THE ROLE OF CODON USAGE IN REGULATING PROTEIN EXPRESSION, STRUCTURE AND FUNCTION

APPROVED BY SUPERVISORY COMMITTEE

Yi Liu, Ph.D. (Mentor)

_____

Benjamin Tu, Ph.D. (Chair)

_____

Joseph Takahashi, Ph.D.

_____

Andrew Zinn, Ph.D.

_____

DEDICATION

To my parents-Jianping Zhou and Xinhua Zhao, and my fianc éEdwin Jiang

THE ROLE OF CODON USAGE IN REGULATING PROTEIN EXPRESSION, STRUCTURE AND FUNCTION

by

MIAN ZHOU

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, 2014

ACKNOWLEDGEMENT

First I want to thank my mentor, Dr. Yi Liu. I appropriate his nice and patient guidance during the past five years. By working with him, I learned a lot of things including experimental design, critical thinking, and how to do scientific presentations. The five-year training in his lab will be critical for my career development in the future.

I would like to thank all member of the Liu lab, both past and present. Dr. Joonseok Cha taught me the basic knowledge and techniques for *Neurospora* circadian clock studies. Jingjing Fu and Dr. Zhipeng Zhou helped with some assays used in Chapter Three. I also want to thank Dr. Qiuying Yang, Dr. Yunkun Dang and Dr. Zhihong Xue for the helpful discussions. In addition, I want to thank Annie Ye and Haiyan Yuan for the technical supports.

I greatly appreciate my dissertation committee members: Dr. Benjamin Tu, Dr. Joseph Takahashi and Dr. Andrew Zinn for their continuous support. Their valuable suggestions and critical insights have played important roles in my projects. Besides, my work was greatly benefited from the collaborations with several other groups. I am thankful to Tao Wang from Dr. Guanghua Xiao's lab for collaborating with me to do those biostatistical analyses. Dr. Jose Barral from UTMB helped to calculate the predicted translation rates, and Dr. Chen from NIBS helped for the mass spec assays.

I would like to express my eternal appropriation towards my parents and my fianc é for their continuous encouragement and support. I would like to thank all my dear friends here in UT Southwestern.

THE ROLE OF CODON USAGE IN REGULATING PROTEIN EXPRESSION, STRUCTURE AND FUNCTION

Publication No. _____

MIAN ZHOU, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2014

Mentor: YI LIU, PhD.

Codon usage bias has been observed in the genomes of almost all organisms and is thought to result from selection for efficient translation of highly expressed genes. Many genes, however, exhibit little codon usage bias. It's not clear whether the lack of codon bias for a gene is due to lack of selection for mRNA translation or it has some biological significance.

The rhythmic expression and the proper function of the *Neurospora* FREQUENCY (FRQ) protein are essential for circadian clock function. However, unlike most genes in

Neurospora, *frq* exhibits non-optimal codon usage across its entire open reading frame (ORF). Optimization of *frq* codon usage results in the abolition of both overt and molecular circadian rhythms. Codon optimization not only increases FRQ expression level but surprisingly, also results in conformational changes in FRQ protein, impaired FRQ phosphorylation, and impaired functions in the circadian feedback loops. These results indicate that non-optimal codon usage of *frq* is essential for maintaining circadian rhythmicity in *Neurospora*.

Interestingly, there is a correlation between codon usage score and FRQ protein structure: the regions that are predicted to be disordered preferentially uses more non-optimal codons. This negative correlation is also found in the proteasome of *Neurospora*, as well in yeast, *Drosophila*, *C. elegans* and *E. coli*. By making a series of *Neurospora* strains with *frq* optimized in different regions, we find that codon optimizations in the predicted disordered regions of FRQ have more prominent effects on FRQ activity and structure. Furthermore, codon optimization of disordered regions in several other *Neurospora* genes results in altered protein degradation rates, suggesting structural changes by codon optimization. Together, these results suggest that codon usage adapts to protein structures and there is a "code" within genetic codons that allow optimal co-translational protein folding.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Zhou, M., J. Guo, et al. (2013). "Non-optimal codon usage affects expression, structure and function of clock protein FRQ." Nature 495(7439): 111-115.

Cha, J., **M. Zhou**, et al. (2013). "CATP is a critical component of the *Neurospora* circadian clock by regulating the nucleosome occupancy rhythm at the frequency locus." EMBO Rep 14(10): 923-930.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

frq (FRQ) – frequency (FREQUENCY)

mRNA – messenger RNA

CAI – codon adaptation index

CBI – codon bias index

ENC – effective number of codons

Fop – frequency of optimal codons

CST – codon selection time

ORF – open reading frame

tAI – tRNA adaptation index

A/DSPS – advanced/delayed sleep phase syndrome

PAS – PER/ARNT/SIM

WC – WHITE COLLAR

ccg – clock controlled genes

per (PER) – period (PERIOD)

CHX – cycloheximide

LS – lineage specificity

CHAPTER ONE

INTRODUCTION

Protein expression levels and protein activities are regulated at four levels. At the transcriptionally level, the participation of various transcription factors regulates mRNA synthesis (Roeder, 1996). Then mRNA levels can be regulated post-transcriptionally, such as alternative splicing and mRNA degradation (Matlin et al., 2005). After proteins are made, they can be subjected to post-translational regulations such as posttranslational modifications and proteasome-mediated degradation are used to control their activity or regulate the protein levels (Ashcroft et al., 1999; Glozak et al., 2005). However, compared with these three levels of regulations, the mechanisms that regulate protein expression at the translational level are much less well studied. Codon usage biases have been observed in almost all organisms. Highly expressed genes prefer to use optimal codons in their open reading frames. It has been long thought that the codon usage bias may regulate protein expression (Gooch et al., 2008; Lavner and Kotlar, 2005; Sharp et al., 1986; Sørensen et al., 1989). Recent studies suggest that codon usage may also play a role in regulating protein folding and activity by regulating protein translation rate (Siller et al., 2010; Spencer et al., 2012; Xu et al., 2013; Zhang et al., 2009). However, the *in vivo* biological significance of this hypothesis has been unclear.

1.1 Codon usage bias

In the nuclear genome, 20 amino acids are coded by more than 60 codons. Most amino acids (except for methionine and tryptophan) are coded by 2-6 synonymous codons. Most synonymous codons differ at the 3rd position of the tri-nucleotide, while some differ by 2-3 different nucleotides. Almost all genomes have their own preferences to use certain synonymous codons against others, and this phenomenon is called codon usage bias (Figure 1-1).

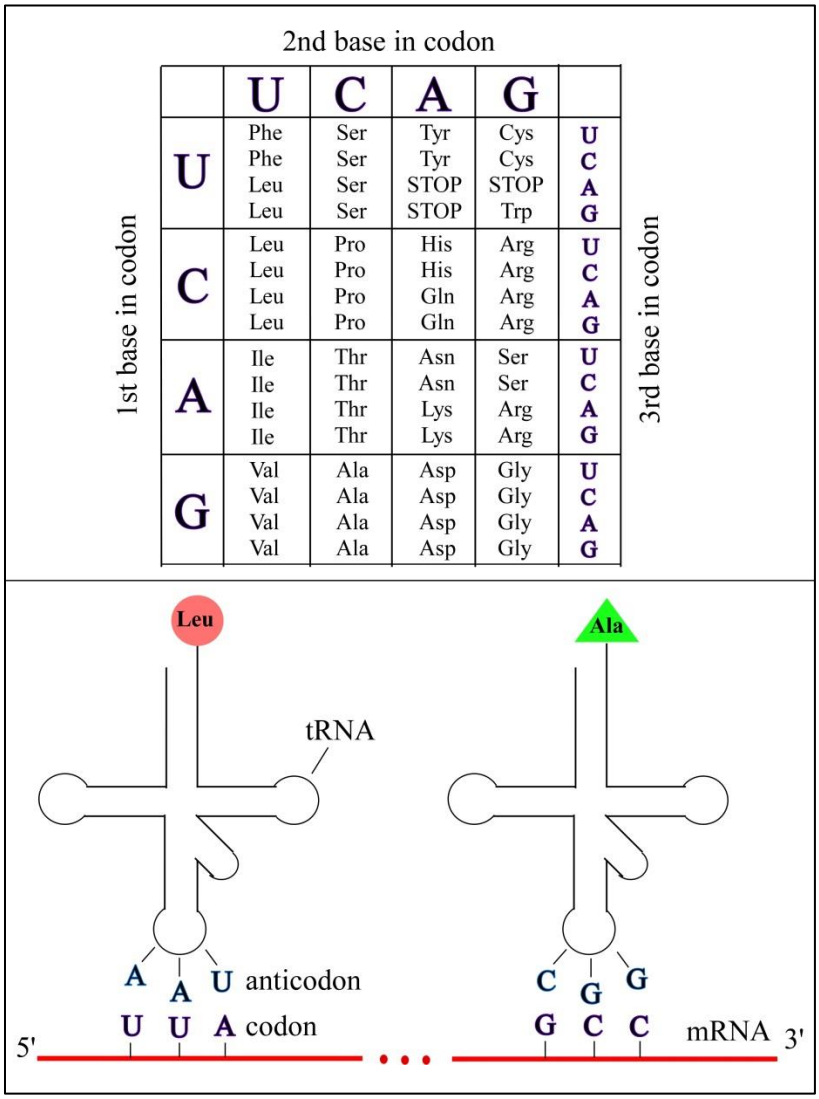


Figure 1-1 Genetic codon and translation.
Top, usually more than one codon encodes the same amino acid or the termination signal; Bottom, codon-anticodon pairing is a key process in translation.

1.1.1 The cause and role of codon usage bias

Selection pressure for efficient translation is thought to be the major cause of codon usage bias in different organisms. As a result, highly expressed genes are encoded predominantly by codons that correspond to highly abundant tRNAs. Codon usage bias is proposed to regulate protein expression by affecting translation rate (Lavner and Kotlar, 2005; Sørensen et al., 1989). Thus codon bias is thought to allow highly expressed genes to be rapidly translated with high fidelity (Akashi, 1994; Drummond and Wilke, 2008). Consistent with this, optimal codon usage has been shown to be functionally important for the expression of several highly expressed genes (Carlini and Stephan, 2003). Although the use of optimal codons for highly expressed genes can be explained by the selection pressure for efficient translation, there are many genes in genomes that exhibit little or no codon usage bias. It is possible that the absence of codon usage bias is due to the lack of selection on translation efficiency for these genes. Alternatively, however, non-optimal codon usage for many genes may also be the result of natural selection and have biological significance.

Codon usage bias has been proposed as a mechanism to regulate protein expression in prokaryotes and some eukaryotes (Gooch et al., 2008; Lavner and Kotlar, 2005; Sharp et al., 1986; Sørensen et al., 1989). Codon optimization is used in multiple heterologous protein expression systems to achieve high levels of protein expression

(Brocca et al., 1998; Gooch et al., 2008). More recently, several bioinformatics and heterologous protein expression studies suggested that codon usage may also regulate co-translational protein folding and activity (Siller et al., 2010; Spencer et al., 2012; Xu et al., 2013; Zhang et al., 2009). For example, when genes of eukaryotic origin are expressed in *E. coli*, synonymous codon substitutions affect their translation speed and protein folding efficiency (Siller et al., 2010; Spencer et al., 2012). In addition, a synonymous SNP in human MDR1 gene is found to alter its substrate specificity (Kimchi-Sarfaty et al., 2007).

1.1.2 Indices used to measure codon usage bias

There are several indices designed to measure codon usage bias. Codon Bias Index (CBI) quantifies the extent to which a gene uses a subset of optimal codons. Its value ranges from -1 (extreme bias to use non-optimal codons) to +1 (extreme bias to use optimal codons). CBI value of 0 indicates a completely random codon choice (Sharp and Li, 1987). Codon adaptation index (CAI) is another index to measure codon bias (Sharp and Li, 1987). By assigning a score (between 0 and 1) to each codon, the CAI score of a gene takes the geometric mean of the scores from all codons, and a higher CAI score indicates a preference to use frequent codons. Effective Number of Codons (ENC) quantifies how far the codon usage of one gene differs from equal codon usage (Wright, 1990), which ranges from 20 (strongest bias) to 61(no bias). Other indices such as tRNA Adaptation Index (tAI) (Reis et al., 2004) and predicted translation rate (Spencer et al., 2012) are calculated based on the cognate tRNA copy number and abundance of each codon. Also there are several new indices which are not designed to measure codon bias but to predict translation rate. One example is Codon Selection Time (CST) (Qian et al., 2012) which is calculated based on ribosome density on mRNA chains.

1.1.3 The mechanism of how codon usage bias functions

It has been proposed that frequently used codons or codons with high cognate tRNA abundance could be recognized by the correct tRNA species more efficiently at ribosome A site, thus accelerating the translation rates (Lavner and Kotlar, 2005; Sørensen et al., 1989). Non-optimal codons may serve as ribosome pausing sites to allow more time for the co-translational folding of the newly synthesized polypeptide chain (Thanaraj and Argos, 1996) (Figure 1-2a). Bioinformatic analyses suggested that optimal codons are more frequently associated with buried residues and structurally sensitive sites (Zhou et al., 2009). And non-optimal codons are enriched in certain secondary structures such as turns and position 2 and 3 in alpha helix (Pechmann and Frydman, 2013). A detailed correlation between codon bias and protein structure is not clear and will be important for the understanding of the role of codon bias in regulating protein folding. Besides codon bias, protein translation rate can be regulated by many other factors such as mRNA structure (Goodman et al., 2013), the existence of anti-SD sequence in prokaryotes (Li et al., 2012), synonymous codon arrangement (Cannarozzi et al., 2010) (Figure 1-2c) and positively charged amino acids (Charneski and Hurst, 2013). A 5' "ramp" hypothesis has been proposed that there are 30-50 non-optimal codons at the N-terminus of many genes, which serve as a "ramp" to reduce ribosome traffic jams during translation initiation (Tuller et al., 2010) (Figure 1-2b). However, a more recent study suggest that the identified high ribosome density at "ramp" regions is an bioinformatics artifact caused by a bias for rapid initiation of smaller genes (Shah et al., 2013). On the other hand, two other studies suggest that the non-optimal codon usage at N-terminus is

due to a selection for reduced mRNA secondary structure rather than non-optimal codons (Bentele et al., 2013; Goodman et al., 2013) (Figure 1-2).

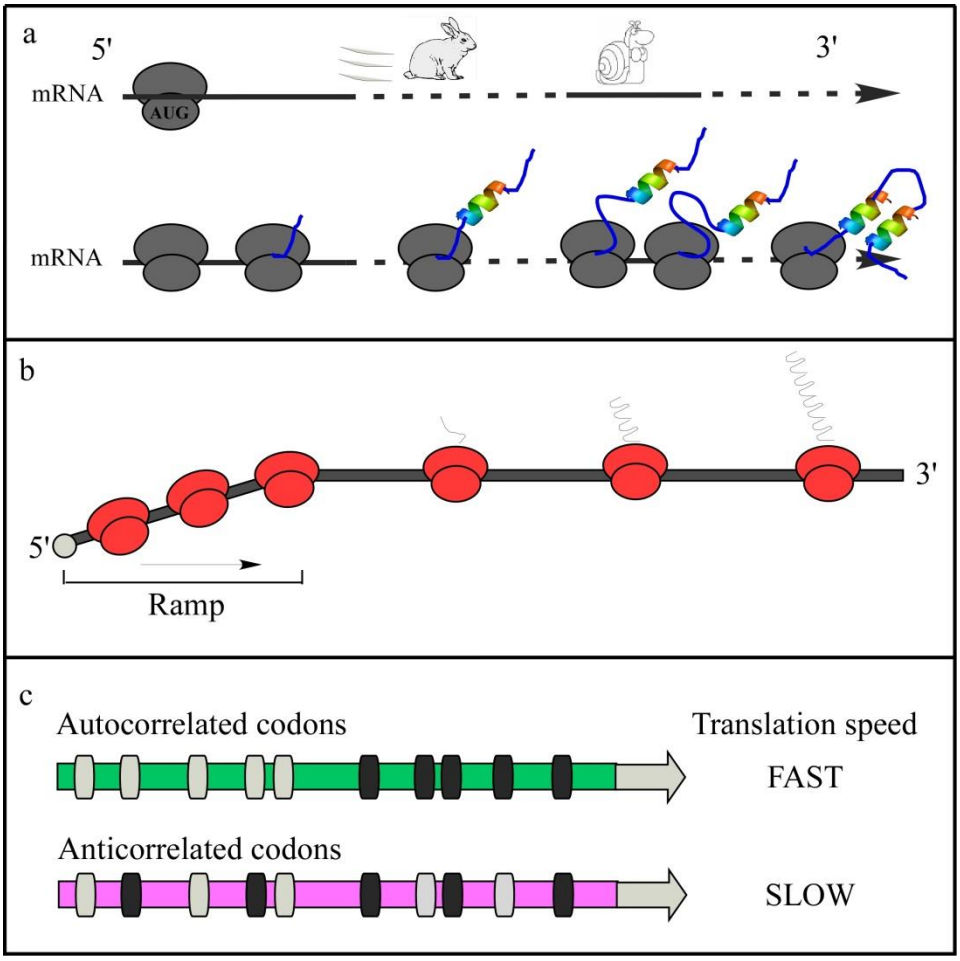


Figure 1-2 Models of how codon usage bias regulates protein translation and folding.
(a) Optimal and non-optimal codons regulate translation rate and nascent polypeptide folding; (b) The "ramp" hypothesis; (c) Codon arrangement affects translation speed.

1.1.4 Codon usage bias in the genome of *Neurospora crassa*

The *Neurospora* genome exhibits a bias against the use of A and T at the wobble base of a codon (Radford and Parish, 1997). We surveyed all predicted protein-encoding *Neurospora* genes and found that the GC content of the 3rd position is 64% with the nucleotide preference of C>G>T>A for almost every codon family. Codon optimization has been shown to enhance expression of heterologous genes in *Neurospora* (Morgan et al., 2003; Segers et al., 2007), suggesting that codon usage is important for regulating protein expression. (Table1-1)

aa	codon	*N.crassa* Observed codon frequency	*N.crassa* Elongation rate (au)	aa	codon	*N.crassa* Observed codon frequency	*N.crassa* Elongation rate (au)
Phe	UUU	0.348	0.333	Glu	GAA	0.344	0.222
	UUC	**0.652**	**1**		**GAG**	**0.656**	**0.852**
Leu	UUA	0.033	0.065	Ser	UCU	0.146	0.172
	UUG	0.180	0.183		**UCC**	**0.245**	**0.517**
	CUU	0.171	0.172		UCA	0.113	0.299
	CUC	**0.323**	**0.516**		UCG	0.177	0.184
	CUA	0.072	0.269		AGU	0.106	0.069
	CUG	0.221	0.194		AGC	0.213	0.207
Ile	AUU	0.314	0.3	Pro	CCU	0.234	0.193
	AUC	**0.594**	**0.9**		**CCC**	**0.347**	**0.579**
	AUA	0.092	0.383		CCA	0.192	0.404
Met	AUG	1	1		CCG	0.226	0.281
Val	GUU	0.232	0.244	Thr	ACU	0.186	0.206
	GUC	**0.417**	**0.731**		**ACC**	**0.411**	**0.619**
	GUA	0.091	0.359		ACA	0.178	0.381
	GUG	0.261	0.192		ACG	0.225	0.254
Tyr	UAU	0.327	0.333	Ala	GCU	0.243	0.226
	UAC	**0.673**	**1**		**GCC**	**0.414**	**0.679**
Stop	UAA	0.316			GCA	0.145	0.405
	UAG	0.263			GCG	0.199	0.202
	UGA	0.421		Cys	UGU	0.306	0.333
Trp	UGG	1	1		**UGC**	**0.694**	**1**
His	CAU	0.391	0.333	Arg	CGU	0.144	0.210
	CAC	**0.609**	**1**		**CGC**	**0.285**	**0.629**
Gln	CAA	0.395	0.267		CGA	0.115	0.267
	CAG	**0.605**	**0.822**		CGG	0.138	0.105
Asn	AAU	0.276	0.333		AGA	0.128	0.171
	AAC	**0.724**	**1**		AGG	0.191	0.114
Lys	AAA	0.225	0.111	Gly	GGU	0.255	0.241
	AAG	**0.775**	**0.926**		**GGC**	**0.404**	**0.722**
Asp	GAU	0.425	0.333		GGA	0.189	0.139
	GAC	**0.575**	**1**		GGG	0.152	0.185

Table 1-1 *Neurospora* codon usage frequencies and the predicted translation elongation rate based on tRNA copy numbers and anticodon-codon interaction. (Table adapted from Zhou et al., 2013)

1.2 Circadian Clock

Circadian clocks are used by most eukaryotic and some prokaryotic organisms to coordinate diverse molecular and physiological activities with daily changes in environmental conditions (Dunlap, 1999). Biological clocks are important in human physiology and mental health, such as the sleep/wake cycles (Crocker and Sehgal, 2010) and endocrine function (Gan and Quinton, 2010). Circadian clock disorder such as advanced/delayed sleep phase syndrome (A/DSPS) and sleep disorders are common clinical cases. Circadian clock principles have been used to optimize therapeutic outcomes. (Hrushesky, 1985) and to treat patients with sleep disorders (Crocker and Sehgal, 2010).

All circadian clocks can be entrained by exogenous signals such as light and temperature, but in the absence of these signals, they maintain an endogenous period that is close to 24 hours. Every circadian clock system can be simplified into three major components: input, oscillator and output (Eskin, 1979). The input pathway senses environment signals and enable the clock to be entrained by external time cues. The oscillator receives information from the input pathway, integrates the information, and generates an endogenous rhythmicity. The output pathway collects timing information from the oscillator, and controls a wide variety of rhythmic activities (Figure 1-3).

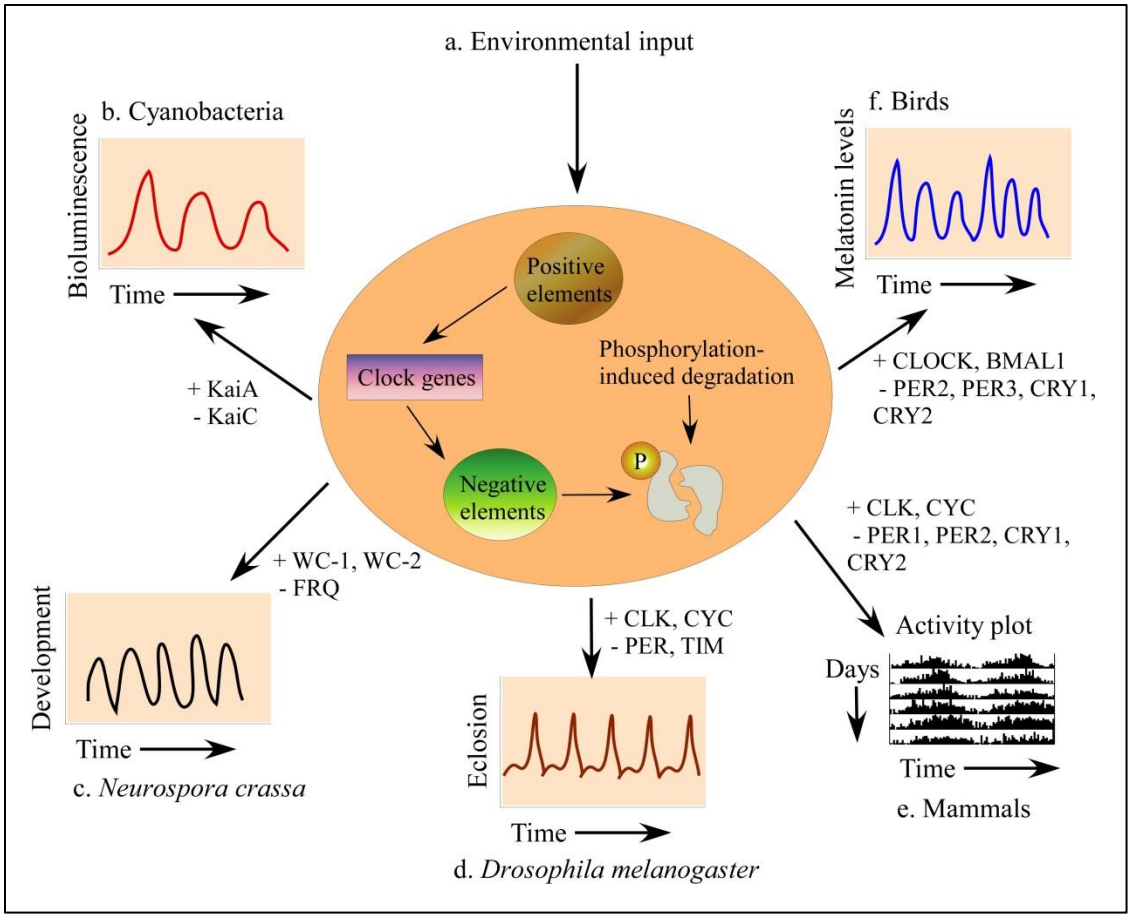


Figure 1-3 Circadian oscillators are controlled through a common mechanism. Schematic diagram showing the positive and negative elements in conserved autoregulatory feedback loops of (b) Cyanobacteria, (c) *Neurospora crassa*, (d) *Drosophila melanogaster*, (e) Mammals and (f) Birds. Circadian oscillators are also influenced by (a) environment inputs. +, positive elements; -, negative elements.

1.2.1 The *Neursoproa* circadian clock

The eukaryotic circadian oscillators consist of autoregulatory circadian negative feedback loops (Bell-Pedersen et al., 2005; Dunlap, 1999; Young and Kay, 2001). Despite the evolutionary distance between the filamentous fungus *Neurospora crassa* and

higher eukaryotes, their circadian oscillator mechanisms share remarkable similarities (Baker et al., 2012; Heintzen and Liu, 2007; Liu and Bell-Pedersen, 2006). In the core *Neurospora* circadian oscillator, FREQUENCY (FRQ) protein is a central component and functions as the circadian negative element with its protein partner FRH (Aronson et al., 1994; Cheng et al., 2005). Two PER-ARNT-SIM (PAS) domain transcription factors, WHITE COLLAR (WC)-1 and WC-2 form a heterodimeric complex that activates *frequency* (*frq)* transcription (Cheng et al., 2001b; Crosthwaite et al., 1997). To close the negative feedback loop, the FRQ-FRH complex inhibits WC complex activity by interacting with the WCs (Cheng et al., 2001a; Denault et al., 2001; He et al., 2006; Schafmeier et al., 2005). The activation, inhibition and reactivation of *frq* transcription results in robust circadian rhythms of *frq* mRNA and FRQ protein levels that control many rhythmic molecular and physiological processes (Garceau et al., 1997). In addition to its role in the negative feedback loop, FRQ also promotes the expression of both WC proteins in an interlocked positive feedback loop that is important for the robustness of the clock (Cheng et al., 2003; Cheng et al., 2001b; Lee et al., 2000). (Figure 1-4).

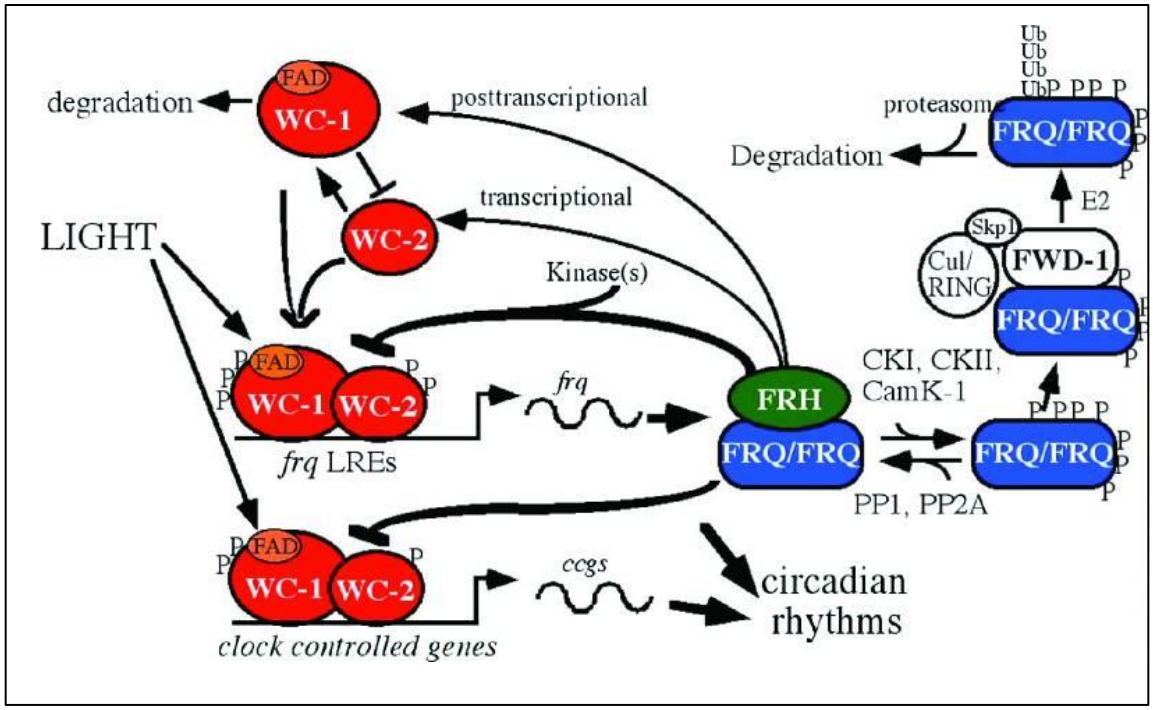


Figure 1-4 Current model of the frq/wc based circadian oscillator in *Neurospora crassa*.
(Figure Adapted from Yi Liu et al., 2006)

1.2.2 Codon usage of clock genes

Unlike those highly expressed genes, many circadian clock elements show non-optimal codon usage. As shown in Figure 1-5, *frq* in *Neurospora*, *KaiBC* in cyanobacteria and human period (per) genes all have large regions of gene with non-optimal codon usage. Codon optimizations on *frq* affect its expression, structure and function, which will be discussed later in chapter 2 and chapter 3. In addition, the non-optimal codon usage of *KaiBC* in cyanobacteria is also a mechanism to enhance organismal fitness at cool temperatures (Xu et al., 2013). Similarly, synonymous polymorphisms in *hPer1 hPer2* ORF has been reported to associate with diurnal preference in human (Carpen et al., 2006; Matsuo et al., 2007)

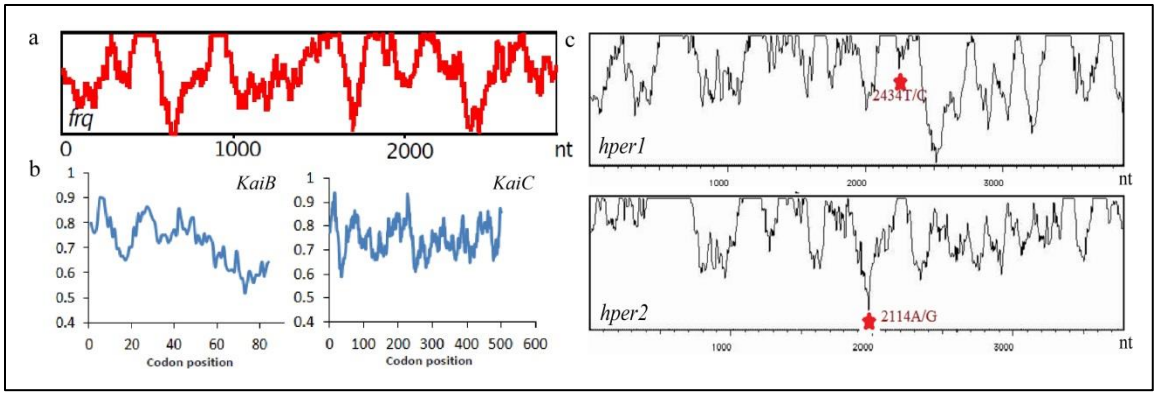


Figure 1-5 Codon score plot of some clock components.
Codon score plot of *frq* in *Neurospora* (a), *KaiB and KaiC* in cyanobacteria (b) and *hper1* and *hper2* in human (c). Stars in C indicate the position of 2 synonymous polymorphisms.

1.3 Summary

As a co-translational mechanism, codon usage bias has been studied in multiple experimental and industrial purposes to achieve optimal protein expression. However, the preference to use non-optimal codons in some genes may also be a result of nature selection and has biological significance. In Chapter Two, I will describe our efforts to determine the role of non-optimal codon usage in *frq* to regulate protein expression, structure and function. Our results suggest that protein folding, which occurs co-translationally, requires protein chaperones and sufficient amounts of time. Codon optimization results in increased translation rates and thus reduces the time available for co-translational folding.

In Chapter Three, I will further describe our study on the correlation between codon usage bias and protein structures. Our results reveal the importance of non-optimal

codons at structurally disordered regions. Such correlations suggest that the codon usage plays important roles in regulating protein structure in eukaryotes.

CHAPTER TWO

NON-OPTIMAL CODON USAGE AFFECTS EXPRESSION, STRUCTURE AND FUNCTION OF CLOCK PROTEIN FRQ

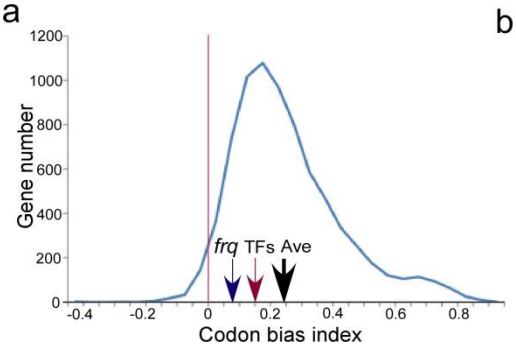
2.1 Introduction

Eukaryotic circadian oscillators consist of autoregulatory circadian negative feedback loops. In the core circadian oscillator of *Neurospora crassa*, FRQ protein is a central component that functions as the circadian negative element with its partner FRH (Baker et al., 2012; Cheng et al., 2005; Heintzen and Liu, 2007). Two transcription factors, WHITE COLLAR (WC)-1 and WC-2, form a heterodimeric complex that activates *frq* transcription (Cheng et al., 2001b; Crosthwaite et al., 1997). The FRQ-FRH complex inhibits WC complex activity by interacting with the WCs (He et al., 2006; Schafmeier et al., 2005). The level and stability of FRQ play a major role in setting period length, phase and clock-sensitivity to environmental signals (Baker et al., 2012; Heintzen and Liu, 2007; Huang et al., 2006). In addition, FRQ promotes the expression of both WC proteins in an interlocked positive feedback loop (Cheng et al., 2001b; Lee et al., 2000).

The protein-coding genes of *Neurospora* exhibit strong codon bias (Figure 2-1a). The third position of almost every codon family in this filamentous fungus has the preference C>G>T>A. Codon optimization enhances expression of a heterologous luciferase gene in *Neurospora (Gooch et al., 2008; Morgan et al., 2003)*. To establish that codon usage bias regulates protein expression, we identified the most abundant *Neurospora* proteins in whole cell extract by mass spectrometry analyses. The genes

encoding the top 100 most-abundant proteins exhibit much stronger codon bias than the rest of the protein coding genes (Figure 2-1b).

a



b

	All genes	Top100	*frq*	*frh*
GC	56%	59%	56%	54%
GC 3rd	64%	76%	**58%**	69%
Fop	52%	78%	**43%**	64%
CBI	0.24	0.64	**0.08**	0.42
ENC	53	34	**57**	44

Figure 2-1 The codon usage profile of *Neurospora crassa* genome.
(a) Distribution of *Neurospora* genes based on their CBI values. Numbers of genes within each 0.05 unit of CBI were counted. The red line indicates CBI value of 0, meaning that codon usage is random. The CBI values of average *Neurospora* genes, *frq*, and the average *Neurospora* transcription factors are indicated. (b) A table summarizing the GC content, GC content at the 3rd position of codons, Fop, CBI and ENC for all predicted *Neurospora* protein-encoding genes, genes encoding the top 100 proteins, *frq* and *frh*.

We classified all predicted *Neurospora* tRNA genes and predicted the relative translation elongation rate for each codon based on tRNA-gene copy numbers, which correlate with tRNA abundance, and the nature of anticodon-codon interactions (Spencer et al., 2012; Tuller et al., 2010). The most preferred codon for each amino acid is always the codon with highest predicted translation elongation rate (Table 1-1). Therefore, to ensure efficient translation of abundant proteins, selection pressure favored a bias for codons that are translated by highly abundant tRNA species.

Many *Neurospora* genes exhibit little or no codon biases (Figure 2-1a). FRQ is a low abundance *Neurospora* protein. Its codon bias index (Bennetzen and Hall, 1982)

(CBI; CBI = 0 indicates completely random codon usage) value of 0.08 indicates that *frq* has little codon bias (Figure 2-1b). A codon usage graph of the *frq* ORF shows that many regions have non-optimal usage, whereas *frh* has good codon usage throughout its ORF (Figure 2-2a).

In this paper, we made a series of *frq* codon optimized *Neurospora* strains, and monitored the expression, structure and activity change of FRQ *in vitro* and *in vivo*. Together, our results reveal that non-optimal gene codon usage of genes is also a result of selection and can have a major role in regulating protein expression and function.

2.2 Materials and Methods

2.2.1 Strains and growth conditions

Neurospora strains used in this study were 87-3 (*bd, a*; clock wild-type), 303-3 (*bd, frq*^{10}*, his-3*) (Cha et al., 2011) and different *frq* N-terminal codon optimized strains created in this study. Strain 303-3 (*bd, frq*^{10}*, his-3*) was used as the host strain for various *his-3* targeting constructs. The *frq*^{10}, bd, wt-*frq* (*frq*^{10} containing a wild-type *frq* gene at the *his-3* locus) strain was used as the control in this study.

Growth conditions have been described previously (Aronson et al., 1994). Liquid cultures were grown in minimal medium (1×Vogel's, 2% glucose). When QA was used, liquid cultures were grown in (10-6M to 10-2M) QA (pH 5.8), 1×Vogel's, 0.1% glucose and 0.17% arginine. Race tube media contained 1×Vogel's, 0.1% glucose, 0.17% arginine, 50 ng/ml biotin, and 1.5% agar. For rhythmic experiments, the *Neurospora* cultures were transferred from LL to DD at time 0 and were harvested in constant

darkness at the indicated time (hours). Calculations of period length were performed as described (Liu et al., 1997).

2.2.2 *frq* codon optimization, codon usage score plot and indices calculation

frq codon optimization was performed for the N-terminal part (1-498 nt) or the middle region (553-1590 nt) of the ORF. Codons were optimized based on *Neurospora crassa* codon usage frequency and the predicted most efficient codon based on tRNA copy numbers. 65 codons were optimized in the m-*frq* construct while 94 codons were optimized in the f-*frq* construct. Sequences surrounding an alternative *frq* 3' splice site in this region were not mutated.

Codon usage score plot was obtained using Codon Usage 3.5 (developed by Conrad Halling) using window size of 35 and logarithmic range of 3. *Neurospora* crassa codon usage frequency table was obtained from http://www.kazusa.or.jp/codon/. To calculate codon bias index (CBI), Frequency of Optimal Codons (Fop), ENC and GC content, codonw in the Mobyle Portal website (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::codonw) was used (Bennetzen and Hall, 1982; Wright, 1990). CBI will equal 1.0 if a gene has extreme codon bias and will equal 0 if codon usage is completely random. If the number of optimal codons is less than expected by random change, the CBI value will be a negative value. *Neurospora* genome sequences were downloaded from the Broad Institute *Neurospora crassa* database (http://www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html). The top 100 abundant proteins were identified by mass spectrometry analyses and ranked by their emPAI values (Ishihama et al., 2005).

2.2.3 Plasmid constructs and *Neurospora* transformation

pKAJ120 (containing the entire wild-type *frq* gene including its promoter and a *his-3* targeting sequence) and pBA50 (containing the wild-type *frq* gene under the control of the qa-2 promoter and a *his-3* targeting sequence) were used as the parental plasmids to create the optimized *frq* constructs (Aronson et al., 1994). Optimized *frq* sequences (synthesized by Genscript) were subcloned into parental plasmids to replace the wild-type *frq* gene to generate m-*frq*, f-*frq* and mid-*frq* constructs. In the m1-*frq* construct, only the codons upstream of the predicted intron branch point were optimized as m-*frq*. For the m2-*frq* construct, only the codons downstream of the intron 3' end were optimized as m-*frq*. The resulting constructs were transformed into strain 303-3 (bd, frq^{10}, *his-3*) by electroporation and targeted to the *his-3* locus (Bell-Pedersen et al., 1996). Homokaryon strains were obtained by microconidia purification.

2.2.4 Protein and RNA analyses

Protein extraction, Western blot analysis, and immunoprecipitation assays were performed as previously described (Cheng et al., 2001a; Garceau et al., 1997; Guo et al., 2009). Equal amounts of total protein (50μg) were loaded in each lane of SDS-PAGE gels (7.5% SDS-PAGE gels containing a ratio of 37.5:1 acrylamide/bisacrylamide). Densitometry of the signal was performed by using Image J.

RNA extraction and qRT-PCR were performed as previously described (Choudhary et al., 2007; Crosthwaite et al., 1995). For qRT-PCR, the primer sequences used for *frq* were 5'-GGAGGAGTCGATGTCACAAGG-3' (forward) and 5'-CACTTCGAGTTACCCATGTTGC-3' (reverse). The *Neurospora* β-tubulin gene was used as an internal control. The primer sequences specific for tubulin were 5'-

GCGTATCGGCGAGCAGTT-3’(forward) and 5’-CCTCACCAGTGTACCAATGCA-3’ (reverse).

frq mRNA secondary structure and folding energy was predicted by mfold program (http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form).

2.2.5 Mass spectrometric analyses and database search

The *Neurospora* proteins were separated on a 4-20% SDS-PAGE gradient gel. The whole protein gel lane was sliced equally into 14 gel blocks from top to bottom. Each gel block was de-stained and then digested in-gel with sequencing grade trypsin (10 ng/μL trypsin in 50 mM ammonium bicarbonate, pH 8.0) at 37 °C overnight. The resulting tryptic peptides from each gel block were extracted with 5% formic acid/50% acetonitrile and 0.1% formic acid/75% acetonitrile sequentially and then concentrated to ~ 20 μl in a CentriVap system (Labconco, Missouri, USA). The extracted peptides from each sample were separated by a home-made analytical capillary column (50 μm × 10 cm) packed with 5 μm spherical C18 reversed phase material (YMC, Kyoyo, Japan). An Agilent 1100 binary pump was used to generate HPLC gradient as follows: 0-5% B in 5 min, 5-40% B in 55 min, 40-100% B in 15 min (A = 0.2 M acetic acid in water; B = 0.2 M acetic acid/70% acetonitrile). The eluted peptides were sprayed directly into a LTQ mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) equipped with a nano-ESI ion source. The mass spectrometer was operated in data-dependent mode (MS scan mass range was from 350 to 2000 Da, the top 5 most abundant precursor ions from each MS scan were selected for MS/MS scans, and dynamic exclusion time was 30 seconds). The mass spectrometric data from all 14 samples were combined and searched against *Neurospora* protein database on an in-house Mascot server (Matrix Science Ltd., London,

UK). The major search parameters were as below: protein N-terminal acetylation and methionine oxidation were included as variable modifications; two missed cleavage sites were allowed; precursor ion mass tolerance was set as 3 Da; fragment ion mass tolerance was 0.8 Da. Only peptides with E-value above 0.01 were retained. The emPAI (Exponentially Modified Protein Abundance Index) (Ishihama et al., 2005) was calculated for each protein by the Mascot software.

2.2.6 Luciferase reporter assay

The luciferase reporter construct (*frq*-luc-bar) was generated by insertion of the 4.7 kb BamHI-NotI fragment of p*frq-luc-I* (a generous gift from Dr. Jay Dunlap)(Gooch et al., 2008) into the corresponding sites of pBARKS1. This construct, which contains the luciferase gene under the control of the *frq* promoter and the bar gene, was transformed into 87-3 (bd, a), wt-*frq* (*frq*^{10}, bd, wt-*frq*), m-*frq* (*frq*^{10}, bd, m-*frq*) and f-*frq* (*frq*^{10}, bd, f-*frq*) strains. Transformants were selected using the basta/ignite (200 μg/mL) resistance conferred by the bar gene.

LumiCycle (ACTIMETRICS) was used for the luciferase assay using a protocol similar to previously described (Gooch et al., 2008). The AFV (autoclaved FGS-Vogel's) medium contained 1×FGS (0.05% fructose, 0.05% glucose, 2% sorbose), 1×Vogel's medium, 50 μg/L biotin, and 1.8% agar. Firefly luciferin (BioSynt L-8200 D-luciferin firefly [synthetic] potassium salt) was added to the medium after autoclaving (final concentration of 50 μM). One drop of conidia suspensions in water were placed on AFV medium and grown in constant light (LL) overnight. The cultures were then transferred to darkness, and luminescence was recorded in real time (DD) using a LumiCycle after one day in DD. The data was then normalized with LumiCycle Analysis software by

subtracting the baseline luciferase signal which increases as cell grows. Therefore, the normalized luciferase signals reflect the amplitude of the rhythm and do not reflect the absolute luciferase signal. Under our experimental condition, luciferasae signals are highly variable during the first day in the LumiCycle and become stabilized afterwards, which is likely due to an artifact caused by the light-dark transfer of the cultures. Thus, the results presented were recorded after one day in DD.

2.2.7 Protein stability assay

The liquid cultures of *Neurospora* strains were grown in LL for 1 day prior to the addition of cycloheximide (CHX, final concentration 10μg/mL). Cells were harvested at the indicated time points.

For the trypsin sensitivity assay, protein extracts were diluted to a protein concentration of 2.5μg/μl. 100μl extracts were treated with trypsin (final concentration 1μg/ml) at 25℃. A 20μl sample was taken from the reaction at each time point (0, 5min, 15min and 30min) after addition of trypsin. Protein samples were mixed with protein loading buffer and resolved by SDS-PAGE. To compare trypsin sensitivity of FRQ from different strains, experiments were performed side-by-side and the protein samples were transferred to the same membrane for western blot analysis.

2.3 Results

2.3.1 N-terminal codon optimization of *frq* increases the protein level, and abolishes the circadian rhythms in *Nerurspora crassa*.

In order to study the role of codon usage in regulating FRQ expression, structure and function, we created two constructs in which the N-terminal end (1-164 aa) of *frq*

was codon optimized. In the m-*frq* construct, only the non-preferred codons were changed, while every codon was optimized in the f-*frq* construct. Predicted stability of RNA secondary structure was not significantly affected by the optimization (Table 2-1). These constructs and the wild-type *frq* construct (wt-*frq*) gene were transformed individually into a *frq* null strain (frq^{10}). Both m-*frq* and f-*frq* strains have significantly higher levels of FRQ proteins in constant light (LL) than that of the wt-*frq* strain (Figure 2-2b). On the other hand, *frq* mRNA levels were comparable in these strains (Figure 2-3a). FRQ is known to up-regulate WC protein levels (Cheng et al., 2001b; Lee et al., 2000). The WC-1 and WC-2 levels, however, were similar in these strains despite the much higher levels of FRQ in the optimized strains (Figures 2-2b, and 2-3b).

	wt-*frq*	m-*frq*	f-*frq*	m1-*frq*	m2-*frq*	mid-*frq*
1-30 nt	-7.10	-7.10	-8.20	-7.10	-7.10	-7.10
1-800 nt	-274.90	-292.50	-272.80	-285.30	-275.80	-264.70
553-1590 nt	-374.60					-339.60
full length	-1135.40	-1160.10	-1148.00	-1136.30	-1142.10	-1130.00

Table 2-1 Predicted free energy values (ΔG(kcal/mol)) of *frq* mRNA using mfold.

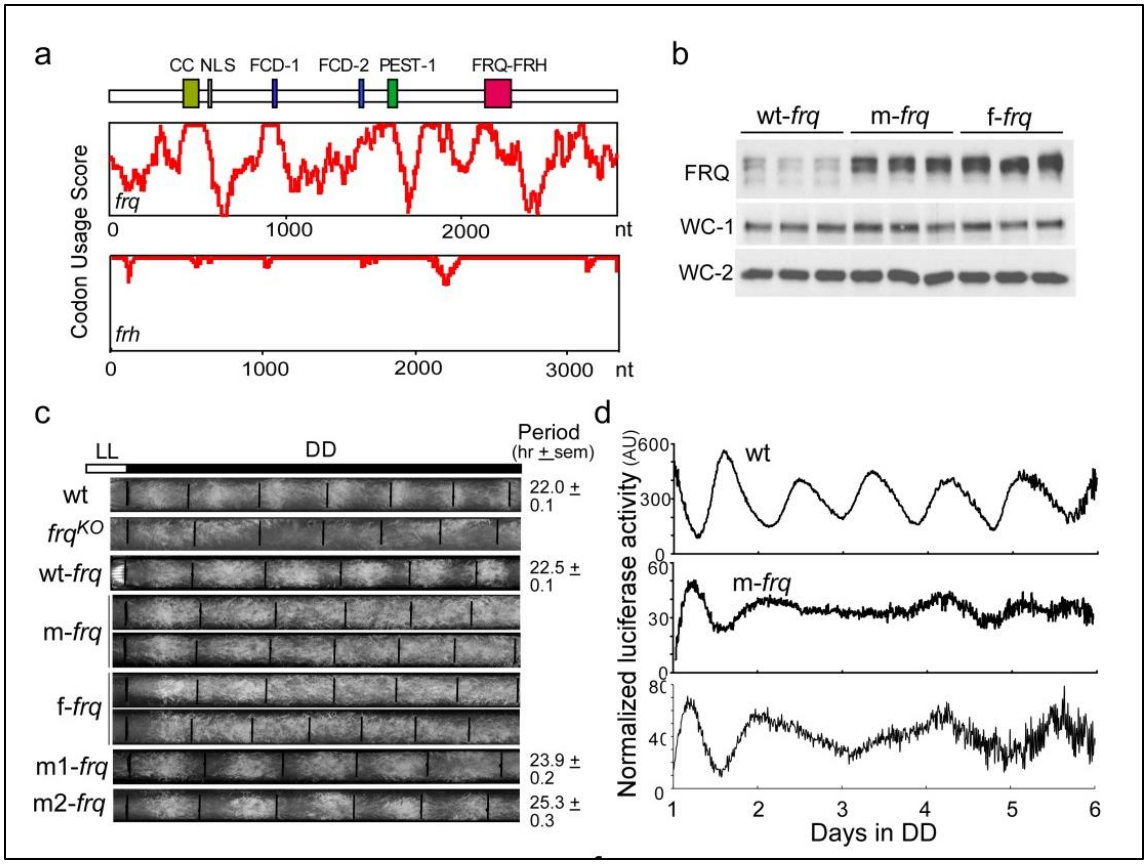


Figure 2-2 Codon optimization of *frq* results in high FRQ expression levels and loss of circadian rhythmicities.
(a) Codon usage score plots of *frq* and *frh* obtained using Codon Usage 3.5. (b) Western blots showing the levels of FRQ and WCs in frq^{10}, wt-*frq* (wt-*frq*); frq^{10}, m-*frq* (m-*frq*) and frq^{10}, f-*frq* (f-*frq*) strains. (c) Race tube analysis showing the conidiation phenotypes in DD. Black lines indicate the growth fronts every 24 h. (d) Luciferase reporter assay showing *frq* promoter activity of the indicated strains after one day in DD.

The wt-*frq* construct was able to fully rescue the arrhythmic conidiation rhythm of the frq^{10} strain in DD (Figure 2-2c), but both of the optimized *frq* strains exhibited arrhythmic conidiation phenotypes; these are not due to the modest changes in the ratios of two alternatively translated FRQ forms, since either form of FRQ alone can support robust rhythms (Colot et al., 2005; Diernfellner et al., 2007).

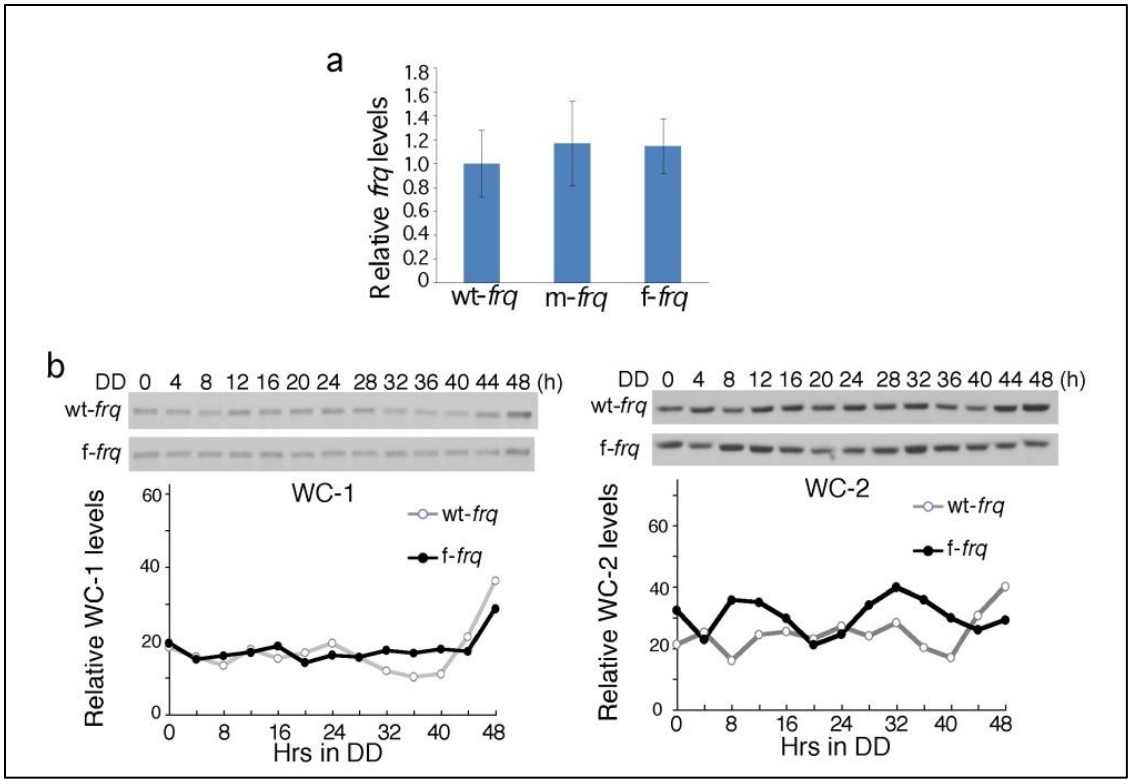


Figure 2-3 The frq mRNA and WC protein levels do not change in codon optimized strains
(a) Quantitative RT-PCR results showing that the levels of *frq* mRNA in LL. Error bars are standard deviation. n=3. (b) Western blot analyses results showing the levels of WC-1 (left panels) and WC-2 (right panels) in the wt-*frq* and f-*frq* strains. *Neurospora* cultures were harvested in DD at the indicated time points.

We created two additional constructs (m1-*frq* and m2-*frq*), in which only the N- or C-terminal segments of the optimized region of m-*frq* were optimized. The frq^{10} transformants carrying either construct exhibited long-period conidiation rhythms and had FRQ levels between those of wt-*frq* and m-*frq* strains (Figures 2-2c & 2-4a). These results suggest that the severe conidiation rhythm phenotypes of the m-*frq* and f-*frq* strains are due to the cumulative effect of codon optimization and are not likely due to mutation of a DNA/RNA element.

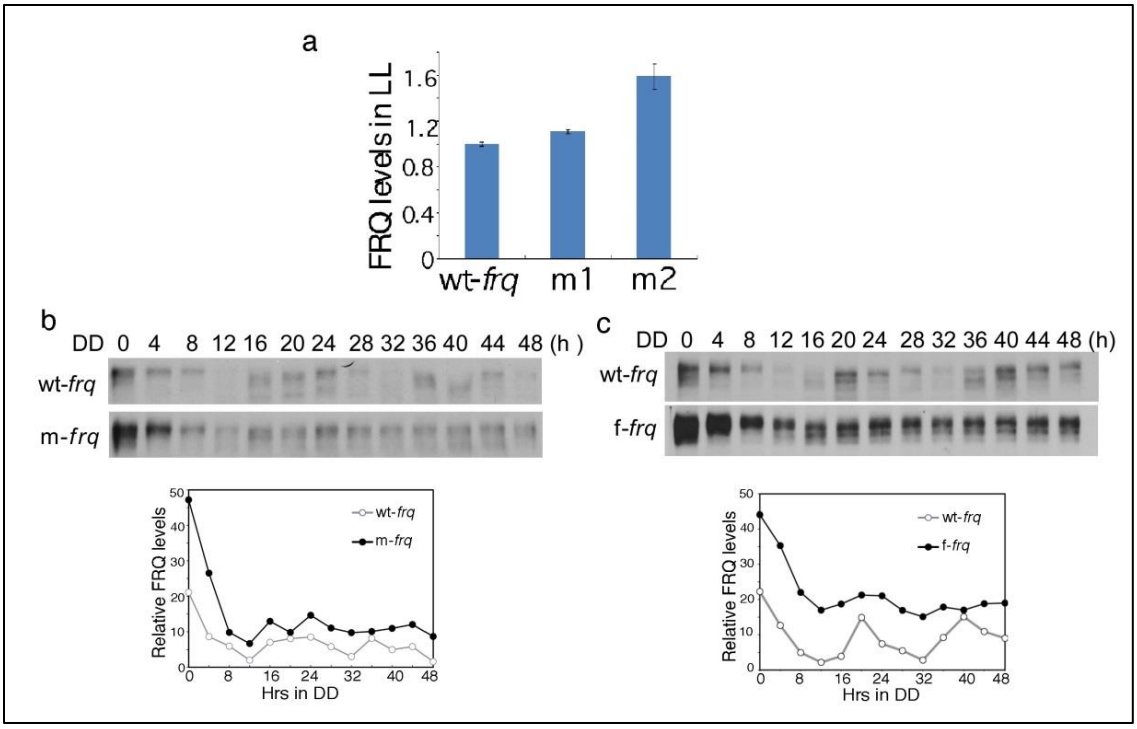


Figure 2-4 Codon optimization of *frq* abolishes its molecular rhythm.
(a) Quantification of western blot results showing the levels of FRQ in the wt-*frq*, m1-*frq* and m2-*frq* strains. Error bars are standard deviation. n=3. (b and c) Western blots and quantification showing loss of FRQ expression rhythms in the codon-optimized strains.

To examine circadian phenotypes at the molecular level, we introduced a luciferase reporter construct that is under the control of the *frq* promoter (Gooch et al., 2008) into wild-type, m-*frq* and f-*frq* strains. As shown in Figures 2-2d, the robust rhythmic luciferase activity was abolished in the optimized strains. FRQ protein levels also lost molecular rhythmicity in the optimized strains (Figures 2-4b & 2-4c): the overall levels of FRQ were high and circadian changes in FRQ abundance and phosphorylation profile were abolished. Together, these results indicate that the non-optimal codon usage of *frq* governs FRQ expression level and is essential for clock function.

2.3.2 FRQ proteins from codon optimized strains have impaired functions in both circadian negative and positive feedback loops.

The loss of circadian rhythms in the optimized strains is surprising because we previously showed that high FRQ levels do not result in arrhythmicity (Cheng et al., 2001b; Liu et al., 1998), suggesting that codon optimization causes defects in FRQ function. FRQ fulfills its function in the circadian negative feedback loop through its interaction with WCs (He et al., 2006). Immunoprecipitation assays showed that the relative amounts of FRQ associated with WC-2 were significantly decreased in both optimized strains (Figure 2-5), suggesting that the FRQ function in the negative feedback loop is impaired in the optimized strains.

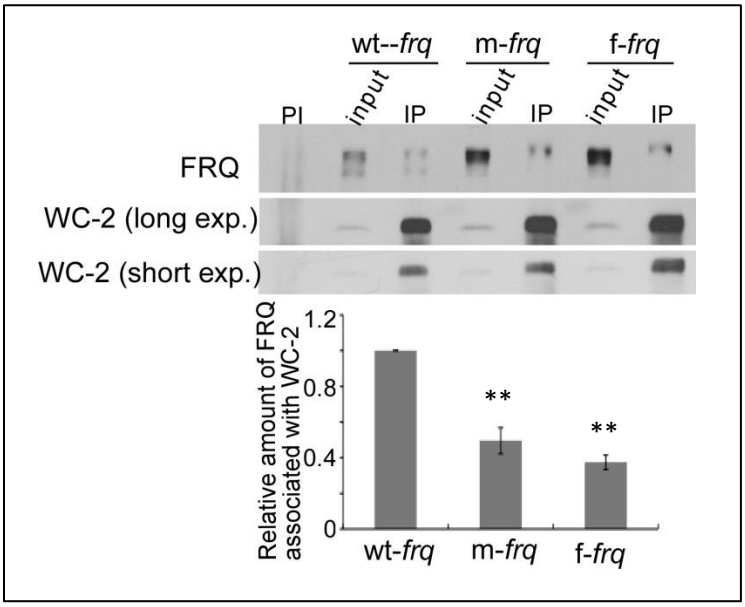


Figure 2-5 FRQ activities in the circadian negative feedback loop are impaired in the *frq* codon-optimized strains.
(a) Immunoprecipitation assay showing that FRQ has a decreased ability to interact with WC-2 in the codon-optimized strains. Densitometric analyses of results from four independent experiments are shown below. PI: pre-immune serum; IP: immunoprecipitation. Error bars indicate standard error (n=9). Asterisks indicate the p value <0.01.

FRQ also acts in a positive feedback loop by promoting WC expression (Cheng et al., 2001b; Lee et al., 2000). Constructs qa-m-*frq* and qa-f-*frq*, in which either m-*frq* or f-*frq* is under the control of a quinic acid (QA)-inducible promoter, were introduced into the *frq*-null strain. As expected, FRQ levels were higher in the qa-m-*frq* and qa-f-*frq* strains than the control qa-*frq* strain at a given QA concentration (Figures 2-6a&b). Induction of FRQ resulted in increased levels of WCs, demonstrating the role of FRQ in the positive feedback loop. At QA concentrations higher than $1x10^{-4}$ M, however, the qa-m-*frq* and qa-f-*frq* strains have lower levels of WC-1 and WC-2 than the qa-*frq* strain despite having higher FRQ levels, indicating that FRQ function in the positive feedback loop is also impaired in the codon-optimized strains.

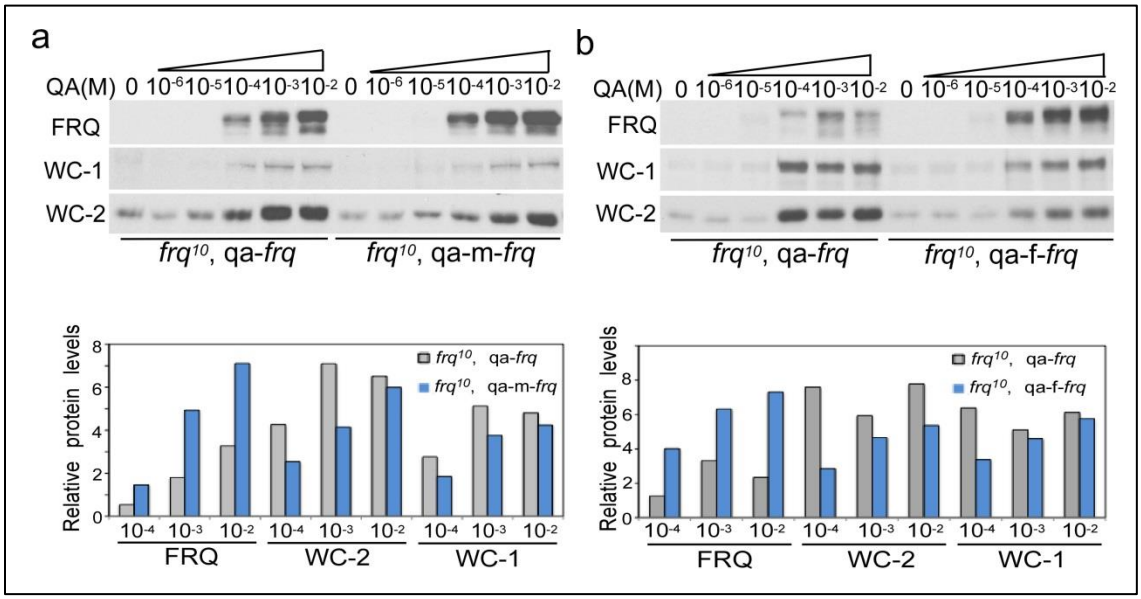


Figure 2-6 FRQ activities in the circadian positive feeedback loop are impaired in the *frq* codon-optimized strains.
(a and b) Western blots showing the levels of WC and FRQ in the indicated strains at different concentrations of QA in LL.

2.3.3 FRQ proteins from codon optimized strains are structurally different from wild-type FRQ.

The impaired FRQ functions despite higher FRQ levels in codon-optimized strains suggest that FRQ's structural conformation is altered. FRQ becomes progressively phosphorylated before its degradation (Heintzen and Liu, 2007). Figure 2-7a shows that, in both m-*frq* and f-*frq* strains, FRQ was less stable than in the wild-type strain after the addition of cycloheximide (CHX). The difference in FRQ stability was most pronounced after 3 hr of CHX treatment, a time when FRQ became hyperphosphorylated.

FRQ from the optimized strains was also less stable after protein extraction in vitro after freeze-thaw cycles. Although FRQ levels in freshly isolated samples were significantly higher in optimized strains, they decreased rapidly after freeze-thaw cycles (Figure 2-7b). In contrast, expression of wt-*frq* in a wc-2^{KO} strain to a level that is comparable to that of the optimized *frq* strains did not affect the freeze-thaw sensitivity of FRQ, indicating that the change in FRQ stability in the optimized strains is not due to high FRQ level or its reduced ability to interact with WCs. Furthermore, limited trypsin digestion showed that full-length FRQ in either optimized strain was more sensitive to trypsin digestion than in the control strain (Figure 2-7c).

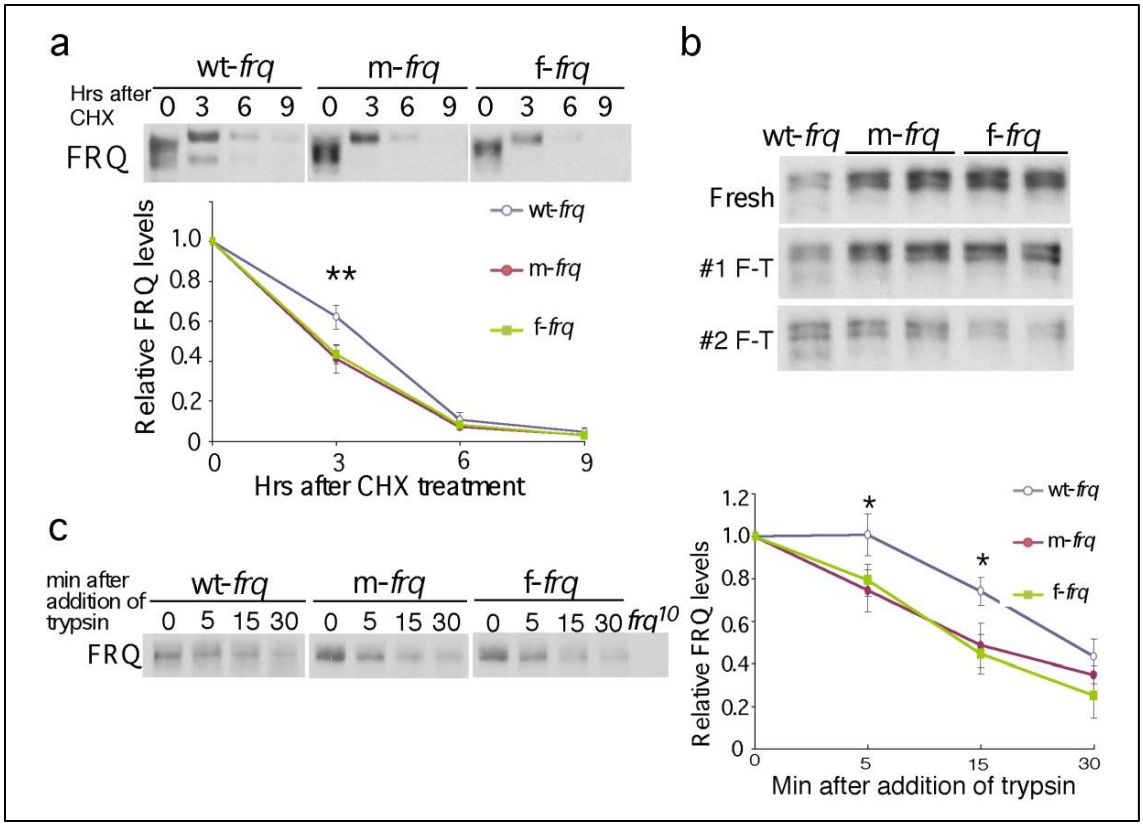


Figure 2-7 FRQ protein in the codon-optimized strains is less stable and more sensitive to trypsin digestion.
(a) Western blots showing FRQ degradation after CHX treatment (10 μg/ml). A longer exposure for the wt-*frq* strain was used so that the FRQ signals at time 0 are comparable in three strains. Densitometric analyses of results of four independent experiments are shown. Error bars, standard deviation. (b & c) Western blots showing sensitivity of FRQ from codon-optimized strains to freeze-thaw cycles (b) and trypsin (1 μg/ml) digestion (c). A longer exposure for the wt-*frq* strain was used in (c). Densitometric analyses of FRQ levels of three independent experiments in are shown. . Two asterisks indicate p value <0.01, and one asterisk indicates p value <0.05.

2.3.4 The structural sensitivity of codon optimized FRQ can be rescued by lower temperature treatment to slow down translation.

We reasoned that the changes in FRQ conformation in the optimized strains are due to an increase of translation rate as a result of codon optimization. Thus, we

examined whether FRQ in the f-*frq* strain can be rescued by decreasing protein translation rate at 18°C because low temperature was known to reduce FRQ expression (Liu et al., 1998) (Figure 2-8a). The low temperature treatment was not successful in restoring the circadian conidiation rhythm of the f-*frq* strain (data not shown), which was not surprising because 18°C is near the low limit of temperature range permissive for rhythmicity (Liu et al., 1997). FRQ in the f-*frq* strain at 18°C, however, is much less sensitive to freeze thaw cycles and to trypsin digestion than at 25°C (Figure 2-8a&b). These results suggest that *frq* codon optimization changes FRQ structure as a consequence of increasing translation rate.

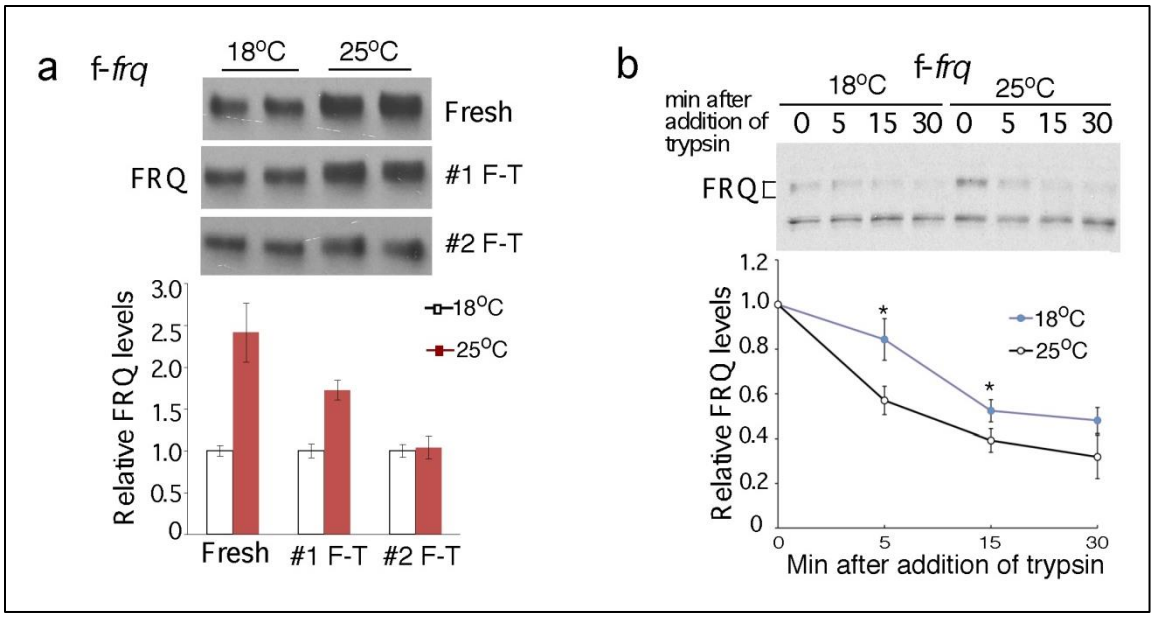


Figure 2-8 Slowing down translation by lower temperature treatment can rescue structural sensitivity of the optimized FRQ.
(d & e) Western blots showing that FRQ from the f-*frq* strain grown at 18°C is more resistant to freeze and thaw cycles (d, n=2) and to trypsin digestion (e, n=4) than that from 25°C. One asterisk indicates p value <0.05.

2.3.5 Codon optimization of a phosphorylated region of FRQ results in impaired FRQ phosphorylation and stabilizes FRQ.

To determine whether the codon-usage effect on FRQ is limited to the N-terminus of FRQ, we created a mid-*frq* strain, in which the middle region (aa 185-530) of *frq* ORF is optimized. This region contains two CK1 interaction domains and most of the phosphorylation sites that are important for FRQ stability and period determination (Baker et al., 2009; Tang et al., 2009). As with m-*frq* and f-*frq* strains, the mid-*frq* strain exhibited arrhythmic conidiation (Figure 2-9a). In the mid-*frq* strain, FRQ levels were high, and FRQ levels and phosphorylation profile were not rhythmic (Figure 2-9b & c). No significant difference in *frq* mRNA was observed (data not show). Importantly, in the mid-*frq* strain, FRQ accumulated in hypophosphorylated forms, was more stable after CHX treatment and was more resistant to trypsin digestion (Figure 2-9b-e). In addition, the CHX-triggered rapid hyperphosphorylation of FRQ in the wt-*frq* strain was abolished in the mid-*frq* strain(Liu et al., 1997). These results strongly indicate that the non-optimal codon usage of *frq* is important for FRQ structural conformation required for proper phosphorylation and stability. The opposite molecular phenotypes of mid-*frq* and the N-terminal optimized strains indicate that changes in FRQ structural conformation in these strains are due to location-specific codon optimization.

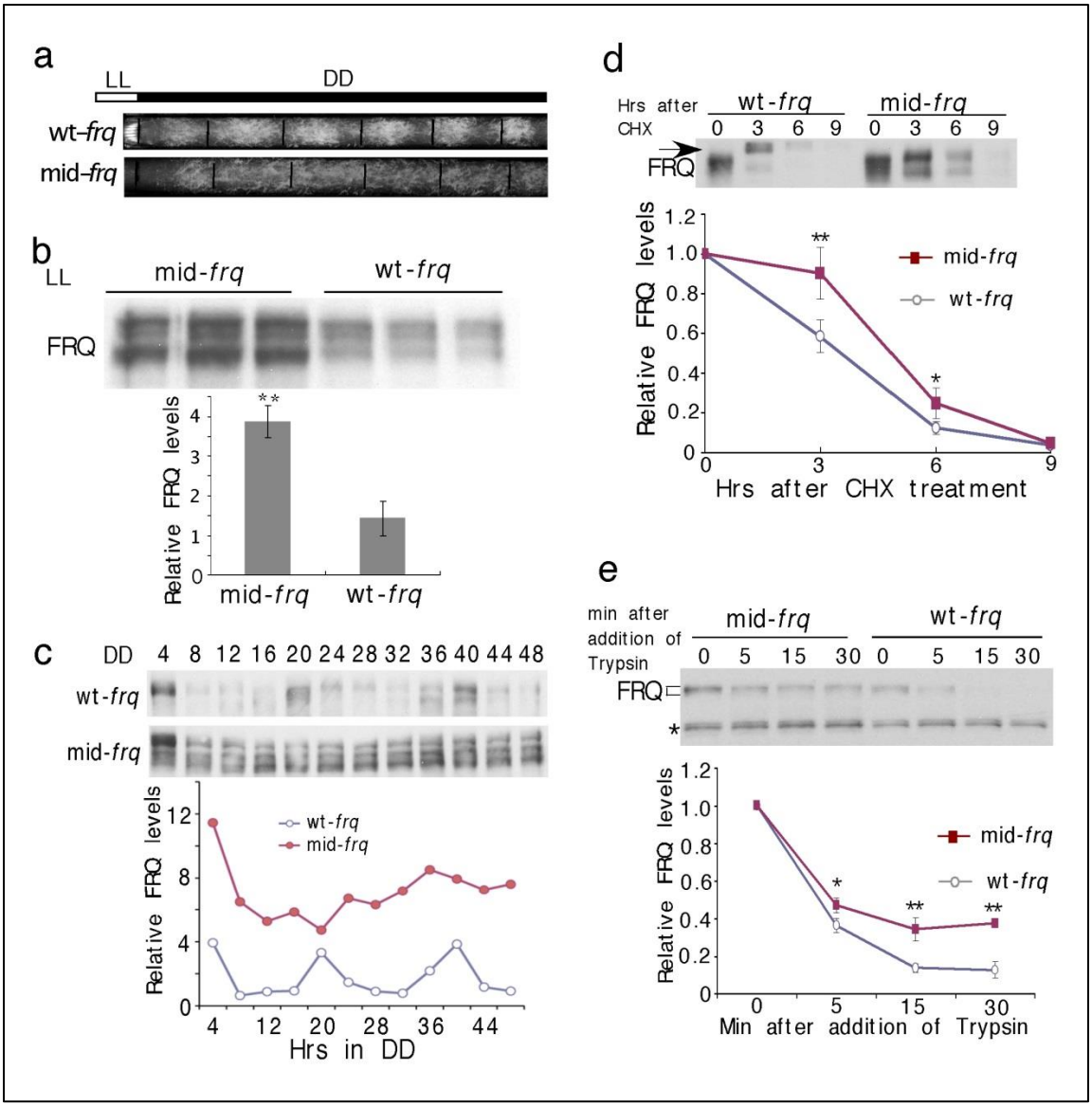


Figure 2-9 Codon optimization of the middle region of FRQ impairs FRQ phosphorylation and stabilizes FRQ.
(a) Race tube analysis showing the conidiation phenoypes of indicated strains in DD. (b) Western blot showing FRQ expression profile in LL. Three independent samples were shown. (c) Western blots showing FRQ expression profile in DD. Densitometric analysis of the result is shown below. (d) Western blot showing the degradation of FRQ after CHX treatment. The arrow indicates the hyperphosphorylated forms of FRQ after the addition of CHX in the wt-*frq* strain. Densitometric analyses from four independent experiments are shown below. (e) Western blot showing reduced sensitivity of FRQ from mid-*frq* to trypsin digestion (2 μg/ml). Densitometric analyses of three independent experiments are shown. Error bars, Standard deviations.

2.4 Discussion

In a companion study, codon optimization of kaiBC clock genes in cyanobacteria results in high Kai protein levels and impaired cell growth at cool temperatures (Xu et al., 2013), suggesting that non-optimal codon usage is a shared adaptive mechanism in both prokaryotes and eukaryotes.

Our study suggests that codon usage regulates not only protein expression level but also its structure and function. Protein folding, which occurs co-translationally, requires protein chaperones and sufficient amount of time (Spencer et al., 2012). Codon optimization results in increased translation rates and thus reduces the time available for folding. Bioinformatics analyses and heterologous protein expression studies previously implicated codon usage as a factor that regulates protein folding (Komar et al., 1999; Siller et al., 2010; Zhou et al., 2009). In addition, a single rare codon in the human MDR1 gene results in altered drug and inhibitor interaction (Kimchi-Sarfaty et al., 2007).

In sharp contrast with the cyanobacterial Kai proteins (Carl Hirschie Johnson, 2011), most of the FRQ protein is predicted to be disordered (Figure 2-10). Interestingly, the known domains of FRQ all have relatively low disorder tendencies and fall in regions where codon usage scores are relatively high, suggesting that a fast translation rate in these well-structured regions does not interfere with protein folding. On the other hand, although the disordered regions may not form stable structures by themselves, they are critical for proper FRQ phosphorylation and stability (Baker et al., 2009; Tang et al., 2009). They may serve as platforms for inter- or intramolecular protein-protein interactions, which may require more time for protein folding than well-structured

domains. Therefore, *frq* non-optimal codon usage should be a mechanism that allows the proper folding of FRQ by reducing translation rate in these predicted disordered regions.

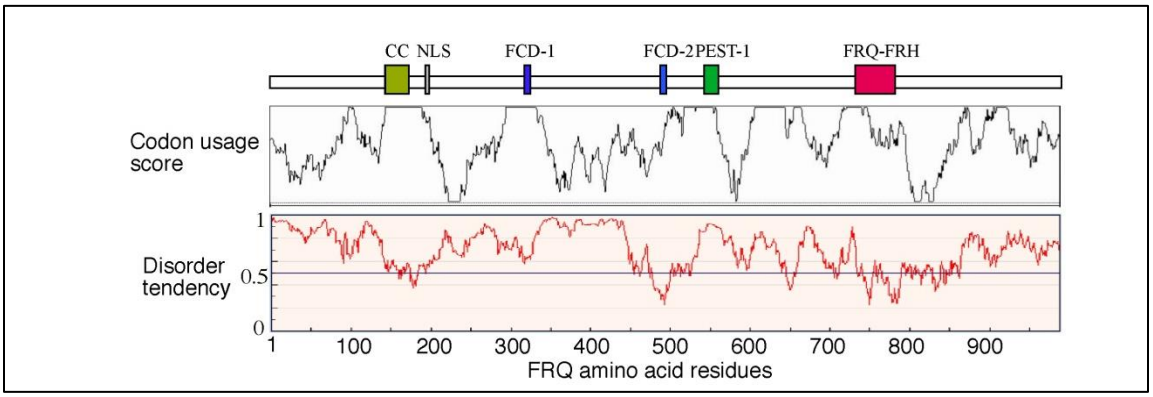


Figure 2-10 Plots shows the location of the previously identified FRQ domains, the frq codon usage score and the predicted disorder tendency of FRQ protein. Codon usage score plot was obtained using Codon Usage 3.5 (window size 35nt, and logarithmic range 3). The codon usage score is highest when codons contain the highest frequency of preferentially used codons in the genome. Protein disorder tendency was determined by IUPred (http://iupred.enzim.hu/). Residues with a score above 0.5 can be regarded as disordered.

CHAPTER THREE

NON-OPTIMAL CODON USAGE IS ASSOCIATED WITH STRUCTURALLY DISORDERED REGIONS AND IS IMPORTANT TO REGULATE THEIR FOLDING

3.1 Introduction

Synonymous codons are not used with equal frequencies, and this is called codon usage bias, which is an essential feature of most genomes (Comeron, 2004; Ikemura, 1985; Sharp et al., 1986). Codon usage bias has been proposed to regulate protein translation rate in both prokaryotes and eukaryotes (Lavner and Kotlar, 2005; Sørensen et al., 1989).

Optimal codons are enriched in highly expressed genes which are thought to be selected for their efficient translation (Akashi, 1994; Drummond and Wilke, 2008). Codon optimization is routinely used when the expression level of an exogenous gene needs to be improved (Gooch et al., 2008). Besides affecting protein expression levels, recent studies suggest a role of codon usage bias in regulating protein folding and function. For example, the presence of a rare codon in human MDR1 gene (C3435T) has been suggested to affect its folding and therefore its substrate specificity (Kimchi-Sarfaty et al., 2007). Studies performed in *E. coli* found that codon bias may regulate translation elongation rates and protein folding efficiency (Siller et al., 2010; Spencer et al., 2012).

We showed that non-optimal codon usage of *Neurospora* clock gene *frequency (frq)* was important to regulate its protein expression level, structure and function (Zhou

et al., 2013). As a core component in *Neurospora* circadian clock, FREQUENCY (FRQ) coordinates diverse molecular and physiological activities with daily changes. *Neurospora* uses a conserved negative feedback loop in eukaryotes to generate its circadian rhythm (Heintzen and Liu, 2007; Liu and Bell-Pedersen, 2006), and FRQ functions as the negative element in this loop with its protein partner FRH (Aronson et al., 1994; Cheng et al., 2005). *frq* transcription is activated by two positive elements WHITE COLLAR (WC)-1 and WC-2 (Cheng et al., 2001b; Crosthwaite et al., 1997). To close the negative feedback loop, FRQ-FRH complex interacts with WC proteins to inhibit *frq* transcription (Denault et al., 2001). FRQ is progressively phosphorylated and then degraded to start a new cycle (He et al., 2006; Schafmeier et al., 2005). Therefore the amount, phosphorylation and stability of FRQ is crucial to determine the period and phase of circadian clock (Garceau et al., 1997; Liu et al., 2000). We found that the domains of FRQ that are predicted to form ordered structures were located at regions with relatively high codon usage score. In contrast, the predicted disordered regions in between preferentially use more non-optimal codon usage. Other groups tried to study the association between codons and protein structures. Claus Wilke's group showed that translationally optimal codons were associated with buried residues and structure sensitive sites (Zhou et al., 2009). Pechmann et al found an enrichment of conserved non-optimal codons in turns using the PDB structure database (Pechmann and Frydman, 2013). However, none of these existing studies have examined the detailed correlation between codon bias and structural disorderness systematically by combining computational analysis and *in vivo* experiments.

Here we compare codon bias profile in *N.crassa* and *S. cerevisiae*, and then determined the correlation between codon bias and protein structures using the *Neurospora* clock gene *frq* as an example. Our biostatistical analyses were used to further examine this correlation in all *Neurospora* genes as well as the other genomes. Our results uncovered a shared negative correlation between codon bias and protein structural disorderness in the five model organisms we studied, and reveal the significance of non-optimal codons in disordered regions in regulating co-translational folding.

3.2 Material and Methods

3.2.1 Strains and growth conditions

Neurospora strains used in this study were 87-3 (*bd, a*; clock wild-type), 303-3 (*bd, frq^{10}, his-3*) (Cha et al., 2011), 301-15 (*bd, his-3*) and various codon optimized strains created in this study. Strain 303-3 (*bd, frq^{10}, his-3*) and 301-15 (*bd, his-3*) were used as the host strain for various *his-3* targeting constructs. The *bd, a* and *frq^{10}, bd,* wt-*frq* strains were used as the wild-type strains in this study.

Growth conditions have been described previously(Aronson et al., 1994). Liquid cultures were grown in minimal medium (1×Vogel's, 2% glucose). Race tube media contains 1×Vogel's, 0.1% glucose, 0.17% arginine, 50ng/ml biotin, and 1.5% agar. For rhythmic experiments, the *Neurospora* cultures were moved from LL to DD at time 0 and were harvested in constant darkness at the indicate time (hours). Calculations of period length were performed as described (Liu et al., 1997; Zhou et al., 2013).

3.2.2 Codon Adaptation Index (CAI), Codon Bias Index (CBI), Effective Number of Codons (ENC), predicted translation rate and protein structural disorderness

CAI and ENC were calculated as previously described(Moriyama, 2001; Sharp and Li, 1987). Codon usage frequency table for *Neurospora* crassa and other species were obtained from http://www.kazusa.or.jp/codon/. For CBI, we used the program codonw in Mobyle Portal website (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::codonw) to calculate (Bennetzen and Hall, 1982). Predicted translation rates were calculated as described by Jose M. Barral's group (Spencer et al., 2012).

IUPRED (http://iupred.enzim.hu) was used with the "long" option to predict and plot protein disorder tendency. By inputting the amino acid sequence of a protein, each amino acid is assigned one disorder tendency score (from 0 to 1, 0 means very ordered and 1 means extremely disordered). Disorderness percentage calculates the percentage of amino acids with IUPRED score higher than 0.5 in one gene, as is suggested by the IUPRED program. PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) was used to predict protein secondary structure. Similar as IUPRED, each amino acid is assigned three scores (H, E and C) representing their tendency to form each of these three structures (alpha helix, beta sheets and coiled coil). Any structures besides helix and sheets will be categorized into coiled coil group by this program.

3.2.3 Correlation between codon usage and protein structural disorderness

ORF sequences of *Neurospora* genes were downloaded from Broad Institute *Neurospora crassa* database (http://www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html). As for *E. coli* (K12), *S. cerevisiae* (S288C), *D. melanogaster* and *C. elegans*, ORF sequences

were downloaded from ecogene database (http://www.ecogene.org), Saccharomyces genome database (http://www.yeastgenome.org), flybase (http://flybase.org) and ensemble (http://useast.ensembl.org/info/data/ftp/index.html), respectively.

For each gene, a simple Pearson correlation coefficient was calculated between CAI vs. IUPRED scores, and CAI vs. PSIPRED scores (H score, E score and C score). Before calculation, the scores (CAI, IUPRED and PSIPRED) were smoothed on a 10aa (amino acid) scale (30aa window size for *E. coli*, *S. cerevisiae* and *D. melanogaster*). For example, the scores for the 1st to the 10th amino acid were averaged and the sores for the 2nd to the 11st amino acid were averaged, etc. In addition, the disorder tendency scores are also smoothed in the same way. For permutated data, codons are shuffled within each gene and a correlation value is calculated using the method above. To delete the disordered regions for calculating correlation between these scores, regions within the smoothed sequences with smoothed disorder tendency scores larger than 0.4 are deleted. The same procedures were carried out for correlation between PSIPRED scores with translation rates.

3.2.4 Codon optimization and codon score plot

frq codon optimization was performed at various ordered and disordered regions. These ordered and disordered regions were identify by FRQ foldable domains (Figure 3-2) and secondary structure predicted by Jpred3 (http://www.compbio.dundee.ac.uk/www-jpred/). For other *Neurospora* genes, codon optimization was performed at regions which were quite disordered and had non-optimal codon usage (Figure 3-9).

Codons were optimized based on *Neurospora* crassa codon usage frequency and tRNA copy numbers. Codon usage score was plotted by software Codon Usage 3.5 (window size 35, and logarithmic range 3).

3.2.5 Plasmid constructs and *Neurospora* transformation

pKAJ120 (containing the entire wild-type *frq* gene under its endogenous promoter and a *his-3* targeting sequence) and pMF272-GFP-5Myc-6His (containing the *gfp* gene under *ccg-1* promoter and a *his-3* targeting sequence; A 5-Myc-6His tag sequence was inserted at the C terminus of *gfp*) were used as the parental plasmids to make wild-type and codon optimized constructs (Cheng et al., 2001b). For *frq*, optimized *frq* sequences were subcloned into pKAJ120 to replace the wild-type sequence. For other *Neurospora* genes, either wild-type or optimized sequences were inserted behind *ccg-1* promoter and fused with *gfp* in pMF272-GFP-5Myc-6His. Wild-type sequences were cloned from *Neurospora* cDNA pool and codon optimized sequences were synthesized by Genscript (http://www.genscript.com/). The resulting constructs were transformed into strain 303-3 (*bd, frq*^{10}*, his-3*) or 301-15 (*bd, his-3*)by electroporation and targeted to the *his-3* locus (Bell-Pedersen et al., 1996).

3.2.6 Protein analyses

Protein extraction, Western blot analysis assays were performed as previously described (Cheng et al., 2001a; Garceau et al., 1997; Guo et al., 2009). Equal amounts of total protein (50μg) were loaded in each protein lane of SDS-PAGE (7.5% SDS-PAGE gels containing a ratio of 37.5:1 acrylamide/bisacrylamide), and then the blot were

developed by chemiluminescence (ECL, Amersham). Densitometry of the signal was performed by using ImageJ.

3.2.7 Freeze and thaw assay

Protein extract was diluted to a total protein concentration of 2.5μg/μl. For each strain, 200μl extract was used in this assay. In each freeze and thaw cycle, the extract was snap freezed by liquid nitrogen, and thawed at 37℃ for 15 minutes. A 20μl sample was taken from the extract before first freeze-thaw, after second freeze-thaw and fourth freeze-thaw, respectively. Each 20μl sample was mixed with protein loading buffer and loaded in each protein lane of SDS-PAGE gel. Western blot was performed to examine FRQ levels at each time point and in different strains (Zhou et al., 2013).

3.2.8 Trypsin sensitivity assay

Protein extract was diluted to a total protein concentration of 2.5μg/μl. 100μl extract was treated with trypsin (final concentration adjusted for different proteins, see figure legends) at 25℃ water bath, with gentle shaking. A 20μl sample was taken from the reaction at each time point (0, 5min, 15min and 30min) after addition of trypsin. Each 20μl sample was mixed with protein loading buffer and loaded in each protein lane of SDS-PAGE gel (12.5%). Western blot was performed to examine target protein levels at each time point and in different strains (Zhou et al., 2013). Experiments were performed side by side and the protein samples were transferred to the same membrane for western blot analysis.

3.3 Result

3.3.1 The codon usage of *N. crassa* genome is much more biased than that in *S. cerevisiae*.

To compare the codon bias profiles of *N. crassa* and *S. cerevisiae* genomes, we calculated the codon bias index (CBI, see methods) for every protein coding gene in each genome. CBI quantifies the extent to which a gene uses a subset of optimal codons (Bennetzen and Hall, 1982). Its value ranges from -1 (extreme bias to use non-optimal codons) to +1(extreme bias to use optimal codons). A 0 CBI value indicates a totally random codon choice. By plotting gene number vs CBI (Figure 3-1a), we find that a majority of *N. crassa* genes have positive CBI values, suggesting their preference to use optimal codons (Figure 3-1a, top panel). Even though there is also a genome-wide bias of codon usage for preferred codons in yeast, the gene number vs CBI profile is significantly shifted towards 0. In addition, the genome average CBI (0.24) for *N. crassa* is much higher than that of much higher than that of *S. cerevisiae* (0.11), suggesting that the codon usage of *Neurospora* genome is much more biased than that of *S. cerevisiae* genome.

To further examine this difference, we used another index named effective number of codons (ENC). ENC measures how far the codon usage of a gene departs from random codon usage (Wright, 1990), and ranges from 20 (strongest bias) to 61(no bias) (Moriyama, 2001). The gene number ratio vs ENC plot (Figure 3-1b) shows that a larger portion of *Neurospora* genes are enriched at smaller ENC groups (32-44) and more yeast genes falls into larger ENC groups (48-52), indicating that the codon usage of the *N. crassa* genome is more biased.

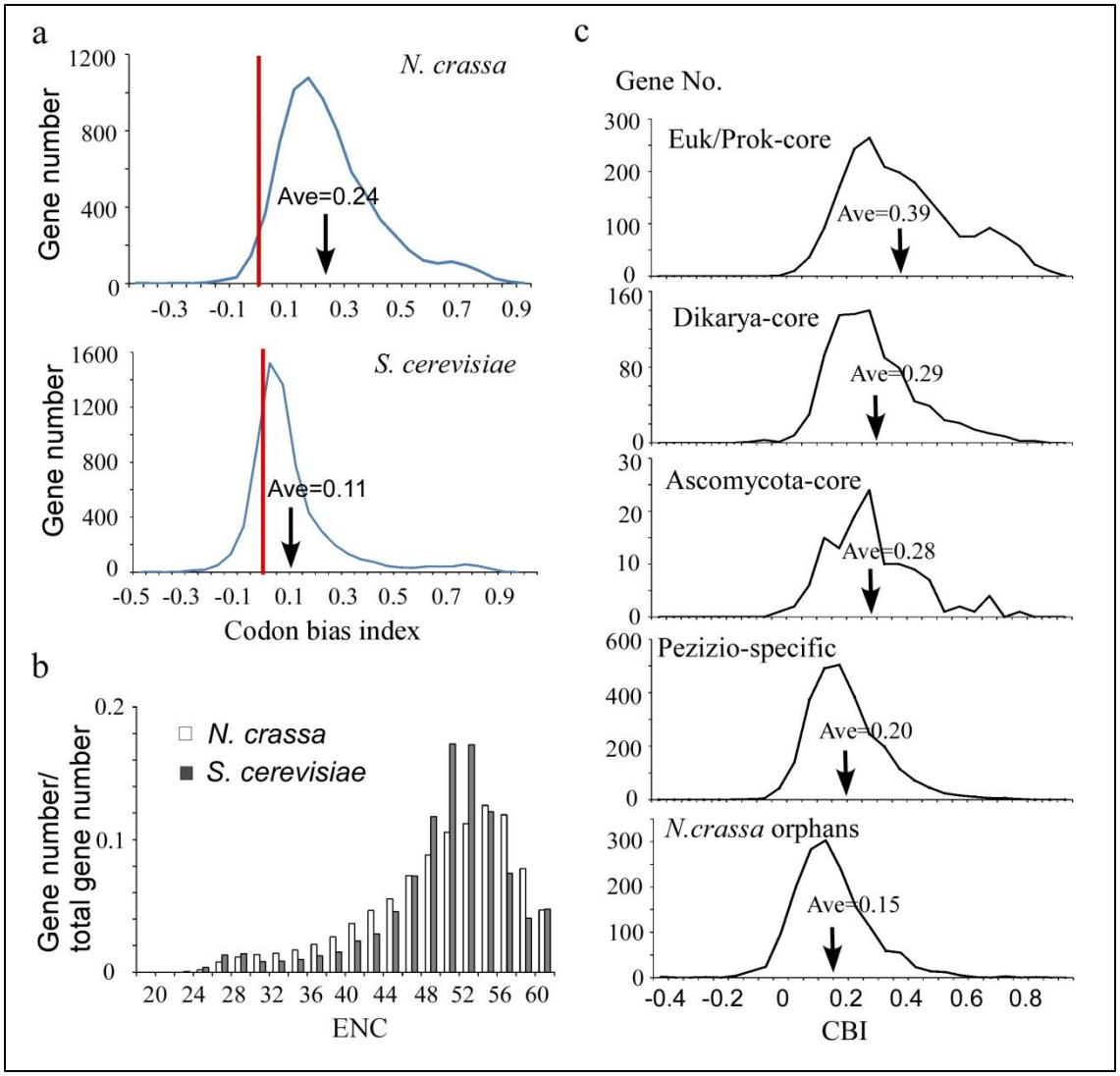


Figure 3-1 The codon usage of *N. crassa* genome is more biased than that in *S. cerevisiae* genome.
(a) Gene number vs codon bias index (CBI) plot in *N. crassa* (top panel) and *S. cerevisiae* (bottom panel) genomes. The red line highlights the position where CBI equals 0. The genome average CBI is 0.24 for *N. crassa*, and 0.11 for *S. cerevisiae*. (b) Effective number of codons (ENC) plot of *N. crassa* and *S. cerevisiae* genomes. The y axis is gene number to total gene number ratio. (c) Gene number vs CBI plot of the five mutually exclusive lineage specificity (LS) groups. The average CBI decreases from conserved groups to lineage specific groups.

Based on sequence conservations, protein coding genes in *Neurospora* could be classified into six mutually exclusive lineage specificity (LS) groups

(Eukaryote/Prokaryote-core, Dikarya-core, Ascomycota-core, Pezizomycotina-specific, *N. crassa*-orphans and Others) based on their phylogenetic distribution(Kasuga et al., 2009). The first five groups (except the Others group) are ranked from highly conserved to lineage specific. In the gene number vs CBI plot of these five groups, we observed a decrease of average CBI (0.29, 0.29, 0.28, 0.20, 0.15, respectively) and a peak shift towards 0 (Figure 3-1c). This correlation suggests that non-optimal codon usage may play an important role in regulating newly evolved genes.

3.3.2 Codon optimizations at predicted disordered regions of FRQ affect FRQ structure and function.

The *Neurospora frequency* (*frq*, NCU02265) gene belongs to the Pezizomycotina-specific group, which is quite lineage specific. In our previous study, we found that codons at N-terminus as well as middle region of *frq* are important to regulate its expression, structure and function (Zhou et al., 2013). Although most of FRQ protein is predicted to be disordered, it's interesting to notice that the known predicted structurally foldable domains all have low disorder tendency and the codon usage score in these regions are relatively high (Figure 3-2) (Zhou et al., 2013). We hypothesized well-structured regions could form easily foldable structures, thus a faster translation rate does not affect its folding. However as for many of the predicted disordered regions of FRQ, they are crucial for FRQ phosphorylation and interaction with other proteins. Since they may not form stable structures easily, a relatively non-optimal codon usage may decrease the translation rate to allow them more time for folding. To test this hypothesis, we made a series of *frq* constructs with different regions being optimized (Figure 3-2). *frq-fcd1*, *frq-fcd2* and *frq-ffd* optimized the two CK1 binding domains and FRH binding domain,

respectively. *frq-d0*, *frq-d1*, *frq-d2* and *frq-d3* optimized the disordered regions in between those foldable domains. By transforming these constructs into a *frq* null strain (frq^{10}), we examined their ability to rescue its arrhythmic conidiation phenotype on race tubes (Figure 3-3). While the wt-*frq* was able to fully rescue the circadian condiation rhythms (period 22.49±0.19h), *frq-d0* failed to rescue the frq^{10} phenotype, resulting in arrhythmic conidiation in DD. On the other hand, *frq-d1* and *frq-d3* strains exhibited long period (23.45±0.40h for *frq-d1* and 23.81±0.50h for *frq-d3*) phenotypes, indicating that the functions of FRQ is partially impaired. In contrast, the other strains are all rhythmic with periods comparable to the wild type. These results suggest that the codon usage at the predicted disordered domains of FRQ play a more important role in regulating FRQ function than that in the well-structured regions.

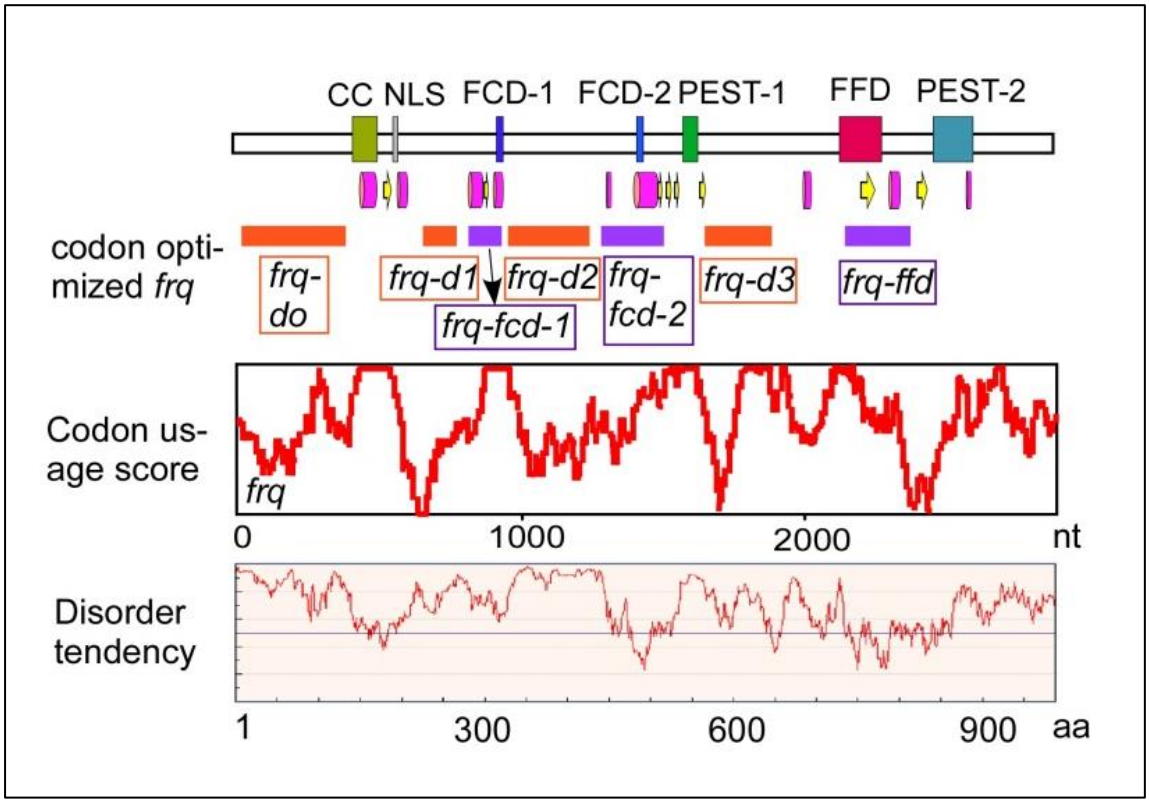


Figure 3-2 The alignment of domains, codon usage score, disorder tendency score and codon optimized constructs on FRQ.
From top to bottom: predicted foldable domains on FRQ (CC, coiled-coil domain; FCD, FRQ-CK1 interaction domain; FFD, FRQ-FRH interaction domain; NLS, nuclear localization signal; PEST, proline (P), glutamic acid (E), serine (S) and threonine (T) domain); Predicted major secondary structures by Jpred3 (pink cylinder, alpha helix; yellow arrow, beta sheet); Various *frq* constructs with codon optimization on structured (purple) and disordered (orange) regions (the length and position of bars represent the optimized position and size); Codon-usage score plot of *frq* obtained using Codon Usage 3.5; Disorder tendency plot of FRQ protein obtained using IUPRED.

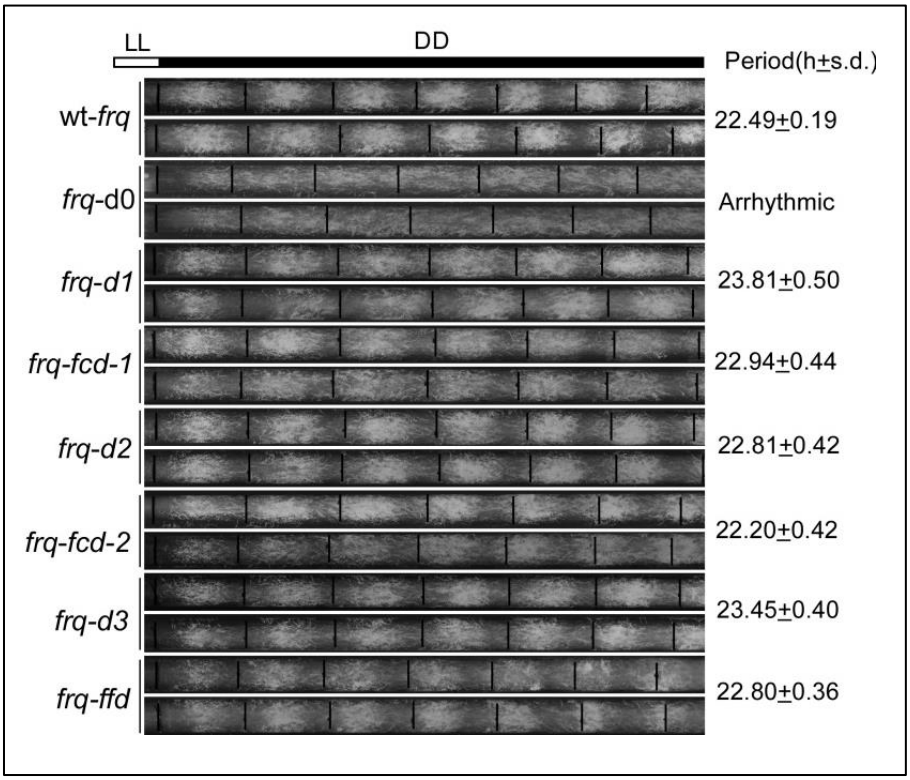


Figure 3-3 Codon optimization on the predicted disordered regions of FRQ affects *Neurospora* circadian rhythm.
Race tube analysis showing the conidiation rhythms and periods of strains with differentially optimized *frq*. Black lines indicate the growth fronts every 24 h. LL, constant light. DD, constant darkness.

To confirm these phenotypes at the molecular level, we examined FRQ protein at different time points in constant darkness after synchronized by light/dark transition (Figure 3-4a, b&c). In the wt-*frq* strain, both FRQ level and phosphorylation profiles exhibited circadian rhythms (Garceau et al., 1997). For the *frq-d0* strain (Figure 3-4a), however, the overall FRQ level was higher and the circadian changes of its amount and phosphorylation profile were both abolished in the second cycle (DD24-DD40). As for *frq-d1* and *frq-d3* strains, the first peaks of FRQ oscillation were delayed when compared with wt-*frq* (Figure 3-4b&c, arrows), which is consistant with their long period phenotypes on race tubes. In addition, the amplitudes of the rhythms of FRQ levels and FRQ phosphorylation profiles were reduced in the second cycle (Figure 3-4b&c, DD32-40h, especially *frq-d3*). Together with the race tube data, these results indicate that the codon usage in predicted disordered regions of FRQ is more important in regulating FRQ function in the clock.

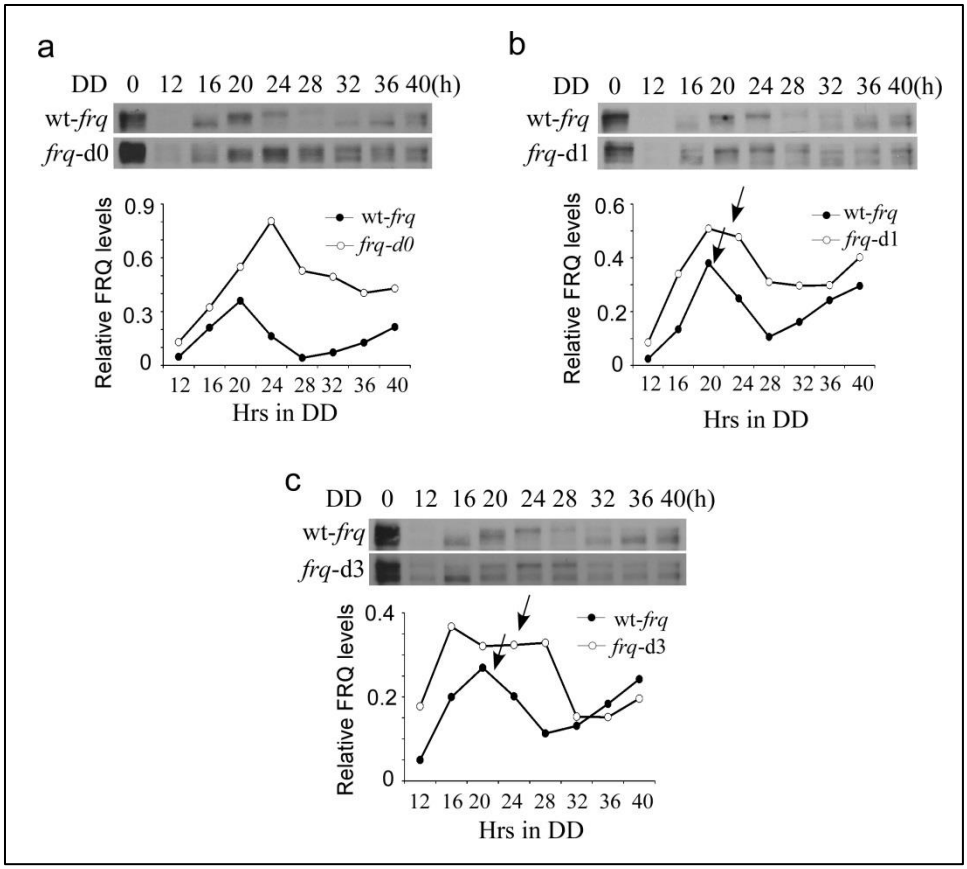


Figure 3-4 Codon optimization on disordered regions of FRQ also affects the molecular rhythm.
Western blots showing the FRQ molecular rhythm in *frq-d0* (a), *frq-d1* (b) and *frq-d3* (c) strains. The lower plots are densitometric analysis of western blot results above using ImageJ. Arrows indicate the positions of first peaks in b and c. FRQ initial amount (DD0) in the wild-type strain is set as 1. The FRQ levels in other strains and other time points are plotted as relative ratio to the wild-type initial level.

To examine the FRQ structural difference in these strains, we performed freeze-thaw and trypsin sensitivity assays. As shown in Figure 3-5a, although FRQ levels from the *frq-d0* strain were much higher than that from wild-type strain in fresh samples, they decreased more rapidly and became comparable with wild-type FRQ after four freeze-thaw cycles. Oppositely, FRQ from the *frq-d3* strain was more resistant to the freeze-

thaw process since it was stable after four freeze-thaw cycles (Figure 3-5b). Consistent with the freeze-thaw assay, limited trypsin digestion showed that FRQ proteins from *frq-d3* strains were more resistant to trypsin digestion (Figure 3-6b) compared with the wt-FRQ, while FRQ from *frq-d0* strain was less stable after trypsin treatment (Figure 3-6a). FRQ from the *frq-d1* strain was also more resistant to limited trypsin digestion (Figure 3-6b), although its stability during freeze-thaw cycles did not differ too much from wt-FRQ (Figure 3-5b). These results suggest that non-optimal codons in disordered regions of FRQ regulate its proper structure at a location specific manner. Codon optimization of different regions can have different effects on protein folding, suggesting that *frq* codon usage adapts to FRQ protein structure.

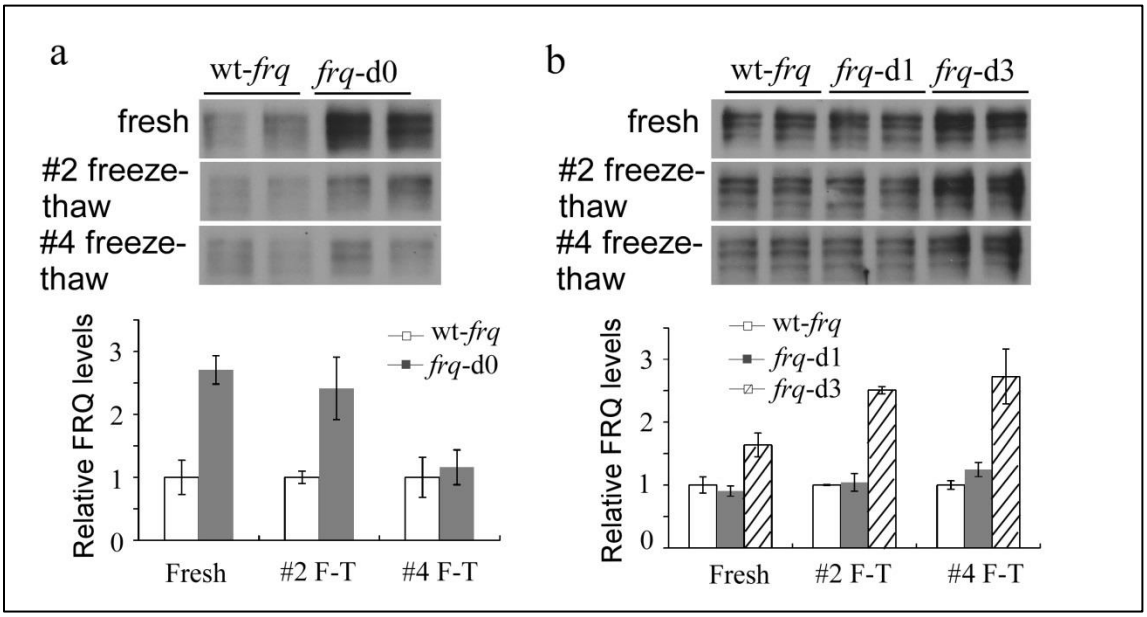


Figure 3-5 Codon optimization on the predicted disordered regions of FRQ changes the FRQ sensitivity to freeze-thaw cycles.
Western blots showing the sensitivity of FRQ from wild-type and codon optimized strains (*frq-d0*(a), *frq-d1* and *frq-d3*(b)) to freeze-thaw (F-T) cycles. Bottom, densitometric analysis of the western blot results above (n=2). The level of wild-type FRQ is set as 1. The relative levels of FRQ in other strains to wild-type are plotted.

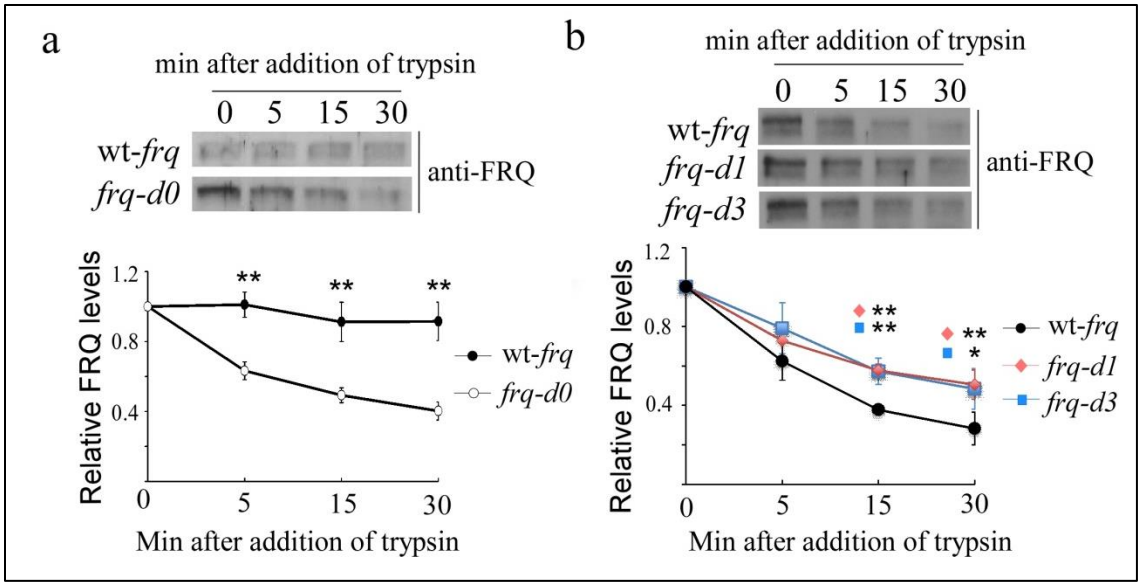


Figure 3-6 Codon optimization on disordered regions of FRQ alters the protein sensitivity to limited trypsin digestion.
Top, western blots showing the sensitivity of FRQ from wild-type and codon optimized strains (*frq-d0*(a), *frq-d1* and *frq-d3*(b)) to trypsin digestion. Trypsin concentration is 0.28μg/ml for *frq-d1*& *frq-d3* and 0.1μg/ml for *frq-d0*. Bottom, densitometric analysis of western blot results above (C, n=9; D, n=4 for *frq-d1* and n=3 for *frq-d3*). *P<0.05, **P<0.01. Error bars denote ±s.d. in A,B,D and ±s.e.m. in C. In each strain, the FRQ initial amount (time point 0) is set as 1. The FRQ levels at other time points are plotted as relative ratio to the initial level.

3.3.3 A genome-wide correlation between gene codon scores and protein structural disorderness in *N. crassa*.

The important role of non-optimal codons at FRQ disordered regions prompted us to examine whether the codon-protein structure relationship we observed in FRQ exists in other *Neurospora* proteins. Codon adaptation index (CAI) is an index to measure codon bias (Sharp and Li, 1987). By assigning a score between 0 and 1 to each codon, the CAI score of a gene takes the geometric mean of the scores for all codons, and a higher CAI score indicates a preference to use frequent codons. We used IUPRED to predict protein

disorder tendency for all predicted *Neurospora* proteins by assigning each amino acid with a disorder tendency score (between 0 and 1, 0 means extremely ordered while 1 means extremely disordered). When we calculated the correlation between CAI and IUPRED scores for every *Neurospora* gene (see methods), we found that there is a strong bias for *Neurospora* proteins to have a negative correlation between CAI and protein disorder tendency scores (Figure 3-7a, 9906 genes, genome average -0.0823), meaning that the predicted disordered protein preferentially uses non-optimal codons. Such a correlation were significantly different from permutated correlations after codons were shuffled randomly within each gene (P value=2.49E-240).

Furthermore, we also examined the correlation between CBI and protein disorderness for *Neurospora* proteins. Since each gene only has one CBI value, we determined the average CBI values in the five different groups of genes based on their percentage of protein sequences that are predicted to be disordered (Figure 3-7b, see methods). Except for the group of genes with extreme disorderness (most of the genes in this group encodes for hypothetical proteins), the average CBI scores increased significantly as proportion of disordered regions in protein increases, suggesting that the predicted disordered regions preferentially use non-optimal codons.

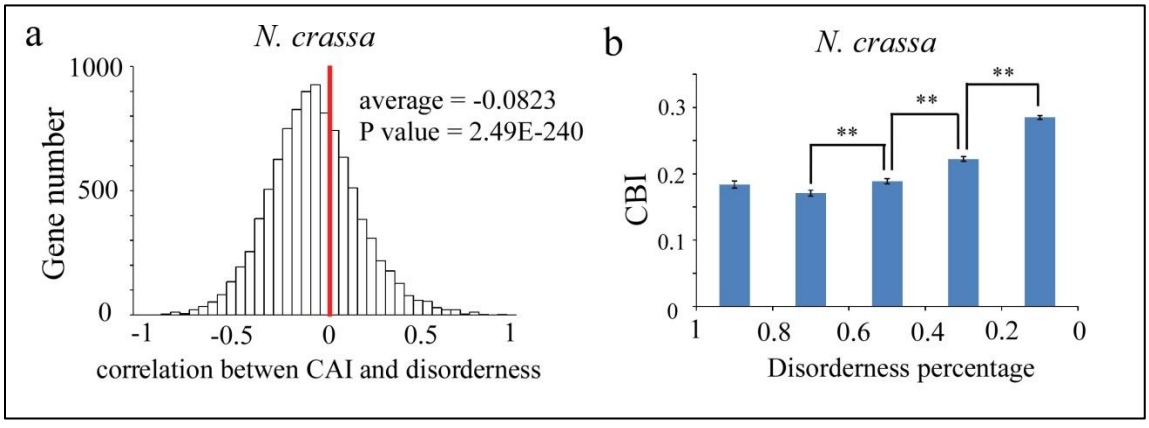


Figure 3-7 Gene codon bias is negatively correlated with the protein structural disorderness in *N. crassa* genome.
(a) The correlation between CAI and disorderness. (b) The disorderness percentage vs CBI plot. *N. crassa* genes were classified into 5 groups based on their disorderness percentage and an average CBI of each group was calculated. For a, genes are classified into 40 bins by correlation scores and plotted against gene numbers. The average correlation and P value are shown on the right. P value was calculated by comparing the actual correlation with permutated correlation. Red lines indicate where correlation equals 0. For b, * means P<0.05 and ** means P<0.01.

To further examine the correlation, we used the PSIPRED program to predict secondary structures of each *Neurospora* protein. For each amino acid, the program will assign three scores: C score, H score and E score, representing its tendency to form each of the three structures: coiled coil, alpha helix and beta sheets, respectively. Structures do not belong to helix and sheets will be categorized into the coiled coil group by this program, including the predicted disordered regions. When we examined the correlation between CAI and each of these three scores, we found a significant negative correlation between CAI and C score (Figure 3-8a, P value=4.4E-149). The negative correlation disappeared after disordered regions were deleted (Figure 3-8b, P value=0.573). This suggests that the negative correlation here is largely caused by non-optimal codons in disordered regions. In addition we also found significant positive correlations between CAI and the other two scores (Figure 3-8c&d, P value=2.47E-79 for H score, 1.30E-09 for E score). The positive correlations suggest that optimal codons are preferentially used in helix and beta sheets regions.

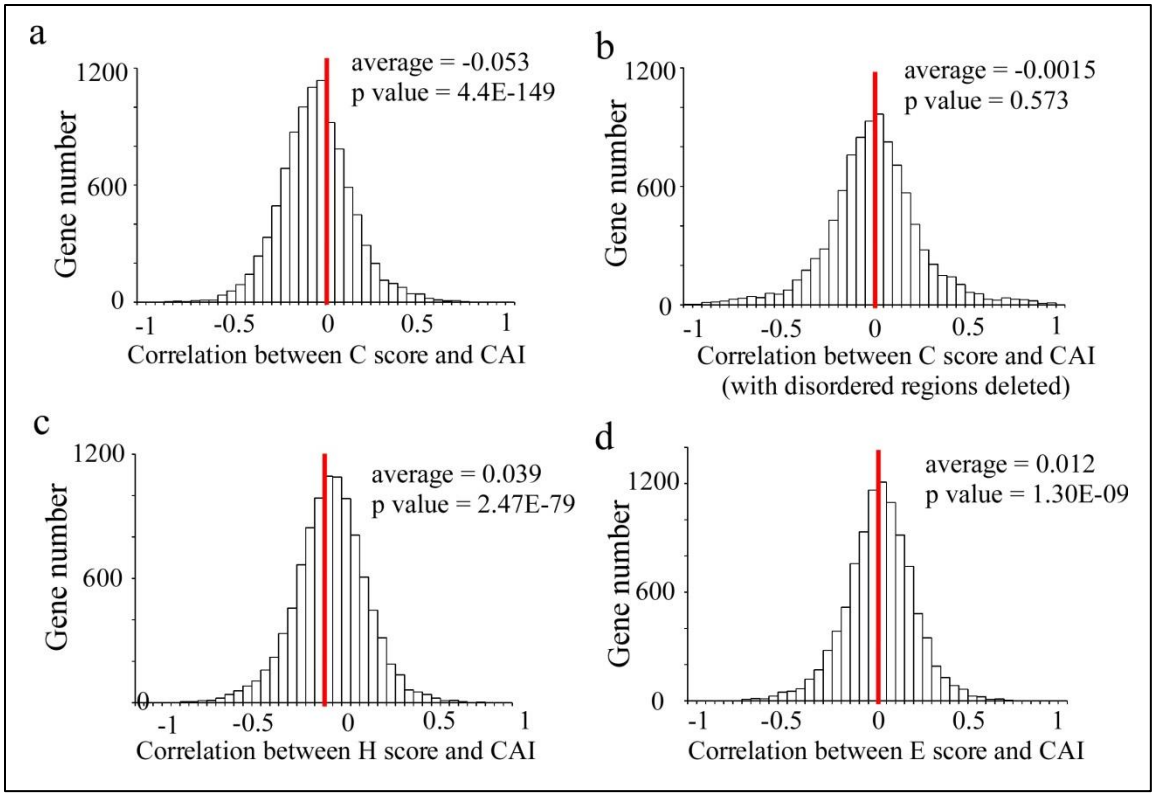


Figure 3-8 Gene codon bias is correlated with the protein secondary structures in *N. crassa* genome.
(a) The correlation between CAI and C score (predicted by PSIPRED, see methods). (b) The correlation between CAI and C score (with disordered regions deleted). (c) The correlation between CAI and H score. (d) The correlation between CAI and E score. Genes are classified into 40 bins by correlation scores and plotted against gene numbers. For each plot, the average correlation and P value are shown on the right. P value was calculated by comparing the actual correlation with 0. Red lines indicate where correlation equals 0.

3.3.4 Codon usage bias affects the expression and folding of other *Neurospora* proteins.

Based on the codon bias profile of *N. crassa* genome and its correlation with protein disorderness, we further examined the role of codon usage bias in regulating protein expression and folding in other genes besides *frq*. Among all the *Neurospora* genes, we picked 7 candidates, each of which had a chunk of disordered regions with

quite non-optimal codon usage. These candidates were *ncu05196, ncu05881, ncu03855, ncu02621, spa1(ncu00627), spa8(03681)* and *spa16(10001)*. SPA proteins are a group of intrinsically disordered proteins which locate at fungal septal pores (Lai et al., 2012). Figure 3-9 shows the CAI scores and disorder tendency scores of these candidates, as well as regions subjected to codon optimization. A myc tag was inserted after both wild-type and codon optimized genes before they were transformed into a wild-type background strain (see methods), and used for western blots to discriminate from endogenous proteins.

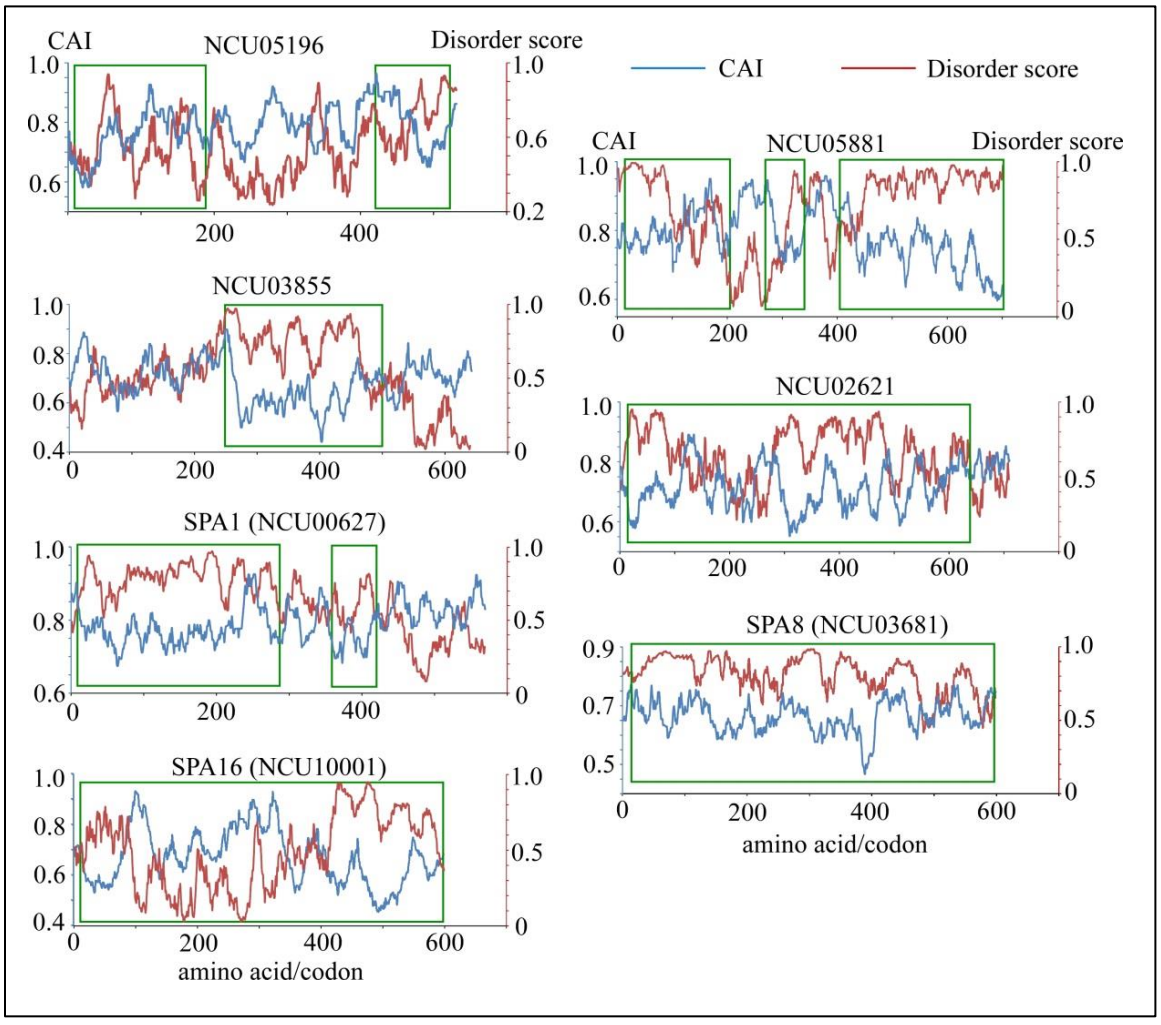


Figure 3-9 CAI and disorder tendency score plot of the 7 *Neurospora* gene candidates.
CAI curves are blue and disorder score curves are red. Green frames indicate the regions subjected to codon optimization.

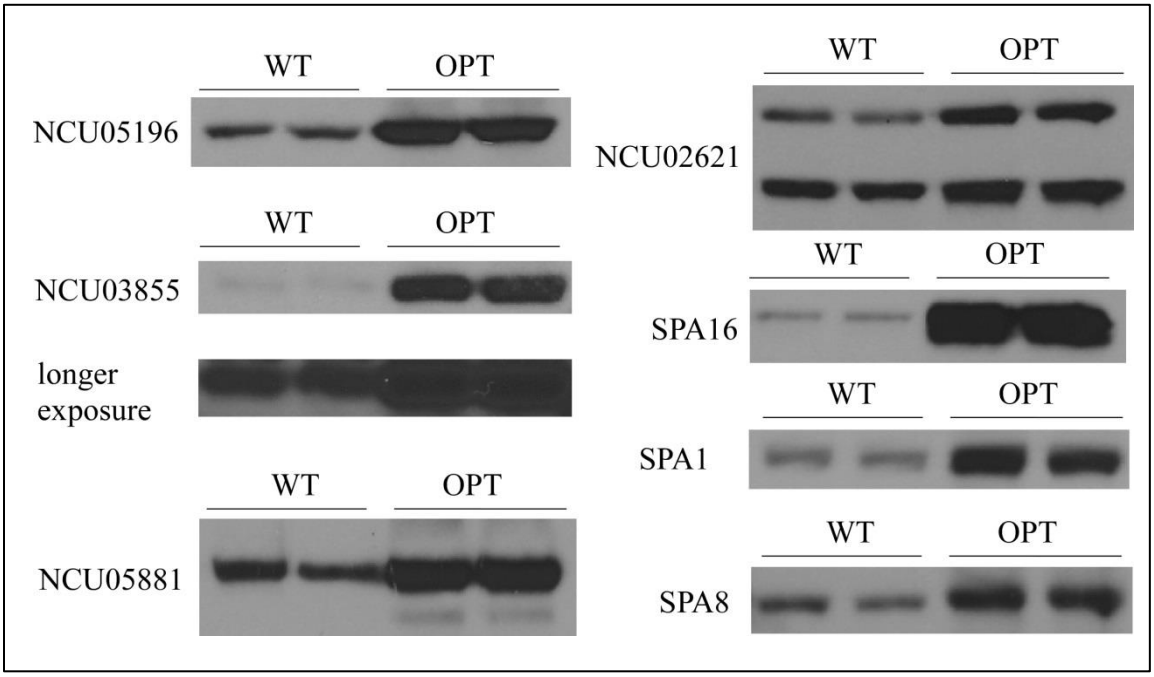


Figure 3-10 Protein expression levels of the 7 *Neurospora* gene candidates we picked.
Protein levels are largely elevated after codon optimization.

After codon optimization, the protein levels in codon optimized strains were largely elevated (Figure 3-10). Four out of seven genes showed altered trypsin sensitivity of their proteins after codon optimization (Figure 3-11 and data not show). NCU05196opt, NCU05881opt and SPA16opt (proteins translated from optimized sequences) became more resistant to limited trypsin digestion than their counterparts NCU05196, NCU05881 and SPA16 (proteins translated from wild-type sequences, Figure 3-11a, b&c). However, being opposite to these three candidates, NCU03855opt was degraded faster than

NCU03855 during trypsin treatment (Figure 3-11d). There results indicate that the effect of codon optimization on different genes could be quite different, depending on the unique structure of the protein and the environment it belongs to.

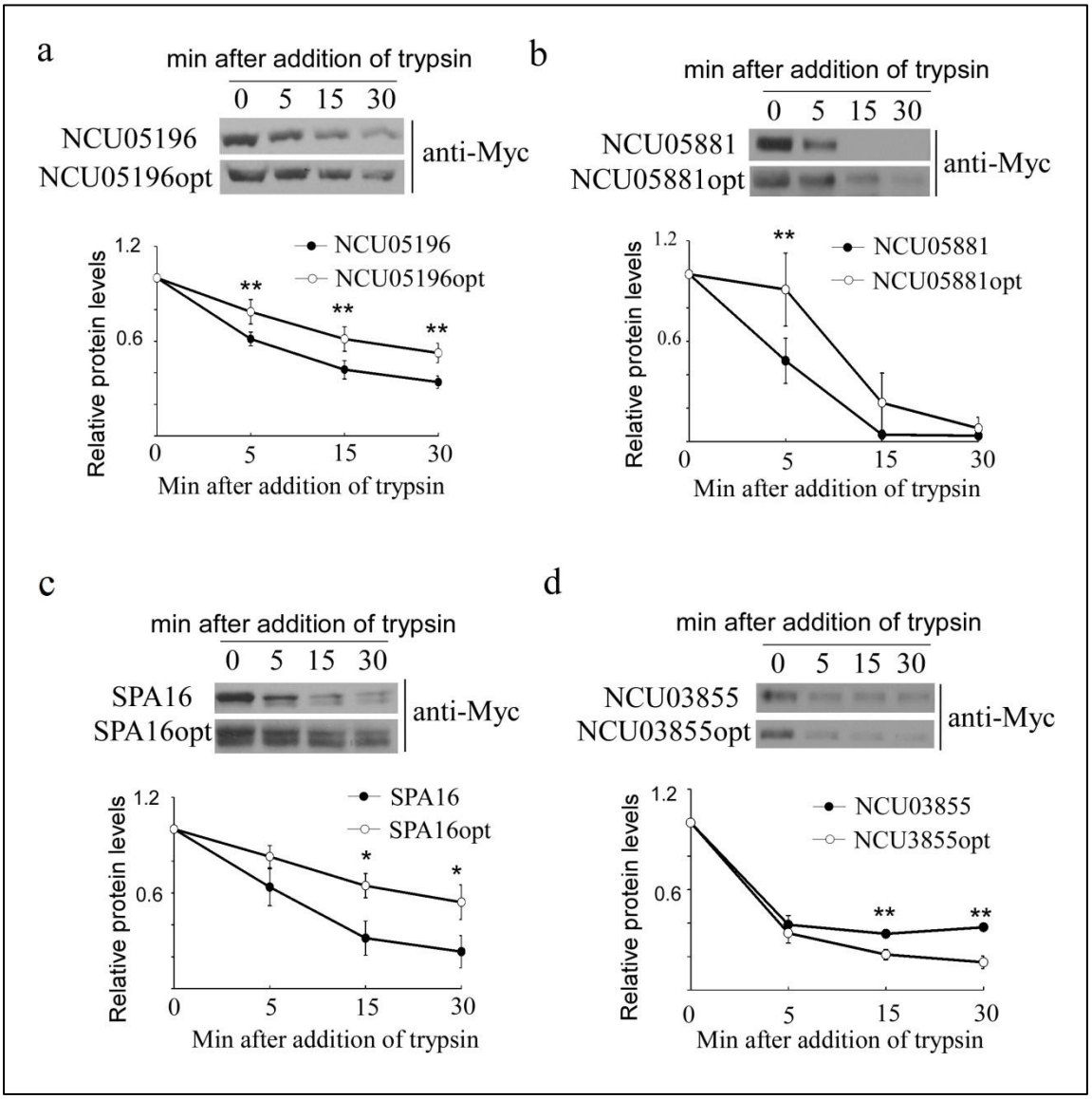


Figure 3-11 Codon optimization of four *Neurospora* gene candidates alters their trypsin sensitivity.
Western blots showing the degradation of indicated proteins from wild-type and codon optimized strains after adding trypsin. Different exposures are used for proteins from wild-type and codon optimized strains. (a) NCU05196 and NCU05196opt, trypsin concentration 2.8μg/ml. (b) NCU05881 and NCU05881opt, trypsin concentration 28μg/ml. (c) SPA16 and SPA16opt, trypsin concentration 0.56μg/ml. (d) NCU03855 and NCU03855opt, trypsin concentration 2.8μg/ml. For each pair, densitometric analysis of the western blots is also shown. For each strain, the initial protein level (time point 0) is set as 1.

3.3.5 The genome-wide correlation between codon usage bias and protein structural disorderness also exists in other prokaryotes and eukaryotes.

Besides *N. crassa*, we also performed biostatistical analysis to look at the correlation between codon bias and protein structural disorderness in other prokaryotic and eukaryotic genomes. Similar to *N. crassa*, we found a negative correlation between CAI and disorder tendency scores in *S. cerevisiae* (6690 genes, genome average -0.0488, p value=1.40E-30), *D. melanogaster* (27244 genes, genome average -0.0483, P value=5.39E-120), *E. coli* (3196 genes, genome average -0.0270, P value=6.71E-05) and *C. elegans* (26095 genes, genome average -0.0608, P value=0) genomes (Figure 3-12).

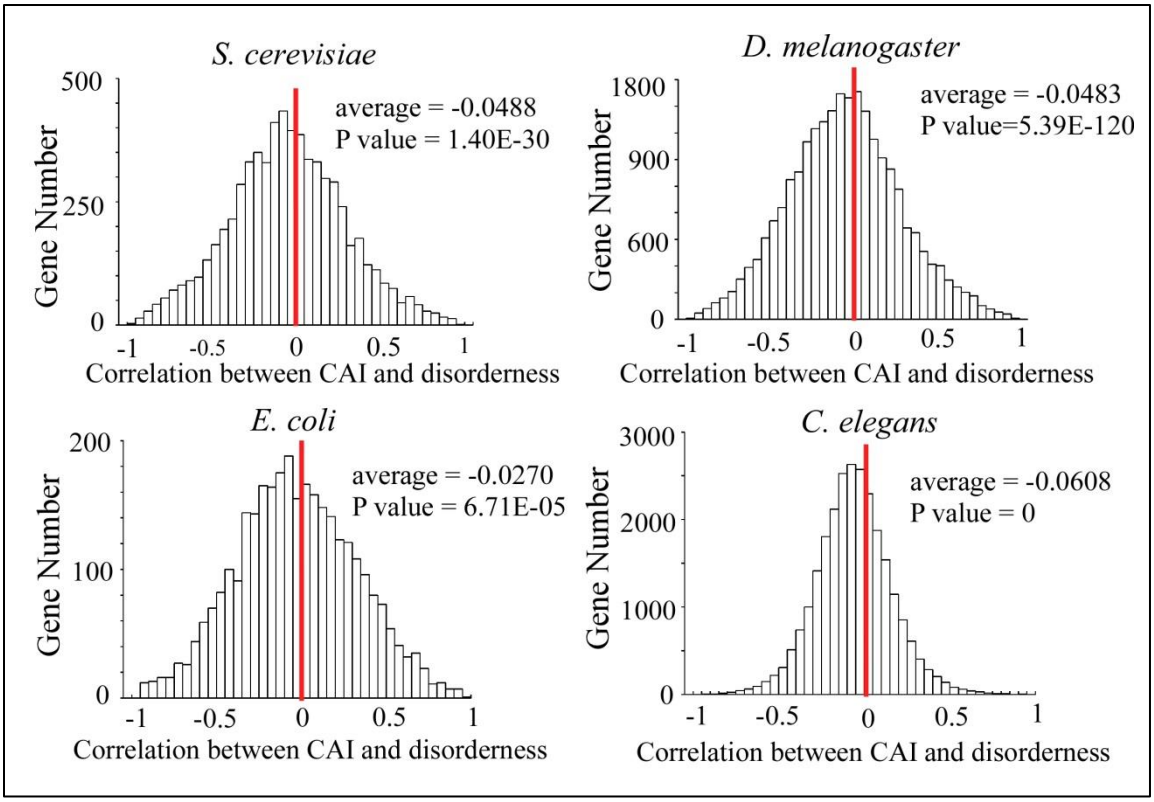


Figure 3-12 Codon bias and protein structural disorderness are also negatively correlated in other four prokaryotic and eukaryotic genomes.
The correlation between CAI and disorder tendency scores in *S. cerevisiae*, *D. melanogaster*, *E. coli* and *C. elegans* genomes. Genes are classified into 40 bins by correlation scores and plotted against gene numbers. For each plot, the average correlation and P value are shown on the right.

Again by using scores for specific secondary structures, CAI score was negatively correlated with C score, and positively correlated with H score in both *S. cerevisiae* and *C. elegans* genomes (Figure 3-13a&b, see plots for genome average and P values). This matches our hypothesis that optimal codons are preferentially used in structured regions while non-optimal codons are favored by disordered regions to slow down the translation rates there. In the other two organisms *D. melanogaster* and *E. coli*, we also found CAI negatively correlated with C scores, but the expected positive correlation between CAI and H scores lost significance. It is possibly due to the noises in score prediction and

calculation. Besides, CAI may not be a perfect index to represent translation rate, which we will talk more in the discussion part.

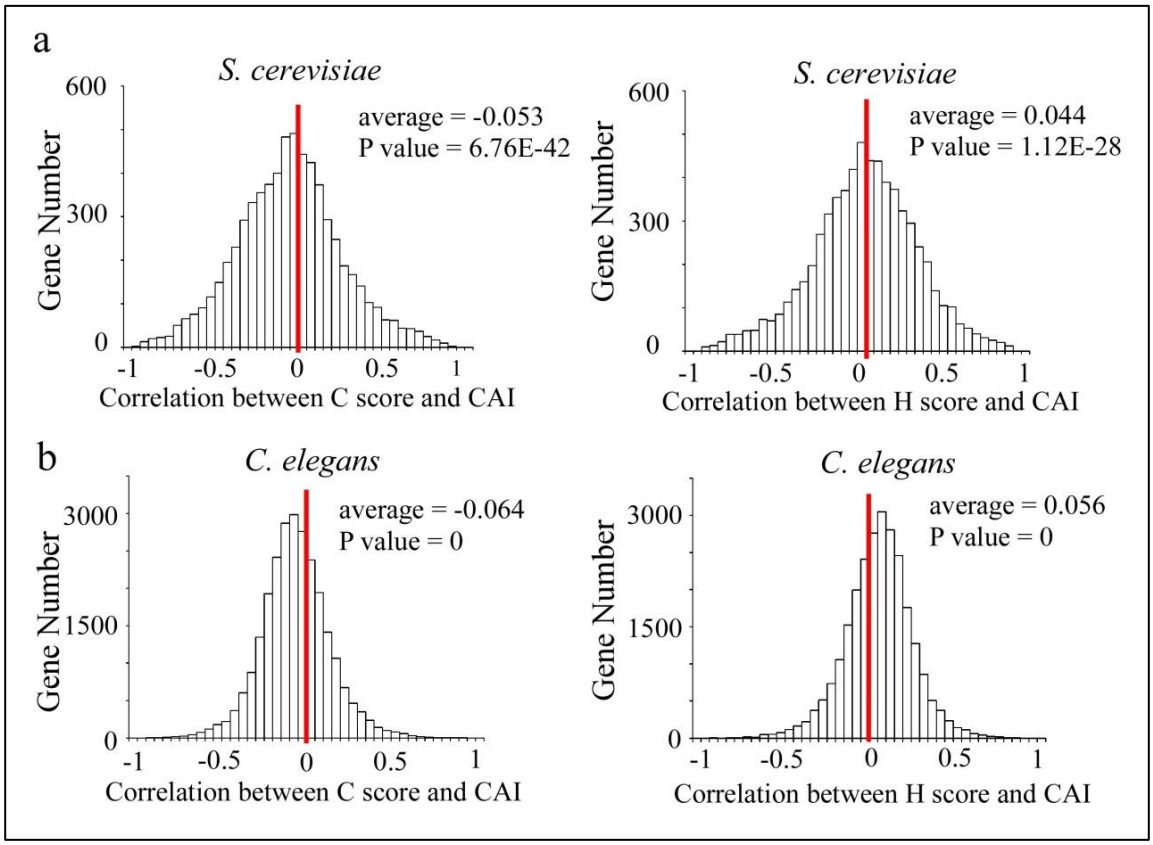


Figure 3-13 Codon bias is negatively correlation with C score and positively correlated with H scores in *S. cerevisiae* and *C. elegans* genomes.
Genes are classified into 40 bins by correlation scores and plotted against gene numbers. For each plot, the average correlation and P value are shown on the right.

3.4 Discussion

We found the enrichment of non-optimal codons at the predicted disordered domains of FRQ, and they play a more important role in regulating FRQ function than codons in the well-structured regions. Besides *frq*, the negative correlation between codon usage bias and protein structural disorderness is genome-wide in *Neurospora*, as well as in other four organisms we examined: *S. cerevisiae, E. coli, D. melanogaster*, and

C. elegans. In addition to CAI and CBI, we also calculated the predicted translation rates with instructions from Jose Barral's group (Spencer et al., 2012). *N. crassa, D. melanogaster,* and *C. elegans* genomes showed significant negative correlations between predicted translation rates and protein structural disorderness. However, the *S. cerevisiae* genome showed a weak positive correlation and *E. coli* genome had no preference towards positive or negative correlations (Figure 3-14). In our hypothesis that slower translation rates at disordered regions are important to help protein folding, one assumption we applied in all analyses is that indices such as CAI, CBI and predicted translation rate explains the real translation rate, at least, to a large extent. This should be true based on various studies on codon bias and protein translation (Lavner and Kotlar, 2005; Sørensen et al., 1989; Spencer et al., 2012). However, not a single index will be perfect in doing this, and each index will share different power in explaining the real translation rate. This explains why the analyses using different indices are not perfectly same, and why we are looking at different indices to study the same question. And the real translation rate is also influenced by other factors such as mRNA secondary structure (Goodman et al., 2013; Shah et al., 2013), anti-SD sequence (Li et al., 2012) and positively charged amino acids(Charneski and Hurst, 2013), which makes the question more complicated.

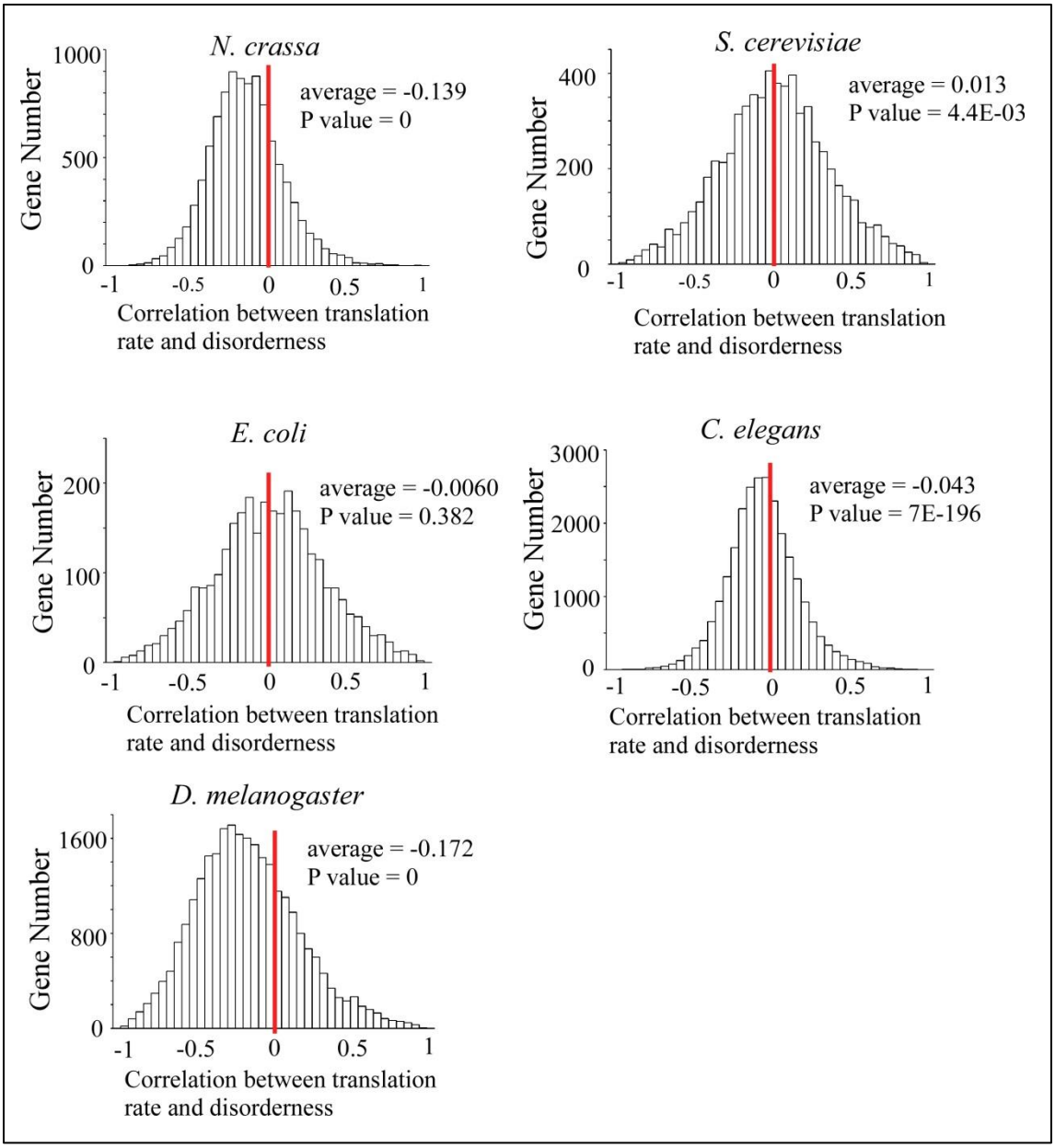


Figure 3-13 The correlation between predicted translation rates and protein structural disorderness in five genomes.
Genes are classified into 40 bins by correlation scores and plotted against gene numbers. For each plot, the average correlation and P value are shown on the right.

Our study is not the first to look at the association between codon usage bias and protein secondary structure. Early in 1990s, both Ding's group and Neidle's group discovered that in mammals synonymous codon families were distributed non-randomly in different protein secondary structure types (Adzhubei et al., 1996; Tao and Dafu, 1998). However, in 2004 Gu et al. denied such a correlation when they analyzed 563 *Homo sapiens* genes and 417 *Escherichia coli* coding genes. More recently, another two groups found that optimal codons are associated with buried residues, structurally sensitive sites (Zhou et al., 2009), and robust protein structures such as alpha helix and beta sheet (Pechmann and Frydman, 2013). Being different from these previous studies, we are the first to manipulate gene codon usage bias in disordered regions of *frq* and other genes *in vivo*, and examine the correlation between codon usage bias and protein structural disorderness genome-widely.

Synonymous mutations are found to be associated with altered protein function. Besides the MDR1 example mentioned in introduction, SNPs in *hper1* and *hper2* are associated with diurnal preferences in human (Carpen et al., 2006; Matsuo et al., 2007). In our study, we showed that codons, especially those in disordered regions, were important for protein folding. This may provide some guidance for solving protein expression issues in both laboratory and industry. If one protein failed to expression after codon optimization, it may be worth restoring codons in important disordered regions. This may also help to explain the mechanisms in some clinical cases in which synonymous SNPs are associated with altered protein activities. Taken together, there is a co-evaluation between codon bias and protein structure and function, which is depicted by a rhythm of translation rate that adapts to protein folding.

CHAPTER FOUR

CONCLUSION AND FUTURE DIRECTIONS

4.1 The role of codon usage in regulating the expression, structure and function of FRQ

Codon optimization has been routinely applied research and industry to increase the expression level of exogenous proteins. However, it's not clear why some genes still prefer a non-optimal codon usage after a long period of nature selection and evolution. Our study here provides a novel insight that non-optimal codon usage serve as a mechanism to regulate co-translational folding by controlling translation speed. Codon optimizations at N-terminus and middle regions of *frq* affect protein structure differently. Since the proper expression level of FRQ is integral for optimal function of the circadian clock, non-optimal codon usage also maintains FRQ expression level within a certain range, which allows the clock function optimally (Figure 4-1).

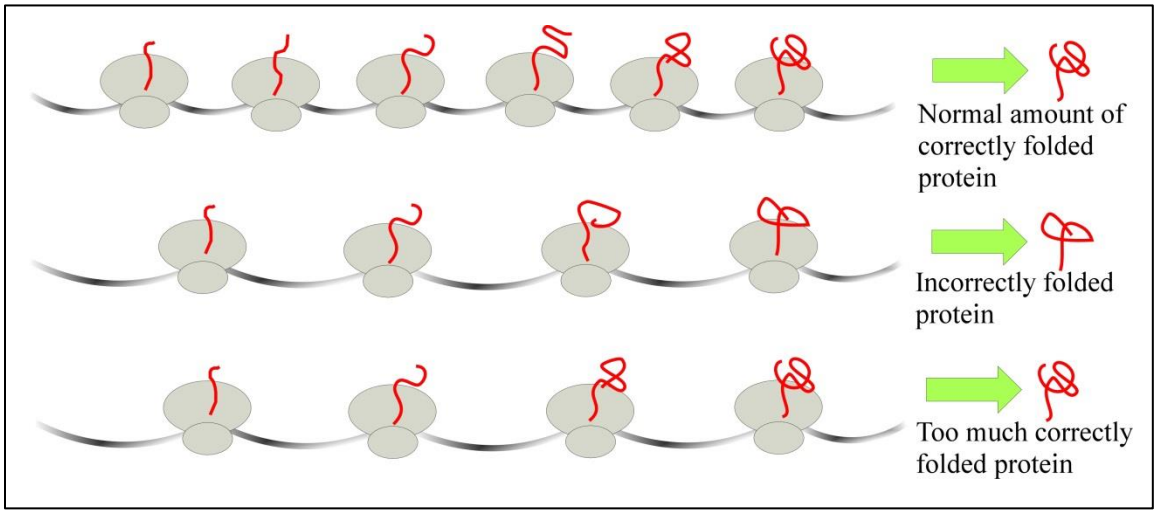


Figure 4-1 A schematic model showing how genetic codons regulate translation. Ribosomes are cellular machines that produce proteins by translating genetic codons on mRNAs to corresponding amino acids. Some synonymous codons are translated faster than others; however, producing more proteins more rapidly is not always advantageous.

There are two other ways to examine the translation rate more directly. One is the cell-free transcription and translation system. By labeling the first amino acid Methionine, translation rate could be reflected by the accumulation of nascent polypeptides (Wang and Sachs, 1997). Thus one of our future directions is to use this system to compare the in vitro translation rates of wild-type and codon optimized *frq* sequences created in Chapter Two, as well as other genes before and after codon optimization. Besides translation rates, the activity of proteins translated from wild-type or codon optimized sequences can also be examined.

Ribosome profiling is useful way to look at ribosome pausing on mRNA chains. Because the density of ribosomes on mRNA is negatively correlated with translation speed there (Qian et al., 2012), this becomes our another future direction. Although some studies reveal no correlation between ribosome density and codon usage bias under nutrient rich conditions (Li et al., 2012), they piled together the density of each specific codon in different genes and distinct positions, which may mask the correlation in a single gene. Thus it's worth examining correlation between codon usage bias and ribosome profiling in each single gene, with a moving window from N-terminus to C-terminus.

4.2 The enrichment and importance of non-optimal codon usage in structurally disordered regions

Based on the hypothesis and proposed model in Chapter Two, non-optimal codon usage in some genes is also a result of nature selection which regulates the expression levels and folding of these proteins. If this model is correct, one prediction could be that non-optimal codon usage is more enriched in structurally disordered regions, because the absence of robust secondary structures makes the peptide folding in these regions more challenging and non-optimal codon usage could decrease the translation rates there to help folding. Therefore codon optimization in these regions will affect protein structure and function more severely. Our statistical analyses performed on five genomes including *N. crassa*, *S. cerevisiae*, *D. melanogaster*, *E. coli*, and *C. elegans* all show the enrichment of non-optimal codons in structurally disordered regions. Furthermore, these enriched non-optimal codons are crucial to determine the proper conformation since codon optimization altered protein structure and function.

In the central dogma of biology, "DNA makes RNA makes protein". Many studies reveal a strong positive correlation between mRNA level and protein amount (Lu et al., 2007). However, this correlation is not perfect. For some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold (Gygi et al., 1999). This suggests that other important factors are also responsible in regulating protein levels, such as codon usage. As a future direction, we are going to examine how important codon usage bias is to regulate protein expression, if compared with other factors such as mRNA levels and protein stability. And besides the co-translational

mechanism, codon usage bias may affect protein expression and activity through the transcription and splicing process (Supek et al., 2014).

BIBLIOGRAPHY

Adzhubei, A.A., Adzhubeib, I.A., Krasheninnikov, I.A., and Neidle, S. (1996). Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. FEBS Letters *399*, 78-82.

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

Aronson, B., Johnson, K., Loros, J.J., and Dunlap, J.C. (1994). Negative feedback defining a circadian clock: autoregulation in the clock gene *frequency*. Science *263*, 1578-1584.

Ashcroft, M., Kubbutat, M.H.G., and Vousden, K.H. (1999). Regulation of p53 Function and Stability by Phosphorylation. Molecular and Cellular Biology *19*, 1751-1758.

Baker, C.L., Kettenbach, A.N., Loros, J.J., Gerber, S.A., and Dunlap, J.C. (2009). Quantitative proteomics reveals a dynamic interactome and phase-specific phosphorylation in the Neurospora circadian clock. Mol Cell *34*, 354-363.

Baker, C.L., Loros, J.J., and Dunlap, J.C. (2012). The circadian clock of Neurospora crassa. FEMS Microbiol Rev *36*, 95-110.

Bell-Pedersen, D., Cassone, V.M., Earnest, D.J., Golden, S.S., Hardin, P.E., Thomas, T.L., and Zoran, M.J. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. Nat Rev Genet *6*, 544-556.

Bell-Pedersen, D., Dunlap, J.C., and Loros, J.J. (1996). Distinct cis-acting elements mediate clock, light, and developmental regulation of the *Neurospora crassa eas* (*ccg-2*) gene. Mol Cell Biol *16*, 513 - 521.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. J Biol Chem *257*, 3026-3031.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Bluthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol *9*.

Brocca, S., Schmidt-Dannert, C., Schmid, R.D., Lotti, M., and Alberghina, L. (1998). Design, total synthesis, and functional overexpression of the Candida rugosa lipl gene coding for a major industrial lipase. Protein Science *7*, 1415-1422.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A Role for Codon Order in Translation Dynamics. Cell *141*, 355-367.

Carl Hirschie Johnson, P.L.S., and Martin Egli (2011). The Cyanobacterial Circadian System: From Biophysics to Bioevolution. Annu Rev Biophys *40*, 143-167.

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

Carpen, J., von Schantz, M., Smits, M., Skene, D., and Archer, S. (2006). A silent polymorphism in the PER1 gene associates with extreme diurnal preference in humans. Journal of Human Genetics *51*, 1122-1125.

Cha, J., Yuan, H., and Liu, Y. (2011). Regulation of the activity and cellular localization of the circadian clock protein FRQ. J Biol Chem *286*, 11469-11478.

Charneski, C.A., and Hurst, L.D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol *11*, e1001508.

Cheng, P., He, Q., He, Q., Wang, L., and Liu, Y. (2005). Regulation of the Neurospora circadian clock by an RNA helicase. Genes Dev *19*, 234-241.

Cheng, P., He, Q., Yang, Y., Wang, L., and Liu, Y. (2003). Functional conservation of light, oxygen, or voltage domains in light sensing. Proc Natl Acad Sci USA *100*, 5938-5943.

Cheng, P., Yang, Y., Heintzen, C., and Liu, Y. (2001a). Coiled-coil domain mediated FRQ-FRQ interaction is essential for its circadian clock function in *Neurospora*. EMBO J *20*, 101-108.

Cheng, P., Yang, Y., and Liu, Y. (2001b). Interlocked feedback loops contribute to the robustness of the *Neurospora* circadian clock. Proc Natl Acad Sci USA *98*, 7408-7413.

Choudhary, S., Lee, H.C., Maiti, M., He, Q., Cheng, P., Liu, Q., and Liu, Y. (2007). A double-stranded-RNA response program important for RNA interference efficiency. Mol Cell Biol *27*, 3995-4005.

Colot, H.V., Loros, J.J., and Dunlap, J.C. (2005). Temperature-modulated Alternative Splicing and Promoter Use in the Circadian Clock Gene frequency. Mol Biol Cell *22*, 5563-5571.

Comeron, J.M. (2004). Selective and Mutational Patterns Associated With Gene Expression in Humans: Influences on Synonymous Composition and Intron Presence. Genetics *167*, 1293-1304.

Crocker, A., and Sehgal, A. (2010). Genetic analysis of sleep. Genes & Development *24*, 1220-1235.

Crosthwaite, S.K., Dunlap, J.C., and Loros, J.J. (1997). *Neurospora wc-1* and *wc-2*: Transcription, photoresponses, and the origins of circadian rhythmicity. Science *276*, 763 - 769.

Crosthwaite, S.K., Loros, J.J., and Dunlap, J.C. (1995). Light-Induced resetting of a circadian clock is mediated by a rapid increase in *frequency* transcript. Cell *81*, 1003 - 1012.

Denault, D.L., Loros, J.J., and Dunlap, J.C. (2001). WC-2 mediates WC-1–FRQ interaction within the PAS protein-linked circadian feedback loop of *Neurospora*. EMBO J *20*, 109-117.

Diernfellner, A., Colot, H.V., Dintsis, O., Loros, J.J., Dunlap, J.C., and Brunner, M. (2007). Long and short isoforms of Neurospora clock protein FRQ support temperature-compensated circadian rhythms. FEBS Lett *581*, 5759-5764.

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

Dunlap, J.C. (1999). Molecular Bases for Circadian Clocks. Cell *96*, 271-290.

Eskin, A. (1979). Identification and physiology of circadian pacemakers. Fed Proc *38*, 2570-2572.

Gan, E.-H., and Quinton, R. (2010). Physiological Significance of the Rhythmic Secretion of Hypothalamic and Pituitary Hormones. In Progress in Brain Research, M. Luciano, ed. (Elsevier), pp. 111-126.

Garceau, N., Liu, Y., Loros, J.J., and Dunlap, J.C. (1997). Alternative initiation of translation and time-specific phosphorylation yield multiple forms of the essential clock protein FREQUENCY. Cell *89*, 469 - 476.

Glozak, M.A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. Gene *363*, 15-23.

Gooch, V.D., Mehra, A., Larrondo, L.F., Fox, J., Touroutoutoudis, M., Loros, J.J., and Dunlap, J.C. (2008). Fully Codon-Optimized luciferase Uncovers Novel Temperature Characteristics of the Neurospora Clock. Eukaryotic Cell *7*, 28-37.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science *342*, 475-479.

Guo, J., Cheng, P., Yuan, H., and Liu, Y. (2009). The exosome regulates circadian gene expression in a posttranscriptional negative feedback loop. Cell *138*, 1236-1246.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. Molecular and Cellular Biology *19*, 1720-1730.

He, Q., Cha, J., He, Q., Lee, H., Yang, Y., and Liu, Y. (2006). CKI and CKII mediate the FREQUENCY-dependent phosphorylation of the WHITE COLLAR complex to close the Neurospora circadian negative feedback loop. Genes & Dev *20*, 2552-2565.

Heintzen, C., and Liu, Y. (2007). The Neurospora crassa circadian clock. Adv Genet *58*, 25-66.

Hrushesky, W.J. (1985). Circadian timing of cancer chemotherapy. Science *228*, 73-75.

Huang, G., Wang, L., and Liu, Y. (2006). Molecular mechanism of suppression of circadian rhythms by a critical stimulus. Embo J *25*, 5349-5357.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution *2*, 13-34.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics *4*, 1265-1272.

Kasuga, T., Mannhaupt, G., and Glass, N.L. (2009). Relationship between Phylogenetic Distribution and Genomic Features in <italic>Neurospora crassa</italic>. PLoS ONE *4*, e5286.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007). A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science *315*, 525-528.

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

Lai, J., Koh, C.H., Tjota, M., Pieuchot, L., Raman, V., Chandrababu, K.B., Yang, D., Wong, L., and Jedd, G. (2012). Intrinsically disordered proteins aggregate at fungal cell-to-cell channels and regulate intercellular connectivity. Proceedings of the National Academy of Sciences *109*, 15781-15786.

Lavner, Y., and Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. Gene *345*, 127-138.

Lee, K., Loros, J.J., and Dunlap, J.C. (2000). Interconnected feedback loops in the *Neurospora* circadian system. Science *289*, 107-110.

Li, G.-W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature *484*, 538-541.

Liu, Y., and Bell-Pedersen, D. (2006). Circadian rhythms in Neurospora crassa and other filamentous fungi. Eukaryot Cell *5*, 1184-1193.

Liu, Y., Garceau, N., Loros, J.J., and Dunlap, J.C. (1997). Thermally regulated translational control mediates an aspect of temperature compensation in the *Neurospora* circadian clock. Cell *89*, 477 - 486.

Liu, Y., Loros, J., and Dunlap, J.C. (2000). Phosphorylation of the Neurospora clock protein FREQUENCY determines its degradation rate and strongly influences the period length of the circadian clock. Proceedings of the National Academy of Sciences *97*, 234-239.

Liu, Y., Merrow, M.M., Loros, J.J., and Dunlap, J.C. (1998). How temperature changes reset a circadian oscillator. Science *281*, 825-829.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotech *25*, 117-124.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol *6*, 386-398.

Matsuo, M., Shiino, Y., Yamada, N., Ozeki, Y., and Okawa, M. (2007). A novel SNP in hPer2 associates with diurnal preference in a healthy population. Sleep & Biological Rhythms *5*, 141-145.

Morgan, L.W., Greene, A.V., and Bell-Pedersen, D. (2003). Circadian and light-induced expression of luciferase in Neurospora crassa. Fungal Genet Biol *38*, 327-332.

Moriyama, E.N. (2001). Codon Usage. In eLS (John Wiley & Sons, Ltd).

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol *20*, 237-243.

Qian, W., Yang, J.-R., Pearson, N.M., Maclean, C., and Zhang, J. (2012). Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet *8*, e1002603.

Radford, A., and Parish, J.H. (1997). The genome and genes of Neurospora crassa. Fungal Genet Biol *21*, 258-266.

Reis, M.d., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research *32*, 5036-5044.

Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends in Biochemical Sciences *21*, 327-335.

Schafmeier, T., Haase, A., Kaldi, K., Scholz, J., Fuchs, M., and Brunner, M. (2005). Transcriptional feedback of neurospora circadian clock gene by phosphorylation-dependent inactivation of its transcription factor. Cell *122*, 235-246.

Segers, G.C., Zhang, X., Deng, F., Sun, Q., and Nuss, D.L. (2007). Evidence that RNA silencing functions as an antiviral defense mechanism in fungi. Proc Natl Acad Sci U S A *104*, 12902-12906.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, Joshua B. (2013). Rate-Limiting Steps in Yeast Protein Translation. Cell *153*, 1589-1601.

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

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

Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C., and Barral, J.M. (2010). Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency. Journal of Molecular Biology *396*, 1310-1318.

Sørensen, M.A., Kurland, C.G., and Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. Journal of Molecular Biology *207*, 365-377.

Spencer, P.S., Siller, E., Anderson, J.F., and Barral, J.M. (2012). Silent Substitutions Predictably Alter Translation Elongation Rates and Protein Folding Efficiencies. Journal of Molecular Biology *422*, 328-335.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324-1335.

Tang, C.T., Li, S., Long, C., Cha, J., Huang, G., Li, L., Chen, S., and Liu, Y. (2009). Setting the pace of the Neurospora circadian clock by multiple independent FRQ phosphorylation events. Proc Natl Acad Sci U S A *106*, 10722-10727.

Tao, X., and Dafu, D. (1998). The relationship between synonymous codon usage and protein structure. FEBS Letters *434*, 93-96.

Thanaraj, T.A., and Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. Protein Science *5*, 1594-1612.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell *141*, 344-354.

Wang, Z., and Sachs, M.S. (1997). Arginine-specific Regulation Mediated by the Neurospora crassa arg-2 Upstream Open Reading Frame in a Homologous, Cell-free in Vitro Translation System. Journal of Biological Chemistry *272*, 255-261.

Wright, F. (1990). The 'effective number of codons' used in a gene. Gene *87*, 23-29.

Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., and Johnson, C.H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature *495*, 116-120.

Young, M.W., and Kay, S.A. (2001). Time zones: a comparative genetics of circadian clocks. Nat Rev Genet *2*, 702-715.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol *16*, 274-280.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J.M., Sachs, M.S., and Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature *495*, 111-115.

Zhou, T., Weems, M., and Wilke, C.O. (2009). Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. Molecular Biology and Evolution *26*, 1571-1580.