor 7 and 9 trafficking.* Immunity, 2011. **35**(1): p. 69-81.

63. Christensen, S.R., et al., *Toll-like receptor 7 and TLR9 dictate autoantibody specificity and have opposing inflammatory and regulatory roles in a murine model of lupus.* Immunity, 2006. **25**(3): p. 417-428.

64. Nickerson, K.M., et al., *TLR9 regulates TLR7-and MyD88-dependent autoantibody production and disease in a murine model of lupus.* The journal of Immunology, 2010. **184**(4): p. 1840-1848.

65. Deane, J.A., et al., *Control of toll-like receptor 7 expression is essential to restrict autoimmunity and dendritic cell proliferation.* Immunity, 2007. **27**(5): p. 801-810.

66. Newman, Z.R., et al., *Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9.* Proceedings of the National Academy of Sciences, 2016. **113**(10): p. E1362-E1371.

67. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli.* science, 2009. **324**(5924): p. 255-258.

68. Pop, C., et al., *Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.* Molecular systems biology, 2014. **10**(12).

69. Tuller, T., et al., *An evolutionarily conserved mechanism for controlling the efficiency of protein translation.* Cell, 2010. **141**(2): p. 344-354.

70. Gingold, H. and Y. Pilpel, *Determinants of translation efficiency and accuracy.* Molecular systems biology, 2011. **7**(1).

71. Presnyak, V., et al., *Codon optimality is a major determinant of mRNA stability.* Cell, 2015. **160**(6): p. 1111-1124.

72. Boël, G., et al., *Codon influence on protein expression in E. coli correlates with mRNA levels.* Nature, 2016. **529**(7586): p. 358.

73. Mishima, Y. and Y. Tomari, *Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish.* Molecular cell, 2016. **61**(6): p. 874-885.

74. Wu, Q., et al., *Translation affects mRNA stability in a codon-dependent manner in human cells.* Elife, 2019. **8**: p. e45396.

75. Zhou, Z., et al., *Codon usage is an important determinant of gene expression levels largely through its effects on transcription.* Proceedings of the National Academy of Sciences, 2016. **113**(41): p. E6117-E6125.

76. Fu, J., et al., *Codon usage regulates human KRAS expression at both transcriptional and translational levels.* Journal of Biological Chemistry, 2018. **293**(46): p. 17929-17940.

77. Radhakrishnan, A., et al., *The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality.* Cell, 2016. **167**(1): p. 122-132. e9.

78. Webster, M.W., et al., *mRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-Not nucleases.* Molecular cell, 2018. **70**(6): p. 1089-1100. e8.

79. Hanson, G. and J. Coller, *Codon optimality, bias and usage in translation and mRNA decay.* Nature reviews Molecular cell biology, 2018. **19**(1): p. 20.

80. Zhou, Z., et al., *Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation.* Elife, 2018. **7**: p. e33569.

81. Xu, Y., et al., *Non-optimal codon usage is a mechanism to achieve circadian clock conditionality.* Nature, 2013. **495**(7439): p. 116.

82. Hershberg, R. and D.A. Petrov, *Selection on codon bias.* Annual review of genetics, 2008. **42**: p. 287-299.

83. Hiraoka, Y., et al., *Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe.* Genes to Cells, 2009. **14**(4): p. 499-509.

84. Quax, T.E., et al., *Codon bias as a means to fine-tune gene expression.* Molecular cell, 2015. **59**(2): p. 149-161.

85. Rudolph, K.L., et al., *Codon-driven translational efficiency is stable across diverse mammalian cell states.* PLoS genetics, 2016. **12**(5): p. e1006024.

86. Hambuch, T.M. and J. Parsch, *Patterns of synonymous codon usage in Drosophila melanogaster genes with sex-biased expression.* Genetics, 2005. **170**(4): p. 1691-1700.

87. Heger, A. and C.P. Ponting, *Variable strength of translational selection among 12 Drosophila species.* Genetics, 2007. **177**(3): p. 1337-1348.

88. Kanaya, S., et al., *Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage*

as assessed by multivariate analysis. Journal of molecular evolution, 2001. **53**(4-5): p. 290-298.

89. Carlini, D.B. and W. Stephan, *In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein.* Genetics, 2003. **163**(1): p. 239-243.

90. Warnecke, T. and L.D. Hurst, *Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster.* Molecular biology and evolution, 2007. **24**(12): p. 2755-2762.

91. Brasey, A., et al., *The leader of human immunodeficiency virus type 1 genomic RNA harbors an internal ribosome entry segment that is active during the G2/M phase of the cell cycle.* Journal of virology, 2003. **77**(7): p. 3939-3949.

92. Wei, J., et al., *The stringency of start codon selection in the filamentous fungus Neurospora crassa.* Journal of Biological Chemistry, 2013. **288**(13): p. 9549-9562.

93. McCallum, C.D., et al., *The interaction of the chaperonin tailless complex polypeptide 1 (TCP1) ring complex (TRiC) with ribosome-bound nascent chains examined using photo-cross-linking.* The Journal of cell biology, 2000. **149**(3): p. 591-602.

94. Kolb, V.A., E.V. Makeyev, and A.S. Spirin, *Folding of firefly luciferase during translation in a cell‐free system.* The EMBO journal, 1994. **13**(15): p. 3631-3637.

95. Brierley, I., *Ribosomal frameshifting on viral RNAs.* Journal of General Virology, 1995. **76**(8): p. 1885-1892.

96. Bazzini, A.A., et al., *Codon identity regulates mRNA stability and translation efficiency during the maternal‐to‐zygotic transition.* The EMBO journal, 2016. **35**(19): p. 2087-2103.

97. Zhao, F., C.-h. Yu, and Y. Liu, *Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells.* Nucleic acids research, 2017. **45**(14): p. 8484-8492.

98. Vogel, H.J., *A convenient growth medium for Neurospora (Medium N).* Microb. Genet. Bull., 1956. **13**: p. 42-43.

99. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.

100. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-169.

101. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.

102. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

103. Peden, J., *Analysis of codon usage: University of Nottingham.* UK: Nottingham, 1999.

104. Spies, N., C.B. Burge, and D.P. Bartel, *3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts.* Genome research, 2013. **23**(12): p. 2078-2090.

105. Espinar, L., et al., *Promoter architecture determines cotranslational regulation of mRNA.* Genome research, 2018. **28**(4): p. 509-518.

106. Apcher, S., et al., *Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway.* Proceedings of the National Academy of Sciences, 2013. **110**(44): p. 17951-17956.

107. David, A., et al., *Nuclear translation visualized by ribosome-bound nascent chain puromycylation.* J Cell Biol, 2012. **197**(1): p. 45-57.

108. Iborra, F.J., D.A. Jackson, and P.R. Cook, *Coupled transcription and translation within nuclei of mammalian cells.* Science, 2001. **293**(5532): p. 1139-1142.