

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Insights into the establishment and continuing evolution of the genetic code

Permalink

<https://escholarship.org/uc/item/11p1d7qr>

Author

Johnson, David Bruce

Publication Date

2010

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Insights into the Establishment and Continuing Evolution of the Genetic Code

A dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy

in

Biological Sciences

by

David Bruce Johnson

Committee in charge:

Professor Lei Wang, Chair
Professor Steven Briggs
Professor Simpson Joseph
Professor Joseph Noel
Professor Kit Pogliano

2010

Copyright

David Bruce Johnson, 2010

All rights reserved

The Dissertation of David Bruce Johnson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

DEDICATION

This doctoral dissertation is dedicated to everyone who helped me scientifically and personally over the past 6 years. If the ensuing degree were to be split among all the people that this encompasses it would be torn into hundreds of pieces.

A special dedication is required for my friend and colleague Greg Macias, who sadly is not with us anymore. He had the uncanny ability to make everyone around him laugh, even when times were tough. I will miss you, especially during these stressful times, and I really wish you could be here now to celebrate all of our graduations.

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	vi
List of Tables	ix
Preface.....	x
Acknowledgements.....	xi
Vita.....	xiv
Abstract of the Dissertation	xv
Chapter 1: Introduction	1
Chapter 1.1: The Genetic Code.....	4
Chapter 1.2: The Early Evolution of the Genetic Code	11
Chapter 1.3: The Later Evolution of the Genetic Code	25
Chapter 1.4: Perspectives.....	33
Chapter 2: The Stereochemical Code	34
Chapter 2.1: Imprints of the Genetic Code in the Ribosome.....	36
Chapter 2.2: Uncovering the Stereochemical Code	63
Chapter 3: The Engineering of Variant Genetic Codes in Modern Organisms	82
Chapter 3.1: Engineering the Genetic Code	84
Chapter 3.2: The Evolution of the Ambiguous <i>E. coli</i> strain DJ.2	100
Chapter 4: Modern Expansion of the Genetic Code	120
Chapter 4.1: Unnatural Amino Acids to Investigate Biological Processes	122
Chapter 4.2: Release Factor One is Non-Essential in <i>E. coli</i>	147
Chapter 4.3: Reassigning the Amber Stop Codon from Nonsense to Sense in <i>Escherichia coli</i>	169
Chapter 5: Conclusions, Summary, and Perspectives.....	214
Chapter 5.1: Early Evolution of the Genetic Code	215
Chapter 5.2: Late Evolution of the Genetic Code.....	218
References.....	222

LIST OF FIGURES

Figure 1.1.1:	The canonical or universal genetic code	10
Figure 2.1.1:	For a subset of amino acids, codons or anticodons are enriched near their respective amino acids in ribosome structures.....	58
Figure 2.1.2:	Determination of the radius cutoff used for all analyses	59
Figure 2.1.3:	Inclusion of Cys in the global analysis dramatically increases the deviation of the randomized codes	59
Figure 2.1.4:	Anticodon-amino acid enrichment observed in ribosome structures is correlated with the canonical genetic code	60
Figure 2.1.5:	Ribosomal anitcodon-amino acid enrichment reveals codon reassignment during code evolution	61
Figure 2.1.6:	Codon block expansion analysis of the other four split blocks	61
Figure 2.1.7:	Analysis of the top 1,000 random codes reveals evidence for codon reassignment during code evolution	62
Figure 2.2.1:	RNA moieties bias the incorporation of isoleucine or valine at ambiguous isoleucyl codons throughout the DJ.2 proteome	76
Figure 2.2.2:	Distribution of significant moieties in the DJ.2 proteomic data set is independent of distance.....	77
Figure 2.2.3:	Distribution of significant moieties in the DJ.2 proteomic data set is independent of position and frame.....	78
Figure 2.2.4:	Stereochemical interaction between an RNA moiety and an amino acid favors the incorporation of the amino acid. Numbered steps are referred to in text	79
Figure 2.2.5:	Introduction of identified anticodon-containing RNA moieties biases amino acid incorporation as model predicts	80
Figure 3.1.1:	The SPI method for the misincorporation of unnatural analogs into a target protein using auxotrophic bacterial strains	99
Figure 3.1.2:	A mutation relaxing the substrate specificity of the ValRS in <i>E. coli</i> leads to a high level of misincorporation of the unnatural amino acid α -aminobutyrate	99

Figure 3.2.1:	Schematics representing the knockout constructs used and described in the text.....	116
Figure 3.2.2:	Intracellular isoleucine concentration in DJ.2, DJ.3, and DJ.5 are comparable to the wild type CU.2 strain	117
Figure 3.2.3:	General scheme for the complementation screen devised to isolate the source of isoleucine contamination in the DJ.2 strain. Details are given in the methods.	118
Figure 3.2.4:	The DJ.2 <i>E. coli</i> strain shows no toxicity to high concentrations of valine after continuous growth of for 83 passages	118
Figure 4.1.1:	Different forms of unnatural amino acids.....	143
Figure 4.1.2:	Structures of unnatural amino acids discussed in text	144
Figure 4.1.3:	Backbone mutations generated by α -hydroxy acids. (a) N-terminal mutation of Leu39 to leucic acid in an α -helix of the T4 lysozyme. (b) Substitution of Leu14 with leucic acid in a β sheet of SNase.....	144
Figure 4.1.4:	Substitution of Lys4 with 2-hydroxyethyl-cysteine in the λ -repressor changes the hydrogen-bonding pattern and DNA substrate specificity from C:G to T:A	145
Figure 4.1.5:	Photolysis of 2-nitrobenzyl caged serine (a) and tyrosine (b) restores the wildtype residues. Photolysis of 2-nitrophenyl glycine (c) cleaves the protein backbone.....	145
Figure 4.1.6:	The known potassium channel blocker, tetra-ethyl ammonium, was linked via an azobenzene group to a cysteine and introduced at specific sites of a K ⁺ ion channel.....	146
Figure 4.2.1:	Knockout of RF1 in a variety of <i>E. coli</i> strains. (a) <i>prfA</i> knockout cassette and procedure used in these experiments. (b) A list of the strains used in this subchapter	164
Figure 4.2.2:	Removal of RF1 in JX1.0 enables the encoding of unnatural amino acids at multiple UAG sites.....	165
Figure 4.2.3:	Top-down mass spectrometry of EGFP purified from pBAD without <i>pActF</i> in the growth media	166

Figure 4.2.4:	Mass spectrometry of various EGFP samples grown with <i>pActF</i> suggests that unnatural amino acid incorporation outcompetes tyrosine, glutamine, and tryptophan	167
Figure 4.2.5:	JX1.0 has a growth defect compared to the parental BL21 (DE3) exacerbated by the presence of an unnatural suppression system and unnatural amino acid.....	168
Figure 4.3.1:	RF1 can be knocked out in a genome-reduced <i>E. coli</i> strain after RF2 is fixed	203
Figure 4.3.2:	Removal of RF1 enables encoding natural or unnatural amino acids at multiple UAG sites in JX3.0.....	204
Figure 4.3.3:	Mass spectrometric analyses of EGFP expressed in JX3.0 show that <i>pActF</i> was selectively incorporated at multiple UAG sites with high fidelity	205
Figure 4.3.4:	Ten UAG sites are simultaneously suppressed with natural or unnatural amino acids in JX3.0.....	206
Figure 4.3.5:	Various unnatural amino acids are incorporated at multiple UAG sites into EGFP In JX3.0	207
Figure 4.3.6:	Endogenous genes ending in TAG extend their peptides to different positions in a mRNA context-dependent manner	208
Figure 4.3.7:	The A293E mutation in RF2 is necessary for survival of the JX3.0 strain but not sufficient for rescue of a tsRF1 strain.....	209
Figure 4.3.8:	Human Histone H3 harboring multiple ActKs is efficiently expressed in JX3.0	210
Figure 4.3.9:	Intact protein mass spectrometry of 1-TAG histone H3 coexpressed with the ActK suppression system.....	211

LIST OF TABLES

Table 1.2.1:	Anticodons and codons are significantly overrepresented in a number of RNA-aptamers	24
Table 2.1.1:	Enrichment values calculated for codons and anticodons	55
Table 2.1.2:	Results for correlation analysis of ribosomal anticodon-amino acid enrichment and the canonical code are consistent regardless of which random code generator was used.	56
Table 2.1.3:	Ribosomal anticodon-amino acid enrichment suggests a two-stage evolution of the code.....	57
Table 2.2.1:	Filtering criteria for autovalidation of database search results	75
Table 3.2.1:	The <i>E. coli</i> strains used in this study.....	115
Table 3.2.2:	Laboratory evolution scheme devised for DJ.2. Only passages in which the ratio of isoleucine to valine was altered are shown. Generations were calculated from the daily measurements of the OD ₆₀₀	115
Table 4.2.1:	Single nucleotide polymorphisms between the derived JX1.0 and the parental BL21 (DE3).....	163
Table 4.3.1:	Filtering criteria for autovalidation of database search results	202
Table 4.3.2:	Protein yields of the various EGFP-TAG constructs expressed in JX2.0 and JX3.0	202

PREFACE

This dissertation is divided into distinct chapters. The first chapter will provide an introduction to the broad themes underlying the evolution of the genetic apparatus. Chapter 2 will focus on the early evolution of the genetic code, while Chapters 3 and 4 will focus on the later stages. Each of these chapters themselves is split into distinct subchapters. The beginning of each chapter will include a few paragraphs to put the chapter in context, briefly describe the contents and the subchapters, and assist the reader throughout the manuscript. Tables and figures will be at the end of each subchapter, and all references are in final section of the dissertation.

ACKNOWLEDGEMENTS

I would first like to acknowledge my professor, Lei Wang, for his support over these many years. I have learned an immense amount during my time in his laboratory. I would also like to acknowledge my committee for their support, guidance, and helpful advice over the past years.

I would also like to acknowledge my family and friends, who have always been supportive of me in every respect. Specifically, my fiancé Amanda has been with me through this entire process, and has helped me in every possible way to reach this milestone.

Finally, I would like to acknowledge the fellow members of the Wang lab, specifically Angela Parrish and Jeffrey Takimoto. We all joined the lab within 6 weeks of each other, and were the first graduate students to do so. We stuck together in good times and bad, and your help, encouragement, challenges, and friendship have helped me get to this final stage in my graduate career.

Chapter 2.1 is a reprint of the material as it appears in the Proceedings of the National Academy of Sciences, USA 2010. Johnson DBF and Wang L (2010) Imprints of the genetic code in the Ribosome. Proc Natl Acad Sci USA 107:8298-303. The dissertation author performed all analyses and experimentation and was the author of the paper.

Chapter 2.2 is a reprint of material being prepared for publication. Johnson DBF, Shen Z, Briggs SP and Wang L. Uncovering the Stereochemical Code. Zhouxin Shen performed the mass spectrometry analyses on for both the full-proteome and EGFP-Z

protein. The dissertation author performed all other computational analyses and experimentation. The dissertation author is the author of the paper.

Chapter 4.1 is a reprint in full of the material as it appears in the Wiley Encyclopedia of Chemical Biology, 2007. Johnson DBF, Takimoto JK, Xu J and Wang L (2007). Unnatural amino acids to investigate biological processes. Wiley Encyclopedia of Chemical Biology. doi:10.1002/97804700486.72.webc585. The dissertation author, J.K. Takimoto, and J. Xu contributed equally to this manuscript. The dissertation author specifically contributed the introduction, methods, and future directions sections in this review.

Chapter 4.2 is a part of a manuscript currently being prepared for submission. Johnson DBF, Xu J, Schultz MD, Schmitz RJ, Ecker JR and Wang L Release Factor One is Non-Essential in *Escherichia coli*. The dissertation author is the principal author of the manuscript and performed all experiments except those described below. Jianfeng Xu was the creator of the JX1.0, JX2.0, and JX3.0 strains and designed the RF1 knockout cassette. Full genome sequencing was performed by Matthew Schultz and Robert Schmitz in Joseph Ecker's laboratory at the Salk Institute. All mass spectrometry experiments were performed by Jessica Read and Wolfgang Fischer in Vincent J Coates Foundation mass spectrometry center at the Salk Institute.

Chapter 4.3 is a reprint in full of a manuscript recently submitted for publication in Nature Chemical Biology. Johnson DBF, Xu J, Shen Z, Schultz MD, Schmitz RJ, Ecker JR, Briggs SP and Wang L Reassigning the Amber Stop Codon from Nonsense to Sense in *Escherichia coli*. This paper was truly a collaborative effort. The dissertation author and Jianfeng Xu contributed equally to this work. Jianfeng Xu was responsible

for the creation of the JX2.0 and JX3.0 strains, as well as the initial efforts to knock out RF1 from DH10 β . The dissertation author performed all other analyses and experiments excluding the mass spectrometry and full genome sequence. All EGFP, YfiA, and SufA mass spectrometry was performed by Zhouxin Shen in the laboratory of Steven Briggs at UC-San Diego, and the histone H3 mass spectrometry was performed by Jessica Read and Wolfgang Fischer in Vincent J Coates Foundation mass spectrometry center at the Salk Institute. Full genome sequencing was performed by Matthew Schultz and Robert Schmitz in Joseph Ecker's laboratory at the Salk Institute. The dissertation author was the writer of the manuscript, excluding specific materials sections.

VITA

2003-2005 Undergraduate Research Assistant, University of Arizona.

2005 Bachelor of Science, University of Arizona

2006-2008 Teaching Assistant, Department of Biological Sciences, University of California-San Diego

2010 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Johnson DBF and Wang L (2010) Imprints of the genetic code in the Ribosome. *Proc Natl Acad Sci* 107:8298-303

Johnson DBF, Takimoto JK, Xu J and Wang L (2007) Unnatural acids to investigate biological processes. *Wiley Encyclopedia of Chemical Biology*
doi:10.1002/97804700486.72.webc585.

Tome ME, Johnson DBF, Samulitus B, Dorr RT and Briehl MM (2006). Glucose 6-phosphate dehydrogenase overexpression models glucose deprivation and sensitizes lymphoma cells to apoptosis. *Anti Redox Sign* 8:1315-1327.

Tome ME, Johnson DBF, Rimsza LM, Roberts RA, Grogan TM, Miller TP, Oberley L and Briehl MM (2005). A redox signature score identifies diffuse large B-cell lymphoma patients with a poor prognosis. *Blood* 106:3594-3601.

FIELDS OF STUDY

Major Field: Biology

Studies in *E. coli* genetics
Professor Lei Wang

Studies in unnatural amino acid mutagenesis
Professor Lei Wang

Studies in computational analysis of structural data
Professor Lei Wang

Studies in Cancer biology
Professor Margaret Briehl

ABSTRACT OF THE DISSERTATION

Insights into the Establishment and Continuing Evolution of the Genetic Code

by

David Bruce Johnson

Doctor of Philosophy in Biological Sciences

University of California, San Diego, 2010

Professor Lei Wang, Chair

The precise mechanisms underlying the evolution of the genetic code have eluded researchers for decades. The structure and order of the code suggests a biological, chemical, or physical impetus directing the codon assignments, however these forces remain mysterious. In addition, other forces continue to shape the code, generating

natural and unnatural variants of the canonical genetic code. My work is split between these two distinct avenues of research:

- 1) Statistical evidence from ribosome structures suggests that a certain subset of amino acids were assigned to their codons via a stereochemical interaction with their respective anticodons. In addition, a novel *in vivo* system relying on the ambiguous decoding of isoleucine in *E. coli* suggests that mRNA sequence context outside of the A-site can influence tRNA selection in the ribosome. The underlying mechanism for this influence may be rooted in stereochemical interactions, as anticodons dominate these contextual moieties.
- 2) Variant codes can give valuable and rare insights into the continued evolution of the genetic code. Mimicking these types of variation in the lab will grant a valuable tool for this field, however full reassignment has yet to be achieved. Two *E. coli* strains were generated (JX1.0 and JX3.0) that harbor no genomic copy of release factor one. The UAG codon is now nonsense in these strains, and no longer encodes for a stop signal. In addition, the expression of an unnatural suppression system leads to efficient suppression of the UAG with an unnatural amino acid, and represents the first engineered organism harboring such a reassignment. Suppression of multiple UAG codons in a single protein is efficient and overcomes a major technological hurdle in the field.

These advances will leave a lasting imprint on these fields, and will help to further understand the evolution of the genetic code. In addition, the novel techniques, strains, and findings described herein may help to answer many remaining questions regarding both the early and late events in the evolution of the genetic code.

Chapter 1: Introduction

In modern organisms the genetic code is responsible for attaching function (in the form of proteins) to the units of inheritance (DNA). Specifically, this occurs through a relationship between a triplet or groups of triples of DNA and amino acids. These provide a blueprint for the construction of proteins inside cells, allowing for a direct and inheritable method for maintaining complex systems. In addition, a fixed genetic code allows for a consistent method for introducing variation, as changes to the units of inheritance will lead to inevitable and precise changes to the proteome. This allows for selection to occur in a precise and directional manner, contributing to the adaptations and selection of organisms that spawned the wide variety of life on this planet.

The genetic code is the most dominant aspect of life on this planet; every organism discovered thus far uses the canonical code, or a slight variation thereof. In addition, the proteins and RNA molecules involved in deciphering of the code are also some of the most conserved throughout all domains of life (Koonin 2003). In fact the catalog of proteins and RNA comprising translation is the most convincing evidence that all life is related and has a common ancestor (Wolf 2007). The genetic apparatus is the common thread that binds all life on this planet.

As the genetic code is central to the establishment of life, understanding its establishment and further development is important to evolutionary biology. An in depth understanding of the canonical code would help to address some difficult questions regarding the evolution of all life on this planet. Although much work has been done to elucidate the mechanisms responsible for deciphering the genetic code in modern organisms, understanding of the history and evolution of the genetic apparatus itself is

still murky. This has a long and ever-changing history, predating the elucidation of the codon assignments themselves. Once these assignments were discovered however, the inherent order in the code became clear. The order in the code has profound implications, suggesting a physical, chemical, or biological impetus driving the establishment of the nearly universal genetic code.

Much of the work in the past 6 decades has focused on explaining such mechanisms. This work is heavily reliant on statistics, conjecture, and theory, with little empirical evidence. The lack of empirical evidence is a direct consequence of the universality of the code itself: no intermediate or alternate organisms exist for study, making direct observations and experimentation difficult. Nonetheless, three theories have come to dominate the field, each helping to address this “notoriously difficult problem” (Crick 1976).

Finally, knowledge of the mechanisms underlying the genetic apparatus has led to some major engineering breakthroughs as well. Many groups have hijacked the genetic code of a variety of organisms to introduce ambiguity on certain codons, make wholesale changes to the code in a modern organism, or to introduce unnatural amino acids into proteins *in vivo*. These technologies have an impact in a variety of fields, including synthetic, structural, molecular, and cellular biology. The true power of these techniques may lie elsewhere, directly mimicking intermediate states of code evolution and giving insights into the evolution of the code. This work is thus cyclical: insights into the evolution of the genetic code will further assist in the continued engineering of the code itself.

Before getting into detail about the elucidation and evolution of the code, some nomenclature must be addressed. Most organisms studied thus far contain the canonical or universal genetic code. Some organisms exist that contain variations to the genetic code (see below), so the code should instead be considered semi- or nearly universal. This terminology will be used throughout this paper, and it is important to note that any time the term canonical or universal genetic code is used, it will refer to the code depicted in **Figure 1.1.1**. Any alterations in this code will be described as variant or unnatural codes.

Chapter 1.1 will focus on the elucidation of the codon assignments and the code itself. I will not get into detail about the workings of the code (i.e. the Central Dogma) as it is hashed out in detail in any biology textbook. The apparent order of the code will also be discussed here, as its discovery immediately followed the elucidation of the codon assignments. Finally, this subchapter will briefly introduce the natural variants in the genetic code important to subsequent discussion.

Chapter 1.2 will focus on the early evolution of the genetic code. The RNA world and the foundations of the genetic apparatus in this setting will also be discussed. This section will also introduce the three major theories underlying the establishment of the code: the stereochemical hypothesis, the adaptive code, and the coevolution theory.

Chapter 1.3 discusses the subsequent evolution of the genetic code, focusing on both natural and unnatural variations. Much of what will be described in this chapter will be further described in later chapters, so this section will only provide a brief overview. This section will also briefly discuss the evolutionary implications of such work.

Chapter 1.1: The Genetic Code

1.1.1: Elucidation of the Core Genetic Apparatus

Many of the principals underlying the transfer of genetic material were solved by the early 1950's. DNA was known to be the inheritable genetic material, the structure was solved, and the relationship between DNA and protein translation were starting to be better understood. It was also known that 20 amino acids made up the proteome of organisms, and that DNA/RNA was composed of four nucleotides. The largest mystery remaining was the direct connection between DNA and protein; how could only four nucleotides encode for the 20 amino acids? There must be some "code" for these connections, but at this point, it remained elusive. Crick called 20 the "magic number" (Crick 1957), and many early theories focused directly on achieving this result. It was suggested early on that this code may be coded in triplet form, as a singlet or doublet code gave too few possibilities (4 and 16 respectively) and a quadruplet gave too many (256) (Gamow 1954). Another major breakthrough in our understanding of the genetic code came in 1952, when it was proposed that amino acids were ordered onto a continuous strand of nucleic acids (Dounce 1952). This represented the first theory implying a direct relationship between units of function and units of inheritance. Early theories relied on direct templating between the structure of DNA and the amino acids themselves to explain this dilemma (Gamow 1954; Dounce 1952), and many of these theories generated the magic number of 20 for the nucleotide-amino acid relationships.

One of the major breakthroughs was when Crick broke from this viewpoint and suggested that there was some form of intermediary between the amino acid and nucleic acid (Crick 1957). Admittedly, his adaptor was rudimentary, coupling a complementary

sequence (the anticodon) to the amino acid directly. These adaptors, in combination with the ingenious “commaless” code led to a solution to the magic number problem. Briefly, this code described a series of triplets that corresponded to one of the 20 amino acids. Crick used triplets because triplets provided enough possibilities (64) with the fewest nucleotides used (there are 16 doublets and 256 quadruplets). A triplet would only code for the amino acid in frame, and any out of frame derivative would not function as a coding triplet. For instance, if an amino acid was specified by the triplet GAC, then the two triplets ACG and CGA could not be used anywhere in the coding sequence because they are contained in two GAC codons in succession (GACGAC). The size of such a code with these restrictions was 20, the magic number. Although not correct in all details it was the starting point for the eventual elucidation of the codon assignments themselves. It proposed the idea of frame and triplets, provided an explanation using a non-overlapping code, and finally proposed the adaptor responsible for the recognition of the codon.

Four years later, Crick used a clever methodology that exploited a mutant library of the rIIB cistron of the T4 bacteriophage to further access the possibility of a triplet code (Crick 1961). This library of mutants abolished the ability of the bacteriophage to infect K and B *E. coli* strains. Critically, this library consisted of mutants that most likely represented insertions or deletions resulting in a frameshift. Recombination of the various mutants showed that only a combination of an insertion and a deletion (+ -) or triple insertions and deletions (+ + + and - - -) resulted in restoration of wildtype (or pseudo-wildtype) phage. Other combinations, such as (++), (- -), (+ - -), and (+ + -), did not restore function. These results suggested that the code was a triplet code (or a

multiple of 3), and not a doublet or quadruplet code. In addition, Crick rightly guessed that the wide variety of mutants obtained were indicative of a degenerate code. He theorized that a frameshift in a nondegenerate code would have a higher probability of introducing a nonsense codon in the region between the initial mutation and suppressor mutation. The large amount of mutants they obtained does not support a nondegenerate code and suggest that the code must have some level of degeneracy.

The codon assignments were eventually mapped out by Nirenberg and Matthaei using an *in vitro* translation system from *E. coli*. When a synthetic polyU RNA was added to the system, a phenylalanine polymer was formed, consistent with UUU encoding phenylalanine and representing the first direct codon assignment of any organism (Nirenberg 1961). Repeating these experiments with polyA and polyC established the lysine and proline assignments as well. The determination of the other assignments proved more difficult, as the ribosome would process non-repeating codons at random, making precise determination difficult. A major technological advancement by Leder and Nirenberg helped to overcome this major hurdle (Leder 1964 p420; Leder 1964 p1521; Nirenberg 1964). In these experiments, only single triplets of synthetic RNA were added to the *in vitro* system, allowing for the trapping of the acylated tRNA corresponding to the synthetic codon in the ribosome. A novel filtration method was then used to separate unbound tRNAs from the ribosome, leaving only the correct tRNA (and accompanying labeled amino acid) to be purified and subsequently assigned. By 1966, all 64 codon assignments were determined, representing the complete map of the genetic code of *E. coli*. Through their work, the triplet code was finally realized, spawning modern molecular biology and leading to another surprising finding.

1.1.2: Order in the Genetic Code

The genetic code is not random, containing a surprisingly large amount of order and structure. The most obvious of these is the degeneracy of the code itself (Woese 1965 p71), with amino acids occupying blocks of triplets rather than a scattered assortment. Those amino acids with four codons occupy a full block only differing in the third position ((NN)N), split blocks occupy two codons differing by a transition of the third nucleotide (NNY and NNR), and the few single codon assignments only split an NNR block. This of course led to the concept of wobble in the codon-anticodon interactions, a key aspect of modern translational machinery. Blocks of codons utilizing wobble is an efficient way of assigning 20 amino acids to 64 codons, requiring fewer adaptor molecules (tRNAs) for translation.

If this were the only source of order in the code, it could be attributed mostly to an increase in translational efficiency, however other sources of order are also evident. The hydrophobicity, chemical properties, charge, and even the general composition of the side chains appear ordered in the universal genetic code. Using either the R_f value (Woese 1965 p1546) or polar requirement (Woese 1966 p966), Woese demonstrated that solvent accessibility and hydrophobicity in general is correlated to the genetic code. Specifically the most hydrophobic amino acids share a U in the second codon position, while the most hydrophilic share an A. Amino acids that share a split block were also similar to each other in their hydrophobicity. Structural similarities are also highly correlated to the genetic code: isoleucine, leucine and valine only differ by a single nucleotide change in the first position, the aromatics tyrosine, phenylalanine, and tryptophan all start with a U, the glutamine/glutamate and asparagine/aspartate pairs only differ by a single nucleotide

change, the basic amino acids arginine, lysine, and histidine only differ by a single nucleotide change, as do the acidic amino acids aspartic acid and glutamic acid, the three stop codons are connected, and finally the structurally similar and hydroxyl containing amino acids serine and threonine only differ by a single nucleotide change at the first position (**Figure 1.1.1**). Overall, the apparent order has profound implications on the evolution and establishment of the genetic code itself. Many of the main theories of the evolution of the genetic code (which will be outlined in Chapter 1.2) focus on explaining this high degree of order.

1.1.3: Natural Variations in the Genetic Code

The genetic code was considered at first to be completely universal (Crick 1968), especially after determining that the genetic code of humans and *E. coli* were identical. Crick suggested that this was a natural consequence of the complexity of modern proteomes. Any changes to the genetic code would lead to a large number of deleterious changes to the proteome of any modern organism and be heavily selected against. This idea was turned on its head when alterations in mitochondrial codes were discovered (Barrell 1979), and further negated upon the discovery of nuclear changes in a wide variety of organisms (see Osawa 1992). Although these changes are still rare, they are much more widespread than originally thought, and with modern sequencing technologies becoming more widespread, more natural variations to the genetic code will inevitably be found. For this work, these variations speak to the evolution of the genetic code in general, and will not be reviewed in detail. For a more complete review of these natural variations see Osawa 1992 and Knight 2001 p49, and specific examples will be covered more in later chapters as well.

These natural variations also have profound implications to the evolution of the genetic code. As they are all derivatives of the canonical code, these changes must have been made in the presence of a complex translational machinery. This makes these alterations evolutionarily distinct from the mechanisms underlying the establishment of the code (and therefore the underlying order). These variants may have evolved via a different path and mechanism. As such, the evolution of the genetic code can be split into two distinct realms of thought: one dealing directly with the establishment of the genetic code which will be discussed in Chapter 2, and the other dealing with the mechanism underlying the continued evolution of the code, which will be discussed in Chapters 3 and 4.

Finally, it is important to discuss the impact that these variations have made on synthetic biology. Numerous synthetic organisms have been created that harbor alternate codes, either with full reassignment (Cohen 1957; Rennert 1963; Wong 1983), or with partial and ambiguous reassignments (Döring 1998; Döring 2001; Pezo 2004; Bacher 2005, Bacher 2007 p6494). A more comprehensive review of these experiments can be found Chapter 3.1. In addition, organisms harboring the site-specific incorporation of a wide variety of unnatural amino acids have been achieved using synthetic means (Wang 2001; Wang 2004). These organisms could prove extremely useful as models for how the genetic code continues to evolve, however, the reassignments in these organisms are incomplete and still in competition with endogenous release factors. In Chapter 4.1, I will review these technologies in greater detail, and in Chapters 4.2 and 4.3, I will discuss the creation of synthetic *E. coli* strains that harbor a complete reassignment of a stop codon to an unnatural amino acid.

UUU UUC	Phe	UCU UCC	Ser	UAU UAC	Tyr	UGU UGC	Cys
UUA UUG	Leu	UCA UCG		UAA UAG	Stop	UGA UGG	Stop Trp
CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg
AUU AUC AUA	Ile	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg
AUG	Met						
GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly

Figure 1.1.1: The canonical or universal genetic code. The order in the code is also shown. The basic (dark blue), acidic (red), small aliphatic (yellow), aromatic (green), and hydroxyl containing (light blue) amino acids are all connected by a single nucleotide change. The stop codons (red text) are also connected by a single nucleotide change.

Chapter 1.2: The Early Evolution of the Genetic Code

This chapter will give an introduction to the theories that explain the origins of the genetic code. These theories have gained more traction in recent years, especially with the wealth of *in vitro* data in support of the stereochemical hypothesis. When I started my thesis, I considered a stereochemical origin to be strongest of the hypotheses (in fact Chapter 2 is solely focused on providing *in vivo* evidence for this theory), but I now take a more relaxed viewpoint with the idea that many different pressures may have shaped the origins of the code. In fact, a key point is that the other two main theories do not actually postulate the original force acting on and dictating the assignments of a primordial code. Therefore, this omission could very well imply that these theories are not mutually exclusive (see Knight 1999).

1.2.1: The RNA World

The establishment of the canonical genetic code most likely occurred sometime around 3 to 4 billion years ago in the heart of what is called “The RNA world” (Gilbert 1986). This hypothetical world consists of RNA-based life dominating the biosphere, with protein- and DNA-based life being minimal or serving a cofactor role (Szathmary 1999). This hypothesized world helps to circumvent the “chicken or the egg” problem associated with DNA-based life: proteins are required for efficient synthesis of DNA, and DNA is required for the efficient synthesis of proteins. RNA is an easy target for an intermediate, simply because in this case it is neither the chicken nor the egg, and was proposed to be such an intermediate early (Crick 1968; Orgel 1968). Almost serendipitously, the intermediate between DNA and protein in modern organisms is RNA. This idea however requires RNA to have some level of catalytic function, and no

evidence was found for this until much later when Kruger discovered the self-splicing *Tetrahymena* rRNA introns (Kruger 1982). Later, it was also discovered that peptidyl-transferase activity in the ribosome was a protein-independent catalysis, relying only on rRNA (Nissen 2000). After these discoveries, the idea of an RNA-based world, coined the “RNA world” by Gilbert in 1986, caught on and is now the dominant theory for the evolution of life.

The RNA world must have RNA based enzymes (ribozymes) that serve a wide variety of functions for an organism. Many of these ribozymes have been supplanted by protein machinery in modern organisms, leading synthetic biology in an attempt to replicate these lost ribozymes. Directed evolution experiments have yielded a number of ribozymes with a wealth of catalytic functions including aminoacyl adenylate synthesis (Kumar 2001), self-aminoacylation (Illangasekare 1999 p5470, Illangasekare 1997), trans-RNA amino acylation mimicking tRNA acylation reactions (Illangasekare 1999 p5470, Illangasekare 1995; Murakami 2003), peptidyl-transferase activity (Zhang 1997; Zhang 1998 p539; Lohse 1996; Illangasekare 1999 p1482), and RNA polymerization (Johnston 2001; Lawrence 2005; Lawrence 2003; McGinness 2002; McGinness 2003). These catalytic functions include many of those required for reproduction (of RNA in this case), and also many required for the potential transition to protein-based life (especially peptidyl-transferase).

The ribosome sits in the middle of the RNA-protein transition. It is comprised of 60-80 proteins in close proximity to the catalytic rRNA (Nissen 2000). This was the major impetus for our analysis of the ribosome in Chapter 2.1. It is important to emphasize this here, because it established the timeline for the evolution of the genetic

code. If the RNA world hypothesis is taken to be true, then the genetic code itself must have evolved in the midst of such a time. This is key in understanding and analyzing any of the major theories explaining the establishment of the genetic code. In addition, one has to take the stance that evolution has no foresight; the sheer complexity of modern translation must have roots outside of translation itself. This is why many believe that amino acids may have served more of a cofactor role early on (Szathamary 1999). This subchapter will not get into detail in the numerous steps that may have led to the evolution of the genetic code in the RNA world. Instead, it will focus on the major force driving the actual assignments of the code itself (and therefore the order as well). However, Wolf and Koonin have a phenomenal paper theorizing and outlining the steps that could lead to the establishment of such a complex process (Wolf 2007).

1.2.2: Early Theories of Code Evolution

The first major theory explaining the origins of the genetic code is the simplest. After evidence was shown that the genetic code of humans and *E. coli* were identical, Crick proposed his famous “Frozen accident” hypothesis (Crick 1968). He postulated that all organisms most likely contained the same genetic code, and that any alterations to such a code would be too detrimental in an organism with a complex proteome. Instead, the codon assignments were not something that was necessarily evolved, but rather accidental. The last universal common ancestor (LUCA) possessed the canonical code through chance, granting it an advantage over other organisms and allowing it to eventually dominate the biosphere. This theory now falls short for a number of reasons. The eventual discovery of variant codes in nature dealt a major blow to this ideology. In addition, this theory fell short in explaining the non-randomness of the code itself. The

order in the code is highly suggestive of a specific driving force in the formation of the code, and is independent of chance events. Many of the theories outlining the early evolution of the genetic code focus solely on explaining this order in some physical, chemical, or biological manner. Natural variants are ignored, mainly because they constitute an alternative pathway of genetic code evolution that belongs with the later theories that will be described in Chapter 1.3.

1.2.2.1: The Stereochemical Hypothesis

Implying a direct relationship and interaction between amino acids and nucleotides is perhaps the earliest theory outlining the evolution of the genetic code, predating the elucidation of the code itself (Gamow 1954). It was expounded upon by multiple sources very early (Woese 1966 p723; Dunnill 1966; Pelc 1966) that differed slightly in their hypotheses. The physiochemical theories suggested that it was copartitioning of amino acids and nucleotides in some matrix that eventually led to the codon assignments (Woese 1966 p723; Woese 1966 p966). Chromatography experiments do show some correlation between the hydrophobicity of amino acids and their anticodons (Weber 1978), which was later found to be significant (Jungck 1978). However, these theories suffer from two major drawbacks: the chromatographic copartitioning is not consistent in a variety of prebiotic conditions (Lehmann 1985), and they posit no clear mechanism for translating the copartitioning into genetic code assignments.

In contrast, the stereochemical theories suggest that there is a direct interaction between amino acids and bases that influenced the genetic code assignments (Pelc 1966; Dunnill 1966). Almost every possible interaction has now been theorized, including

codons (Pelc 1966), anticodons (Dunnill 1968; Ralph 1968), reverse codons (Root-Bernstein 1982 p885; Root-Bernstein p895), and a four-nucleotide complex (Shimizu 1982). These theories were quickly mired in problems including that Pelc built his models for codon interactions backwards (Crick 1967). In addition, as the mechanisms of translation became clearer, it was obvious that any of these interactions did not persist in modern translation. However, this idea is not a major drawback, as the absence of such interactions in modern systems does not negate the possibility of their involvement earlier. Modern translation may have shuffled or usurped these interactions for those that are stronger, leaving little to no evidence of their previous existence.

In vitro systems have been exploited to further access the possibilities of codon or anticodon interactions with their respective amino acids. The multitude of these experiments performed mainly by the Yarus group was most likely spurred by a coincidental finding in the structure and function of self-splicing group I introns. Normally the splicing mechanisms recognize and bind to guanosine, however due to similar structures and hydrogen bonding patterns, arginine is also bound (Yarus 1988). Interestingly, the binding specificity is conferred by a conserved arginine codon in close proximity to several other arginine codons (Yarus 1989; Yarus 1991). Natural evidence such as this is rare and anecdotal, and numerous selection pressures over millions of years of evolution have shaped the active site of the self-splicing intron. Instead, work was shifted to using the SELEX method for evolving RNAs with specific functions (Ciesiolka 1996). These evolved RNAs, termed aptamers, have brought a rebirth in this field by allowing the synthesis and analysis of RNAs that specifically bind amino acids. Once these aptamers are formed, precise biochemistry and statistics can be applied to

determine if codons or anticodons confer specificity for binding their respective amino acids.

A wide variety of amino acid-binding RNAs have been derived using the SELEX methodology (extensively reviewed in Yarus, 2005; Yarus 2009), presenting a testable hypothesis relating to a fundamental inference of the stereochemical hypothesis. If codons or anticodons have a specific affinity for their respective amino acids, they should be enriched in the binding sites of these structures. Yarus recently reviewed the wealth of data for testing this hypothesis (see Yarus 2009). Briefly, the binding sites of aptamers for arginine, glutamine, isoleucine, and tyrosine were significantly enriched with codons for these amino acids. In addition, the binding sites of aptamers for arginine, histidine, isoleucine, leucine, phenylalanine, tryptophan, and tyrosine were significantly enriched with anticodons (see **Table 1.2.1** and references therein). Combining these statistical measurements using Fisher's method reveals that codons and anticodons are both highly enriched in these aptamers ($P = 5.3 \times 10^{-45}$ and 2.1×10^{-46} respectively). Yarus suggests that this evidence points directly to a "stereochemical era" during the evolution of the genetic code (Yarus 2009) in which RNA binding sites for amino acids dictated their eventual assignments in the canonical genetic code.

Although this *in vitro* statistical evidence is extremely robust, it does suffer from a variety of predicaments. It is difficult to logically justify evidence for both codons and anticodons playing a role in codon assignments. The proposed evolutionary steps between these interactions and the genetic code are already tenuous (see Wolf 2007), and inclusion of both codons and anticodons necessitates two independent mechanisms. Although two mechanisms are possible (one may have even morphed into the other

throughout the evolution of the code), it is a logically difficult pathway to codon assignments. One key point relies on the results for arginine, isoleucine, and tyrosine. These amino acids are special in that their codons (CGN, AUN, and UAY respectively) contain their anticodons (NCG, NAU, and RUA). This overlap may be skewing the statistics for either codons or anticodons as all three of these amino acids are represented in both groups. For instance, if we remove these amino acids from the codon data set from the Yarus 2009 review (specifically CGC, CGG, AUU, UAU, and UAC), only three significant relationships remain (AGG, CAA, and UUG). If we do the same for the anticodon set, we are left with 6 significant enrichments (AUG, GUG, UAG, AAA, GAA, and CCA). Although this data presents a potential logical dilemma, this may be solely due to such overlaps.

These aptamers are complex large RNA molecules that have no real basis in modern systems, representing another issue with the SELEX results. For one, this suggests that a complex RNA world had to already be present for the establishment of the genetic code. This is not too pressing of an issue, but it does have some implications on the broader theory and the pathways of actual codon assignments. In addition, these molecules are derived *in vitro* and may not represent the actual processes of evolution that have shaped the genetic code. The *in vitro* data is also in conflict with the *in vivo* data for a variety of amino acid binders that do not directly support the stereochemical hypothesis (Yarus 1998; Yarus 2009). It is important to mention that these persistent dilemmas do not negate the robust statistical data; instead they raise numerous unanswered questions that need to be addressed before a “stereochemical era” for genetic code expansion can be widely accepted.

1.2.2.2: Adaptation of the Code

The next two theories postulate that the code has undergone changes throughout evolution and that the order in the code is a consequence of the underlying forces of these changes. The first of these theories suggests that natural forces have selected the canonical code due to advantageous changes in the code. This can be separated further into two possibilities. The first is that an initial code was alone or dominant in the biosphere and was then adapted throughout evolution. In contrast, there could have been numerous possible codes that were selected upon, and the canonical code outcompeted its challengers due to the some additional level of robustness in the canonical code. Both of these possibilities raise a similar and interesting point. For the first, where is the code going, and what pressures is it adapting to? More importantly, how do these changes explain the obvious structure in the code itself? For the second possibility we have to ask why did this code survive and dominate; what made it so much stronger? We know that the code can change due to the wide variety of variant codes, so why isn't one of these variants the canonical code?

Adaptation and optimization of the genetic code for reduction of errors from translation or mutation was an immediate explanation for the apparent order. Any order in the code, specifically having similar amino acids separated by only one nucleotide change, would limit the effects of mutations (Sonneborn 1965; Zuckerkandl 1965) or translational errors (Woese 1965 p71). Basically, these theories postulate that the canonical code won out because it was better at handling errors. Any organism harboring such a code would be more able to survive problems with translation. These pressures were even more apparent early, as the code must have been formed prior to the existence

of the modern machinery responsible for maintaining translational fidelity and of replication of the organism itself. Finally, the modern code is thought to have expanded from a smaller primordial code (Crick 1968), with newer amino acids being added to an already functioning code. A pressure to minimize any changes to the current proteome of the organism would have influenced the placement of these amino acids. A code in which similar amino acids were placed nearby would have been more robust at handling such mutational events. It is important to mention however, that a major change occurred in a variant organism that does not adhere to such stipulations (Kawaguchi 1989; Sugita 1999).

This idea that the canonical code is optimized was directly tested early on by a Monte Carlo simulation using 200 random codes. It was found that almost none of these random codes minimized changes in the polarity of the amino acids after a mutation in the 1st or 3rd base better than the canonical code (Alff-Steinberger 1969). The same report suggested that the code was better at minimizing changes based on other criteria as well, including polar requirement, isoelectric point, pK_a of the carboxyl group, molecular weight, number of dissociating groups, and alpha-helix forming abilities. All of these criteria would be important in minimizing the effects of mutations from a wide variety of changes. With the advance of technology, these experiments became larger in scope, with Haig and Hurst running a Monte Carlo with 10,000 codes. They found that the code was optimal in minimizing changes in polar requirement or hydropathy, and not with molecular volume or isoelectric point (Haig 1991; Haig 1999). Freeland expanded on this by weighing for the natural transition/transversion bias and found only one code more optimal than the genetic code at reducing changes in polarity from a pool of one

million random codes (Freeland 1998). Finally, 100-fold fewer better random codes were found upon using a matrix that measures the frequencies of amino acid substitutions in distantly related proteins (Ardell 1998; Freeland 2000).

An interesting point that has been raised by others (Wong 1980; Di Giulio 1989) is that the optimality of the code is not perfect. This is important to mention here, as we also found that the code is not at the theoretical maximum in our results (Chapter 2.1). Our results however, are from a noisier *in vivo* data set that contains many other causes for such imperfection. Even if it was only applied to the adaptive theories, this critique is a little shortsighted. It is important to realize that natural selection would have chosen a code that was better than others, or that simply worked. It does not necessarily have to be the best code, just the best available. As there are 10^{84} possible permutations in a 64-codon genetic code coding for 20 amino acids, the sheer odds of obtaining a perfectly optimal code is beyond astronomical. In addition, further mutational events that could lead to a more optimized code may be difficult to achieve in modern organisms (or even primitive ones). Even though there are variant codes, changes to the genetic code are not easy to achieve or overcome. The canonical code may therefore be at a local minimum and any further optimization is simply not achievable or necessary. This idea does not in any way negate the possibility that the code has undergone an adaptive evolution, or that it was selected for from a series of primitive competitors.

1.2.2.3: The Coevolution Theory

The last theory that attempts to explain the origins and evolution of the genetic code is the coevolution theory originally proposed by Wong in 1975. This theory also assumes that the modern genetic code evolved through a more primitive smaller code,

which is a logical assumption (Crick 1968). A key difference and challenge, especially to the adaptation theories, is that expansion of the code occurred through infiltration of newer amino acids via a biosynthetic relationship. In this case, the order and the apparent optimization is simply a signature of these historical alterations to the genetic code. Specifically, Wong posits that new amino acids are first added through a biosynthetically related pathway. This method could be similar to that of Gln/Asn and Glu/Asp in which a modification of the latter on the tRNA gives rise to the former. Regardless, Wong postulates that amino acids are added to the genetic code through a purely biosynthetic relationship and these “product-precursor” pairs would only be separated by a single nucleotide change (again, just as Asp/Glu are separated by Asn/Gln by a single nucleotide change) (Wong 1975). The placement and expansion of the genetic code in this manner would necessitate an apparent order or optimization, as similar amino acids are in general biosynthetically related.

Wong presented 8 specific product precursor pairs: Ser-Trp, Ser-Cys, Val-Leu, Thr-Ile, Gln-His, Phe-Tyr, Glu-Gln, Asp-Asn. He then determined the general probability for each of these 8 pairs having codons of the product connected to at least one codon of the precursors and compared it to what was actually observed. He combined each experiment using Fisher’s method and found that the overall probability of observing a similar amount of overlap by chance to be 0.0002 (Wong 1975).

Wong’s analyses in combination with the naturally occurring examples of the coevolution theory (i.e. the Asp/Glu – Asn/Gln relationships) present a strong case, however, there are some major flaws in his analyses (Amirnovin 1997). Amirnovin comprised a codon correlation score to more quickly analyze random codes for any

correlations similar to what was found by Wong. Analysis of the same 8 product-precursor pairs revealed that the probability of finding a random code with a similar correlation as the canonical code was 0.1%, slightly more probable than Wong's original assessment. It was further noted that two of these pairs (His-Gln and Leu-Val) seemed to dominate the results, and their removal pushed the chance of a random code having a higher correlation to 3.6%. This demonstrated that the selection of product-precursor pairs was critical to the overall result of the analyses. To address this potential problem (as product-precursor selection is chosen by the researcher, and thus may represent a bias), Amirnovin chose a list of product-precursor pairs using the biosynthetically related amino acids in *E. coli*. This new set, dictated by a modern organism, showed no increased correlation with 34% of random codes producing a higher correlation score (Amirnovin 1997).

Although Amirnovin's use of the codon correlation score is contested (Di Giulio 1999; Di Giulio 2000), his critique of Wong's 8 original pairs reveals a fundamental flaw. For instance, one of his product precursor pairs requires a metabolic process to be reversed. The Thr-Met pair actually has a common intermediate in homoserine, and conversion between the two would actually require the reversal of two enzymatic steps that are coupled to ATP hydrolysis. In addition, Wong never addresses the issue that no tRNA can distinguish between NNU and NNC, and as such NNY is cannot be split. We address this problem in Chapter 2.1, and the exclusion of splitting the NNY yields no difference in our analyses. Fixing these problems reveals that Wong's original probabilities of chance alone having a similar level of connected product-precursor pairs increases to 16.8% (Ronneberg 2000). Including Amirnovin's expanded pairings from *E.*

coli raises the probability even further to 62% (Ronneberg 2000). Although some codon assignments may have occurred through such processes, there is no statistical support that such events molded the entire genetic code.

1.2.2.4: Are the Early Evolution Theories in Conflict?

I want to end this subchapter by suggesting that these three seemingly disparate theories are not mutually exclusive. This idea is not new, and has been championed by Knight (1999). In fact, the adaptive and coevolution theories really do not posit the establishment of the code. Instead, they attempt to explain the apparent order in the code via stepwise alterations from a more primitive code. Therefore, it is entirely plausible that stereochemistry shaped the original code, which was further optimized through adaptation, with some additions occurring through a coevolution process. Our results outlined in Chapter 2.1 also support a multivariable solution to the evolution of the genetic code, and as such, I will revisit this idea in Chapter 5.

Table 1.2.1: Anticodons and codons are significantly overrepresented in a number of RNA-aptamers. This data is collected from numerous sources, and is adapted from Table 1 in Yarus 2009. The probability values represent the probability that the triplets are not overrepresented in the binding sites. References for each aptamer set is listed, however much of this data is unpublished outside of the Yarus 2009 review.

Amino Acid	Codon	Prob	Anticodon	Prob	References
Phenylalanine	UUU	0.98	AAA	0.012	(1)
	UUC	0.98	GAA	5.5×10^{-5}	
Isoleucine	AUU	8.0×10^{-110}	AAU	1	(2)
	AUC	1	GAU	1	
	AUA	1	UAU	3.2×10^{-131}	
Histidine	CAU	0.87	AUG	0.01	(3)
	CAC	0.12	GUG	1.6×10^{-8}	
Leucine	UUA	0.98	UAA	-	(4)
	UUG	0.029	CAA	0.71	
	CUU	-	AAG	0.95	
	CUC	0.99	GAG	0.25	
	CUA	0.3	UAG	0.006	
	CUG	0.3	CAG	-	
Glutamine	CAA	0.042	UUG	0.97	(5)
	CAG	-	CUG	0.65	
Arginine	CGU	0.92	ACG	0.85	(6)
	CGC	0.0014	GCG	0.0018	
	CGA	0.59	UCG	2.8×10^{-6}	
	CGG	3.4×10^{-4}	CCG	0.78	
	AGA	0.72	UCU	0.71	
	AGG	1.2×10^{-20}	CCU	0.71	
Tryptophan	UGG	1	CCA	2.7×10^{-6}	(7)
Tyrosine	UAU	0.026	AUA	6.0×10^{-6}	(8)
	UAC	0.0041	GUA	0.002	
Total		5.3×10^{-45}		2.1×10^{-46}	

(1) - Illangasekare 2002

(2) - Majerfeld 1998; Legiewicz 2005; Lozupone 2003

(3) - Majerfeld 2005 226

(4) - Unpublished results, see Yarus 2009

(5) - Unpublished results, see Yarus 2009

(6) - Many aptamers, see Knight 1998; Yarus 2005; Yarus 2009

(7) - Majerfeld 2005 5482

(8) - Mannironi 2000

Chapter 1.3: The Later Evolution of the Genetic Code

The genetic code is not “frozen” and is capable of change. As of this writing, the term “Alternative Genetic Code” yields some 2000 references on Pubmed. The sheer volume of these organisms and changes implies an active mechanism (or mechanisms) for conferring these changes. These mechanisms, unlike those discussed in Chapter 1.2, must have occurred under the modern pressures imposed by the translational machinery. These variant codes are a valuable resource, representing evidence of the further evolution of the genetic apparatus. This subchapter will give a brief overview of these variants, and the two main theories that attempt to explain how such alterations are achieved. Finally, I will introduce the concept of unnatural amino acid incorporation. Synthetic biology has taken cues from evolutionary biology in order to further expand the genetic code for our own gains. This process is cyclical, and any major findings from such synthetic organisms further increases our understanding of the complex mechanisms underlying the vast alterations to the proteome of an extant organism.

1.3.1: Natural Alterations - Shuffling

Crick’s “Frozen accident” hypothesis experienced its first major setback when it was found that vertebrate mitochondria and nuclear genetic codes differed (Barrell 1979). Originally cast aside as an artifact, most likely from the genome minimization in mitochondria or an artifact of endosymbiosis, the idea of a flexible code became more prevalent when nuclear variants were also found (reviewed in Osawa 1992). With the advances in genomic sequencing and projects such as Venter’s Global Ocean Sampling Expedition, the number of variant codes should continue to rise. This section will focus

on the variant codes in which codons have been shuffled or reassigned. Latter sections will focus on the natural and unnatural expansion of the code.

1.3.1.1: Variant Codes

This introduction is by no means a complete review of this ever-expanding field. There are two main reviews that go into greater detail into this topic (Osawa 1992; Knight 2001 p49). There are many changes in the mitochondria of numerous lineages, and due to the inevitable effects of genome minimalization, I will not discuss these in great detail. These changes, although numerous and potentially insightful (Knight 2001 p299), have had a major pressure shaping their evolution that differs from the scope of the work presented here. As such, I will be focusing on the changes that have occurred in the nuclear codes of a variety of organisms.

There are many examples of organisms that have undergone changes to their nuclear codes. The most extensive changes are in the Ciliates, which have undergone a wide variety of changes including reassignment of the UAR stop codons to glutamine, and the UGA stop codon to cysteine or tryptophan (Lozupone 2001). Other lineages have alterations as well, including diplomads (UAR to glutamine) (Keeling 1997), *Mycoplasma* species (CGG to nonsense) (Oba 1991), *Micrococcus* species (AUA and AGA to nonsense) (Kano 1993), *B. subtilis* (UGA to tryptophan), *Candidatus Hodgkinia cicadicola* (UGA to tryptophan) (McCutcheon 2009) and finally many species of the genus *Candida* (CUG to serine) (Sugita 1999). There are some consistencies in this group of changes; many changes occur multiple times in different lineages, and many target stop codons for reassignment. Stop codons may represent an easy target for reassignment, which will be discussed in much greater detail in Chapter 4. Finally, many

of these organisms are unique in one way or another. Some have reduced genomes, (*Mycoplasma* and *Candidatis*) although this is not a prerequisite for change. Many also have a high AT/GC genome content that have helped to limit the usage of certain codons (*Mycoplasma*, *Candidatus*, and *Micrococcus*) although this too does not appear to be an absolute requirement. Taken together, these organisms suggest multiple routes to fixation of non-standard genetic codes. This raises numerous questions, such as what are the potential forces of change? Are some codons more prone to change (this seems very likely in mitochondria (see Knight 2001 p299))? What causes fixation of such changes? And finally, how do organisms, especially those with large genomes adapt to such changes?

1.3.1.2: Pathways of Change

The natural variations in the canonical code present clues into the mechanisms that underlie genetic code evolution. Changes in minimal genome organisms and mitochondria suggest that a smaller genome may be more prone to changes. A codon with minimal usage in a small genome would in theory produce fewer potential deleterious mutations introduced into the proteome. This idea is echoed in organisms without small genomes as well, specifically in those that have undergone millions of years of AT or GC pressure. *Micrococcus luteus* has been suggested to have lost two codons (UAA and AUA) due to a high genomic GC content (Kano 1993). The codon capture hypothesis expounds on these ideas and proposes that codons can be eliminated from an entire genome by AT or GC-mutational pressure, and subsequently reassigned to a different meaning when the pressure reverts and the codon reappears (Jukes 1987; Osawa 1989). This theory can be applied to mitochondrial changes as well. Intense

pressure to minimize the genome of mitochondria removed codons from the genome. In many cases, this removal remains, while in others it was followed by reassignment. This may also help to explain the large number of variant codes that target stop codons. Stop codons are the rarest of codons, only used once per gene, and their removal would take fewer mutational events.

In contrast, the ambiguous intermediate theory claims that reassignment occurs through mutations in or duplication of tRNAs that cause a codon to have an ambiguous meaning (Schultz 1994). If this change was biochemically small or even beneficial, it may eventually become fixed, replacing the original meaning. The most interesting aspect of this theory is that there exists an organism that appears to be in a transitional phase. Many species within the genus *Candida* have reassigned the leucine CUG codon to serine (Kawaguchi 1989; Sugita 1999), but in certain species this reassignment is still taking place (Suzuki 1997). A mutation in the anticodon of a serine tRNA enables it to recognize the CUG codon (Yokogawa 1992), and in some species the CUG codon encodes serine while retaining some level of leucine incorporation (Suzuki 1997). This pivotal finding gave substantial support to the ambiguous intermediate theory and was the first to demonstrate the flexibility of nuclear genetic codes.

The codon capture and ambiguous intermediate theories are both supported by genomic data and may not be mutually exclusive. The reduction in the use of a codon, as postulated by the codon capture hypothesis, would greatly diminish the deleterious effects of an ambiguous codon. But just how rare does a codon need to be in order for reassignment? *Candida* suggests that it does not need to be too rare, as nearly 30,000 CUG codons are present in the ancestors of this organism (Miranda 2006), raising

numerous questions about the mutability of the genetic code. The prevalence of the CUG codon in ancestors of *Candida* suggests that this was not a small alteration in the genetic code, but in fact affects a large number of proteins. This substitution is not trivial either, as serine is smaller than leucine, has substantial polarity changes, and adds additional functionality to positions in the code once reserved for the functionally inert leucine. The adaptive evolution theories of the genetic code hypothesize that the standard code was specifically selected to avoid changes such as these. How did *Candida* survive these changes, as well as natural selection, which was presumably working against these types of changes? It is believed that the survival of *Candida* is due to an induction of natural stress response proteins (Santos 1999). When the tRNA conferring the leucine/serine ambiguity was transferred to *S. cerevisiae*, a similar response was induced, suggesting that the statistical protein population caused the increase in the stress response. *Candida* cells were thus likely primed to survive many different high stress conditions, outcompeting and outgrowing their wildtype counterparts.

The ancestors of LUCA and LUCA itself may have had an ambiguous codon table (Crick 1968). The ambiguity would allow an organism to survive droughts of amino acids that at the time could not be biochemically synthesized. Similar to *Candida*, an organism with ambiguous codons in these harsh conditions would be more adaptable to change and thus survive rapidly changing conditions. However, as the ability to biosynthesize different products evolves, so must the code. As organisms gained some freedom from the constraints of their environment, natural selection was able to act directly on the code and make it as efficient as possible. To presume that this selection has reached a terminal point is shortsighted. As more complete genomes are sequenced,

more variants of the code will inevitably be found. A major drawback in relying on natural variant codes for the study of genetic code evolution is that each variant code is simply a snapshot of where the genetic code is in evolution or where it has gone. Efforts to mimic many of these changes have led to interesting and sometimes surprising results (Santos 1999; Pezo 2004; Bacher 2007 p1907; Bacher 2007 p6494), but have fallen short of full reassignment. Chapters 3.1 and Chapter 4 will address our efforts to contribute to these endeavors.

1.3.2: Natural Alterations – Expansions

One of the long-standing unanswered questions regarding the evolution of the genetic code is why there only 20 amino acids. Why are useful moieties such as glycosyl, phosphate, and sulfate groups absent considering their usefulness in biological processes? Post-translational modifications of proteins and amino acids are widespread throughout biology (Uy 1977). Post-translational modification is an obvious route for introducing such changes, as it allows precise alterations without changes to the proteome on a global scale. However, the multitude of variant codes suggests that the code is flexible, and the introduction of another amino acid should be feasible. The introduction of this amino acid should also confer some new functions leading to a potential advantage. One response to this question may be that the genetic code formed from what was primarily available, relying on the presence of prebiotic compounds until complex synthesis machinery evolved. If this were the case, why are many of the prebiotic amino acids found by Miller absent in the code (Miller 1953; Miller 1976; Miller 1987)?

This question was largely unanswered until 1991 when it was discovered that selenocysteine was added to the code cotranslationally (Bock 1991). Selenocysteine is

synthesized from a serine that is pre-acylated on a suppressor tRNA (Schon 1989) and is inserted into an in-frame stop codon. Selenocysteine may therefore represent an expansion of the genetic code through coevolution, as outlined by Wong (Wong 1975). Suppression in this case is controlled by an mRNA-based element and various protein factors, limiting incorporation of selenocysteine to only select UGA sites (Berry 1991). The 22nd amino acid, pyrrolysine, was found a decade later in archaea (Srinivasan 2002). Similar to selenocysteine, pyrrolysine is inserted at an in-frame stop codon (UAG), however it has a specific tRNA/synthetase pair responsible for its incorporation (Polycarpo 2004). It is still unclear if pyrrolysine incorporation is complete, or if it is still in competition with endogenous release factors (Zhang 2005).

There are two major conclusions to be drawn from the incorporation of selenocysteine and pyrrolysine. The first is that they, like many of the other natural variations, target stop codons. In addition, pyrrolysine is affectively aminoacylated even if the anticodon of the tRNA is mutated (Ambrogelly 2007; Yanagisawa 2008), suggesting that it *could* have been inserted via another codon, but UAG was apparently preferred. This again suggests that stop codons are a preferred target for genetic code variation and expansion. This theme is consistent with the theorized mechanisms outlined above; a rare stop codon would either be easier to reduce in number for capture, or would introduce fewer errors resulting from ambiguity. This has been exploited by other unnatural expansion schemes (see the next section), and was further expanded by us (see Chapter 4). Secondly, it is important to mention that these amino acids do not represent a complete change to the genetic code (the verdict on pyrrolysine is still out).

A full reassignment of a codon for expansion has yet to be achieved through either natural or synthetic means.

1.3.3: Unnatural Alterations

Modern molecular and synthetic biology has allowed for the direct manipulation of the genetic code in modern organisms. These methods have been utilized in a variety of studies, but focus on two major avenues of research. In the first, precise alterations are made in the genomes of modern organisms to alter the genetic code itself. Much of this work is solely focused on mimicking the pathways of genetic code evolution and the creation of strains useful in answering these questions. In that regard, these synthetic organisms make up for what nature is lacking, allowing the study of adaptation to such changes in real time. The second avenue of research is focused on exploiting our knowledge of the genetic code and the translational apparatus to affect changes that benefit research as a whole. These techniques rely on the incorporation of a wide variety of unnatural amino acids into proteins for structural, functional, and physical studies. Advances to these techniques now allow for the site-specific insertion of unnatural amino acids *in vivo* in a wide variety of organisms. As much of this dissertation focuses on utilizing both of these techniques, a further in depth review of both of these fields will be in their respective chapters.

Chapter 1.4: Perspectives

Many important questions regarding the translational apparatus and the genetic code have been answered in the past 7 or so decades, however some major questions linger. There remains little to no evidence regarding the original establishment of the code itself. This field has had some recent excitement due to the wealth of *in vitro* aptamer data from the Yarus group, however much remains to be discovered. In addition, the natural code variants give valuable evidence for how the genetic code adapts and changes through time. However, these organisms only provide a glimpse into the actual changes, and provide little to no evidence of the adaptation events underlying such changes. Understanding these events would be of considerable value outside the evolutionary biology field, as these changes may represent some of the most drastic evolutionary events imposed on an organism. Attempts to mimic these changes have led to valuable and rare empirical data, however they all fall short of mimicking full reassignment. Only with full reassignment can the entire adaptation events be well understood.

Theories outlining the evolution of the genetic code can be split into early and late varieties. The work described here addresses both of these avenues with early theories discussed and analyzed in Chapter 2. Late evolution theories can themselves be split into two different areas, differing mainly in their approach. Mimicking of the natural processes will be discussed in Chapter 3, and the synthetic expansion of the code will be discussed in Chapter 4.

Chapter 2: The Stereochemical Code

This section focuses on the evolution of the genetic code before LUCA (the last universal common ancestor). This places us in the middle of the RNA world, about 4 billion years ago. As outlined in the introduction, there are three theories that delve into this complex topic. One of which is the stereochemical hypothesis that postulates that triplets of RNA nucleotides have a specific and efficient affinity towards amino acids. These interactions may have left an imprint on the code by directly establishing the relationship between amino acid and the interacting RNA (in this case, through fixation of the interacting RNA as a codon or anticodon). This theory has drawn my interest for a number of reasons. The chemical nature of these interactions puts it at an advantage over others that rely on adaptation and optimization. Adaptation and optimization requires an established system to optimize, and these theories lack explanations of the principals underlying this establishment. In addition, the evolution of such a complex apparatus has to adhere to the principal that natural selection and evolution lacks foresight. The stereochemical hypothesis posits a plausible scenario for the establishment of the genetic code that can logically avoid this dangerous pitfall (Wolf and Koonin go into this into great detail in their 2007 review). Finally, research into the stereochemical hypothesis has experienced a boon in recent years thanks to the well-established SELEX systems outlined in the introduction. As a consequence, more and more work reliant on theoretical grounds alone appears to use this theory as a plausible scenario for code evolution.

The stereochemical hypothesis still suffers from internal ambiguity in many of the *in vitro* aptamer experiments, and also a general lack of any compelling *in vivo* evidence

for the existence of these interactions. Why do some amino acids prefer binding sites enriched in codons, others anticodons, and many none at all? Having some amino acids not involved could be a logical consequence of continued evolution of the code from core starting elements, and these amino acids may not have relied on stereochemical interactions for their modern placement. All of these conflicts could be irrelevant, as a major unanswered question also remains: even if these interactions existed could they have been strong enough to drive code evolution?

Chapter 2.1 introduces the stereochemical code discovered in modern ribosomes, putting the stereochemical hypothesis in a meaningful *in vivo* context. The results from these analyses address some of the major dilemmas in the stereochemical hypothesis, introduce a new model of studying these interactions, and raise numerous interesting questions.

Chapter 2.2 extends on these theories using a novel ambiguous codon system. The DJ.2 *E. coli* strain was generated that harbors ambiguous isoleucine codons. The ambiguous genetic code in DJ.2 reveals that all isoleucine codons are not decoded equally, and that neighboring context, specifically in the anticodons of isoleucine and the invading amino acid valine, can influence tRNA selection. These data suggest that stereochemical interactions may still exert a role in modern translation.

Chapter 2.1: Imprints of the Genetic Code in the Ribosome

2.1.1: Abstract

The establishment of the genetic code remains elusive nearly five decades after the code was elucidated. The stereochemical hypothesis postulates that the code developed from interactions between nucleotides and amino acids, yet supporting evidence in a biological context is lacking. We show here that anticodons are selectively enriched near their respective amino acids in the ribosome, and that such enrichment is significantly correlated with the canonical code over random codes. Ribosomal anticodon-amino acid enrichment further reveals that specific codons were reassigned during code evolution, and that the code evolved through a two-stage transition from ancient amino acids without anticodon interaction to newer additions with anticodon interaction. The ribosome thus serves as a molecular fossil, preserving biological evidence that anticodon-amino acid interactions shaped the evolution of the genetic code.

2.1.2: Introduction

The origin and evolution of the genetic code is a critical transition in the evolution of all modern organisms (Szathmary 1995). Understanding why the genetic code evolved to its modern form is as, if not more, important as knowing the code itself (Woese 1965 p71), as it is central to understanding major evolutionary breakthroughs. However, the universality of the code is also its downfall with regards to studying its formation, as no organisms exist containing a primitive or intermediate genetic code for comparison. Although multiple hypotheses have been proposed to explain why codons are selectively assigned to specific amino acids (Knight 1999; Yarus 2005), empirical data are extremely rare and difficult to obtain (Woese 1966 p723; Majerfeld 1998; Mannironi 2000;

Legiewicz 2005), leaving many theories in the realm of conjecture. One theory addressing this challenging question is the stereochemical hypothesis, which postulates that the genetic code developed from interactions between anticodon or codon-containing polynucleotides and their respective amino acids (Woese 1966 p723; Dunnill 1966). This theory is supported by RNA aptamer experiments in which RNA molecules evolved to bind specific amino acids *in vitro* are enriched with anticodon and codon elements for the amino acid (Majerfeld 1998; Mannironi 2000; Legiewicz 2005; Majerfeld 2005 p226; Majerfeld 2005 p5482). Codons for arginine have also been found to confer binding specificity for arginine in the self-splicing group I introns (Yarus 1989). Nonetheless, the biological relevance of these aptamers and introns to the genetic code evolution is unknown, and no further *in vivo* data exist to support this hypothesis.

If chemical or physical interactions between nucleotides and amino acids did influence the evolution of the genetic code, relics of this evolution may be present in modern cells. In particular, we may uncover such imprints within RNA-binding proteins and RNA-protein interactions. The ribosome presents an excellent model for the study of these potential interactions, as it is a large ribonucleoprotein complex with some 50 proteins interacting extensively with the ribosomal RNAs (rRNAs) for stability and structure (**Figure 2.1.1**) (Ban 2000). In addition, the ribosome emerged from an early evolution stage of life to establish the translation of the genetic code before the last universal common ancestor (LUCA) (Fox 2004), and thus is more likely to preserve relics of the underlying force driven the formation of the code. A comprehensive analysis of ribosome structural data will help to reveal if such interactions survive in the ribosome and if they correlate with the canonical genetic code. The correlation, if

established, would provide empirical *in vivo* data from modern organisms for understanding the origin and evolution of the genetic code.

2.1.3: Materials and Methods

2.1.3.1: Ribosome Structural Analysis

Ribosome structural data was analyzed by using new Perl scripts. If any atom of an amino acid was within 5 Å of an rRNA nucleotide, this was considered a hit and the amino acid and surrounding rRNA sequence was put into a separate array. The arrays from all four structures were compiled and used for all subsequent analyses.

The selection of 5 Å as the cutoff was the result of balancing two conflicting parameters: number of interactions and interaction specificity. RNA-amino acid interactions would be more specific within a closer distance, and thus a smaller radius would allow more accurate analysis. However, the number of interactions would become limiting if the radius is too small, leading to the exclusion of certain amino acids from robust statistical analyses. To determine the cutoff, global correlation analysis (see below) was performed on data generated from radii ranging from 4 to 10 Å. The lower limit was set by the resolution of the ribosome structures themselves. As expected, an increased radius led to less significant results, with only 4-6 Å radii having a *P*-value < 0.05 (**Figure 2.1.2**). Radii of 4 or 5 Å had very similar statistics; however, the limited data from 4 Å led to the exclusion of three amino acids. Data generated from a 5 Å radius only excluded cysteine (see below), and was therefore used for all subsequent analyses.

Populations containing every possible 3 or 4 nucleotide RNA sequence were extracted from the data using new Perl scripts and used to calculate probabilities. The

probability for finding a certain sequence near a specific amino acid was then calculated for all anticodon and codon-containing sequences for all 20 amino acids. The probability (P_X where X is the amino acid in question) of finding a particular set of codon or anticodon-containing sequences within 5 Å of an amino acid versus all other sequences was calculated. For those amino acids with multiple codons or anticodons, the results were additive. For example, Phe in the canonical code has two codons, UUU and UUC. P_F would be the number of times UUU and UUC appears within 5 Å of Phe residues divided by the number of total 3 nucleotide sequences (NNN) within 5 Å of Phe. A normalization factor (N_X where X is the amino acid in question) was then calculated using the exact same calculations, only for the 19 remaining amino acids. The calculation for N_F in this case would be the number of times UUU and UUC appear within 5 Å of the other 19 amino acids divided by the number of total 3 nucleotide sequences (NNN) within 5 Å of the 19 amino acids. The final enrichment value for Phe would then be $E_F = \frac{P_F}{N_F}$. This enrichment value thus represents the increase or decrease in likelihood of finding a particular set of 3-nucleotide sequences near a particular amino acid. If the value is 1, then there is no increased likelihood of finding these sequences near a particular amino acid, compared to the other 19 amino acids.

A two-tailed chi-square analysis was then used to determine if any codon or anticodon-containing sequences were more likely to be found near amino acid X versus the other 19 amino acids. If there were less than 10 hits for any given sequence, the statistical analysis is less robust and was thus not determined. All amino acids showing enrichment were then sorted by their P -value for combining results via Fisher's method.

The combination of the largest set of amino acids that satisfied $P < 0.05$ was determined and used in the subsequent Monte Carlo simulations.

2.1.3.2: Random Code Generation

Each of the three methods for generating 1,000,000 randomized codes used a series of new Perl scripts. The RAND generator was essentially as described by Novozhilov (2007), with a few minor modifications. In the RAND and NNY generators we did not maintain the positioning of the stop codons, and each was assigned independently. These generators put more emphasis on maintaining the number of codons per amino acid, and less on maintaining the exact structure of the canonical code. In addition, Ile and Met were not treated as a single block and were assigned independently. Those amino acids occupying a full block (4 codons) were first assigned a position in the 16 possible locations using Perl's random number generator. Ile was then assigned by first treating it as occupying two codons. It was placed randomly in any of the remaining NNY or NNR blocks. Following placement, the third Ile codon was randomly placed in the corresponding neighboring block (i.e. if Ile was assigned to AAY, then the third position would be AAA or AAG). Next, the amino acids with two codons were randomly assigned to any remaining NNY or NNR blocks, followed by those that occupy a single codon and the three stop codons. Those amino acids with 6 codons were split up into a four-codon block and a two-codon block, which were assigned independently.

The NNY generator was designed after the RAND generator, but with modifications to ensure that no NNY block was split. Ile was first randomly paired with Met, Trp or a stop codon, and then the entire block was randomly assigned codons using

Perl's random number generator. There was no restriction on which NNR the third Ile codon occupied. The 4 remaining single codon-containing amino acids and stop codons were randomly paired and assigned to a random NNR. The remaining 4-codon and 2-codon amino acids were then assigned as described in the RAND generator.

The SYN generator was designed as in Haig (1991). With this generator, we did not alter the position of the stop codons, as the impetus behind the creation of these random codes is different from those above. This generator assumes that the exact structure of the code is critical to its function, and in theory, any major deviation would make the randomized codes more unrealistic and unfit. To create these codes, each block as it exists in the canonical code was assigned a number. Each of the 20 amino acids was then randomly assigned to a block to ensure that each amino acid was represented at least once. The remaining three blocks were then randomly assigned any of the 20 amino acids, completing the randomized codes.

2.1.3.3: Monte Carlo Simulations

For all comparisons, average enrichment was chosen over chi-square analysis because it is less likely to be influenced by a single member of the population. Average enrichment was also much faster to calculate and thus facilitated the running of the numerous simulations required. Finally, results between the two analyses should be very similar regardless, as the two numbers themselves are significantly correlated ($P < 0.0001$) using the non-parametric Spearman calculation.

For the specific subset analysis, only those amino acids determined from the canonical code using the Fisher's method were analyzed for each randomly generated code. The enrichment values for the codons or anticodons as dictated by the randomized

code were calculated for each of the amino acids. The average enrichment was calculated and compared to that of the canonical code. Those codes with a higher average enrichment were considered more correlated with the ribosome structural data.

For both the global and optimal subset analysis, all amino acids were analyzed from each randomly generated code. The enrichment of either codons or anticodons as dictated by the randomized code were calculated and sorted. For the global analysis an average enrichment value for the 22 independent calculations was obtained and compared to the canonical code. Those amino acids with 6 codons were split into their respective 4- and 2-codon blocks and analyzed independently. Cys was removed because it contained the fewest number of surrounding rRNA hits, leading to an increase in the percent error of the randomized codes (**Figure 2.1.3**). In addition, Cys rarely met the >10 hits cutoff imposed on the statistical analysis in the majority of the random codes. For the optimal subset, those amino acids that had the highest 8 enrichment values for codons or 11 for anticodons were used for each randomized code. Therefore, each randomized code may have a different subset of amino acids used for comparison. Again, those codes with a higher average global or optimal enrichment are considered more correlated with the ribosome structural data.

Each amino acid in a shared block was expanded individually and the global Monte Carlo analysis was performed and compared to the standard global anticodon analysis before expansion. Any difference in the number of more correlated random codes was used as a comparison to determine if expansion led to a higher correlation. In the analysis of the shared codon blocks, one amino acid was expanded to occupy the full codon block while the other amino acid sharing the block could be either analyzed at its

positions in each code or excluded from the analysis. Exclusion of the sharing amino acid makes direct comparison of random codes difficult, because the identity of the sharing amino acid may vary in different codes. For instance, in the case of the canonical code, the expansion of Ile would measure the enrichment of all NAU anticodons for Ile and exclude the calculations for Met. In random codes using the RAND generator Ile is generally paired with Met, Trp, or a stop codon. In these codes, the excluded amino acid could be any of those possibilities. Exclusion of different amino acids makes comparison of two codes ambiguous and potentially incorrect, as the makeup of the calculations would be inconsistent. For these reasons, we calculated the sharing amino acid at its positions in each code. This led to one or two anticodons being used twice for each calculation. For instance, in the Ile expansion analysis of the canonical code, Ile occupied the full NAU block while Met occupied CAU. This allowed us to never exclude an amino acid and to make more accurate comparison of random codes.

2.1.4: Results

2.1.4.1: A Subset of Amino Acids are Enriched with Codons and Anticodons in Ribosome Structures

We examined whether codons (or anticodons) assigned to each amino acid by the canonical genetic code were enriched near the respective amino acid in the ribosome. Structures of ribosomes from 4 species: 1 archaeobacterium and 3 eubacteria (Selmer 2006; Borovinskaya 2007; Wilson 2008; Schmeing 2005) were analyzed. Many amino acids were found to contact their codon (or anticodon) sequences in rRNAs (**Figure 2.1.1B**). A codon (or anticodon) enrichment value for an amino acid was calculated from the probability of finding a block of triplet codon (or anticodon) sequences within 5 Å of

the amino acid relative to other 19 amino acids in the ribosomal structures (**Table 2.1.1A**, see Methods). To determine if a global trend of codon or anticodon enrichment was significant, independent statistical tests of each enriched amino acid were combined using Fisher's method for each analysis. Using a statistical cutoff of $P < 0.05$, 11 amino acids were found to be significantly enriched with anticodons ($P = 0.039$) and 8 with codons ($P = 0.045$) (**Table 2.1.1B and C**). Two amino acids (Leu and Ile) are present in both categories. These results demonstrate that, for a subset of amino acids, codons or anticodons are enriched near their respective amino acids in the ribosome.

2.1.4.2: Monte Carlo Simulations Show Enrichment is Correlated to the Canonical Code

Is the codon (or anticodon)-amino acid enrichment observed in the ribosome correlated with the canonical genetic code, or can similar results be obtained with random codes? We applied a series of Monte Carlo simulations to determine if the amino acid-codon assignment in the canonical genetic code is better at explaining the enrichment than the assignments in random codes. We first analyzed the probability of finding a better code using only those amino acids in **Table 2.1.1B and C**. One million random codes were generated. The average enrichment values for both the anticodon and codon enriched amino acids were determined using the ribosome structural data for each code, and were compared to those determined from the canonical code (1.23 for anticodon-enriched and 1.17 for codon-enriched amino acids, see Methods). The distribution of the average enrichment values in the random codes is shown in **Figure 2.1.4A** for codons (left panel) and anticodons (right panel). For comparison, the average enrichment value for the canonical code is denoted by an arrow in each of the histograms. For the 11

amino acids enriched with anticodons and 8 enriched with codons, 99.9555% and 99.9716% of random codes have lower average enrichment than the canonical code, respectively.

The next challenge in these analyses is to demonstrate that the enrichments in the specific subsets are not due to random chance. To investigate whether correlations can also be found in random codes with different subsets of amino acids enriched with codons (or anticodons) in the ribosome, we next performed a global correlation analysis and an optimal subset correlation analysis. In the global correlation analysis, all amino acids (excluding Cys due to too few Cys available at the surface of ribosomal proteins to interact with rRNA, see Methods) were taken into consideration to calculate the average enrichment values based on the ribosomal structural data. For the enrichment of anticodons, 99.0225% of randomized codes produced a lower average enrichment than the canonical code (**Figure 2.1.4B**, right panel). In contrast, the enrichment of codons no longer showed correlation with the canonical code, with only 54.5447% of random codes producing a lower average enrichment value (**Figure 2.1.4B**, left panel). In the optimal subset correlation analysis, we used the subset of 8 amino acids most enriched with their codons and of 11 amino acids with their anticodons in the ribosome for each randomly generated code. The identity of amino acids in the subset thus varied with the code. Distributions for these analyses were plotted and compared to the canonical code for codons and anticodons as before (**Figure 2.1.4C**). Results are similar to the global correlation analysis, with 71.0076% and 95.1925% of random codes having lower enrichment of codons and anticodons, respectively. These two analyses suggest that the codon enrichment with the specific subset of amino acids cannot be attributed to more

than random chance, as a significantly large portion of random codes generate better results with either an optimal set of 8 amino acids or the entire set. The anticodon enrichment of the canonical code proves to be stronger than a significant portion of random codes in all analyses, demonstrating a strong correlation between the ribosomal anticodon-amino acid enrichment and the canonical code that cannot be attributed to random chance alone. Such a correlation suggests that amino acid-anticodons interactions could have contributed to the organization of the canonical genetic code, supporting the stereochemical hypothesis.

2.1.4.3: Correlations Consistent Regardless of Random Code Generator

Several methods have been devised to generate random genetic codes for analyzing code properties (Freeland 1998; Novozhilov 2007; Haig 1991), and different generators have yielded different results in some cases (Novozhilov 2007, Ronneberg 2000; Amirnovin 1997). To determine if our findings were independent of the code generation methods, we employed all three methods in our correlation analyses. The first method (RAND method) maintains the number of codons per amino acid as dictated by the canonical code and randomizes the positions of the amino acids (Novozhilov 2007) (see Methods). The second method (NNY method) is similar to the RAND method but imposes the restriction of not splitting NNY blocks (Ronneberg 2000). No evidence exists in non-standard genetic codes that suggest translation can distinguish between codons occupying a NNY block (Knight 2001 p49), and thus it has been proposed that NNY codons should not be split when generating random codes (Ronneberg 2000). The third method (SYN method) maintains the exact block structure of the canonical code and randomizes the amino acid in each block (Haig 1991). We generated one million codes

using each method to perform the correlation analyses, and found that the results are consistent regardless of which method was employed (**Table 2.1.2**). The RAND method produced more optimized codes and less significant probabilities. As such, the RAND generator was used for all analyses described, as it was the most statistically stringent method.

2.1.4.4: Analysis of Ribosome Data Reveals Evidence for Codon Capture Events

In the canonical code, 8 blocks of four codons that differ only at the third position are each shared by two amino acids or by amino acids with stop signals, while the other 8 blocks are each occupied by a single amino acid. Codon block sharing has implications on codon reassignment (Knight 2001 p49). To explore whether ribosomal anticodon-amino acid enrichment can reveal evidence for such potential evolutionary events, we used global correlation analysis to re-analyze the code by expanding each amino acid in the split blocks to occupy the entire codon block. The average enrichment value of each expansion of the canonical code was calculated and compared to that of each random code after expansion of the same amino acid. The number of “better” codes, which have a higher average enrichment value and are thus more correlated to the ribosome structure data than the canonical code, was determined before and after expansion of a particular amino acid.

Upon expansion to occupy a full codon block, Leu, Ile, His, and Lys all markedly improved the correlation of the expanded code with the ribosomal anticodon-amino acid enrichment, and thus had a large decrease in the percentage of better random codes (**Figure 2.1.5**). In contrast, the other amino acid in each block (Phe, Met, Gln, and Asn respectively) had no such large decrease after expansion. Remarkably expansion of Phe

showed the opposite effect, greatly increasing the number of better random codes by 40%. The other four shared codon blocks did not show such contrast in change of better random codes (**Figure 2.1.6**). These results suggest that Leu, Ile, His, and Lys may have once occupied the entire codon block, and specific codons were later conceded to the other amino acids. Specifically, the data provide evidence for the capture of AUG from Ile to Met, consistent with what is considered a key codon capture event (Trifonov 2004; Jukes 1973), as well as the first evidence that Phe may have captured the UUY codons from Leu. Our data also suggest that a subset of codons of Lys and His were reassigned to Asn and Gln, respectively.

Although the correlation between the canonical code and the ribosomal anticodon-amino acid enrichment is strong, some random codes were found to have a higher average enrichment value in the global correlation analysis. Analysis of amino acid placements in the top 1000 random codes correlated with the ribosomal enrichment data may also reveal further evidence for reassignments of amino acids in code evolution. For instance, Ile shows a strong preference for the Met codon AUG in 44% of the top 1000 random codes (**Figure 2.1.7A**). Consistently, CAUC, which contains the corresponding anticodon CAU, was found significantly enriched within 5 Å of Ile than the other 19 amino acids ($P = 0.0003$) in the ribosome. This result further suggests that Ile may have initially occupied the AUG codon and Met captured it later. In another example, 26% of the top 1000 random codes place Pro at the AAN position (**Figure 2.1.7B**) due to the significant enrichment of the UU moiety within 5 Å of Pro in the ribosome ($P = 0.0091$). Proline thus may have undergone a shift from AAN to its current CCN positions in the canonical code, for which no previous evidence exists.

2.1.4.5: The Eleven Amino Acids Enriched with Anticodons are New Additions to the Genetic Code

Evidence of codon reassignment events have been previously limited to changes in nuclear and mitochondrial codes of certain organisms (Knight 2001 p49; Osawa 1992). The ribosomal data shown here presents novel evidence for the existence of these evolutionary events that drive the establishment of the canonical code. The reassignment of amino acids to different codons during evolution also helps to explain why only a subset of amino acids are involved in amino acid-anticodon enrichment in the ribosome. Our finding on ribosomal anticodon-amino acid enrichment, when compared with the consensus chronology of amino acids built on 60 criteria (Trifonov 2004), revealed a distinct stage for the evolution of the canonical code. Except Asp, all other amino acids that are significantly enriched with their anticodons in the ribosome are at the later stage of amino acid expansion (**Table 2.1.3**), starting from Leu (ranked No. 9) to Trp (No. 20). For the 2 amino acids at the later stage without apparent anticodon enrichment, Cys is too rare on ribosomal protein surface for statistical analysis, leaving Asn as the only inconclusive amino acid. Lysine-rRNA interactions are the second most abundant amino acid-RNA interaction in the ribosome, due to abundant ionic interactions with the RNA backbone, which masks the enrichment of Lys to its anticodons. Although with an apparent weak enrichment, Lys showed a marked decrease in the number of better random codes when expanded to occupy the full AAN block (**Figure 2.1.5**). We therefore tentatively place lysine in the population of those amino acids with an anticodon interaction. The two-stage distribution suggests that ancient amino acids implied in the Miller-Urey experiment (Miller 1976) were fixed into the genetic code

through means other than amino acid-anticodon interaction or that such interactions were masked in the ribosome due to later codon reassignment upon addition of new amino acids. Newer amino acids, on the other hand, were added to the genetic code mainly through the anticodon interaction.

2.1.5: Discussion

The canonical genetic code is one of the most dominant aspects of life on this planet, and thus studying the origin of the genetic code is critical to understanding the evolution of all life. In this study, we tried to address this long-standing and difficult question using structural information of ribosomes from bacteria and archaea, 2 of the 3 life domains. We found that anticodons are selectively enriched near their respective amino acids in ribosome structures and that such enrichment is correlated with the canonical genetic code. Previously codons and anticodons have been found enriched in the binding sites in RNA aptamers selected to bind amino acids (Yarus 2005; Yarus 2009), and conserved arginine codons have been identified to bind Arg specifically in group I self-splicing introns (Yarus 1989). Such experiments have provided precious, hard to obtain empirical data, demonstrating the association of coding triplets with amino acids. Our findings corroborate such association but place the anticodon-amino acid interactions in a meaningful biological context: modern ribosomes, the machinery responsible for translating the genetic code. Although selective pressures have continued refining the ribosome during evolution, our results suggest that the essence of the primitive RNA-amino acid interaction remains at the heart of modern ribosomes (Woese 2001). More importantly, our data establish a direct connection between the anticodon-amino acid association and the genetic code, filling a critical gap that has not been

addressed using any means before. Our findings thus represent the first comprehensive *in vivo* evidence supporting the tenets of the stereochemical hypothesis for code origin.

The original stereochemical hypothesis proposes that the genetic code is shaped by interactions between amino acids and cognate coding triplets, which can be codons and/or anticodons. Whether codons or anticodons played the role in the origin of the genetic code has remained elusive (Szathmary 1999; Wolf 2007). In the *in vitro* experiments of selecting RNA aptamers to bind amino acids, both codons and anticodons are found significantly enriched in the cognate amino acid binding sites. Specifically, Arg and Ile are found to associate with both codon and anticodons, whereas His, Phe, Trp, and Tyr are associated with anticodons only (Yarus 2009). It is difficult to interpret the stereochemical complementarity for amino acids enriched with both codons and anticodons in RNA aptamers. In the arginine binding site of the self-splicing group I introns, only codons are found to interact with Arg (Yarus 1989). In contrast, our results clearly indicate that it is the anticodon-amino acid association, but not the codon-amino acid association, correlated with the genetic code. This conclusion advances the stereochemical hypothesis by pinpointing anticodon-amino acid interaction as the potential drive for code formation. In addition, it will help our understanding of the origin of the translation system. It has been proposed that one critical step is the evolving of amino acid-binding RNAs under the selection pressure of amino acid accumulation (Wolf 2007). The specific binding mechanism determines the correspondence between amino acids and cognate triplets. These RNAs subsequently acquire autocatalytic aminoacylation to covalently couple the amino acid, and eventually become tRNAs. The anticodon-amino acid binding fits naturally with the evolution and function of the tRNA,

which contains the anticodon. Further study based on the anticodon-amino acid association will shed new light onto the predecessors and elementary steps that eventually led to the emergence of translation.

There are two distinct periods for the evolution of the genetic code, before and after the LUCA. The canonical code emerged before the LUCA; all life following the LUCA inherited this canonical code nearly-ubiquitously, yet minor deviations from the canonical code have been found in some nuclear and mitochondria lineages (Osawa 1992). Such code variants suggest that the genetic code is still evolving and codon meanings can be reassigned in extant organisms. Had codons also been reassigned during the establishment of the canonical code before LUCA? Both the adaptation theory and the coevolution theory hypothesize codon reassignment as a necessary step for the formation of the canonical code. The adaptation theory predicts that the code underwent optimization to minimize the impact of errors of translation (Freeland 2003), while the coevolution theory predicts that the code was expanded through the addition of biosynthetically-related new amino acids, which took codons from the precursor amino acids (Wong 1975). However, because primitive life before LUCA and LUCA itself do not exist today for comparative study, evidences for codon reassignments during the canonical code formation are lacking. Our results obtained through analyzing ribosomal structural data and its correlation with the canonical code now provide such evidences. Specifically, we found evidence suggesting that a subset of codons originally set for Leu, Ile, His, and Lys were reassigned to Phe, Met, Gln, and Asn respectively, and that Pro was reassigned from AAN to the current CCN codons. In comparison to the codon reassignments found in post-LUCA organisms (Knight 2001 p49), the pre-LUCA codon

reassignments we identified show overlaps and distinctions: the reassignment events from Ile to Met and from Lys to Asn have also been observed in mitochondria and nuclear lineages, whereas those from Leu to Phe, from His to Gln, and for Pro have never been identified in extant life. It should be noted that the codon reassignments identified in this study may not cover all possible reassignment events. We could not analyze the reassignment between stop codons and sense codons because of the nonexistence of a stop entity in proteins, whereas the reassignment of stop codons to amino acids is often seen in mitochondrial and nuclear code variants.

It is hypothesized that the coding process started with a set of primitive amino acids and that others were added until the total of 20 was reached. What amino acids were the primordial ones and how the new amino acids were added to the genetic repertoire are unknown and under debate (Yarus 2005; Novozhilov 2007; Ronneberg 2000; Di Giulio 2009). In addition, if RNA-amino acid interactions influenced the organization of the code, the pervasiveness of these interactions, and the interplay between such interactions and other potential driving forces are also unknown. A novel finding from our study is the temporal order in which anticodon-amino acid interactions may have taken place. We demonstrate that the canonical code may have evolved through two distinct stages. The formation of the early code, consisting of the prebiotically available amino acids, was not influenced by the anticodon-amino acid interactions. Once some critical events occurred, for instance, the establishment of primitive translation, these interactions became relevant and induced a watershed for code expansion, allowing newly available amino acids a route into the genetic code.

A major challenge in studying the origin of the genetic code is that many questions are out of reach of direct experimentation. Empirical data, although limited in the past, has yielded landmark insights. The spark-tube experiments confirmed that amino acids could have been produced abiotically in the atmosphere of the early earth (Miller 1976; Miller 1987), and *in vitro* RNA aptamer selections provided solid evidences for the association of coding triplets with cognate amino acids (Yarus 2009). Our usage of ribosome structural data further demonstrates the usefulness of expanding the theoretical inputs to a biological context. This methodology leads to robust results and novel findings, underscoring the potential created by empirical evidence in the study of code origin and evolution.

Table 2.1.1: Enrichments values calculated for codons and anticodons. (A) Relative enrichment values of codons and anticodons near their respective amino acids in ribosome structures from four different species. (B) Eight amino acids showed significant enrichment of codon- containing rRNA. (C) Eleven amino acids showed significant enrichment of anticodon- containing rRNA.

A			B		
Amino Acid	E_x (Codons)	E_x (Anticodons)	Amino Acid	<i>P</i> -value	$-2\ln(P)$
Ala	0.912	0.951	Ile	0.0225	7.588
Arg (CGN)	<u>1.034</u>	<u>1.014</u>	Leu (UUR)	0.0652	5.461
Arg (AGR)	0.929	0.998	Thr	0.0816	5.012
Asn	0.796	0.987	Glu	0.2409	2.847
Asp	0.989	<u>1.320</u>	Gly	0.2998	2.409
Gln	0.946	<u>1.234</u>	Arg (CGN)	0.4420	1.633
Glu	<u>1.149</u>	0.596	Lys	0.6437	0.881
Gly	<u>1.058</u>	1.006	Tyr	0.6513	0.858
His	0.904	<u>1.078</u>	Sum (χ^2)		26.69
Ile	<u>1.437</u>	<u>1.283</u>	<i>P</i> -value		0.0451
Leu (UUR)	<u>1.403</u>	<u>1.119</u>	C		
Leu (CUN)	1.059	<u>1.050</u>	Amino Acid	<i>P</i> -value	$-2\ln(P)$
Lys	<u>1.033</u>	0.955	Phe	0.0118	8.879
Met	0.731	<u>1.152</u>	Trp	0.1052	4.504
Phe	0.862	<u>1.384</u>	Asp	0.1118	4.382
Pro	0.980	0.993	Ile	0.1169	4.293
Ser (UCN)	0.954	0.988	Gln	0.1182	3.852
Ser (AGY)	0.984	1.042	Tyr	0.1457	1.462
Thr	<u>1.156</u>	0.923	Leu (UUR)	0.4814	1.196
Trp	0.667	<u>1.640</u>	His	0.5498	1.046
Tyr	<u>1.130</u>	<u>1.274</u>	Leu (CUN)	0.5962	1.046
Val	0.916	0.922	Arg (CGN)	0.7596	0.055
			Met	0.7834	0.488
			Sum (χ^2)		34.92
			<i>P</i> -value		0.0395

Table 2.1.2: Results for correlation analysis of ribosomal anticodon-amino acid enrichment and the canonical code are consistent regardless of which random code generator was used. One million random codes were generated from each generator, and Monte Carlo analysis was performed using both the specific and optimal subset of amino acids. The percentage of codes with a lower average enrichment than that of the canonical code is shown.

	Specific Subset	Optimal Subset
RAND	99.9555%	95.1925%
NNY	99.9971%	95.7479%
SYN	99.9996%	97.8008%

Table 2.1.3: Ribosomal anticodon-amino acid enrichment suggests a two-stage evolution of the code. The chronological order of amino acid addition to the code is shown as well as the amino acids that are thought to have been prebiotically available in the Miller- Urey experiments. The amino acids with significant enrichment of anticodon-containing rRNA in ribosome structures are later additions to the genetic code.

Order	Amino Acid	Miller-Urey	Anticodon
1	G	+	
2	A	+	
3	D	+	+
4	V	+	
5	P	+	
6	S	+	
7	E	+	
8	T	+	
9	L	+	+
10	R		+
11	N		
12	I	+	+
13	Q		+
14	H		+
15	K		+
16	C		N/D
17	F		+
18	Y		+
19	M		+
20	W		+

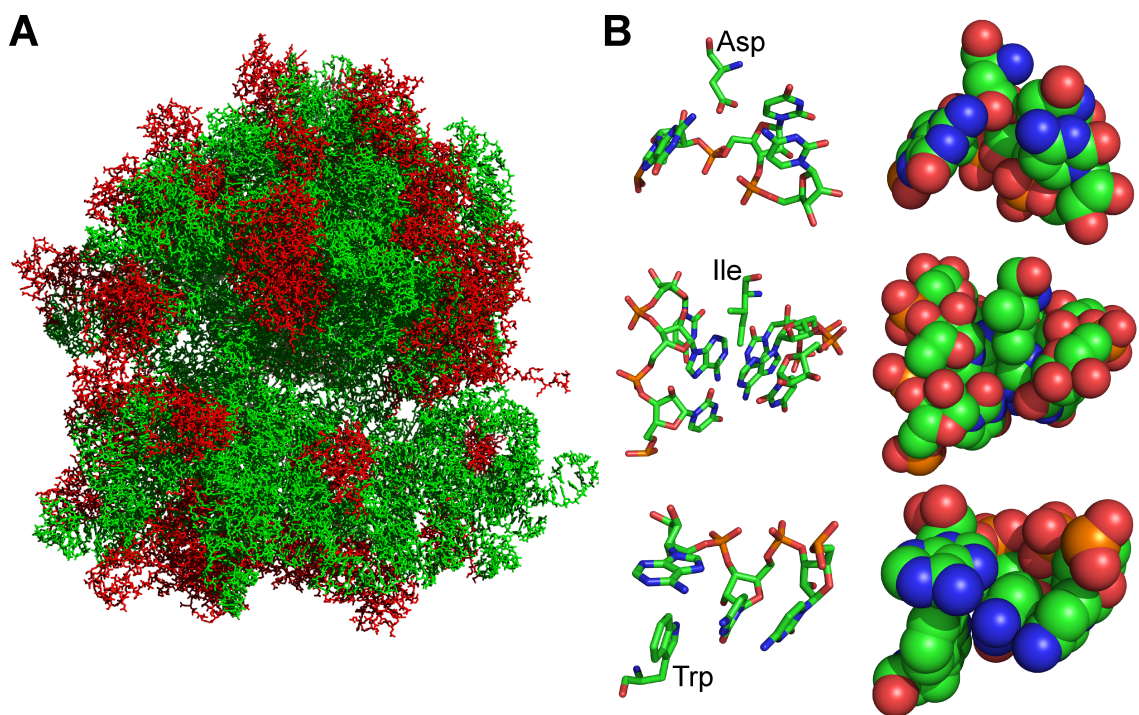


Figure 2.1.1: For a subset of amino acids, codons or anticodons are enriched near their respective amino acids in ribosome structures. **(A)** The structure of *Thermus thermophilus* ribosome with proteins highlighted in red and rRNAs in green (adapted from PDB 2J00 and 2J01, Selmer 2006). **(B)** Examples of amino acid contacting its anticodon rRNA shown in the stick (left) and spherical (right) representation. Top: Asp with GUC; Middle: Ile with 2 UAU; Bottom: Trp with CCA (The top and middle panel are adapted from *Escherichia coli* ribosome, PDB 2QBA, Borovinskaya 2007; the bottom panel is adapted from *Haloarcula marismortui* ribosome, PDB 1VQO, Schmeing 2005).

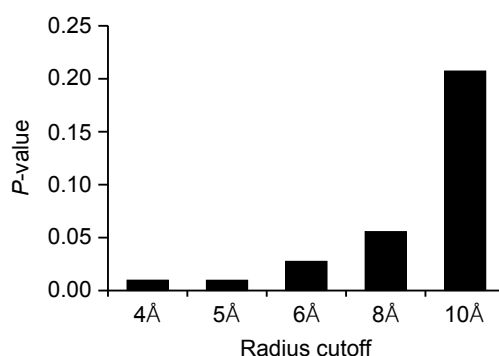


Figure 2.1.2: Determination of the radius cutoff used for all analyses. Amino acid-rRNA interactions in ribosome structures were defined using radii ranging from 4 to 10 Å. Global correlation analysis of ribosomal anticodon-amino acid enrichment and the code was performed on 10^6 random codes using these 5 data sets to determine the *P*-values corresponding to each radius. Similar analysis of ribosomal codon-amino acid enrichment and the code was not statistically significant.

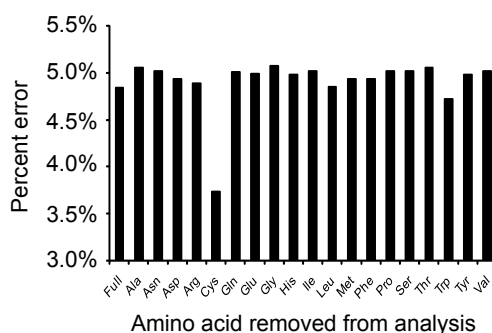


Figure 2.1.3: Inclusion of Cys in the global analysis dramatically increases the deviation of the randomized genetic codes. Each amino acid was excluded from global analysis of ribosomal anticodon-amino acid enrichment individually and the percent error of the average enrichment values from 10,000 codes was calculated. Only the removal of Cys shows a marked effect on the percent error in the populations.

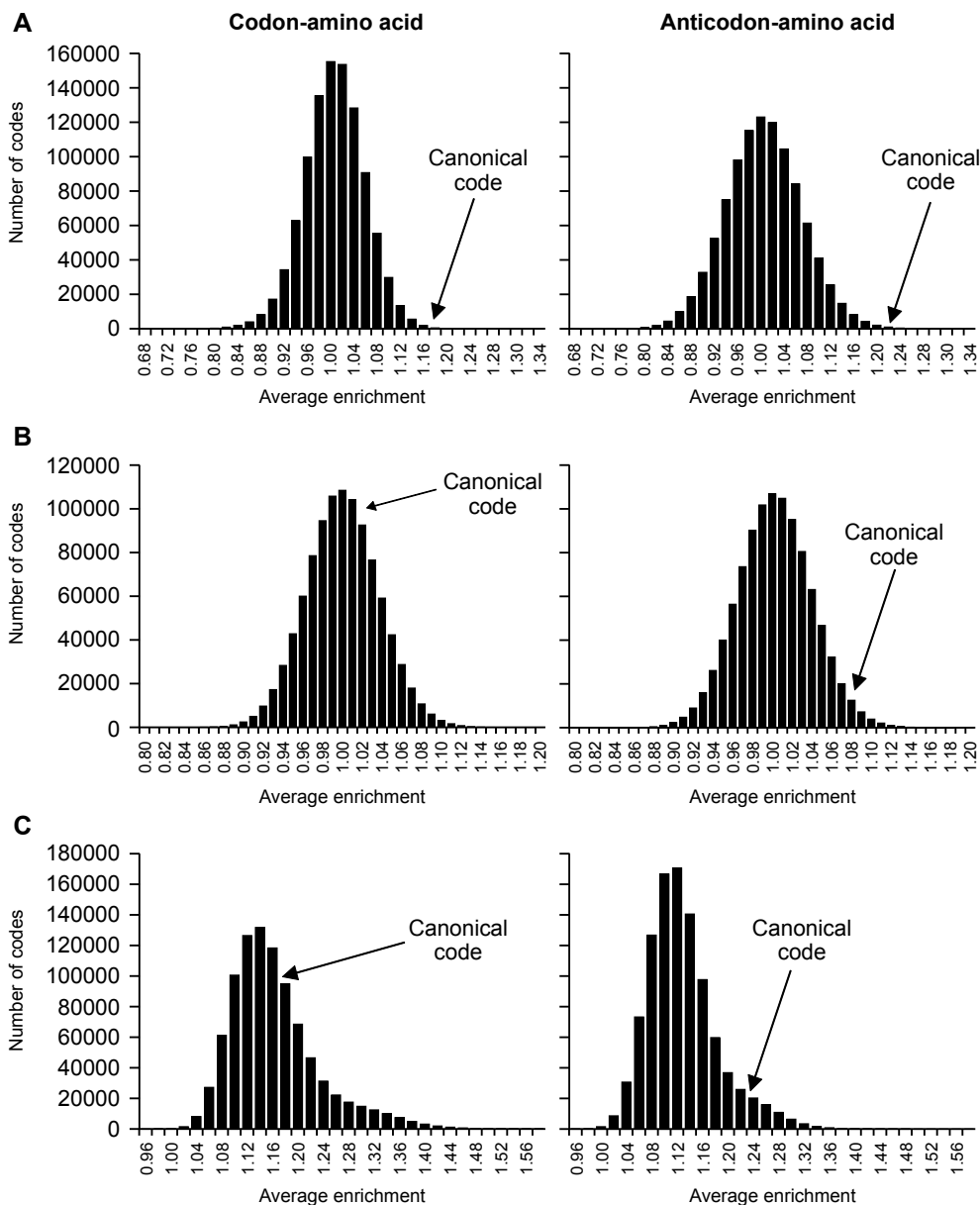


Figure 2.1.4: Anticodon-amino acid enrichment observed in ribosomal structures is correlated with the canonical genetic code. Monte Carlo simulations were used to analyze the canonical code and random codes for codon- or anticodon-amino acid enrichment in the ribosomal structures. The distributions of the average enrichment value of codon (left) or anticodon (right) for 10^6 random codes are shown with the arrow pointing to where the canonical code stands. The correlation analyses were performed using (A) a specific subset of amino acids that is optimal for the canonical code as in **Table 2.1.1B** and C (B) all amino acids except cysteine, and (C) a subset of amino acids optimal for each code.

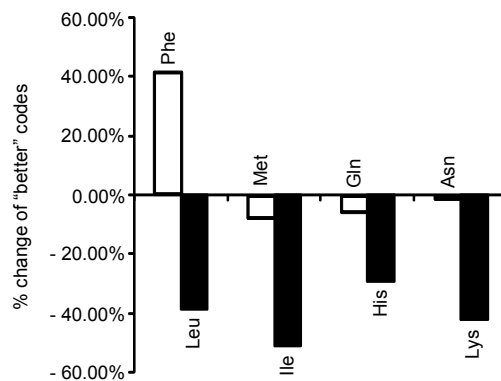


Figure 2.1.5: Ribosomal anticodon-amino acid enrichment reveals codon reassignment during code evolution. Four amino acids in the shared codon blocks (Leu, Ile, His and Lys), when each expanded to occupy the entire codon block, markedly improved the correlation of the code with the ribosomal anticodon-amino acid enrichment. The number of “better” codes, random codes with higher global average enrichment, was significantly reduced. The other amino acid in the shared codon blocks showed no such reduction when expanded.

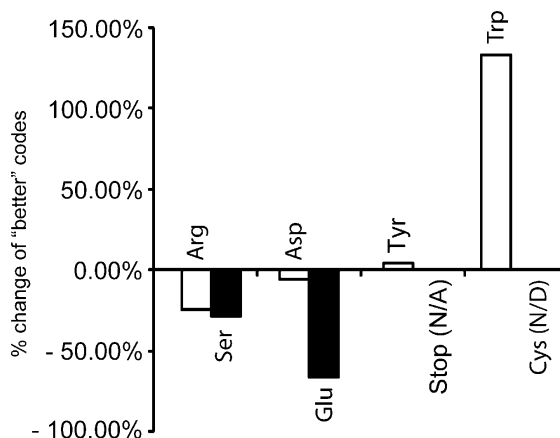


Figure 2.1.6: Codon block expansion analysis of the other four split blocks. There was no difference between Arg and Ser after expansion. The Asp-Glu codon block appeared to show a similar phenomenon as those amino acids in Figure 2.1.5, as Glu had a large reduction in the number of better codes. However, Glu had an extremely low anticodon enrichment value (0.596) (**Table 2.1.1C**), which increased to 0.837 upon Glu expansion to occupy the entire GAN block. A <1 enrichment value suggests no significant enrichment of anticodons near Glu in the ribosomal structures. Therefore, we concluded that the finding described in Figure 2.1.5 was not applicable to the Asp-Glu codon block. Expansion of the stop signal was not applicable because of the nonexistence of a stop entity in proteins, and Cys expansion could not be determined because of its too few presence at ribosomal protein surface.

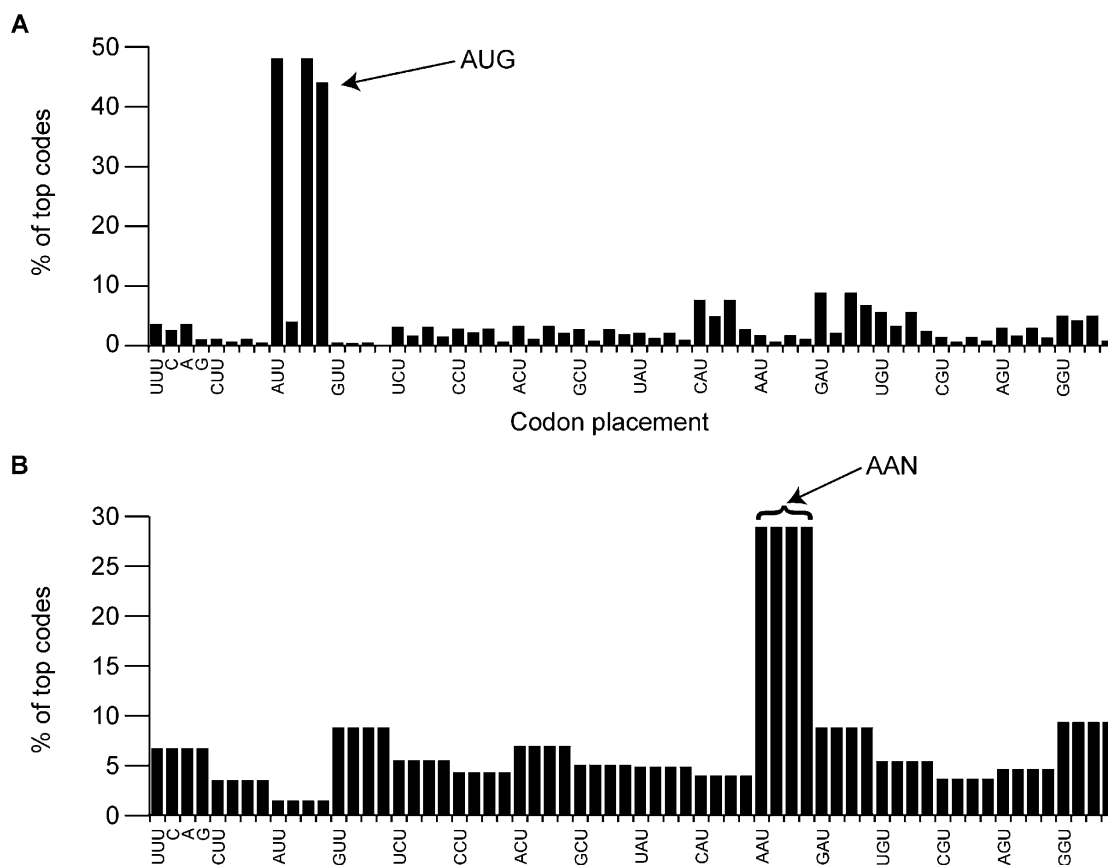


Figure 2.1.7: Analysis of the top 1,000 random codes reveals evidence for codon reassignment during code evolution. The probability of finding an amino acid at a specific location in the 1,000 random codes with the highest global average enrichment values was determined (see the Methods for how Ile and Pro are placed in the random codes). (A) The placement of Ile in the top 1,000 codes reveals a strong preference for the methionine codon AUG because of the strong enrichment of Ile near the corresponding anticodon CAU in the ribosome. (B) The placement of Pro in the top 1,000 codes reveals a strong preference for the AAN codon block due to the ribosomal enrichment of Pro near the UU moiety.

Chapter 2.2: Uncovering the Stereochemical Code

2.2.1: Abstract

The establishment of the genetic code is a critical transition in the evolution of all modern organisms (Szathmary 1995). One of the dominant theories addressing this difficult question is the stereochemical hypothesis, which postulates that the genetic code developed from interactions between nucleotides, particularly triplet codons or anticodons, and their respective amino acids (Dunnill 1966; Pelc 1966; Woese 1966). RNA molecules evolved in vitro to bind various amino acids are enriched in their respective codons or anticodons (Majerfeld 1998; Mannironi 2000; Legiewicz 2005; Majerfeld 2005 p226; Majerfeld 2005 p5482). However, in vivo evidence that these interactions are strong enough to influence the genetic code in a meaningful biological context is lacking. Here we show that RNA-amino acid interactions affect the translation of an *E. coli* strain with ambiguous decoding of isoleucyl codons. Incorporation of isoleucine and valine attached to an otherwise identical tRNA is selectively biased by nearby (± 24 nt) anticodons for respective amino acids. We also found that corresponding anticodons are enriched near isoleucine and valine residues in ribosome structures, suggesting a role of such interactions in ribosome formation and in the RNA to protein transition. Our results demonstrate that anticodon-amino acid interactions still manifest effects in modern cell translation, strongly suggesting their past contribution to codon assignment during code origin and evolution. In addition, our ambiguous codon methodology will enable the determination of all possible RNA-amino acid interactions to elucidate the stereochemical code, revealing the primordial code from which the modern code evolved.

2.2.2: Introduction

Shortly after the elucidation of the genetic code, Woese claimed that the question “why” the genetic code evolved to its modern form is more important than knowing the code itself (Woese 1965). This bold claim is grounded in the fact that the code is the most dominant aspect of life on this planet, and that understanding its roots and development is at the centre of understanding any major evolutionary breakthrough. The universality of the code is also its downfall with regards to studying its evolution, as no intermediate codes exist for comparison. Although multiple hypotheses have been proposed to explain why codons are selectively assigned to specific amino acids (Knight 1999; Yarus 2005), empirical data are rare (Woese 1966 p723; Majerfeld 1998; Mannironi 2000; Legiewicz 2005; Majerfeld 2005 p226; Majerfeld p5482) and direct experimentation on the genetic code in a live organism is out of reach.

If the genetic code evolved through stereochemical interactions of RNA and amino acids (Dunnill 1966; Pelc 1966; Woese 1966), relics of this evolution may be present in modern cells. Provided that these interactions are strong enough to exert an influence in modern organisms, we may be able to uncover the effect by perturbing the code of a modern cell. Evidence of this in nature would provide the first in vivo support that such an interaction could shape the evolution of the genetic code. The determination of the stereochemical code, the compilation of all possible RNA-amino acid interactions, would reveal the nature of a primordial genetic code from which the modern code evolved.

2.2.3: Materials and Methods

2.2.3.1: Synthesis of the DJ.2 *E. coli* Strain

The DJ.2 *E. coli* strain was created from CU.2 (Umbarger 1951) through the replacement of the genomic *IleRS* gene with one harboring an 11-alanine substitution in the editing domain (Pezo 2004) using established procedures (Datsenko 2000). The parental CU.2 strain contains a single nucleotide change in the *ilvE* gene involved in isoleucine synthesis, abolishing its function. To prevent potential reversion of this mutation, the entire *ilvE* gene was replaced with a tetracycline resistance cassette in DJ.2. DJ.2 cells were serially grown over a span of six months to yield the derived DJ.2 p83 strain, which has an increased valine misincorporation rate (see Chapter 3). Cells were grown in glycerol minimal media containing high concentrations of valine (1.8 mM) with minimal isoleucine (9 μ M) to increase the misincorporation ratio of valine in DJ.2.

2.2.3.2: Quantitative Mass Spectrometric Analysis of the Whole DJ.2 Proteome

For quantitative mass spectrometric analysis of the whole-proteome, cells were grown in glycerol minimal medium supplemented with 1.8 mM valine. All proteins from these cells were prepared and profiled using quantitative whole-cell proteome mass spectrometry, which yielded the percent of isoleucine incorporation at each isoleucine codon. Briefly, DH10 β and DJ.2 *E. coli* cells were harvested, washed in PBS buffer and lysed in 250 μ L 2% RapiGest (Waters Corp.) and Hepes buffer (10mM Hepes, 150mM NaCl, pH 7.2) by a Branson Sonifier 450 Ultrasonic Homogenizer. Protein solutions were diluted four times in 50mM Hepes buffer (pH 7.2), and reduced and alkylated using 1 mM Tris(2-carboxyethyl)phosphine (Fisher, AC36383) at 94°C for 5 minutes and 2.5 mM iodoacetamide (Fisher, AC12227) at 37°C in dark for 30 minutes, respectively. The proteins were digested with 1:50 trypsin (Roche, 03 708 969 001) overnight.

iTRAQ derivatization of DH10 β (iTRAQ 114 and 115) and DJ.2 (iTRAQ 116 and 117) was performed after trypsin digestion. Briefly, 50 μ g of each sample was reacted with one tube of iTRAQ reagent (Applied Biosystems) in 70% isopropanol at a pH of 7.2 for 2 hours at room temperature, and then dried in a vacuum concentrator. Peptides were dissolved in water and samples were combined. Samples were incubated at 4 °C overnight with 1% TFA (pH 1.4) to precipitate RapiGest, and centrifuged at 16,100 x g for 15 minutes. Supernatant was collected and centrifuged through a 0.22 μ M filter.

Automated 2D nanoflow LC-MS/MS analysis was performed using LTQ tandem mass spectrometer (Thermo Electron Corporation, San Jose, CA) employing automated data-dependent acquisition. The detailed LC-MS/MS method can be found in our published work (Tanner 2007; Castellana 2008).

The full MS scan range of 400-2000 m/z was divided into 3 smaller scan ranges (400-800, 800-1050, 1050-2000) to improve the dynamic range. Both CID (Collision Induced Dissociation) and PQD (Pulsed-Q Dissociation) scans of the same parent ion were collected and followed by 4 pairs of CID-PQD MS/MS scans of the most intense ions from the parent MS scan. A dynamic exclusion of 1 minute was used to improve the duty cycle of MS/MS scans.

The raw data was extracted and searched using Spectrum Mill (Agilnet, version A.03.02). The CID and PQD scans from the same parent ion were merged together. MS/MS spectra with a sequence tag length of 1 or less were considered as poor spectra and discarded. The rest of the MS/MS spectra were searched against the NCBI (National Center for Biotechnology Information) RefSeq protein database (version 21, January 2007) limited to E. coli (16,324 sequences). The enzyme parameter was limited to full

tryptic peptides with a maximum miscleavage of 1. All other search parameters were set to SpectrumMill's default settings (carbamidomethylation of cysteines, iTRAQ modification, +/- 2.5 Da for precursor ions, +/- 0.7 Da for fragment ions, and a minimum matched peak intensity of 50%). A concatenated forward-reverse database was constructed to calculate the in-situ false discovery rate (FDR). Proteins with shared peptide(s) were grouped together into protein groups. MS/MS spectra were validated using the criteria list in **Table 2.2.1**.

2.2.3.3: Quantitative Mass Spectrometric Analysis of EGFP-Z Proteins

For quantitative mass spectrometric analysis of EGFP-Z proteins, plasmids expressing the EGFP-Z fusion gene were transformed into DJ.2 and DH10 β cells. Cells were grown in minimal media containing 1.8 mM valine to mid-log phase. After 4 hours of induction with Isopropyl β -D-1-thiogalactopyranoside, cells were lysed with BPER (Pierce, Rockford, IL) and protein purified by using nickel-NTA resin (Qiagen, Valencia, CA). Percentage of isoleucine incorporation at the target isoleucine site in the linker region was determined by quantitative mass spectrometry. A two-tailed unpaired t-test was used to determine if mutant sequences had significantly different valine misincorporation compared to the wildtype sequence.

Purified EGFP-Z fusion proteins were digested and iTRAQ labeled using the same protocol as above. iTRAQ labeled peptides were mixed and analyzed by 1-D nanoflow LC-MS/MS using an acetonitrile gradient of 0 to 80% for 120 minutes.

2.2.3.4: DJ.2 Whole-Proteome Data Analysis

All data analysis described below was performed using a series of new Perl scripts with exception to any statistical analyses.

Unusable peptide spectra were first excluded from the data set. Peptides containing a total intensity less than 1000 calculated from all 4 iTraQ markers were considered too low and excluded. An iTraQ marker (iTraQ 118) not used in peptide labeling was used as reference for background intensity. If the total peptide spectral intensity calculated from iTraQ 114-117 was less than 10-fold that of iTraQ 118, the spectrum was considered to have high background and excluded. Finally, each peptide was labeled with two iTraQ markers per strain, and peptides with higher than 25% error between the two markers were excluded.

The remaining peptides were split into populations containing 0, 1, or multiple isoleucine residues. Those containing zero isoleucine residues were compiled for each individual protein present in the array. For each protein, the sum of intensities for individual iTraQ markers (114 - 117) was calculated from all peptides belonged to that protein. An intensity ratio of DH10 β to DJ.2 was then determined from these summations. This set of ratios, compiled for each individual protein using only the non-isoleucine containing peptides was used as the normalization factors for subsequent analysis.

Valine misincorporation was then determined for peptides containing a single isoleucine residue. Peptides containing 2 or more isoleucines were not analyzed, as precise quantitation of valine misincorporation for each residue would be impossible using these methods. A ratio of DH10 β to DJ.2 iTraQ intensities was calculated for each peptide. For the peptides containing multiple spectra in the data set, the ratios for these peptides were calculated individually and then averaged. The average ratio was then used for subsequent calculations. This ratio was normalized using the normalization

factor for the specific protein containing this peptide. This final ratio was used to calculate the isoleucine identity of all 496 isoleucine-containing peptides within the array. Peptides can have greater than 1.00 isoleucine identity due to compounded error from the calculations.

Once isoleucine identity was calculated for each peptide, genomic nucleotide sequences corresponding to each isoleucine-containing peptide sequence were obtained, as described in text. All statistical analyses employed the nonparametric Mann-Whitney test, as these populations were non-Gaussian as determined from a D'Agostino and Pearson omnibus test. Significance threshold was set at $P < 0.01$ to limit potential false positives.

2.2.4: Results

2.2.4.1: The Ambiguous *E. coli* Strain DJ.2 Shows Non-Normal Distribution of Valine Misincorporation Across the Proteome

To investigate the effect of RNA-amino acid interaction on translation in live cells, we developed an ambiguous codon strategy (**Figure 2.2.1a**). An otherwise identical tRNA was charged with two different amino acids, and contextual effects of surrounding mRNA sequence on the incorporation of the amino acid at codons cognate to the tRNA was determined. We constructed an *E. coli* strain (DJ.2), in which the isoleucyl-tRNA synthetase (IleRS) gene was replaced with a mutant IleRS harboring mutations in the synthetase editing domain. These mutations abolish the editing activity of IleRS, leading to mischarging of isoleucyl-tRNAs (tRNA^{Ile}s) with valine in vivo (Pezo 2004; Hendrickson 2000 p8180). During translation the probability of utilizing an isoleucine-charged tRNA^{Ile} (Ile-tRNA^{Ile}) over a valine-charged tRNA^{Ile} (Val-tRNA^{Ile})

should be similar for all isoleucyl codons. However, quantitative whole-cell proteome profiling of this strain using mass spectrometry showed a broad distribution of valine misincorporation at isoleucyl codons with many peptides enriched in either valine or isoleucine (**Figure 2.2.1b**). This result suggests that Ile-tRNA^{Ile} and Val-tRNA^{Ile} are used selectively at different isoleucyl codons during translation.

2.2.4.2: Valine Misincorporation is Dependent Upon RNA Context

To determine if RNA context could potentially bias the aa-tRNA^{Ile} selection, we searched 24 nucleotides on both the 5' and 3' flanks of each isoleucine codon for every possible 3, 4, and 5-nucleotide sequence moiety in all frames. Populations containing one or more instances of a given sequence moiety were compiled, and the median isoleucine incorporation was determined from the whole-cell proteome profiling. Six moiety populations showed a significant shift in median isoleucine incorporation compared to the entire population (**Figure 2.2.1c**). Moieties UUA, ACAC and UGGUU correlated with higher valine incorporation while moieties CAUG, GCCCG and GAUCC with higher isoleucine incorporation. Interestingly, moieties ACAC, CAUG and GAUCC contain anticodon elements for the enriched valine (anticodon NAC) and isoleucine (anticodon NAU), respectively. Also, moieties UGGUU and GAUCC contain codon elements for the enriched valine (codon GUN) and isoleucine (codon AUH), respectively. These data suggest that the presence of short RNA moieties near an isoleucyl codon can somehow bias the incorporation of either isoleucine or valine at the isoleucine site, and the RNA moieties are dominated by anticodon/codon of the enriched amino acid.

The incorporation bias cannot be explained by known mechanisms involving amino acids or mRNA context with the translational apparatus. The identity of the esterified amino acid on aa-tRNA can contribute to the binding of aa-tRNA to EF-Tu and to the ribosomal A site (Dale 2004; Dale 2009), which could lead to the preferential incorporation of one amino acid over another. However, the reported differences between isoleucine and valine recognition by EF-Tu is small (Dale 2004), and we observed that both isoleucine and valine were preferred by their corresponding moieties. The incorporation bias should not be caused by isoleucine or valine reaction rate preference for peptide bond formation, as there was no correlation between the incorporation bias and the P-site amino acid identity. Consistently, it has been shown that peptide bond formation with different peptidyl- and aminoacyl-tRNAs takes place at similar rates regardless of the amino acid (Wohlgemuth 2008). The existence of the RNA moieties greater than 6 nt away (**Figure 2.2.2**) from the target isoleucyl codon negates the role of codon-pair or triplet bias, which explains codon pairing preferences limited to only one or two codons away (Buchan 2006; Moura 2007). The relatively small size of these moieties excludes the possibility of forming major secondary structures to selectively deliver one aa-tRNA into the ribosome (Berry 1991). Finally, the finding that these moieties can be out of frame, upstream, and downstream (**Figure 2.2.3**) argues against nascent peptide effects or resultant reading-rate differences between Ile-tRNA^{Ile} and Val-tRNA^{Ile}.

We propose that the incorporation bias is due to a direct stereochemical interaction between an RNA moiety and the cognate amino acid (**Figure 2.2.4**). This interaction causes a rise in local concentration of the amino acid **1**, which in turn

enhances nearby charging of tRNA^{Ile}s with this amino acid by the mutant IleRS **2**. Subsequent binding to EF-Tu **3** and A-site selection **4** would carry this bias from the tRNA^{Ile} to the ribosome and thus to the protein. Direct interaction between the RNA moiety and the aa-tRNA is less likely, because the aminoacylated CCA end of the aa-tRNA is bound by EF-Tu making the esterified amino acid largely inaccessible (Nissen 1995). This model explains how the identified moieties could influence translation both up and downstream, in or out of frame, and at various distances. It has been suggested that when ribosomes pack in polyribosomes on mRNA, an unacylated tRNA released from the E site of the preceding ribosome can increase the local concentration of cognate tRNA for the adjacent ribosomes (Yusupova 2001). Local enrichment of cognate amino acids by RNA moieties can synergize with the cognate tRNA to facilitate in situ aminoacylation and selection of correct aa-tRNA into the ribosomal A site, which can modulate the translation of certain mRNAs.

2.2.4.3: The Introduction of Anticodon-Containing RNA Biases Valine

Misincorporation in an Exogenous Peptide

The whole-proteome translation data suggest that anticodon-containing RNA moieties selectively interact with amino acids. To validate this and our proposed model, we placed the identified anticodon-containing RNA moieties near an isoleucyl codon in an exogenous gene and determined their influence on the translation of the target isoleucine codon in the DJ.2 strain. We constructed a gene to express the enhanced green fluorescent protein (EGFP) fused to the Z-domain of protein A via a flexible linker region containing a single isoleucyl codon (**Figure 2.2.5a**). The RNA moieties were introduced into the linker near the isoleucyl codon to test their effect. These constructs were

expressed in the DJ.2 strain, and purified proteins were analyzed with quantitative mass spectrometry using corresponding proteins expressed in the unambiguous DH10 β strain as control. The EGFP-Z protein purified from DJ.2 contained 90.6% isoleucine at the isoleucyl codon in the linker region. As predicted, the introduction of ACAC, which contains the valine anticodon, markedly increased valine misincorporation and thus decreased the isoleucine content to 61.3% (**Figure 2.2.5b**). In addition, when the isoleucine codon-containing CAUG or GAUCC moiety was introduced, the isoleucine content increased to 100% and 98.1%, respectively. The target isoleucine site is at a flexible linker flanked by two stable proteins. More importantly, the GAUCC was introduced without changing the peptide sequence. Therefore, the incorporation bias detected here excludes the possibility that the preference for isoleucine or valine is due to protein stability or turnover. These data corroborate the whole-proteome results and demonstrate that anticodon-amino acid interaction affects protein translation *in vivo*.

2.2.5: Conclusions

This work argues the existence in modern cells of relics from a stereochemically-evolved genetic code. The surprising finding that interactions between specific RNA moieties and amino acids are robust enough to influence modern translation strongly suggests they could have played active roles in determining the initial form and subsequent development of the genetic code. Given the unique role of the ribosome in bridging the RNA and protein world (Crick 1968; Gilbert 1986), selective RNA-amino acid interactions observed in ribosome structures may be remnants of a time when such stereochemical interactions played an increasingly important role and contributed directly to the establishment of ribosome and translation. That identified RNA moieties were

dominated by anticodons for the enriched amino acid supports the tenet of the stereochemical hypothesis (Dunnill 1966). The ambiguous codon methodology described here should enable the expansion of this work to other amino acids and organisms, which will further define the scope of these interactions and establish a complete stereochemical code. A stereochemical code will give us a previously unapproachable glimpse into the primordial code, and afford a starting basis to investigate the code expansion and reassignment, critical processes required for life to crawl out of the RNA world.

Table 2.2.1: Filtering criteria for autovalidation of database search results.

	1+ peptide	2+ peptide	3+ peptide
Single peptide hits	>12.5	>13.4	>18.4
Two or more peptide hits	>7.8	>7.3	>10.4

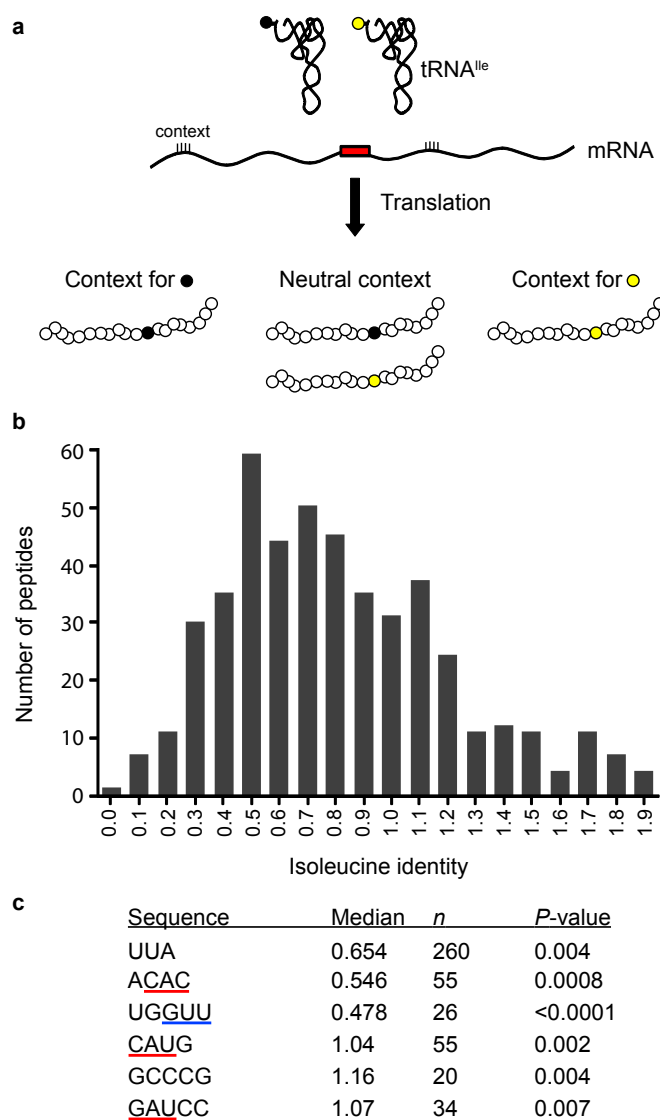


Figure 2.2.1: RNA moieties bias the incorporation of isoleucine or valine at ambiguous isoleucyl codons throughout the DJ.2 proteome. **a**, An ambiguous codon strategy to investigate potential RNA-amino acid interaction *in vivo*. A tRNA is charged with two different amino acids (black and yellow circle) to decode its cognate codons (red box). If RNA moieties (hash marks) affect the translation of the ambiguous codons, preferentially incorporated amino acid is identified to reveal the correlation. **b**, Isoleucine identity varies markedly at the isoleucyl codons ambiguously coding for isoleucine and valine in 496 peptides from the *E. coli* DJ.2 proteome. **c**, Moieties found to enrich isoleucine or valine at the ambiguous isoleucyl codons in the DJ.2 whole proteome mass spectrometric data. The median of isoleucine identity of the whole population is 0.778. Anticodons for the enriched amino acid are underlined in red and codons in blue.

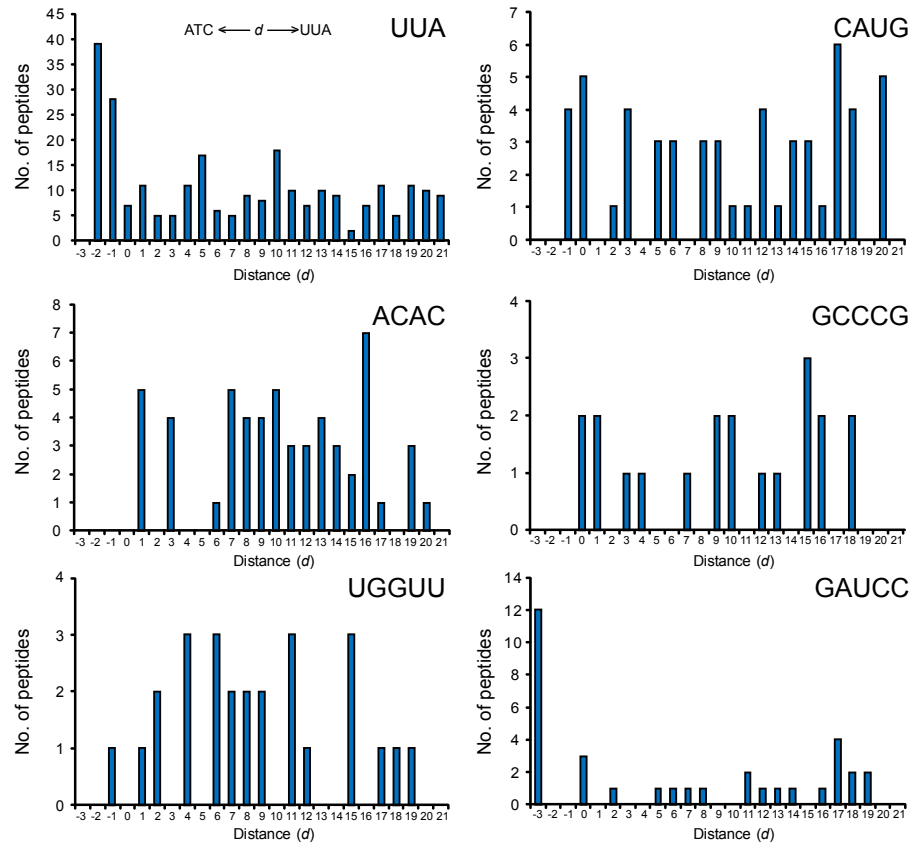


Figure 2.2.2: Distribution of significant moieties in the DJ.2 proteomic data set is independent of distance. The distribution of moiety-containing peptides based on distance from isoleucine codon is shown. Distance (d) is measured as number of nucleotides between the isoleucine codon and the moiety. Upstream and downstream distances were combined. The relatively high usage of UUA at a distance of -2 and GAUCC at a distance of -3 is because each moiety contains an isoleucine codon ($[AUU|ANN|$ and $[NNG|AUC|CNN|$).

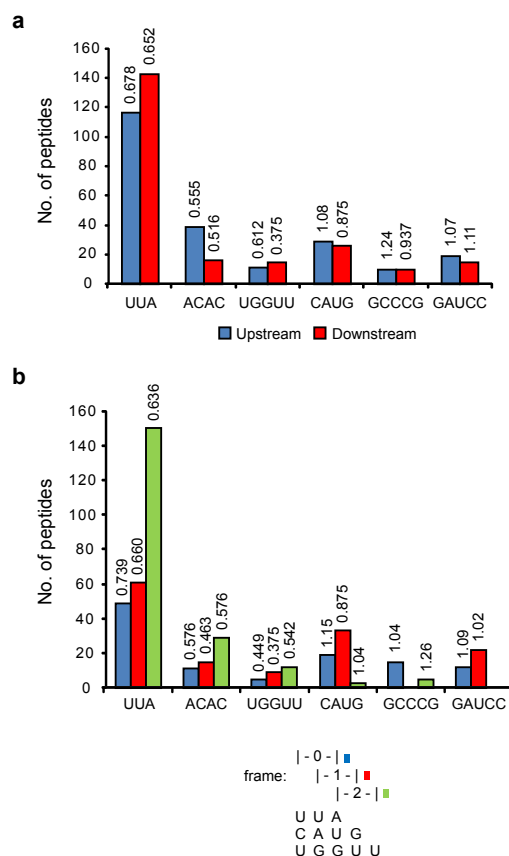


Figure 2.2.3: Distribution of significant moieties in the DJ.2 proteomic data set is independent of position and frame. **a**, Number of peptides containing each moiety found upstream or downstream of the isoleucine site. Medians of each population are shown above bars. **b**, Moieties are found in all three reading frames. CAUG is used less frequently in reading frame 2 as it introduces a UGN in frame, which codes for relatively rare cysteine, tryptophan or stop. UUA usage in frame 2 is dominated by the large number of instances in which the isoleucine codon AUU itself is followed by an A ([AUU|ANN]). Medians are shown above graphs.

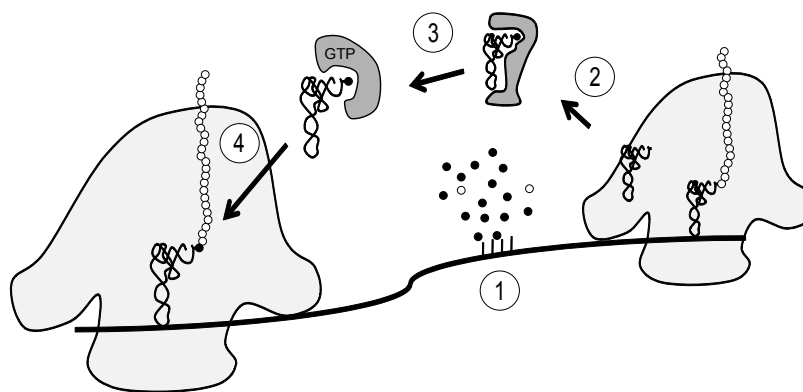


Figure 2.2.4: Stereochemical interaction between an RNA moiety and an amino acid favors the incorporation of the amino acid. Numbered steps are referred to in text.

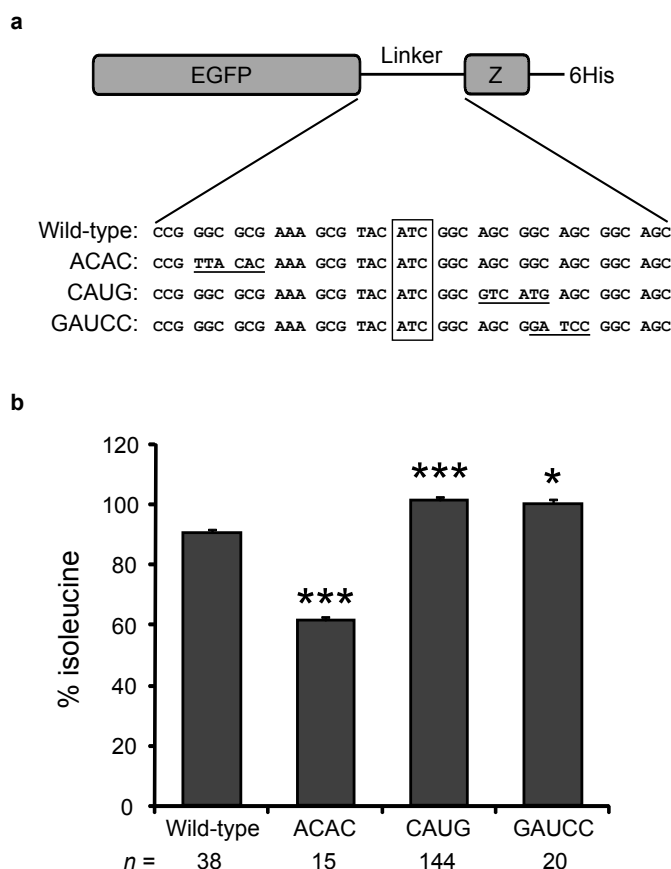


Figure 2.2.5: Introduction of identified anticodon-containing RNA moieties biases amino acid incorporation as model predicts. a, EGFP-Z protein construct with the wildtype and mutant sequences shown. The target isoleucyl codon (boxed) in the linker region is centred in the sequence. Introduced moieties are underlined. A GU dinucleotide mutation was made before CAUG and a UU before ACAC to keep the stability of the mutant peptide linker close to the wildtype linker. GAUCC could be introduced without amino acid change. b, Isoleucine content determined by quantitative mass spectrometry for the target isoleucine site in the linker region. Error bars show s.e.m. P -value = 0.018 for * and < 0.0001 for ***.

Chapter 2.1 is a reprint of the material as it appears in the Proceedings of the National Academy of Sciences, USA 2010. Johnson DBF and Wang L (2010) Imprints of the genetic code in the Ribosome. *Proc Natl Acad Sci* 107:8298-303. The dissertation author performed all analyses and experimentation and was the author of the paper.

Chapter 2.2 is a reprint of material being prepared for publication. Johnson DBF, Shen Z, Briggs SP and Wang L. Uncovering the Stereochemical Code. Zhouxin Shen performed the mass spectrometry analyses on for both the full-proteome and EGFP-Z protein. The dissertation author performed all other computational analyses and experimentation. The dissertation author is the author of the paper.

Chapter 3: The Engineering of Variant Genetic Codes in Modern Organisms

This chapter moves from the establishment of the canonical code in the RNA world described in chapter two into the subsequent (and modern) evolution of the genetic code. Natural variations from the universal genetic code were discovered early and are becoming more numerous with the advancement of genome research. The discovery of these organisms was antagonistic to the “frozen accident” (Crick 1968) theory and opened up many areas of research into how and why these changes may have occurred. The general principal underlying the “frozen accident” (that changes to the genetic code in a modern or established organism would result in hundreds or thousands of potentially damaging mutations in the proteome and thus be deleterious) is still sound, and much of the research spawned from these discoveries attempted to understand just how an organism can cope with such dramatic changes.

With these organisms now available, one might think that these questions would be more approachable to evolutionary and molecular biologists. Unlike the “notoriously difficult” problem regarding the establishment of the code itself, these questions have systems readily available (Crick 1976). However, these issues are proving to be just as difficult to answer and are still marred in conjecture and theory. The source of this dilemma is in that organisms harboring these non-standard genetic codes are at an endpoint. Only one natural example exists of an organism that is still in the process of altering its genetic code (Suzuki 1997) and thus can give us insights into the numerous questions regarding this dramatic alteration. This field will remain in its current state without further systems to study, specifically those in which the genetic code is in a ambiguous or altered state.

That is where synthetic and engineering-based approaches come in. Engineering-based approaches would allow for the construction of organisms undergoing these changes. In fact, numerous labs have used these techniques to study the survival mechanisms of cells in an ambiguous state (Pezo 2004; Bacher 2005; Bacher 2007 p1907; Bacher 2007 p6494) demonstrating just how permeable the genetic code truly is.

This chapter focuses on synthetic engineering of modern organisms, such as *E. coli*, to generate strains useful in evolutionary studies and previously unavailable in nature. Much of this work can be split into two distinct categories: those that use *E. coli* as a model for understanding natural events that have helped shape the modern code, and those that use it as a model or technological tool for the artificial expansion of the genetic code. The work described in this chapter attempts to create and study such a system for the first category, while work focusing on the latter, more technologically based, is described in Chapter 4.

Chapter 3.1 provides an overview of the various methods for engineering variant genetic codes in modern cells. Specifically it focuses on global alterations brought out by a variety of methods. Unnatural amino acid mutagenesis will not be covered in this section, as it is covered in detail in Chapter 4.1.

Chapter 3.2 describes our efforts to generate an *E. coli* strain that replaces isoleucine residues with valine throughout the entire proteome. In the end, these efforts did not lead to our desired final product, however the resultant strain and adaptation principals may be beneficial for the design of further experimentation. In addition, the resultant strain, DJ.2, was used for the studies previously described in Chapter 2.2.

Chapter 3.1: Engineering the Genetic Code

3.1.1: Abstract

Structural, functional, and genetic studies of the genetic code have been ongoing since its elucidation nearly 6 decades ago. The precise mechanisms underlying the evolution of the genetic code itself, however, remain a mystery. The root of this problem lies in the universality of the code; no primitive or intermediate species exist for comparison. Synthetic and engineering based approaches can bypass this roadblock through the synthesis of such organisms. This subchapter focuses on the numerous efforts to create synthetic organisms for evolutionary, synthetic, and molecular biology. Biologists have been rewriting and expanding the genetic code of organisms for decades. At first this process relied on organism's inherent promiscuity, however, understanding of the mechanisms underlying such global changes have allowed more precise alterations. The use of the auxotrophic, mutated, and ambiguous strains described here have yielded organisms with substantial alterations to their genetic codes, including complete replacement of canonical amino acids with a structurally related unnatural amino acid. These studies have yielded otherwise rare empirical evidence regarding the continued evolution of the genetic code. Further engineering and research into the resultant organisms will lead to valuable insights into all aspects of the evolution of the genetic code and the adaptive process underlying such drastic changes.

3.1.2: Introduction

All living organisms are confined to the same 20 amino acids as building blocks for the construction of proteins. These canonical amino acids, while sufficient for a wide range of diverse life, are limited in their chemical or structural properties and

functionality. This limitation is seen in modern organisms that have evolved numerous methods of adding various functional groups by post-translational modification of amino acids or by using cofactors (Uy 1977). The number of substantial alterations directly encoded in the genetic code remains small and are currently limited to selenocysteine and pyrrolysine, which require dedicated elements for incorporation and thus are not universal (Ambrogelly 2006). Evidence exists for the presence of many non-proteinogenic amino acids in the prebiotic world (Miller 1976). Some amino acids such as α -aminobutyrate are present in very high levels in the classic spark-tube experiments performed by Miller (1953). Although many of these amino acids are structurally or functionally similar to a canonical amino acid, it is unclear why life chose to exclude them from modern translation, especially if their relative abundance was higher than other canonical amino acids. Why is that the 20 canonical amino acids in particular were chosen through Darwinian evolution? This long-standing question lies at the heart of the debate as to how the genetic code evolved to its current form.

This question is also inherently difficult to address directly, as no intermediate organisms exist today that can help to elucidate how the modern genetic apparatus was evolved. Perhaps such organisms can be recreated through engineering efforts. The creation of these organisms may help to recapitulate the process of evolution inside a testable environment, granting the rare empirical evidence pertaining to the establishment of the genetic code. Engineering of a variant genetic code, on the most basic level, involves altering the amino acid repertoire that an organism has at its disposal. This can be a replacement of a canonical amino acid with an unnatural analog, shuffling of the canonical amino acids, or an expansion of the repertoire through the addition of an

unnatural amino acid that will not affect the utilization of common amino acids. In a way, these methods could mimic the probable natural expansion of a primordial code to the current form or the evolutionary paths discovered in nature regarding the addition of amino acids such as pyrrolysine. Therein lies the utility of these systems and experiments: they can simultaneously create novel synthetic life forms while helping to elucidate the mysteries of the evolution of the core genetic apparatus. This approach is extremely powerful, as the interplay between synthetic and evolutionary biology is cyclical: any major findings from one avenue of research can greatly facilitate research in the other. In addition, these engineering efforts can help to answer such questions as to how malleable the modern genetic code is and how receptive it is to new amino acids, thus forecasting the future of the genetic code.

3.1.3: Challenges in Engineering Unnatural Genetic Codes

There are two major obstacles that need to be overcome in order to engineer variant genetic codes. The first is an engineering obstacle related to the infiltration and modification of the endogenous translational machinery. In recent decades, much effort has been placed in understanding how the genetic apparatus maintains a strict fidelity in translation. A critical step in maintaining specificity hinges on the function of aminoacyl-tRNA synthetases (herein called synthetases) (Woese 2000). These enzymes are responsible for charging the correct amino acid onto its corresponding tRNA, a process that is performed with great accuracy even amongst amino acids that differ by as little as a single methyl group (Ibba 2000). The high substrate specificity of synthetases turns out to be the bottleneck in engineering variant genetic codes, as other components and processes in protein biosynthesis appear to be less stringent. It has been known for

almost five decades that misacylated tRNAs can be active in protein translation (Chaperville 1962). The ribosome itself is remarkably flexible as to what amino acids it allows in translation, with many different unnatural amino acids being successfully incorporated to date (Hohsaka 2002; Mendel 1995; Xie 2005). More recent work has even described mutations in the rRNA that relax the ribosome's specificity to allow D-amino acids (Dedkova 2003). Elongation factor-Tu (EfTu), which delivers acylated tRNAs to the A-site of the ribosome, is also promiscuous as its level of discrimination seems to be small in comparison to that of the synthetases (LaRiviere 2001).

The second and probably larger obstacle is whether an organism can cope with an variant genetic code. Proteins have evolved for billions of years under the stringency of the modern genetic apparatus. Major changes may prove difficult, as folding and function of proteins can be severely impaired. While proteins have reasonably high tolerance for mutations at single sites (Hellenga 1992; Zabin 1991), a variant genetic code will result in many mutations per protein and 1000's of global changes to the proteome, thus imposing an enormous challenge to the host organism. Nature suggests that it is possible, as numerous organisms have been discovered that survived such drastic changes (Knight 2001 p49).

3.1.4: Infiltrating the Genetic Code Through Synthetase Promiscuity

Even when the understanding of the genetic code was in its infancy, it was found that unnatural analogs of amino acids could be mistakenly incorporated into proteins. The physiochemical aspects of the code allowing these infiltrations were not clear until many years later. It is now known that many synthetases are somewhat promiscuous for close analogs of the cognate amino acid (Budisa 2004). This may be partly due to the

rarity of these unnatural amino acids, as selection against misincorporation would have never been required in the absence of these amino acids. When these unnatural analogs are added to the growth media, they infiltrate the translational apparatus and are incorporated into proteins at sites for the cognate amino acid. Such replacement could disrupt protein function to the detriment of the organism, explaining the antimicrobial effects of numerous canonical amino acid analogs (Kohn 1990; Lemeignan 1993; Richmond 1962).

Exploiting the inherent promiscuity of the synthetases enables the global replacement of one common amino acid with an unnatural analog that is structurally close. An obvious hurdle is that it is impossible to make full replacement of critical amino acid residues in essential genes. The extent of incorporation of the unnatural analog varies at different sites – only in rare cases can quantitative substitution be achieved since it is difficult to completely deplete the endogenous amino acids inside the cell. However, these early experiments laid the foundations for future engineering methodologies. Retrospectively, they also gave a first glimpse into the malleability of the genetic code, suggesting that it may not be as “frozen” as once thought (Crick 1968).

3.1.5: The Use of Auxotrophic Bacterial Strains

Evidence that the genetic code is flexible came as early as the 1950's when unnatural amino acid analogs were found to infiltrate the translational apparatus of rats (Levine 1951), *Tetrahymena* (Gross 1955), and later *Staphylococcus aureus* (Richmond 1959). Numerous other unnatural analogs were found to be inhibitory to the growth of *E. coli* (Cohen 1959). It was thought at the time that the inhibitory effect was due to the incorporation of these analogs into proteins and thus abolishing their functions, a

speculation later found to be correct (Richmond 1962). These studies were the first to show that the established genetic code could be altered. However, the overall theme from these experiments was that these changes were too toxic to cellular growth and thus full-scale misincorporation of the unnatural analog was improbable.

A major breakthrough came in 1957 when methionine (Met) was globally replaced by selenomethionine (SeMet) in *E. coli* cells (Cohen 1957). In this experiment, an *E. coli* strain auxotrophic for Met was used to completely replace Met with SeMet. The cells were able to fully replace their requirement for Met with SeMet while maintaining exponential, albeit slower growth than if grown in Met alone. These results were critical for several reasons: it demonstrated that cellular viability was not drastically harmed by a replacement of an amino acid that has more than 38,000 instances in the genome (Nakamura 2000), it showed the first example of the experimental design required for such a replacement, and it demonstrated the flexibility of the methionyl-tRNA synthetase (MetRS). MetRS has been further exploited for the incorporation of a large set of unnatural amino acids (Kiick 2000; van Hest 1998). This work has also had a large impact on the broader biological community as it helps solve the phase problem in X-ray crystallography (Hendrickson 2000 p637), demonstrating that such engineering experiments can hold tremendous breakthroughs to other areas of biological studies.

The next few examples of global replacement used fluorinated amino acids. Fluorine substitution of hydrogen should minimally alter the size and structure of the amino acid, but will greatly change its electronic properties. In one experiment, a Leu-auxotrophic *E. coli* strain was adapted to grow in trifluoroleucine alone, with complete replacement of leucine in the proteome (Rennert 1963). This experiment was considered

to be feasible because of the use of the functionally inert leucine. Later, fluorinated tryptophan (FTrp) analogs were shown to be specific inhibitors of *E. coli* growth, due to their detrimental incorporation into proteins (Browne 1970). These analogs were also found to support growth of Trp-auxotrophic *E. coli* strains grown in limiting amounts of tryptophan. Global misincorporation rates in these experiments were calculated to be around 75%, nearing complete reassignment (Browne 1970, Pratt 1975).

A procedure names Selective Pressure Incorporation (SPI) limits the auxotrophic strain method to incorporate unnatural amino acids into specific proteins rather than globally, minimizing deleterious effects on the host cell (Minks 2000; Yoshikawa 1994). The SPI method utilizes the same basic principles as described above, but the misincorporation is limited to a single target protein through more controlled fermentation techniques. Specifically, an auxotrophic host strain carrying an inducible plasmid with the protein of interest is grown in limiting amounts of the canonical amino acid until growth arrest is reached (**Figure 3.1.1**). Upon induction of the plasmid, an amino acid analog is added to the media to replace the now limited canonical counterpart. This unnatural analog hijacks the host-cell machinery and is incorporated at positions coding for the depleted canonical amino acid. The host cells machinery is mostly unaffected by this as the cell is relying on *older* proteins containing the canonical amino acid, while the newly made protein contains the unnatural amino acid at a high level.

A variety of unnatural analogs of tyrosine, methionine, tryptophan, phenylalanine, leucine, and isoleucine have been incorporated into proteins using this approach and applied to protein structural and functional studies (Minks 2000; Link 2003). Work has also being done to specifically target these replacements to a single codon using a codon-

specific tRNA (Kwon 2003). Changes made by SPI are only transient and apply to the protein of interest, and thus no persistent variant genetic code can be established in the cell. Nonetheless, knowledge on how these unnatural amino acids infiltrate the translational apparatus and affect protein stability and folding can facilitate the future design of unnatural organisms.

3.1.6: Laboratory Evolution

Tryptophan is considered a good target for replacement due to its relatively low use throughout the proteome (~20,000 in *E. coli*) and that its role in catalysis is limited (Nakamura 2000). Fluorinated tryptophan (FTrp) would have nearly insignificant changes to its structural properties, but would alter the electrostatics and potentially cause toxicity. In an ambitious experiment, Wong thought that a species of bacteria would be able to fully substitute tryptophan if given selective pressure to do so (Wong 1983). A Trp-auxotrophic *B. subtilis* strain was grown in plates containing various concentrations of 4-FTrp. Larger colonies were selected, mutagenized, and then negatively selected for growth on the canonical Trp. After two rounds of this selection, the HR15 strain was isolated that grew exponentially in 4-FTrp but with linear and inhibited growth in the canonical Trp. To date, no other experiment of this sort has been reported that achieves complete substitution coupled to a reversal in preference between a canonical amino acid and its unnatural analog. For example, a similar experiment was later attempted in *E. coli* (Bacher 2001). After a similar serial growth and evolution experiment with an *E. coli*, the preference for tryptophan over 4-FTrp persisted. Whole-cell proteomic characterization was not possible when the classic Wong experiment was performed, however a latter report shows that proteins in these cells exhibit a lack of fluorescence

(Bronskill 1988). As 4-FTrp has a loss of fluorescence in comparison to the canonical Trp, this is highly suggestive that the replacement is global.

This pioneering work demonstrates that the once thought “frozen” nature of the code is actually flexible in an experimental setting, and that the preference for an amino acid could change as well. This work also provides support for Wong’s coevolution theory of the genetic code. This theory postulates that as the ability to synthesize new amino acids arrived to more primitive organisms, so did their incorporation into the genetic apparatus (Wong 1975). Specifically, this could have been facilitated through the hijacking of already present synthesis and translation machinery if the new amino acid was structurally or biosynthetically similar to an already existing amino acid. In this case, the unnatural amino 4-FTrp represents a “new” amino acid invading the existing Trp machinery. With mutational selection possibly mimicking a selective pressure in nature, this new amino acid displaces the old one and alters the genetic code of that organism. This work also pioneered the combination of using auxotrophic strains with laboratory evolution and selection techniques, expanding the research of the genetic code to new avenues. In fact, many of the modern techniques for engineering unnatural codes rely on laboratory evolution at some point in the process.

3.1.7: Engineering Ambiguity into the Genetic Code

Once it was understood how unnatural analogs entered into modern translational machinery, more direct, rational methods were used to introduce changes into the genetic code of modern day organisms. Global changes to the genetic code can be made through direct alteration of the endogenous cellular components involved in protein biosynthesis. With these engineering based approaches, evolutionary biologists can address previously

unapproachable questions, such as how to capture a codon by natural or even unnatural amino acids and what role ambiguity played and continues to play in the shaping of the genetic code. Unnatural amino acids that deviate from the cognate amino acid can infiltrate the genetic code after mutation and relaxation of synthetase specificity. Alternatively, tRNAs can be mutated to decode non-cognate codons, resulting in the incorporation of non-cognate amino acids at these codons.

Relying on the inherent promiscuity of the otherwise highly specific synthetases severely limits substitutions to close structural analogs. Changes as small as one atom could prevent the misincorporation. For example, the prolyl synthetase misincorporates thiaproline but not selenaproline (Budisa 1998). To overcome this limitation, numerous efforts have been made to relax the substrate specificity of the aminoacyl-tRNA synthetase. The A294G mutation in the *E. coli* PheRS increases the size of the substrate-binding pocket and results in the acylation of tRNA^{Phe} with *p*-chlorophenylalanine (*p*-Cl-Phe) (Ibba 1994). An *E. coli* strain harboring this mutant PheRS allows the substitution of phenylalanine with *p*-Cl-Phe in firefly luciferase (Ibba 1995). The incorporation of this unnatural amino acid into the luciferase almost completely abolished luciferase activity, suggesting that *p*-Cl-Phe is toxic to cells in such a way as to disrupt protein function. This mutant PheRS was later used to incorporate *p*-bromo-, *p*-iodo-, *p*-cyano-, *p*-ethynyl-, and *p*-azidophenylalanine as well as pyridylalanine (Kirshenbaum 2002). Similarly, an F130S substitution near the amino acid binding site of *E. coli* tyrosyl-tRNA synthetase allows azatyrosine to be incorporated more efficiently than tyrosine (Hamano-Takaku 2000).

Strains harboring these synthetases with a relaxed substrate affinity and grown in media containing both the canonical and unnatural amino acid, have a population of proteins containing a statistical distribution of misincorporation to be produced. This imposes ambiguity on the codon block for the cognate amino acid, potentially mimicking a theorized path for the evolution of the genetic code proposed in the “ambiguous intermediate” hypothesis (Schultz 1994). This avenue of investigation into the evolution genetic code is potentially useful if expanded to other amino acids, however, substrate discrimination is difficult to achieve for some amino acids such as isoleucine, as a binding pocket for isoleucine would also fit the smaller valine. To maintain fidelity, some synthetases employ a secondary catalytic site that has proofreading activity to hydrolyze mischarged amino acids. The cooperation of this editing site with the active site in a double-sieve mechanism allows discrimination between extremely similar amino acids differing by as little as a single methyl group (Fersht 1979). Synthetases with impaired proofreading activity can also be used to incorporate unnatural amino acids into proteins by allowing close structural analogues of the cognate amino acid to escape the editing function.

Much of this work has focused on the well-studied mechanism of editing in the synthetases for the small aliphatic amino acids isoleucine, leucine, and valine. Each of these synthetases contains a conserved editing domain (Lin 1996; Starzyk 1987), that when abolished causes mischarging of the respective tRNAs both *in vitro* and *in vivo* (Döring 2001; Hendrickson 2000 p8180; Karkhanis 2007; Pezo 2004; Zhai 2005). Mutations in ValRS that abolish editing were first selected for using a selective pressure imposed on *E. coli* cells harboring a mutant thymidylate synthase gene. In this mutant

gene, a codon for an essential catalytic Cys residue is mutated to a Val codon (Döring 2001). Mischarging of tRNA^{Val} with cysteine by a mutant ValRS would restore thymidylate synthase activity allowing growth on media lacking supplemented thymine. Using this method mutations inside the CP1 editing domain of ValRS were found. The unnatural amino acid α -aminobutyrate was able to partially rescue a growth defect in an *E. coli* strain auxotrophic for valine harboring a selected mutant of ValRS (**Figure 3.1.2**). Proteomic-wide substitution of Val by α -aminobutyrate was 24%, a rather significant amount considering *E. coli* harbors nearly 100,000 valine codons (Nakamura 2000). When these strains were grown in non-auxotrophic backgrounds, both threonine and α -aminobutyrate were found to be toxic to cells due to their infiltration of the valine pathway.

The level of misincorporation of the unnatural amino acid α -aminobutyrate in mutant *E. coli* cells is significant for understanding the evolution of the genetic code. α -aminobutyrate is found in high amounts in experiments mimicking pre-biotic environments (Miller 1953; Miller 1976), and why it is excluded in modern translation is a mystery. Only a single nucleotide change in gene encoding ValRS allows the infiltration of this unnatural amino acid (Döring 2001). The relative ease of this mutational event suggests the following scenario: the genetic code was once ambiguous, allowing organisms to survive periods of intense drought of amino acids; only in recent evolution did the code become fixed in such a way as to preferentially encode valine over α -aminobutyrate. In addition, that a modern organism such as *E. coli* can survive such a change to its proteome is significant. These results suggest that it may be experimentally feasible to completely reassign an amino acid or codon with appropriate selective

pressures. Complete reassignment would be a great leap forward towards the goal of engineering organisms with variant genetic codes.

Studies exploiting IleRS mutants have advanced our knowledge of how an ambiguous genetic code and statistical protein population could provide an evolutionary advantage even with detrimental effects. An editing defective mutant of IleRS was identified to misincorporate valine and other unnatural amino acids such as norvaline and norleucine at isoleucyl positions (Hendrickson 2000 p8180; Pezo 2004). A strain harboring this mutant IleRS showed inhibited growth in a variety of media, and an increased sensitivity to unnatural amino acids and antibiotics (Bacher 2005). The ambiguity at isoleucyl codons introduced into *E. coli* produces a marked decrease in survivability, supporting the idea that modern organisms are at such a level of complexity that changes to their genetic code would be unobtainable (Crick 1968). However, when this strain was grown in limiting supplies of isoleucine, a growth advantage was conferred over a wildtype *E. coli* strain (Pezo 2004). Another organism, *A. baylyi*, exhibits the same phenomenon (Bacher 2007 p6496). This demonstrates that an ambiguous code may create a selective advantage during times of amino acid starvation, and supports that ambiguity is a feasible path for introducing major alterations into the genetic code.

Another example of an ambiguous genetic code is found in nature. Species within the genus *Candida* have undergone a substantial alteration to their genetic code in the reassignment of CTG from leucine to serine (Kawaguchi 1989). This reassignment appears to have occurred through a long process of ambiguity that may still be present in certain species (Suzuki 1997). Transfer of the mutant tRNA causing this alteration in *C.*

albicans to the distant relative *S. cerevisiae* induced misincorporation of serine at leucyl codons introducing a statistical protein population (Silva 2007). *S. cerevisiae* harboring this tRNA induced a novel set of proteins involved in stress response leading to increased fitness in a wide variety of stresses (Santos 1999). The authors propose that the decreased fitness induced from the introduction of an ambiguous code was offset by an increased survivability to stress in the environment. Once this niche was created for these intermediate organisms, they were allowed to thrive in the evolutionary competition and thus facilitated the complete reassignment of their genetic code. Introduction of this variant code in *S. cerevisiae* helps to mimic the evolutionary pathways of *Candida*, and glimpse into processes of evolution that are otherwise unobtainable.

The introduction of genetic code ambiguity in organisms has led to a number of important findings facilitating both engineering and evolutionary studies. These studies help to recapture the progress and pathways of organisms as they undergo significant alterations to their genetic code. These can be direct correlations, such as with *Candida* or through indirect means, such as the multitude of work performed in *E. coli*. Ambiguity has also played a large role in understanding how far the envelope can be pushed. Classical reflections that the genetic code should be rigid and inflexible are unsupported by the large misincorporation levels seen in the isoleucyl (Pezo 2004) and valyl (Döring 2001) ambiguity experiments. Understanding the flexibility of the genetic code in an organism such as *E. coli* is critical for knowing the level of engineering that will be required for whole-scale alterations and a future permanent new code.

3.1.8: Conclusions

The engineering of variant genetic codes greatly increases our understanding of the evolution of the code itself. Although some organisms may be at an intermediate stage for code evolution (Susuki 1997), repeating these processes in the laboratory would afford a more detailed and comprehensive understanding. By monitoring organisms undergoing these changes, adaptation to the alterations in their genetic codes can be directly investigated. This would give a glimpse into the history of the code and show how intermediate organisms may have survived through Darwinian selection. In addition, synthetic organisms can also give a glimpse into the natural progression of the genetic code, and where it may be headed in the distant future. Finally, organisms with a variant genetic code may have major advantages from their increased or altered repertoire of chemical moieties. Studying this may lead to major breakthroughs in evolutionary population studies.

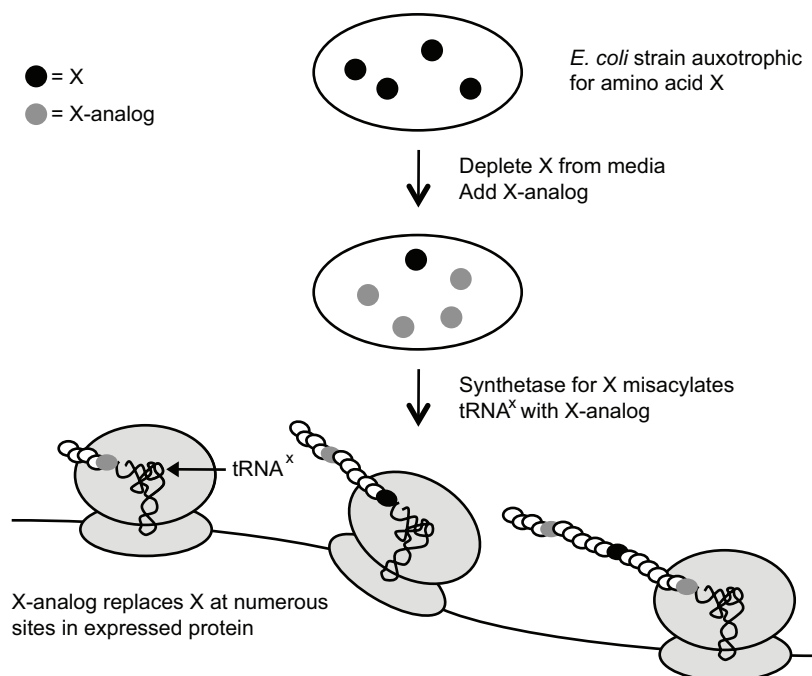


Figure 3.1.1: The SPI method for the misincorporation of unnatural analogs into a target protein using auxotrophic bacterial strains.

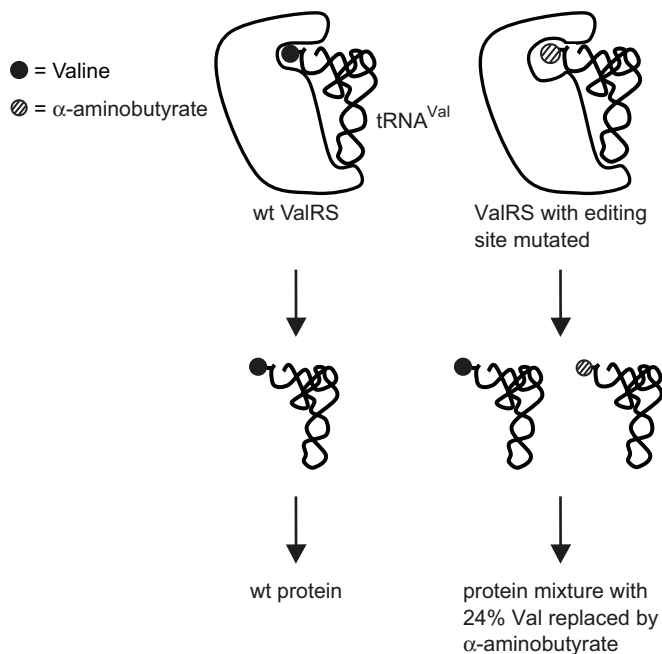


Figure 3.1.2: A mutation relaxing the substrate specificity of the ValRS in *E. coli* leads to a high level of misincorporation of the unnatural amino acid α -aminobutyrate.

Chapter 3.2: The Evolution of the Ambiguous *E. coli* Strain DJ.2

3.2.1: Abstract

Many theories postulate that the modern genetic code evolved from a more primitive code containing less than 20 amino acids. This process of amino acid addition is difficult to study in detail, as no primitive organism exists today. Here we show the synthesis and evolution of an isoleucine auxotrophic *E. coli* strain, DJ.2, that harbors ambiguous isoleucine codons. This strain was serially passaged for 6 months while gradually reducing the amount of isoleucine in the growth media. Full proteome mass spectrometry of this evolved strain reveals that ~25% of isoleucine residues throughout the proteome are replaced with valine. Although replacement of isoleucine with valine is incomplete and ambiguous, the increased level of replacement of one canonical amino acid with another is unprecedented. Laboratory evolution of this strain induced a number of novel adaptation events, including a reduction of the valine toxicity known in these ambiguous strains (Pezo 2004) and the recovery of isoleucine synthesis via a novel mechanism. This strain may prove valuable for future adaptation studies mimicking the hypothesized pathway to genetic code infiltration of an invading amino acid, or the process of generating an alternative genetic code through ambiguous intermediates. This strain has also proved valuable in uncovering the stereochemical code, which is outlined in Chapter 2.2.

3.2.2: Introduction

The genetic code is universal, in that all organisms studied thus far use the standard genetic code or a similar derivative. This is striking as it implies that the code

was established before the last universal common ancestor (LUCA) and was relatively unchanged throughout the splitting of the three main branches of life. The persistence and universality of the code has inspired the desire to understand why it was preferred by natural selection to any other options. However, the origins and subsequent evolution of the code remain elusive.

The earliest theory defines the code as a static entity in which changes would be too deleterious, since a single change in the genetic code leads to mutations throughout the proteome (Crick 1968). This “frozen accident” was challenged when variant codes were found in the mitochondria of metazoans (Barrell 1979). Initially thought to be artifacts from endosymbiosis, variant mitochondrial codes are now considered direct evidence that the genetic code has evolved due to the sheer number and variety of changes discovered thus far (Knight 2001 p49). Some alterations, such as reassigning UGA from a stop to a tryptophan codon, have occurred separately in different lineages. Variant nuclear codes were eventually discovered (reviewed in Knight 2001 p49), directly challenging the idea of a static code. Even though these variant codes argue against a “frozen” code, understanding how these organisms are capable of overcoming the deleterious effects involved in codon reassignment remains an important question. It also makes one wonder just how flexible the genetic code truly is?

Species that use a variant of the standard genetic code give clues into possible models explaining how the genetic code evolved since LUCA. The codon capture hypothesis proposes that codons can be eliminated from the entire genome by AT or GC-mutational pressure, and subsequently reassigned to a different meaning when the pressure reverts and the codon reappears (Jukes 1987; Osawa 1989). *Mycoplasma*

capricolum has an AT-rich genome, in which the arginine codon CGG has been unassigned to a blank codon (Oba 1991). *Micrococcus luteus* is suggested to have lost the use of at least two codons (UUA and AUA) in translation due to high genomic GC content (Kano 1993). The ambiguous intermediate theory claims that reassignment occurs through mutations in or duplication of tRNAs that place ambiguity on specific codons (Schultz 1994). If this change was biochemically small or even beneficial, it may eventually become fixed in the genetic code and replace the original meaning. For a detailed description of these two theories, see Chapter 1.3.1.2.

The ancestors of LUCA and LUCA itself may have had an ambiguous codon table (Crick 1968). The ambiguity would allow survival in droughts of amino acids that could not be biochemically synthesized. An organism with ambiguous codons in these harsh conditions would be more adaptable to change and survive rapidly changing conditions. As more complete genomes are sequenced, more variants of the code will inevitably be found, and each variant code represents a snapshot of where the genetic code is in some point in its evolution. Establishing a system to mutate the genetic code in a laboratory setting on a reasonable time scale will enable the evolution of the genetic code to be studied in real time. It may also yield empirical evidence supporting the differing theories underlying genetic code evolution, and evidence for where the genetic code is heading.

To understand the dynamic features of the genetic code and its evolution process in full spectrum, we propose to evolve an extant organism, *E. coli*, and to study in real time whether and how the organism adapts to changes to its genetic code. We propose that the genetic code is more malleable than previously described, and that creating an

ambiguous code state is the key to altering the code. An ambiguous code state is an unstable state of the genetic code that is able to survive certain pressures due to an increased stress response. Pressure to remove this instability is strong, and thus this state may be more prone to alternate pathways of fixation, thereby introducing changes into the genetic code. We will apply pressure directly onto the components executing the code, a different approach from previous efforts in which the code was perturbed indirectly and inefficiently. Efficient and direct pressure would accelerate the evolutionary process and thus enable this process to be studied on the laboratory time scale. The directed evolution of the genetic code in a living organism would give insights into not only how the code was established before LUCA, but also the dynamics and future expansion of the modern code.

3.2.3: Materials and Methods

3.2.3.1: Chemicals

All chemicals used in this study were purchased from Sigma (St. Louis, MO), unless otherwise noted.

3.2.3.2: Strain Construction

All pathway and gene information for *E. coli* K12 derived strains was obtained through the EcoCyc online database (Keseler 2009).

The DJ.1 *E. coli* strain (See **Table 3.2.1** for a list of strains used in this study) was created from CU.2 (Umbarger 1951) through the replacement of the genomic *ileS* gene encoding the isoleucyl-tRNA synthetase (IleRS) with one harboring an 11-alanine substitution in the editing domain (Pezo 2004) using established procedures (Datsenko 2000). Briefly, a kanamycin resistance gene was added to the 3' end to a mutated form

of the *ileS* gene using overlapping PCR (**Figure 3.2.1a**). The kanamycin cassette also contains a 51-nucleotide overhang on the 3' end that is homologous to the final 51 nucleotides of the *ileS* gene. CU.2 cells harboring the pKD46 red recombinase plasmid were electroporated with 200ng of the knockin cassette and Kan^R colonies were selected. The *ileS* gene from Kan^R colonies was sequenced to verify the presence of the 11-alanine mutation in the editing domain of the IleRS.

The DJ.2 strain was synthesized from DJ.1 by complete replacement of the *ilvE* gene with a tetracycline resistance cassette (**Figure 3.2.1b**). First, 51 nucleotide overhangs completely homologous to the corresponding genomic sequences of the flanking regions of *ilvE* were appended to tetracycline resistance gene. DJ.2 cells harboring the pKD46 plasmid were electroporated with 200ng of the tetracycline cassette and cells were screened for tet^R. Knockout was verified via PCR of the genomic loci. Replacement of the *ilv* operon or the *avtA* gene with a Cm^R cassette in DJ.2 p83 was performed as described for *ilvE* to yield DJ.5, DJ.3 respectively. Replacement of the *ilv* operon in DH10 β (Invitrogen, Carlsbad, CA) was also performed as described for *ilvE* to yield DJ.6.

The DJ.18 strain was synthesized in order to remove all possible sources of isoleucine contamination or redundancies (**Figure 3.2.1c**). A modified version of gene gorging (Herring 2003; Tischer 2006) was used to remove each set of genes with no permanent marker insertion. For each set of genes or operons to be deleted a cassette was generated that contained a Cm^R cassette flanked on both sides with I-SceI cut sites and a repeating element 75 nucleotides long (**Figure 3.2.1c** blue and green boxes respectively). The repeating element served as part of the 5' homologous region to facilitate

recombination. Finally, a 3' homologous overhang was appended to the cassette. Each cassette was electroporated into the DJ.2 strain (or derivative) harboring the pKD46 plasmid. Positive knockouts were identified with PCR and the pATBSR plasmid was then introduced (a tetracycline derivative of pACBSR) (Herring 2003). Induction of this plasmid results in the excising of the Cm^R cassette and repair of the introduced double stranded break, leaving a scarless and complete removal of the target gene. The resultant strain was then used for the next round of deletions. Using this method, the region comprising the *ilvGMEDAYC*, *ilvHI*, *ilvBN*, and *avtA* genes were knocked out of DJ.2 in succession to give the final DJ.18 strain.

3.2.3.3: Continuous Culture of DJ.2

The DJ.2 strain was serially passaged over the span of six months in liquid minimal media (M9 salts, 0.8g/L glucose, 2mM MgSO_4 , 0.1mM CaCl_2 , 10mg/L Thiamine, and 0.001% FeSO_4) supplemented with essential amino acids (100mg/L each of leucine, methionine, arginine, threonine, and serine; 200mg/L each of histidine, phenylalanine, proline, tryptophan, and tyrosine) and 5 $\mu\text{g}/\text{mL}$ of tetracycline. DJ.2 cells from previous days were diluted to an OD_{600} of 0.001 and then grown for 24 hours at 37°C. The ratio of isoleucine to valine was altered every seven passages to gradually reduce isoleucine from the growth media (**Table 3.2.2**). Cells were also mutagenized with ethane methyl sulfonate (EMS) weekly in parallel. For mutagenesis, cells were grown to an OD_{600} of 0.5, harvested, washed in 0.067M phosphate buffer pH6.8, and resuspended in 5mL's of phosphate buffer containing 0.2M of EMS. Cells were incubated at 30°C for one hour, washed 2 times in phosphate buffer, and recovered overnight in minimal media.

3.2.3.4: Assay of Intracellular Isoleucine Concentration

To determine intracellular concentrations of isoleucine, a novel growth screen was developed. CU.2 and DJ.2 cells were grown overnight in minimal media with and without isoleucine respectively. The OD₆₀₀ was adjusted to one, and five milliliters of each strain was harvested, washed 2 times in PBS buffer, and resuspended in 50μL PBS plus 50μL of 13N HCl. Cells were incubated overnight at 110°C, and the solution was neutralized with NaOH. Fresh CU.2 cells were diluted in minimal media without isoleucine supplemented with 1 or 7μL of each HCl-treated sample and grown overnight at 37°C. Growth of each CU.2 sample was assayed after 48 hours and compared to that of a standard curve (**Figure 3.2.2a**) generated with known concentrations of isoleucine using the same procedure. Results were normalized to wildtype CU.2 cells grown in the presence of isoleucine.

3.2.3.5: Complementation Screens

To investigate the source of the isoleucine contamination in the DJ.2 strain, a complementation screen was used (**Figure 3.2.3**). DJ.2 cells were grown, harvested, and genomic DNA was prepared using the DNeasy kit from Qiagen (Valencia, CA). 5μg of genomic DNA was digested with *Bam* HI, *Hind* III, or *Pst* I (New England Biolabs, Ipswich, MS) and then ligated into pBSK2+ (Stratagene, Wilmington, DE). Ligated products were directly transformed into DJ.6, and plated onto minimal media plates without isoleucine. Potential clones were picked, grown in liquid culture, minipreped, and sequenced to determine the identity of the insert inside pBKS2+.

3.2.3.6: Valine Toxicity Assays

Valine toxicity tests were determined essentially as described (Pezo 2004). CU.2, DJ.2 p.0, and DJ.2 p.83 cells were diluted to an OD₆₀₀ of 0.001 in minimal media containing 100μM of isoleucine and leucine and various concentrations of valine. The OD₆₀₀ of each sample was determined after growth for 48 hours in 37°C. Measurements were normalized to the highest reading to yield a normalized growth yield.

3.2.3.7: Comparative Full-Proteomic Mass Spectrometry

Comparative full proteome mass spectrometry was performed as described in the methods section of Chapter 2.2. As that methodology is fully detailed there, it will not be repeated here. The median isoleucine identity derived in those experiments was used to calculate the median valine misincorporation in the 500 peptides described.

3.2.4: Results

3.2.4.1: Construction of the DJ.2 Strain

In order to introduce ambiguity at the isoleucine codons, we mutated the editing domain of the isoleucyl-tRNA synthetase (IleRS). Mutations in the IleRS editing domain in *E. coli* cells lead to mischarging of tRNA^{Ile} with valine generating a statistical protein population containing valine substitutions at numerous different locations in each protein (Pezo 2004). The *E. coli* strain harboring the mutant IleRS was shown to have significant growth defects in low concentrations of isoleucine and an enhanced mutational rate in aging bacteria (Bacher 2007 p1907). We first replaced the genomic copy of IleRS in the *E. coli* strain CU.2 with the mutant IleRS using established procedures (Datsenko 2000). The DJ.1 strain at first exhibited a growth dependence on isoleucine that was quickly reverted after several generations in serial culture. The mutation in the CU.2 strain leading to isoleucine auxotrophy is known to be in the *ilvE* gene, which encodes a

branched-chain amino transferase responsible for the final step in isoleucine, valine, and leucine synthesis (from CGSC). Amplification and sequencing of this gene revealed a W147Am(tag) mutation in *ilvE* producing a truncated product. As expected, the DJ.1 growth phenotype mapped to this position, with a reversion of the stop codon to the original tryptophan. To prevent further reversion events, the *ilvE* gene cassette was replaced in full with a tetracycline resistant cassette (**Figure 3.2.1b**). The resultant strain, DJ.2, behaves similarly to a wildtype strain in rich media with only minimal retardation of growth rate. However, this strain does not grow when shifted to media lacking isoleucine.

3.2.4.2: Passage of DJ.2 Reveals Novel Growth Phenotype in Media Without Ile

To induce an ambiguous intermediate state in DJ.2, the balance of valine misincorporation at isoleucine codons was first tuned by increasing ratio of valine to isoleucine in the growth media. As expected, the DJ.2 strain could not survive in media with high levels of valine, most likely due to the deleterious effect of valine misincorporation (Pezo 2004). To circumvent this problem, the strain was continuously grown in minimal media while slowly raising the valine to isoleucine ratio over numerous generations (**Table 3.2.2**). Cells were mutagenized regularly in parallel to randomly introduce mutations into the genome. This may not be necessary, however, as this strain already have an increased mutation rate in limiting amounts of isoleucine (Bacher 2007 p1907).

Serial passage of DJ.2 was stopped after 84 passages over 6 months, yielding approximately 652 generations of continual growth. At this point, essentially no isoleucine was being added to the growth media (final concentration of 0.15 μ M). Cells

were assayed for their ability to grow without isoleucine, and under these conditions were found to grow to an OD₆₀₀ of ~0.15 within 24 hours. This phenotype was seen even under continual growth in media lacking isoleucine, suggesting that carryover of any residual isoleucine was not the cause of growth. At this point, cells were collected and subjected to full proteome mass spectrometry to assay if isoleucine was present in the proteome. The median misincorporation percentage of 500 peptide fragments throughout the proteome was found to be 23.1% (See Chapter 2.2). Although valine misincorporation at isoleucine residues was not complete, it is nearly twice as high as previously reported (10-15%) (Pezo 2004). This suggests that there is either an endogenous source of isoleucine contamination, or that some amount of residual isoleucine is present in the purified amino acids used to supplement the minimal media. Regardless, this phenotype was still derived through experimentation, as the parental strain at day zero is unable to grow in the same media lacking isoleucine.

3.2.4.3: Intracellular Ile Concentration in DJ.2 is Comparable to Parental Strain

To accurately determine the intracellular concentration of isoleucine in these strains, a novel growth assay was developed using the parental strain CU.2 (see methods). As this strain is completely dependent on isoleucine for growth, any isoleucine derived from DJ.2 cells would enable growth. From this assay, DJ.2 was determined to have an internal isoleucine concentration that is ~60% of wildtype cells (**Figure 3.2.2b**). This represents a mild reduction in isoleucine, suggesting an overall high level of isoleucine contamination in the DJ.2 strain.

3.2.4.4: *ilvHI* Redundancy is the Source of Ile Contamination

This isoleucine contaminant could be from a number of sources. Due to the reversion of the *ilvE* gene in DJ.1, an obvious first target would be a similar enzyme or enzymes involved in the synthesis of the branched-chain amino acids. To address this issue, we generated a number of different knockout cassettes for possible targets. Specifically, we generated strains in which the *ilv* operon (DJ.5), or *avtA* (DJ.3) gene were removed. The *ilv* operon knockout removes the *ilvX ilvM, ilvE, ilvD, ilvA, ilvY*, and *ilvC* genes as well as the *ilvG_1* and *ilvG_2* pseudo genes involved in the synthesis of the branched-chain amino acids abolishing all known avenues for protein synthesis. The *avtA* gene encodes the valine aminotransferase that catalyzes the same reaction as *ilvE* but is specific for valine synthesis. Both knockouts derived from DJ.2 retained their ability to grow in media lacking isoleucine, and a high level of internal isoleucine contamination (**Figure 3.2.2b**). These data suggest that the derived growth phenotype and isoleucine contamination is not due to any direct biochemical bypass of the *ilvE* gene as it persists without the entire synthesis pathway.

If the source of isoleucine contamination is a mutation in a gene in DJ.2, it should be transferable and complement an isoleucine auxotroph. Genomic DNA was harvested from DJ.2, digested with a variety of enzymes (see methods), and ligated into the pBSK2+ vector. Ligation products were transformed into DJ.6 (see methods) and cells were grown on minimal media plates supplemented without isoleucine (**Figure 3.2.3**). After 48 hours of growth only one colony was found that could grow efficiently on media lacking isoleucine. The cells were grown, miniprepmed, and the insert in pBSK2+ was sequenced. The operon containing the genes *ilvH* and *ilvI* was found inside pBSK2+ without any mutations. *ilvI* and *ilvH* are one of three sets of redundant genes involved in

the synthesis of isoleucine and valine (the other two being *ilvB/ilvN* and the two pseudo genes *ilvG_1/G_2/ilvM* which were knocked out in the *ilv* knockout). To demonstrate that *ilvHI* is necessary for DJ.2 growth in media lacking isoleucine, the DJ.18 strain was assayed for growth in media lacking isoleucine. DJ18 is a strain derived from DJ.2 that lacks many of these genes, including the *ilv* operon, the *avtA* gene, and the *ilvHI* and *ilvBN* redundancies (**Figure 3.2.1c**). This strain, unlike DJ.2, can no longer grow in media lacking isoleucine. As the knockouts of the *ilv* operon and *avtA* genes from DJ.2 did not have this effect (in DJ.5 and DJ.3 respectively), it must be attributed to the *ilvHI* and *ilvBN* redundant genes, suggesting that these redundancies are involved in generating the isoleucine contamination in the DJ.2 strain.

3.2.4.5: High Levels of Valine are no Longer Toxic to the Evolved DJ.2 Strain

Even though the evolved DJ.2 strain has a source of isoleucine contamination, the high misincorporation rate of valine throughout the proteome suggests that this strain may be beneficial in the study of intermediate states of code evolution. Numerous adaptation events were seen in this strain that may contain some insights into code evolution. The DJ.2 strain at passage 0 exhibits a severe toxicity to high concentrations of valine, which is a similar response to published results (Pezo 2004). This toxicity is abolished in DJ.2 p83 cells grown in valine concentrations as high as 6mM (**Figure 3.2.4**). This alteration can be traced back to around passage 41, suggesting that the continued presence of a high concentration of valine (1.7mM at passage 41) leads to an adaptive response to the invader amino acid. As the valine toxicity is most likely attributed to an increase in proteomic misincorporation (as the amino acid histidine

showed no such response), further analysis of the genetic changes underlying this response uncover novel mechanisms for dealing with such stresses.

3.2.5: Conclusions

Many theories postulate that amino acids were gradually added into an earlier minimal code. The mechanisms underlying these additions are difficult to study, as no organism has been found to use less than 20 amino acids. Every organism found has undergone millions of years of evolution in the presence of these 20 amino acids, and thus have evolved dependence. However, through manipulation of an extant organism, a state mimicking that of an earlier minimal code could be achieved. The evolved *E. coli* strain DJ.2 demonstrates the difficulties in achieving these goals. The amino acid isoleucine is most likely an early to mid-addition to the genetic code (Trifinov 2004) containing 80,000 instances in the proteome of *E. coli* K12 (Nakamura 2000). We reasoned that the limited catalytic role of isoleucine coupled to the modest alteration to valine might ease the transition of *E. coli* in making such a drastic change to its proteome. Millions of years of coevolution of this organism and isoleucine metabolism has made even this *modest* change difficult to achieve. DJ.2, not surprisingly, has taken the path of least resistance and found new methods of synthesizing isoleucine, or using scarce amounts of isoleucine more efficiently.

Pinpointing the source of the isoleucine contamination leads to a rather perplexing result. The *ilvHI* and *ilvBN* redundancies are obvious targets for sources of contamination, but they appear to be independent of the rest of the synthesis pathway of isoleucine. *IlvHI* was able to recover a deficiency in the DJ.6 strain lacking the entire synthesis pathway, including three enzymes that are downstream of *ilvHI*. Unlike *ilvHI*

these genes have no known redundancies, suggesting that some other pathway is involved that is yet to be identified in DJ.2. The source of this contamination will need to be identified before further evolution of the DJ.2 strain and its derivatives.

Our initial goals, although ambitious and fraught with the difficulties of long-term laboratory evolution of *E. coli*, may still be achievable. Given our findings, a more appropriate strain may be derived that will help to achieve our desired outcome. Eliminating the redundancies in the branch-chain amino acid synthesis machinery, coupled with a more relaxed timeline may allow precise control of this pressures exerted on the evolving organism. More unknown hurdles will most likely be encountered and remedied throughout the evolution process.

Regardless of the success of such experiments, the adaptation process throughout this timeline can be studied in great detail and compared to natural variations such as those found in *Candida* (Kawaguchi 1989; Sugita 1999). The DJ.2 strain's resistance to valine toxicity is one such example of such adaptation events. It was shown previously that high concentrations of valine in the growth media retard growth in these ambiguous organisms (Pezo 2004). In contrast, evolved DJ.2 cells show no retardation of growth in valine concentrations as high as 6mM, demonstrating some level of adaptation to misincorporation. The 23.1% misincorporation rate of DJ.2 p83 cells is nearly twice as high as previous reports (Pezo 2004), also suggesting some level of adaptation. Further experimentation is required to explain this phenomenon, but it does appear that some novel adaptation occurred in the evolution of DJ.2 allowing it to survive such intense pressures.

The DJ.2 strain described here may prove valuable for future studies of the evolution of the genetic code, even without reaching our initial goals. Indeed, further analysis of the proteomic mass spectrometry results generated herein was critical in the work described in the previous chapter. The DJ.2 strain's high misincorporation rate allowed us to more accurately assess the spectrum of valine misincorporation and generate useful data. The phenomenon described in chapter 2 would have been more difficult to tease out of a data set generated from another strain.

Table 3.2.1: The *E. coli* strains used in this study.

Strain	Genotype	Parental	Source
DH10 β	See source	K12	Invitrogen
CU.2	<i>ilvE12 ilv-2025</i>	K12	1,2
DJ.1	<i>ilvE12 ilv-2025 ileS(mAla)</i> ³	CU.2	This work
DJ.2	<i>ilv-2025 ileS(mAla)::Kan⁴ ΔilvE::tet⁴</i>	DJ.1	This work
DJ.3	<i>ilv-2025 ileS(mAla)::Kan ΔilvE::tet ΔavtA::cat⁴</i>	DJ.2 p.83	This work
DJ.5	<i>ileS(mAla)::Kan ΔilvGMEDAYC::cat</i>	DJ.2 p.83	This work
DJ.6	<i>ΔilvGMEDAYC::cat</i>	DH10 β	This work
DJ.18	<i>ileS(mAla)::Kan ΔilvGMEDAYC ΔilvHI ΔilvBN ΔavtA</i>	DJ.2 p.83	This work

1 - Strain was recieved from the Coli Genetic Stock Center at Yale

2 - Umbarger, 1951

3 - This represents the 11 amino acid substitution in the editing domain of *ileS*

4 - Kan - Kanamycin resitance, tet - tetracycline resistance, cat - chloramphenicol resistance

Table 3.2.2: Laboratory evolution scheme devised for DJ.2. Only passages in which the ratio of isoleucine to valine was altered are shown. Generations were calculated from the daily measurements of the OD₆₀₀.

Passage	[Ile]	[Val]	Ratio (V:I)	Generations
0	139 μ M	1.55mM	12 : 1	0
7	73 μ M	1.63mM	22.4 : 1	57
13	49 μ M	1.65mM	33.6 : 1	110
20	37 μ M	1.67mM	44.8 : 1	172
27	15 μ M	1.69mM	112 : 1	234
35	7.6 μ M	1.700mM	224 : 1	302
42	6.1 μ M	1.700mM	280 : 1	355
50	3.0 μ M	1.704mM	560 : 1	417
58	1.5 μ M	1.704mM	1120 : 1	477
66	0.76 μ M	1.705mM	2239 : 1	532
72	0.30 μ M	1.706mM	5598 : 1	575
79	0.15 μ M	1.707mM	11197 : 1	625
83	0.15 μ M	1.707mM	11197 : 1	653

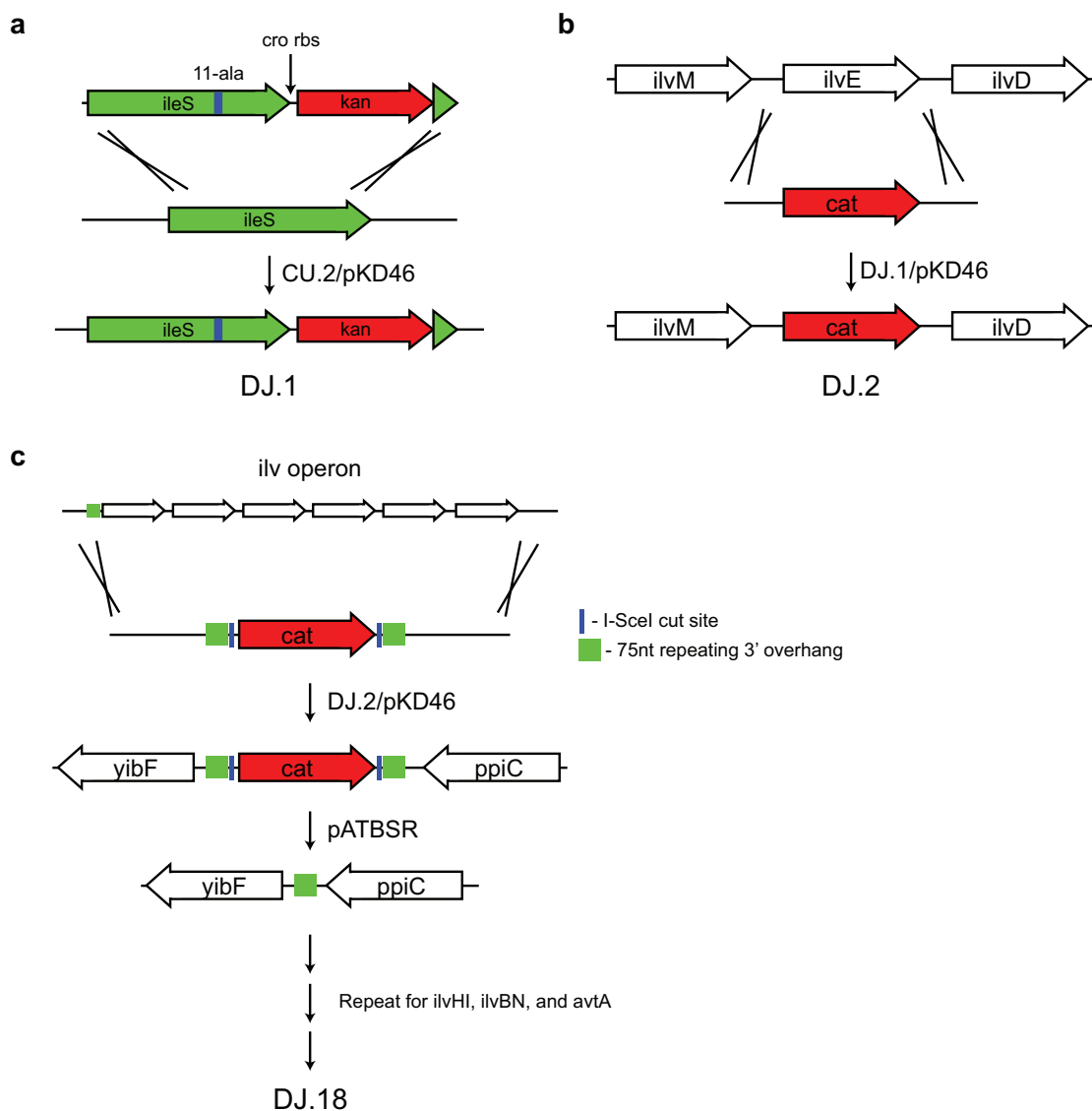


Figure 3.2.1: Schematics representing the knockout constructs used and described in the text. **(a)** Scheme for inserting the 11-alanine mutation (shown in blue) in the editing domain of *ileS*. **(b)** General method for gene replacement with the *cat* gene conferring chloramphenicol resistance. Shown is the knockout cassette generated for *ilvE*, but the same method was used for the *ilv* operon and *avtA*. **(c)** Creation of the *DJ.18* strain relied on multiple gene gorging steps. The *ilv* operon, *ilvHI*, *ilvBN*, and *avtA* were scarlessly removed from *DJ.2* in succession to yield the final *DJ.18* strain.

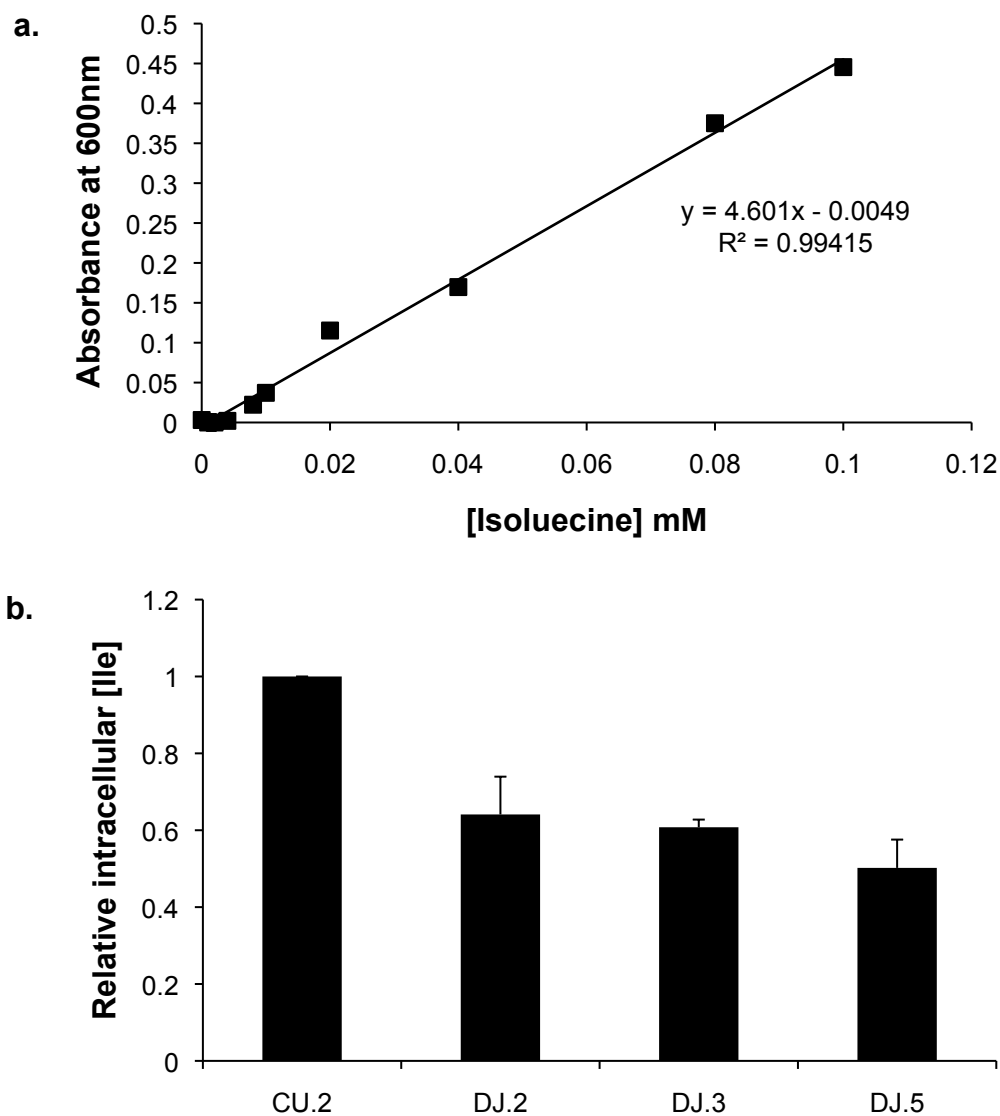


Figure 3.2.2: Intracellular isoleucine concentration in DJ.2, DJ.3, and DJ.5 are comparable to the wildtype CU.2 strain. **(a)** The *E. coli* strain CU.2 demonstrates a linear pattern of growth dependence on the concentration of isoleucine added to the growth media. **(b)** The DJ.2, DJ.3, and DJ.5 strains show a significant amount of intracellular isoleucine contamination, albeit lower than the wildtype CU.2 strain.

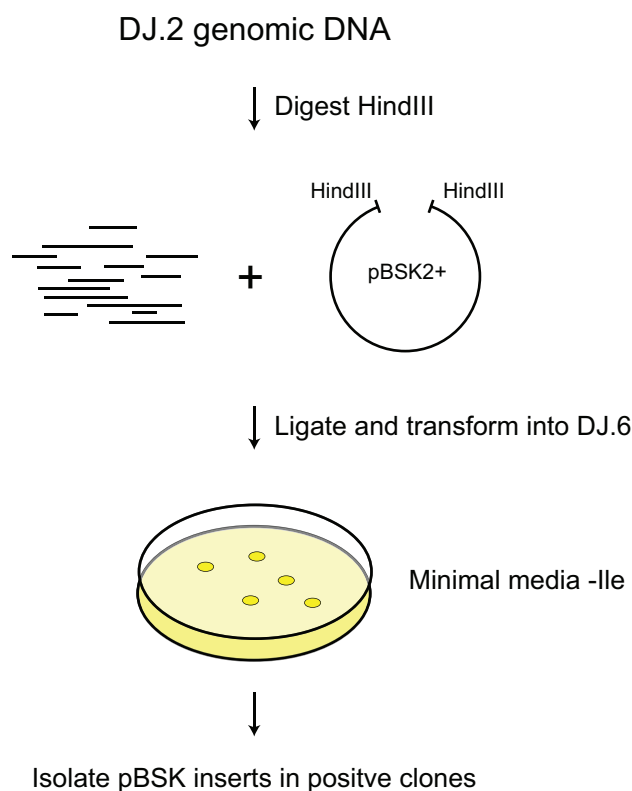


Figure 3.2.3: General scheme for the complementation screen devised to isolate the source of isoleucine contamination in the DJ.2 strain. Details are given in the methods.

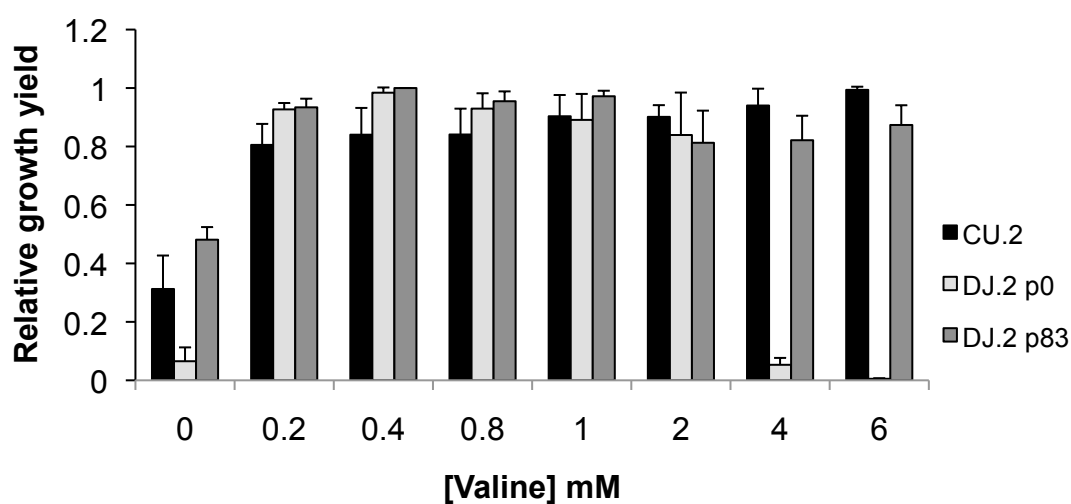


Figure 3.2.4: The DJ.2 *E. coli* strain shows no toxicity to high concentrations of valine after continuous growth of for 83 passages. DJ.2 p0 represents the strain prior to continuous growth and DJ.2 p83 represents the derived strain from continuous culture.

Chapter 3.2 couldn't have been accomplished without the help of Dr. Zhouxin Shen who performed the full-proteome mass spectrometry in the laboratory of Dr. Steven Briggs. CU.2 was obtained from the Coli Genetic Stock Center at Yale University.

Chapter 4: Modern Expansion of the Genetic Code

This chapter continues the theme from Chapter 3, but focuses on a different approach. Chapter 3 was directed at engineering variant codes through the existing synthetases or tRNAs. As such, the variations in these organisms are focused on shuffling, replacement, and addition of amino acids via the hijacking on sense codons. Nature shows us that this may not be the best approach. Of the natural variations to the genetic code, reassignment of stop codons appears to be the preferred avenue. This appears to be a more logical solution to the inherent problems associated with mass alterations to the proteome. Stop codons are by definition rare, used only once per gene, and their rarity would potentially minimize harmful or lethal effects to the proteome.

Targeting stop codons may more accurately represent the history of the genetic code as well. New additions to the genetic code either had to capture sense codons or stop (nonsense) codons in order to infiltrate the existing code. Chapter 2.1 suggests that capture of sense codons could have been the method for some amino acids, but that may not be the whole story. With the apparent natural inclination of targeting stop codons for reassignment, ignoring the role that nonsense or stop codons may have played in the expansion of the genetic code is shortsighted.

This chapter addresses the feasibility of such an approach, and applies it to advance the engineering technology available to synthetic and molecular biology. The methodology here is slightly different than the previous chapter; the targets themselves have changed and we are expanding the code instead of replacing an amino acid with another. Instead of engineering changes to endogenous synthetases or tRNAs, here we

target the endogenous release factors responsible for recognition of stop codons. We also introduce an orthogonal tRNA/synthetase pair in order to artificially expand the genetic code with an unnatural amino acid (UAA). Unlike many of the previous reports using this UAA technology, some of the strains described here have complete reassignment of a stop codon, mimicking key events in the evolutionary history of the genetic code. An additional benefit to this work is the generation of a novel, stable *E. coli* strain capable of the efficient incorporation of numerous unnatural amino acids into a single protein. This strain should prove invaluable to the synthetic production of unnatural proteins for functional, biological, and structural studies.

Chapter 4.1 briefly reviews the field of unnatural amino acid mutagenesis. Although this subchapter is not directly related to the evolution of the genetic code, we apply these techniques towards our end goals. This chapter will give a basic overview of these technologies as they are used in depth in the following subchapters.

Chapter 4.2 introduces our system for reassignment of the UAG stop codon in *E. coli* cells. The results from these studies demonstrate the genetic code of an extant organism is amenable to major alterations. These studies mimic the addition of an additional amino acid to the genetic code, and provide new strains for adaptation and evolutionary studies.

Chapter 4.3 describes in detail the generation and characterization of the novel *E. coli* strain JX3.0 which harbors no genomic copy of release factor one. JX3.0 has vastly improved yields of unnatural amino acid incorporation similar to those described in 4.2, however JX3.0 is healthy and clonable. A novel mutation found in release factor two allows for this valuable growth phenotype.

Chapter 4.1: Unnatural Amino Acids to Investigate Biologic Processes

4.1.1: Abstract

To circumvent the constraint imposed by the 20 canonical amino acids on the study of protein structure and function, various chemical and biosynthetic methods have been developed to incorporate unnatural amino acids into proteins. Unnatural amino acids now can be genetically encoded in living cells in a manner similar to that of common amino acids, which expands site-directed mutagenesis to diverse novel amino acids. The use of unnatural amino acids grants researchers a multitude of chemical and physical properties that cannot be found in the normal genetic repertoire, which significantly improves their ability to manipulate proteins and protein-involved biologic processes. Changes have been tailored into proteins to dissect accurately the contribution of hydrogen bonding, hydrophobic packing, cation- π interaction, and entropy to protein stability, as well as to precisely examine the structural and functional role of crucial residues. Unnatural amino acids also enable the introduction of new chemical reactivities, biophysical probes, mock posttranslational modifications, photoactive groups, and numerous other functionalities for the modification and regulation of protein activities. These studies not only reveal fundamental information of protein structure and function but also explore new means for generating novel protein properties and controlling biologic events.

4.1.2: Introduction

Conventional site-directed mutagenesis of specific amino acids currently is the preferred method for investigating various structural and functional characteristics of proteins. A serious limitation to this methodology is the constraint of using the 20

canonical amino acids fixed by the universal genetic code. This constraint lies in the limited chemical and physical properties of these amino acids, which hinder the ability to make precise alterations. For instance, modification of an amino acid such as glutamine is limited because only asparagine has similar characteristics. For amino acids such as proline, no analogous amino acid exists in the genetic repertoire, which makes it difficult to investigate the role of this amino acid in specific processes without abolishing it completely. Breaking this limitation would enable in-depth investigation of the principles underlying protein structure and function as well as the engineering of novel protein properties and cellular functions. In the past decade, great progress has been made in incorporating unnatural amino acids into proteins to harness their extensive and powerful capabilities. Here, we will introduce unnatural amino acids with a brief overview of various methods for incorporating them into proteins and we will present examples that illustrate how unnatural amino acids have impacted a wide array of research that investigates biologic systems.

4.1.3: Unnatural Amino Acids

Common amino acids consist of an amino group, a carboxyl group, a hydrogen atom, and a side chain all attached to the C $^{\alpha}$ in the L configuration. Analogs with altered side chains, or those that deviate from other features, are generally called unnatural amino acids (**Figure 4.1.1**). The most widely used group of unnatural amino acids is the group in which the C $^{\alpha}$ side chain is changed. Variation of unnatural side chains is diverse and can range from structural analogs of canonical amino acids to those with specific chemical moieties, such as reactive functionalities and reporter groups for biophysical characterization. Modification of the amino group results in changes in the peptide

backbone. For example, changing the amino group into a hydroxy or sulfhydryl group converts the endogenous amide bond between two residues into an ester and thioester link, respectively. Such changes make the resultant analog no longer an “amino” acid but an α -hydroxy acid and a thio acid. An analog with aminooxy replacing the amino group also has been incorporated into protein biosynthetically (Eisenhauer 2002). The amino group also can be alkylated with different moieties to form amino acids that contain secondary amines (Ellman 1992). Another category of unnatural amino acids is α,α -disubstituted amino acids, whose α -hydrogen is replaced by an additional side chain. Moving the amino group away from the α carbon leads to extended β - or γ -amino acids, which can be compatible with protein biosynthetic machinery as well (Hartman 2006). Finally, D-amino acids, which are mirror images of the L counterparts, have been introduced selectively into functional proteins to study structural characteristics (Valiyaveetil 2004).

4.1.4: Methodology

Although this review is mainly focused on the use of unnatural amino acids in the investigation of biologic systems, a brief background of the methodology of incorporation is useful in understanding the power and application of this technology. A more comprehensive coverage of various methods can be found in (Wang 2004) and the references contained therein.

4.1.4.1: Chemical Approaches

Global alterations to certain amino acids can be done *in vitro* through chemical modification of their exposed reactive side chains (Means 1990). The selectivity of chemical modification relies on the differences in chemical reactivity of amino acid

functional groups. Judiciously selected chemicals will react with specific amino acids only, which allows chemical changes to be applied to that amino acid alone. Typical modification involves the thiol group of Cys, the ϵ -amino group of Lys, the carboxylate group of Asp and Glu, and the N-terminal amino group. The hydroxy group of Ser and Thr can be oxidized selectively when Ser and Thr are at the N-terminus of the protein (Geoghegan 1992). Side chains of Tyr and Trp can be modified selectively with transition metal catalysts (Antos 2004; Tilley 2006). Initial applications of this method focused on determining the functional roles of certain amino acid species for biologic activity, but they have expanded to other applications, including biophysical probe tagging, chemical cross-linking, and the conjugation of various synthetic functionalities. However, site-specific alterations are difficult using this approach because the chemical will react with all accessible target amino acids if more than one such amino acid exists in the protein. In addition, chemical modification must be done *in vitro*, which affects only those side chains that are solvent-accessible.

Another method to introduce unnatural amino acids into a polypeptide chain is through complete chemical synthesis (Kent 1988). The predominantly used method, stepwise solid-phase peptide synthesis (SPPS), attaches the C-terminal amino acid to a solid support, and amino acids are added one at a time to the N-terminus. A clear advantage of chemical synthesis is that it enables the accurate introduction of unnatural amino acids at any site in a protein. The number of unnatural amino acids that can be introduced is limited only to the size of the chain, and chains of entirely unnatural amino acids can be produced using this method. Chemical synthesis is useful particularly for the incorporation of isotopic labels and unnatural amino acids that are toxic to cells or

incompatible with the translational machinery. However, construction of a polypeptide chain using even the most advanced chemical synthesis techniques is daunting when confronted with the construction of an entire protein, as these methods currently are limited to approximately 100 amino acids (Kent 1988).

Semisynthetic protein ligation methods, in which two or more protein fragments of recombinant or synthetic origin are chemically ligated to make the full-length protein (Wallace 1995), overcome the size limitation of SPPS. Among these methods, the native chemical ligation strategy couples peptide fragments to form a native peptide link, which leaves no chemical artifacts behind (Dawson 1994; Muir 1998). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and thus is incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein fragment (Giriat 2003). This method, like SPPS, has the power to introduce various unnatural structures that are synthetically accessible. However, it requires appropriate sites for cleavage and ligation, and it becomes cumbersome for internal sites in large proteins. Microinjection of either the *in vitro* ligation product for *in vivo* studies or the synthetic fragment for *in vivo* ligation can be a drawback to this method as well.

4.1.4.2: Biosynthetic Approaches

Methods that use the endogenous cellular machinery to introduce unnatural amino acids into proteins are not limited by protein size and will facilitate the investigation of biologic processes *in vivo*. A general *in vitro* biosynthetic method allows for the site-

specific incorporation of unnatural amino acids into proteins (Noren 1989). In this method, a suppressor tRNA is chemically acylated with an unnatural amino acid, and the codon of interest in the target gene is mutated to the amber stop codon, TAG. When added to cell extracts that support transcription and translation, the suppressor tRNA recognizes and selectively incorporates the attached unnatural amino acid in response to the UAG in the transcribed mRNA. Using this method, a variety of unnatural amino acids have been incorporated into proteins, regardless of position or protein size, and have been applied to a large number of problems in protein chemistry (Cornish 1995). Besides the amber stop codon, rare codons and extended codons also have been used to specify the unnatural amino acid (Hohsaka 2001). An extension of this method involves the microinjection of the chemically acylated tRNA and UAG-containing mutant mRNA into *Xenopus* oocytes (Nowak 1995). The endogenous oocyte protein synthesis machinery supports translation and incorporation of the unnatural amino acid. This method enables the structure-function studies of integral membrane proteins, which are generally not amenable to *in vitro* expression systems (Dougherty 2000). A purified *in vitro* translation system that consisted of only ribosomes, initiation factors, elongation factors, mRNA, and tRNAs preloaded with desired amino acids was used to incorporate simultaneously several unnatural amino acids into peptides in response to sense codons (Forster 2003). By reassigning the meaning of codons, this system ultimately may allow the synthesis of peptides and proteins that contain multiple unnatural amino acids. The drawback to these methods lies in the chemical acylation of the suppressor tRNA, which is demanding technically and can exclude certain unnatural amino acid from attachment. In addition, acylated tRNA is consumed stoichiometrically and cannot be regenerated in

cells or cell extracts, which leads to low expression of the target protein.

Multisite incorporation of unnatural amino acids by using cellular machinery has been achieved in auxotrophic bacterial strains (Budisa 2004) and in mammalian cells (Suchanek 2005). This method relies on the idea that aminoacyl-tRNA synthetases can mischarge unnatural amino acids that are close structural analogs of the cognate amino acids, although with very high substrate specificity. An unnatural amino acid analogous to a canonical counterpart is introduced into a bacterial strain that is incapable of producing the natural amino acid or into mammalian cells that are deprived of the natural amino acid. The translational machinery then replaces the natural amino acid with its analog in all proteins. The incorporation efficiency of unnatural amino acids can be improved by increasing the expression level of the synthetase (Tang 2001) and by introducing mutations that relax the substrate specificity of the aminoacylation domain (Ibba 1994) or attenuate the proofreading function of the editing domain of certain synthetases (Döring 2001). However, this strategy is limited because it is restricted to global replacement of one amino acid with an analog and does not allow specific single alterations with a specific protein.

It would be ideal to genetically encode an unnatural amino acid in a manner similar to that of common amino acids, which would enable site-directed mutagenesis in living cells with unnatural amino acids. A general method to expand the genetic code to include unnatural amino acids was developed. It involves the generation of a new tRNA-codon-synthetase set that is specific to the unnatural amino acid and does not crosstalk with other sets for common amino acids (Wang 2001). The new synthetase is evolved to charge specifically an unnatural amino acid onto the new tRNA. This tRNA recognizes a

codon that does not encode any common amino acids (e.g., a stop codon or an extended codon). When expressed in cells, the new tRNA-synthetase pair enables the unnatural amino acid to be site-specifically incorporated into proteins at the unique codon with high fidelity and efficiency. This method allows the use of unnatural amino acids in the investigation of biologic systems in an *in vivo* setting. It may be possible to generate stable cell lines or transgenic animals capable of inheriting such alterations for long-term studies. However, toxic unnatural amino acids and those incompatible with the protein biosynthesis machinery cannot be incorporated using this approach.

4.1.5: Applications of Unnatural Amino Acids

Unnatural amino acids enable the structural, chemical, and physical properties of the building blocks of proteins to be customized according to needs. Such tailored changes have contributed to our understanding of the fundamental questions of protein chemistry on the molecular and atomic level, have been used to modify and enhance protein properties, and are being exploited to control protein activities to investigate various biologic processes and to create novel biologic functions.

4.1.5.1: Protein Stability

There are many factors contributing to protein stability, including hydrogen bonding, hydrophobicity, packing, and conformational entropy, among others. It is difficult to access individual contributions by using conventional mutagenesis because changing one common amino acid to another often alters several properties at a time. For example, mutagenesis to disrupt hydrogen bonds, usually by deleting one member of a hydrogen-bonded pair, will leave an unpaired hydrogen donor or acceptor and/or alter local solvation and packing interactions, all of which may lead to protein destabilization.

To determine the effect of side-chain hydrogen bonding on protein folding, Tyr27 in staphylococcal nuclease (SNase) was replaced with several isosteric, fluorinated tyrosine analogs (analogs 1 to 3) (**Figure 4.1.2**) (Thorson 1995). These unnatural amino acids were designed to gradually increase the strength of the Tyr27-Glu10 hydrogen bond while minimizing the steric and electronic perturbations associated with deleting one hydrogen-bonding member. The stability constants K_{app} of the corresponding mutants were found correlative with the pK_a of the hydroxyl group in the tyrosine analogs. This result provides strong evidence that intramolecular side-chain hydrogen bonds preferentially stabilize the folded state of a protein relative to the unfolded state in water.

α -Hydroxy acids have been used to study the contribution of the backbone hydrogen bonds to protein stability (**Figure 4.1.3**). The replacement of a common amino acid with an α -hydroxy acid that contains the same side chain effectively substitutes a good hydrogen-bond acceptor (the amide carbonyl group) with a considerably weaker one (the ester carbonyl group) in a conservative manner and disrupts a potential backbone hydrogen bond because the ester link cannot serve as a hydrogen-bond donor as does the NH. α -Hydroxy acids were incorporated at the N-terminus, the middle, and the C-terminus of the α -helix 39-50 of T4 lysozyme (Koh 1997). At the N-terminus and the C-terminus, where only one hydrogen-bonding interaction is perturbed, the ester substitution destabilizes the protein by $0.9 \text{ kcal mol}^{-1}$ and $0.7 \text{ kcal mol}^{-1}$, respectively. In the middle of the helix, where such substitution perturbs two hydrogen bonds, the protein is destabilized by $1.7 \text{ kcal mol}^{-1}$. In another study, Leu14 in an antiparallel β sheet of SNase was replaced with leucic acid (Chapman 1997). This amide-to-ester change decreases the stability by $1.5\text{-}2.5 \text{ kcal mol}^{-1}$. Altogether, these results convincingly show

that both side-chain hydrogen bonds and main-chain hydrogen bonds significantly contribute to protein stability.

To examine the importance of the packing interaction in the core of a protein, Leu133 in T4 lysozyme was replaced with a series of analogs with extended or shortened alkyl side chains (analogs 4 to 7) (Mendel 1992). Leu133 lies along the edge of the largest cavity in the interior of T4 lysozyme, which makes it possible to change the bulk of the side chain with minimal concomitant strain. Incorporation of (S,S)-2-amino-4-methylhexanoic acid (analog 4) and (S)-2-amino-4-cyclopentylpropanoic acid (analog 5) stabilizes T4 lysozyme by $0.6 \text{ kcal mol}^{-1}$ and $1.24 \text{ kcal mol}^{-1}$, respectively, which indicates that the increased bulk of buried hydrophobic residues can enhance protein stability. During protein folding, the cyclic amino acid 5 will lose less conformational entropy than does 4. That the 5-containing mutant is more stable than the 4-containing mutant suggests that side-chain entropy also affects protein stability. As expected, when the side chain of Leu133 is shortened systematically, as in unnatural amino acids 6, 7, and alanine, the protein becomes increasingly less stable.

Another method for increasing hydrophobicity while minimizing structural perturbation is to replace hydrocarbons with fluorocarbons. Using solid-phase synthesis, L-5,5,5,5',5',5'-hexafluoroleucine 8 was substituted for seven core leucine residues in a 30-residue peptide that can form homodimeric coiled coils. Hydrophobic side chains of the core residues pack against each other in the coiled coil. Fluorination of these side chains increased the hydrophobicity and raised the melting temperature of the homodimer from 34°C to 82°C (Bilgicer 2001). In addition, fluorocarbons are insoluble in hydrocarbons at room temperature and, thus, form a fluorous phase by interacting with

other fluorocarbons. When a disulfide-bound heterodimer of the hexafluoroleucine core peptide and a leucine core peptide was allowed to undergo disulfide exchange, the peptides self-sorted into homodimers (Bilgicer 1992).

This fluorous effect could lead to a novel protein-protein recognition. In another report, six leucine residues in the hydrophobic core of an antiparallel 4- α -helix bundle were replaced by hexafluoroleucine 8 (Lee 2006). The free energy of the unfolding of the mutant peptide increases by 0.3 kcal mol⁻¹ per residue when the two central leucines are substituted and by an additional 0.12 kcal mol⁻¹ per residue when the outer leucines are replaced, which confirms that hydrophobic packing stabilizes proteins.

4.1.5.2: Protein Structure and Function

Unnatural amino acids can be designed to elucidate the functional role of a residue that is misinterpreted by or remains ambiguous to conventional mutagenesis and other methods. For example, Glu43 is important for the catalytic activity of SNase because its replacement by Asp and Gln significantly decreases the catalytic efficiency. Previous structural and mutagenesis studies suggested that Glu43 functions as a general base to activate a water molecule for hydrolyzing the phosphodiester bond of DNA. However, substitution of Glu43 with either homoglutamate (analog 9) or (S)-4-nitro-2-aminobutyric acid (analog 10) yielded mutant enzymes with kinetic constants similar to those of wide-type SNase (Judice 1993). Because these two unnatural amino acids are isoelectronic and isosteric to glutamate but a much poorer base, such substitution would decrease SNase activity if Glu43 were a general base during catalysis. In addition, the X-ray crystal structure of the homoglutamate mutant showed that the carboxylate side chain of this residue occupies a position and orientation similar to that of Glu43 in the wildtype

enzyme. Therefore, Glu43 may play a structural role instead and serve as a bidentate hydrogen-bond acceptor to fix the conformation of the neighboring loop.

Proline is unique among the natural amino acids in that its α -nitrogen is part of a pyrrolidine ring. The proline residue disrupts main-chain hydrogen bonding; it cannot serve as a hydrogen-bond donor because of the lack of a backbone NH moiety. Also, proline forms *cis*-peptide bonds at a frequency (5%) much higher than any other natural amino acids (<0.1%). In ion channels, Pro often is conserved at crucial sites, such as Pro221 in the nicotinic acetylcholine receptor (nAChR) and Pro256 in the 5-hydroxytryptamine-3A receptor (5-HT₃AR). To probe which feature of Pro is functionally significant, α -hydroxyl acids (analogs of Gly, Val, and Leu) were incorporated at these sites, which all produced mutant receptors with properties similar to the wildtype receptor (England 1999; Dang 2000). In contrast, incorporation of canonical amino acids Gly, Ala, or Leu yielded nonfunctional receptors. Because α -hydroxyl acids similarly lack the NH moiety for backbone hydrogen bonding and the nature of side chains does not affect receptor activity, these results suggest that the functional importance of the conserved Pro in both receptors is to remove backbone hydrogen bonding.

Another conserved proline residue of the 5-HT₃AR, Pro308, has been shown to be indispensable for channel gating using conventional mutagenesis. However, substitution of this Pro with α -hydroxy acids produced nonfunctional receptors, which suggests that the lack of backbone hydrogen bonding is not the key to the proper function of this Pro. Interestingly, proline analogs that strongly favor the *trans*-conformer (analogs 11 and 12) produced no gating response, but those that favor the *cis*-conformer

(analogs 13 and 14) yielded highly sensitive channels. Moreover, a linear energy correlation was observed between the *cis-trans* energy gap of the proline analogs and the receptor activation (Lummiss 2005 p248). This study strongly suggests that the critical role of Pro308 is to provide the switch that interconverts the open and closed states of the channel through *cis-trans* isomerization.

Cation- π interaction is a noncovalent electrostatic interaction between a cation and the electrons in π orbitals, which plays an important role in protein structure, binding, and catalytic function. The energetic contribution of this interaction to proteins could not be measured accurately with conventional mutagenesis because no positively charged natural isosteres exist for common amino acids. To engineer a cation- π interaction in SNase, Val74, which occupies a hydrophobic pocket composed of one tyrosine side chain and two phenylalanine side chains, was replaced with the positively charged S-methylmethionine (analog 15). Another mutant was made by replacing Val74 with homoleucine (analog 16), which is isosteric to S-methylmethionine. Comparison of the thermodynamic stability of these two mutant proteins showed that the magnitude of cation- π interaction is about $2.6 \text{ kcal mol}^{-1}$ (Ting 1998).

Several aromatic amino acids have been identified near the agonist-binding site of the nAChR, which suggests that cation- π interactions may be involved in binding the quaternary ammonium group of the agonist acetylcholine. A series of progressively fluorinated tryptophan derivatives (analogs 17 to 20) were incorporated at α Trp149. Because fluorine is an electron-withdrawing group, substitution of H with F in the aromatic ring weakens the cation- π interaction. *Ab initio* quantum mechanics was used to predict the cation- π -binding abilities of the fluorinated tryptophans, and the calculated

binding energy has a linear relationship with receptor activation by the agonist (Zhong 1998). Such correlations were not observed for other aromatic residues, which suggests that the cation- π interaction indeed exists for agonist binding and pinpoints it to α Trp149. This interaction was shown later as a general binding pattern between the Cys-loop superfamily of neurotransmitter receptors, such as the 5-HT_{3A} receptors and the γ -aminobutyric acid receptors, and their cationic ligands or substrates (Lummiss 2005 p993).

The site-specific introduction of biophysical probes into proteins has proven extremely powerful in revealing subtle changes of proteins with high spatial resolution. The carbon-deuterium (C-D) bond absorbs at $\sim 2100\text{ cm}^{-1}$, which is within the transparent IR window ($\sim 1800\text{-}2700\text{ cm}^{-1}$) of proteins and, therefore, makes it easily observable by IR spectroscopy. The inherently fast timescale of IR spectroscopy also provides high temporal resolution. Therefore, unnatural amino acids with C-D bonds are excellent probes of protein folding and dynamics. Absorptions at different frequencies indicate the existence of multiple intermediates, and an increased line width of the absorption shows increased flexibility of the local environment. Amino acids containing C-D bonds were incorporated at different positions throughout cytochrome *c* (cyt *c*) by using semisynthetic approaches (Sagle 2006). By characterizing the absorption frequencies and line widths of the C-D bonds of these residues, it was found that no significant difference exists in the flexibilities of the oxidized and reduced states of cyt *c*. The data also show that parts of the protein exist in dynamic equilibrium with locally unfolded states and that cyt *c* is less stable than previous studies suggest.

Another infrared probe, *p*-cyano-L-phenylalanine (*p*CNPhe, 27), has been

genetically encoded in *E. coli* and used to examine different ligand-bound states of the heme group in myoglobin (Schultz 2006). The stretching vibration of the nitrile group of *p*CNPhe has strong absorption and a frequency (ν_{CN}) at $\sim 2200 \text{ cm}^{-1}$, which falls in the transparent window of protein IR spectra. A substitution of *p*CNPhe was made for His64, which is at the distal face and close to the iron center of the heme group in myoglobin. In the ferric myoglobin, when the Fe(III) ligand was changed from water to cyanide, ν_{CN} shifted from 2248 cm^{-1} to 2236 cm^{-1} , which indicates a less polar active site. In the ferrous myoglobin, a ν_{CN} absorption at 2239 cm^{-1} was observed for the linear Fe(II)CO complex, and the bent Fe(II)NO and Fe(II)O₂ complexes showed a ν_{CN} absorption at 2230 cm^{-1} . These results demonstrate that the nitrile group is a sensitive probe for ligand binding and for local electronic environment.

Small fluorescent probes sensitive to various environmental changes have the great potential for monitoring many biologic events as a complementary reporter for the widely used fluorescent proteins. For example, L-(7-hydroxycoumarin-4-yl)ethyl-glycine (CmrGly, 28) has been incorporated into holomyoglobin to study its local unfolding (Wang 2006). CmrGly was incorporated at position Ser4 in helix A and at position His37 in helix C, respectively. The coumarin fluorescence intensity increases with solvent polarity. When the Ser4CmrGly mutant was unfolded with 2M urea, its fluorescence increased 30%, which indicates that helix A is disordered. In contrast, the fluorescence intensity of the His37CmrGly mutant did not change significantly until the urea concentration was raised to 3 M. These results suggest helix C and helix A unfold at different times and concentrations of the denaturing agent.

4.1.5.3: Modification and Regulation of Protein Activity

Green fluorescent protein (GFP), whose chromophore is autocatalytically formed by the tripeptide Ser65-Tyr66-Gly67, has become one of the most important *in vivo* markers for biologic studies. An aromatic amino acid at position 66 is necessary for fluorescence generation. To determine how the spectral properties of GFP could be altered by this residue, tyrosine analogs bearing different substituents at the *para* position of the phenyl ring (analogs 21 to 24) were used to replace Tyr66 (Wang 2003 p174). The absorbance and fluorescence emission maxima of mutant GFPs are all blue-shifted, spanning the range from 375 to 435 nm and 428 to 498 nm, respectively. The wavelengths of the maxima increase in the order of bromo, iodo, methoxy, hydroxyl, amino, and deprotonated hydroxyl group. This shifting trend is consistent with the electron-donating ability of the substituents. In another experiment, Trp66 of the enhanced cyan fluorescent protein was replaced with L-4-aminotryptophan (analog 25) (Bae 2003). The electron-donating amino group significantly red-shifts the fluorescence emission by 69 nm, which changes the color from cyan to gold.

Comparison of the *p*-methoxy-Phe (22) mutant GFP with wildtype GFP also provides direct evidence for the peak assignment of GFP. Wildtype GFP has two absorbance maxima at 397 nm and 475 nm, which are believed to correspond to a neutral chromophore (phenol of Tyr66) and an anionic chromophore (phenolate anion of Tyr66), respectively. Excitation at either absorbance peak leads to a single fluorescence emission centered at 506nm, which corresponds to the anionic chromophore in the excited state (Tsien 1998). Picosecond spectroscopy revealed that the excited neutral chromophore should emit at 460nm (Chattoraj 1996). The absence of 460 nm emission in wildtype GFP suggests that an excited state proton transfer process is involved. Substitution of the

hydroxyl group of Tyr with a methoxy group removes the possibility of deprotonation and proton transfer. Indeed, when Tyr66 is replaced with *p*-methoxy-Phe, only one absorbance maximum at 394nm is observed, which is close to the absorbance maximum of the neutral chromophore of wildtype GFP. Moreover, only one emission maximum at 460nm is detected for this mutant, which corroborates the ultrafast spectroscopic results (Wang 2003 p174).

The specificity of nucleic acid-binding proteins relies greatly on the hydrogen bonding between protein polar atoms and nucleic acid bases. Unnatural amino acids that can change isosterically the hydrogen-bonding pattern have been exploited to alter the substrate specificity. The λ -repressor recognizes the C:G pair at position 6 in the operator site O_L1, and Lys4 of the λ -repressor is crucial for this recognition. The ϵ -NH₂ group of Lys4 forms hydrogen bonds with the carbonyl group of Asn55 and the 6-oxo group of the guanine, which function as two hydrogen bond donors. Substitution of Lys4 with isosteric *S*-(2-hydroxyethyl)-cysteine changes the ϵ -NH₂ to the -OH group, which now should accept hydrogen bonding from the amino group of adenine while preserving hydrogen bonding with Asn55 as a donor (**Figure 4.1.4**). In fact, after the unnatural amino acid was introduced into the λ -repressor through site-directed mutagenesis and chemical modification, the binding specificity was switched from the C:G to T:A base pair (Maiti 2005).

The chirality of D-amino acids has been harnessed for pharmaceutical purposes. D-peptide ligands should be resistant to proteolytic degradation and thus are more desirable as drugs. However, large libraries of D-conformers cannot be encoded genetically and expressed for selection. A method termed mirror-image display solved

this problem in an intriguing way (Schumacher 1996). An L-peptide library is encoded genetically and displayed on the phage surface, and peptides of this library are selected by the target protein that is synthesized using all D-amino acids. The identified L-peptide then is resynthesized using D-amino acids, which should interact with the target protein of the natural handedness for reasons of symmetry. This approach has been used successfully to identify D-peptides that bind the Src homology 3 domain of c-Src and the HIV-1 gp41 protein (Schumacher 1996; Eckert 1999).

Unnatural amino acids that mimic posttranslational modifications can control protein functions. For example, protein phosphorylation regulates many signal transduction pathways and is a reversible process catalyzed by various phosphatases and kinases. The dynamic change of the phosphorylation status of a protein makes it difficult to study the effect of this modification in detail. The generation of metabolically stable phosphoproteins would be useful to dissect the function and to direct signal transduction. Unnatural amino acid *p*-carboxymethyl-L-phenylalanine (*p*CMF, 26) is a non-hydrolyzable analog of phosphotyrosine and was found capable of mimicking the phosphorylated state of Tyr. This capability was demonstrated in a model phosphoprotein, the human signal transducer and activator of transcription-1 (STAT1). STAT1 has only a weak affinity for DNA, but during phosphorylation of Tyr701, STAT1 forms a homodimer and strongly binds a DNA duplex that contains M67 sites. The mutant STAT1 with Tyr701 substituted with *p*CMF also bound the M67-containing DNA duplex tightly, which suggests that *p*CMF could replace phosphotyrosine in the generation of constitutively active phosphoproteins (Xie 2006).

The development of photoactive amino acids provides researchers with an

extremely useful tool not only to probe biologic function but also to control spatially and temporally a variety of biologic processes. One strategy is to attach a suitable photoremovable protecting group to the amino acid, which renders the amino acid inactive. Photolysis releases the caging group and converts the amino acid to an active form, which generates abrupt or localized changes to the target protein. The 2-Nitrobenzyl derivative is the most prevalent form for caged compounds. For example, the conserved Ser1082 at the upstream splice junction of the self-splicing DNA polymerase of *Thermococcus litoralis* was substituted with *o*-(2-nitrobenzyl)serine (**Figure 4.1.5a**). The full-length precursor protein underwent protein splicing only when the unnatural residue was reverted back to wildtype Ser during photolysis (Cook 1995). In other examples, *o*-nitrobenzyltyrosine (**Figure 4.1.5b**) was used to replace Tyr93 or Tyr198 in the α subunit of the nAChR. These two Tyr residues are highly conserved for agonist binding. Millisecond flashes of light at 300-350nm decaged the protected tyrosines and produced abrupt increments of currents that were conducted by the ion channel (Miller 1998). Also, *o*-Nitrobenzyltyrosine has been incorporated at the essential Tyr503 site of β -galactosidase to activate its enzymatic activity by using light both *in vitro* and in *E. coli* (Deiters 2006). Mutation of the active-site cysteine residue in the proapoptotic protease caspase 3 to *o*-nitrobenzylcysteine led to a catalytically inactive enzyme, whose activity could be restored by photocleavage (Wu 2004). In addition to caging the active side chains, the 2-nitrobenzyl group has been harnessed also to cleave the protein backbone photo-chemically. 2-Nitrophenyl glycine (**Figure 4.1.5c**) was introduced into sites of the signature disulfide loop of the nAChR. Irradiation at 360nm resulted in site-specific backbone lesion and an almost complete loss of nAChR activity

(England 1997).

Photolysis of a caged amino acid residue is an irreversible process. Reversible modulation can be achieved with the photochromic azobenzene compounds. Azobenzene undergoes a reversible *cis-trans* isomerization: The more stable *trans*-isomer can be converted to the *cis*-isomer during illumination at 320-340nm, and the *cis*-form can revert to *trans*-form either thermally or by irradiation at >420nm. The resultant change in geometry and/or dipole of the compound can be used for regulating protein activity. For example, a known K⁺ channel blocker, tetra-ethyl ammonium, was linked via an azobenzene group to a cysteine that was introduced at specific sites of a K⁺ ion channel (**Figure 4.1.6a**). When the azobenzene group isomerizes between the extended -orm and the shorter *cis*-form in response to specific wavelengths of light, the structural change moves the blocker into or out of channel-blocking position and, thus, opens and closes the ion channel, respectively (Banghart 2004). Such photomodulation can be used to control neuronal activity noninvasively. The azobenzene group has been encoded genetically in the form of phenylalanine-4'-azobenzene (AzoPhe). AzoPhe was incorporated at the Ile71 site of the *E. coli* catabolite activator protein, a transcriptional activator. Its binding affinity for the promoter sequence decreased fourfold after irradiation at 334 nm (**Figure 4.1.6b**), which converts the predominant *trans* -AzoPhe to the *cis*-form. The isomerized *cis*-AzoPhe then was switched back to the *trans*-state by irradiation at >420nm, after which the affinity of the protein for the promoter was completely recovered (Bose 2006).

4.1.6: Future Directions

The examples summarized here are only representative and by no means

comprehensive. Many unnatural amino acids now can be incorporated, but simply have yet to be used in the investigation of biologic function. Unnatural amino acids that contain photocross-linkers, biophysical probes, chemical moieties with unique reactivities, and posttranslational modifications, among many others, have much promise in their capabilities. The use of these amino acids will expand the capabilities of probing protein structure and function as well as protein-involved biologic processes. The methodology of incorporation is advancing as well. It may be possible to encode genetically unnatural amino acids in many other cell types and organisms. The incorporation of multiple unnatural amino acids simultaneously by using extended codons may enable more complex investigations to be performed.

Additional work using unnatural amino acids can lead to the design and synthesis of novel and diverse biologic functions. By incorporating specific chemical moieties and physical characteristics into proteins, new protein properties may be discovered and used. Such exploration can be attempted either rationally or combinatorially. Diversities of protein libraries would be increased greatly by the addition of only a few unnatural amino acids, which may enhance the probability of discovering proteins that contain novel properties and functions. It is easy to see how unnatural amino acids can be extended into the pharmaceutical industry to create more efficient therapeutics. Finally, the creation of a sustainable organism that is capable of using unnatural amino acids will enable the investigation of the evolution of the genetic code on this planet.

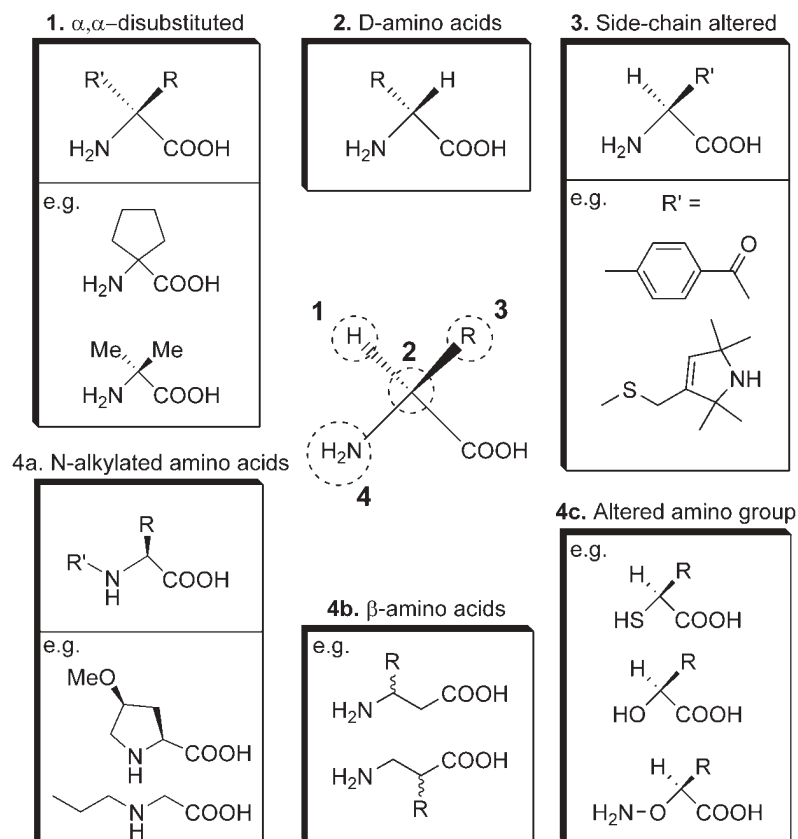


Figure 4.1.1: Different forms of unnatural amino acids.

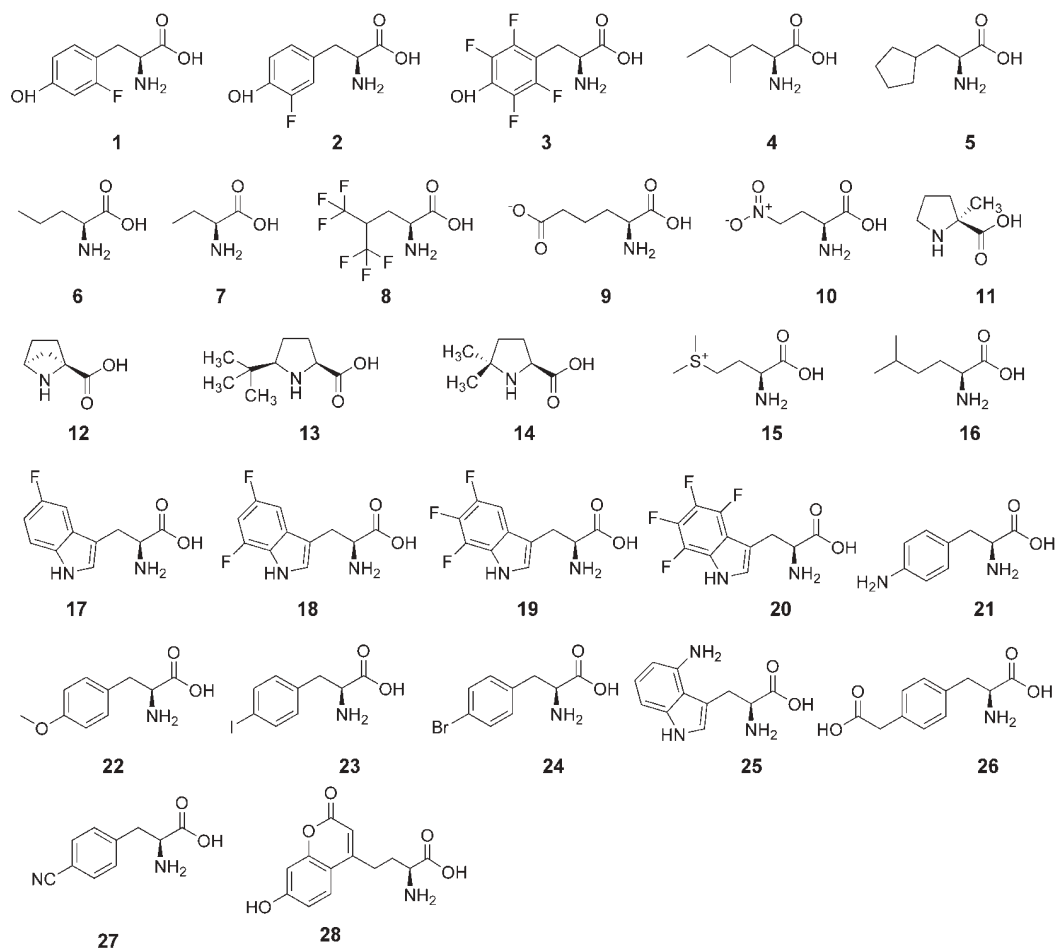


Figure 4.1.2: Structures of unnatural amino acids discussed in the text.

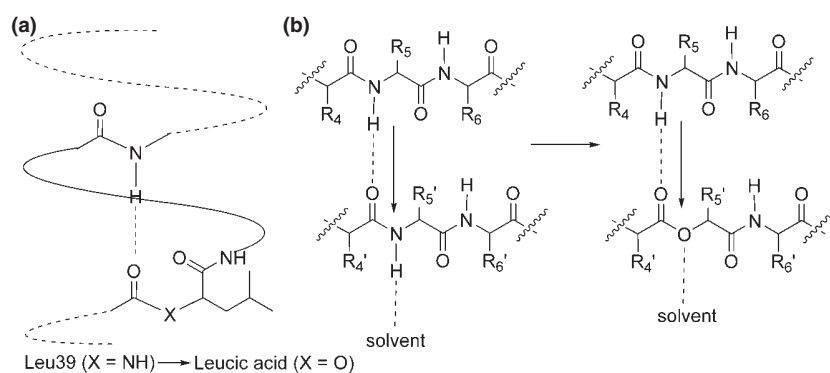


Figure 4.1.3: Backbone mutations generated by α -hydroxy acids. (a) N-terminal mutation of Leu39 to leucic acid in an α -helix of the T4 lysozyme. (b) Substitution of Leu14 with leucic acid in a β sheet of SNase.

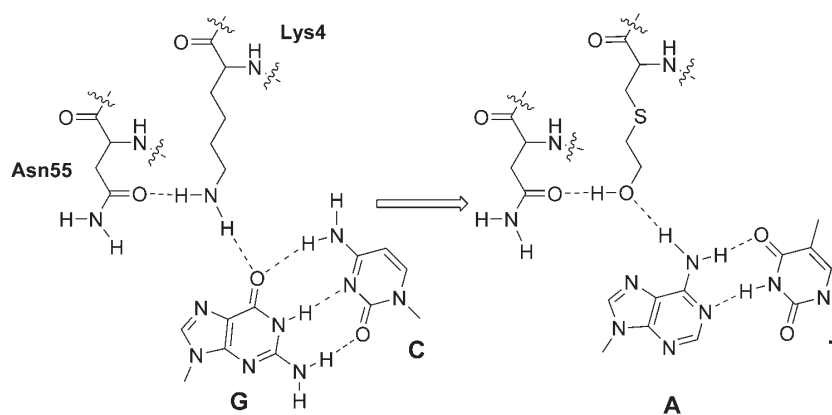


Figure 4.1.4: Substitution of Lys4 with 2-hydroxyethyl-cysteine in the λ -repressor changes the hydrogen-bonding pattern and DNA substrate specificity from C:G to T:A.

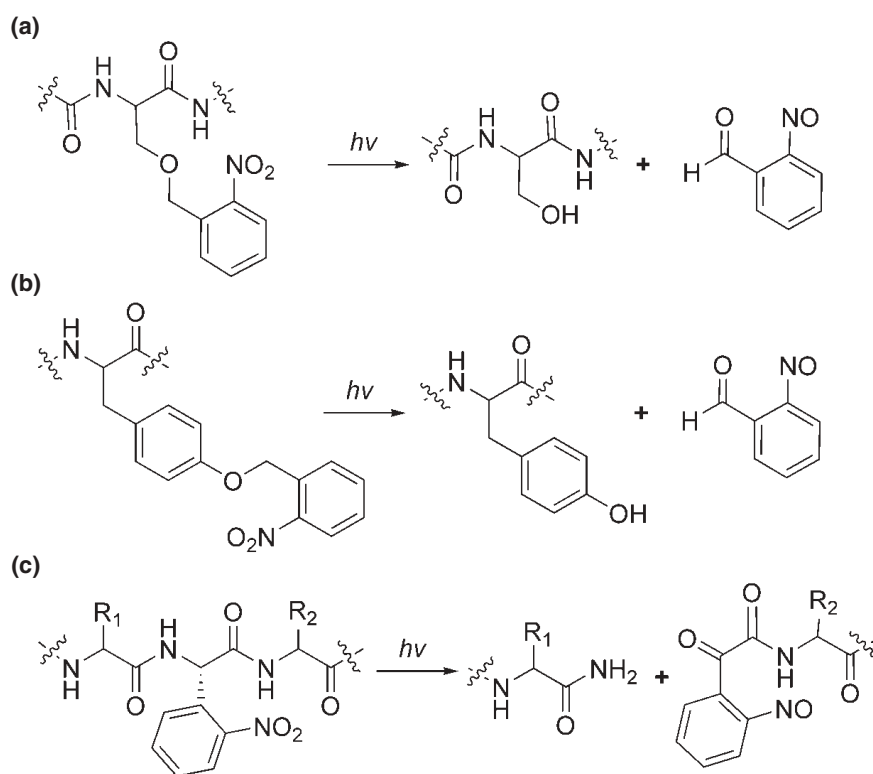


Figure 4.1.5: Photolysis of 2-nitrobenzyl caged serine (a) and tyrosine (b) restores the wildtype residues. Photolysis of 2-nitrophenyl glycine (c) cleaves the protein backbone.

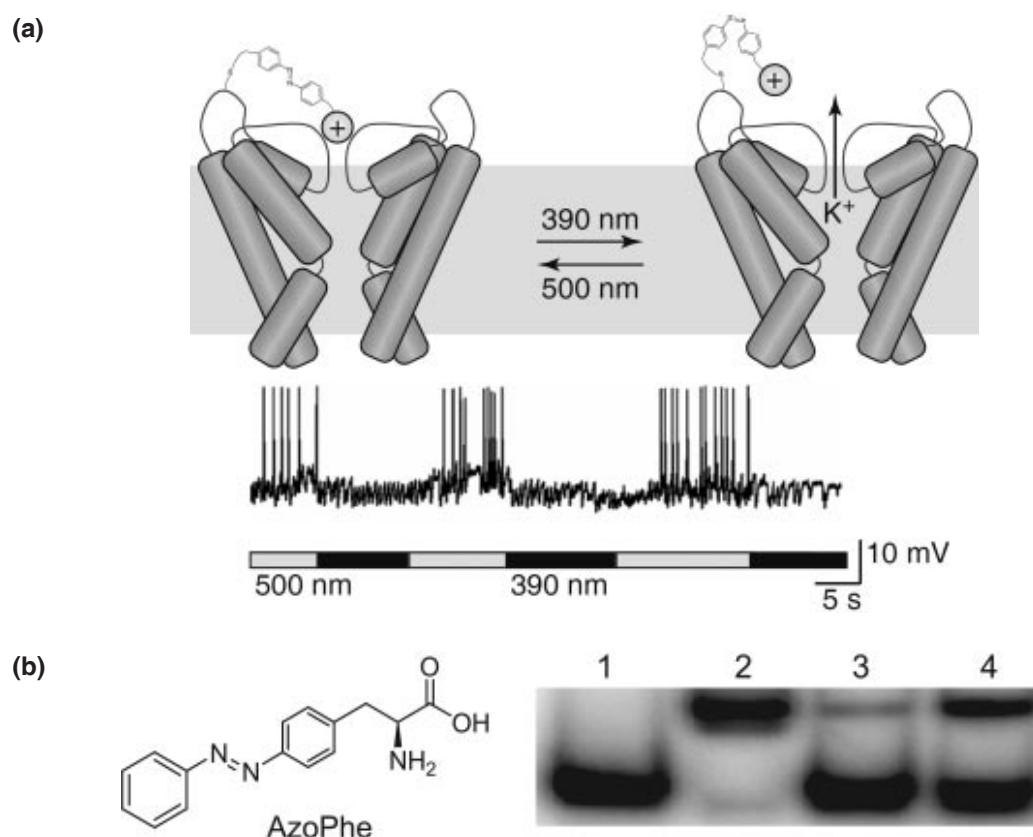


Figure 4.1.6: The known potassium channel blocker, tetra-ethyl ammonium, was linked via an azobenzene group to a cysteine and introduced at specific sites of a K⁺ ion channel. (a) The geometrical change resultant from the cis-trans isomerization of azobenzene moves an ion channel blocker in and out of the ion channel to close and open the ion channel, respectively. Such activity was used to modulate the spontaneously firing hippocampal neurons. The firing frequency is significantly decreased when the azobenzene is in the cis-form after irradiation at 390 nm. Normal firing behavior is restored during irradiation at 500 nm (reprinted from (Gandhi 2005), Copyright 2005, with permission from Elsevier). (b) Structure of phenylalanine-4'-azobenzene (AzoPhe) in trans-form and gel mobility shift assay to determine the binding affinity of the catabolite activator protein (CAP) to the lactose promoter DNA fragment (reprinted with permission from (Bose 2006), Copyright 2006, American Chemical Society). Lane 1, DNA only. Lane 2, DNA+wildtype CAP. Lane 3, DNA+CAP with AzoPhe incorporated at residue 71 (after irradiation at 334 nm). Lane 4, DNA+CAP with AzoPhe incorporated at residue 71 (before irradiation at 334 nm). Substitution of Ile71 with trans-AzoPhe in CAP results in a fourfold decrease of the binding constant K_b of the CAP for its promoter sequence. Photoirradiation at 334 nm partially converts the trans-AzoPhe to the cis-form and decreases the K_b by another fourfold. The latter affinity loss can be completely recovered after irradiation at > 420 nm, which switches the cis-form back to the predominant trans-state.

Chapter 4.2: Release Factor One is Non-Essential in *E. coli*

4.2.1: Abstract

There are many variant codes found in nature that rely on the reassignment of stop signals. These organisms provide valuable insights in the continued evolution of the genetic code but fall short in providing examples of the complex adaptation events underlying alterations to the genetic code. The synthesis of an organism harboring such a change will provide valuable insights to evolutionary and synthetic biology and allow the study of such events in real time. Here we show the synthesis of an *E. coli* strain with release factor one deleted. This strain, unlike those previously published, harbors an unconditional knockout of release factor one with no complementary genomic mutations. As a consequence of RF1 removal, UAG stop codons are blank in this strain, and expression of EGFP harboring internal stop codons results in the incorporation of tyrosine through anticodon:codon mismatches in the ribosome. The addition of an unnatural suppression system grossly outcompetes this endogenous tyrosine incorporation and yields high levels of protein production from EGFP genes harboring up to 10 internal TAGs. Overall, this strain represents an intermediate state in code evolution, allowing in-depth investigation into such drastic changes for the first time. Finally, this strain allows for the efficient incorporation of proteins harboring multiple unnatural amino acids, overcoming a major hurdle in that field.

4.2.2: Introduction

The semi-universal genetic code is not a “frozen accident” as once thought and is capable of undergoing complex changes in a modern organism (Crick 1968). Organisms harboring altered mitochondrial or nuclear codes have been found (Knight 2001 p49),

and more will likely be discovered with the development and application of efficient genomic sequencing techniques. Although the existence of these organisms helps to address the malleability of the code itself, they also offer general insights into cellular adaptation to such a drastic change to their proteomes. For instance, there is an underlying theme to many of the natural variations found in the genetic code, as organisms seem to prefer to reassign or replace stop codons (Knight 2001 p49). Stop codons are rare, only occurring once per gene, and reassignment to a sense codon would not cause deleterious mutational events in catalytic sites of proteins. Reassignment of a stop codon would extend the C-termini of many genes, and may cause these misfolded extended proteins to be degraded by the cellular machinery. These deleterious effects may be small in comparison to other global changes including the reassignment of one sense codon to another, leading to an increase in the of the observed frequency of variant codes with the reassignment of stop codons.

Organisms harboring a variant genetic code give valuable insights into the evolution of the genetic code, however they are at an evolutionary end-point. We can obviously learn much about the initial mutational events underlying the changes (Kawaguchi 1989) but the entire process itself remains elusive. Many questions remain: how does an organism survive the intense selective pressure against changes to the genetic code? How do these changes become fixed in time? Is the adaptation quick, or does it take many generations? Is there an actual benefit to cellular health from these changes? To help address these difficult questions, we would need to study an organism actually undergoing these changes in real time. Indeed, experiments meant to mimic these events have led to valuable insights addressing many of these questions (Santos

1999; Pezo 2004; Bacher 2007 p1907; Silva 2007), but they too fall short of full reassignment. Here we will attempt to generate such an organism. Specifically, we will attempt to knock out the Release Factor 1 from *E. coli* in an attempt to capture the UAG stop codon.

E. coli contains two release factors that recognize and terminate translation in response to the three stop codons. Release factor one (RF1) is responsible for recognition of the UAG and UAA stop codons, while Release Factor 2 (RF2) recognizes UAA and UGA (Craigie 1985). UAG is the least used codon in *E. coli*, and would most likely represent the *easiest* codon to capture from the comparatively minimal effects from its reassignment. UAG is solely recognized by RF1, and removal of RF1 would presumably turn UAG into a nonsense codon in *E. coli*. RF1 is currently considered essential in *E. coli* (Gerdes 2003), however these experiments have been no means been exhaustive. They have only been focused on *E. coli* strains derived from K12, and not on the numerous other strains available. As these strains are derived, mutated for laboratory and cloning use, and grown now for decades, they may not solely represent the biology of *E. coli*. In addition, these experiments fail to separate whether removal of RF1 is infeasible due to a cellular necessity for efficient stoppage at UAG codons, or whether RF2 is insufficient for the recognition of all UAA and UGA stop codons upon removal of RF1 (>92% of all stop codons in *E. coli*). This critical distinction is important, as the latter explanation may be overcome through further manipulation of RF2, allowing the removal of RF1.

The process of reassigning a stop codon has implications towards synthetic biology as well, specifically in the generation of an organism with an expanded genetic code. A large number of unnatural amino acids have been efficiently incorporated in a

variety of organisms, giving valuable insights to a wide variety of biological questions (Wang 2004; Wang 2009). These technologies rely on the introduction of an orthogonal tRNA/synthetase pair responsible for the incorporation of the unnatural amino acid in response to an internal UAG stop codon. Even with the most efficient synthetases however, the reassignment is incomplete due to competition with the endogenous release factors. Removal of RF1 from *E. coli* would remove this competition, increasing the efficiency of such systems and creating the first synthetic expanded genetic code in an extant organism.

4.2.3: Materials and Methods

4.2.3.1: Strain Construction

Removal of the gene encoding RF1 (*prfA*) was attempted using a knockout cassette via established procedures (Datsenko 2000). Briefly, 51 nucleotide overhangs homologous to the regions immediately 5' and 3' of *prfA* were appended to the gene encoding chloramphenicol acetyltransferase (**Figure 4.2.1a**). One microgram of this cassette was then electroporated into the strains listed in **Figure 4.2.1b** harboring the pKD46 plasmid expressing the red-recombinase system. Cm^R clones were screened for knockout by genomic PCR (**Figure 4.2.1a**) and any positive clones were verified by genomic sequencing.

MDS42 and DH10 β derivatives harboring an alanine at position 246 in *prfB*, the gene encoding RF2, were constructed as follows. A knockin cassette was first generated containing the *prfB* gene from BL21 (DE3) transcriptionally coupled to a Kan^R cassette. The Kan^R cassette was flanked on the 3' end by a 51-nucleotide region homologous to the 3' end of the endogenous *prfB* gene. One microgram of this cassette was

electroporated into MDS42 and DH10 β harboring the pKD46 plasmid. Kan^R clones were screened using PCR and sequence verified for mutation of position 246 to alanine.

4.2.3.2: Genomic Sequencing

Genomic DNA from BL21 (DE3) and JX1.0 was harvested and purified using a Qiagen DNeasy kit. One μ g of genomic DNA was used to prepare DNA libraries for sequencing. Genomic DNA was fractionated using the Covaris S2 System (Applied Biosystems, Foster City, CA) using the following parameters: cycle number = 6, duty cycle = 20%, Intensity = 5, cycles/burst = 200 and time = 60 seconds. Fractionated DNA was purified using a Qiagen PCR minielute purification kit. Purified DNA was repaired using the Epicentre End-It Repair Kit (Madison, WI) and purified using a Qiagen minielute column. Purified DNA was A-tailed using dATP and Klenow (3'-5' exo-) from New England Biolabs (NEB, Ipswich, MA) and then purified with a Qiagen minielute column. Purified DNA was then ligated overnight with Illumina genomic DNA adapters using T4 DNA Ligase from NEB and purified using a Qiagen minielute column. The ligated DNA was run on a 2% agarose gel and size selected to remove adapters. Gel extraction was performed on the gel slice using Qiagen minielute gel purification kit. Purified DNA was PCR amplified using 1 μ L of ligated DNA and Phusion Taq from NEB and size selected from a 2% agarose gel.

Genomic DNA libraries were sequenced using the Illumina Genome Analyzer II (Illumina, San Diego, CA) as per manufacturer's instructions. Sequencing of genomic DNA libraries was performed up to 82 cycles. Image analysis and base calling were performed with the standard Illumina pipeline (Firecrest v1.3.4 and Bustard v.1.3.4).

Sequence alignments and SNP analysis were performed using the SHORE package (Ossowski 2008) according to the documentation provided with the software. In brief, the *E. coli* BL21 (DE3) reference genome was preprocessed into a SHORE acceptable format. Next, FASTQ files for each sample were converted to a SHORE flat file format. Reads were mapped using Genomemapper contained within the SHORE package using the following parameters $-n\ 4$, $-g\ 3$. From this analysis the only deletion different between the two strains was the *prfA* gene.

4.2.3.3: Plasmid Construction

All plasmids were assembled by standard cloning methods and confirmed by DNA sequencing. pAIO plasmids containing an EGFP gene with different TAG codons were synthesized as the following: EGFP cassettes with an N-terminal His6 tag, containing TAG's at various positions were created using overlapping PCRs. The following sites were used: Y182 for 1-TAG; Y39 and Y182 for 2-TAG; Y39, Y182 and Y151 for 3-TAG; Y39, K101, D102, E132, D133, K140, E172, D173, D190 and V193 for 10TAG. These cassettes were first cloned into pBP-Blunt (Biopioneer, San Diego, CA), and then digested and ligated into pBK-AIO vectors containing the orthogonal tRNA^{Tyr}_{CUA} and the *M. jannaschii* TyrRS (Wang 2001) or the LW1RS (Wang 2003 p56) using *Spe* I and *Bgl* II. pBAD vectors containing the 1 or 3-TAG containing EGFP genes were constructed by inserting the EGFP constructs described using the *Nco* I and *Hind* III cut sites.

4.2.3.4: In-Cell Fluorescence

In-cell fluorescence intensity was determined using a FluoroLog-3 (Horiba Jobin Yvon). *E. coli* colonies were picked and grown 16 hours with or without unnatural

amino acids. Cells were washed two times in PBS buffer, and diluted in PBS to an OD₆₀₀ of 0.1. The emission spectrum of EGFP was measured using an excitation wavelength of 488 nm scanning from 503 to 560 nm. Fluorescence intensity of each sample was compared using the intensity at the maximal emission at 511 nm.

4.2.3.5: Western Analyses

E. coli cells containing EGFP expression plasmids were grown at 37°C for 16 hours, harvested, washed 2 times with PBS and diluted to an OD₆₀₀ of 0.1 in PBS. One milliliter of cells was collected and resuspended in 100µl Blue Juice (Qbiogene, Carlsbad, CA) incubated for 10 minutes at 95°C, and proteins were separated on 12% or 15% SDS polyacrylamide gel. After transfer, EGFP was detected using the HRP-conjugate penta-His antibody (Qiagen, Valencia, CA). All blots were developed using the pico chemiluminescence kit (Thermo Scientific, Rockford, IL) according to manufacturer's specifications.

4.2.3.6: Protein Purification

For EGFP preps, 500mL cultures were grown for 16 hours with or without unnatural amino acid. Cells were pelleted, resuspended in lysis buffer (10% glycerol, 50mM Tris pH8.0, 500mM NaCl, 5mg/mL lysozyme, DNase, and 10mM β-Mercaptoethanol) and sonicated for 4 cycles at 90% power with a duty cycle of 50. Cell lysate was collected after centrifugation at 12,000 x g for 30 minutes. Lysate was then added to 500µl of pre-equilibrated Ni-NTA resin (Qiagen, Valencia, CA) washed with 50 column volumes of wash buffer (50mM Tris, 500mM NaCl, 20mM Imidazole, pH8) and then eluted in three 1ml fractions of elution buffer (same as wash, but 250mM Imidazole). Purified EGFP was buffer exchanged to 50mM Tris buffer containing

500mM NaCl (pH 8.0) using Microcon Ultracel YM-10 spin columns (Millipore, Billerica, MA), and further purified using a Sephadex-200 size exclusion column on a UPC-900 FPLC (GE healthcare, Piscataway, NJ). Peak fractions were analyzed by SDS-PAGE and pooled for further analysis. All protein concentrations and total yields were determined using the Bio-Rad protein assay kit (Hercules, CA) according to manufacturer's specifications.

4.2.3.7: Mass Spectrometry

Trypsin-digested protein was analyzed by LC electrospray ionization MS as described (Schubert 2009; Herrera 2009). Briefly, samples were loaded onto a capillary column with integrated spray tip (75 μ m I.D., 10 μ m tip, New Objective, Woburn, MA), which was packed in-house with C₁₈ reversed phase material (Zorbax SB-C₁₈, 5 μ m particle size, Agilent, Santa Clara, CA) to a length of 10 cm. The reversed phase elution was achieved by of a linear gradient of 0-60% acetonitrile in 0.1% formic acid within 60 minutes at a flow rate of 300 nl/min. The eluate was introduced into a Thermo LTQ-Orbitrap mass spectrometer (ThermoFisher, Waltham, MA) via a nano-spray source. Mass spectrometric analysis was conducted by recording precursor ion scans at a resolution of 60,000 in the Orbitrap Fourier-transform analyzer followed by MS/MS scans of the top 5 ions in the linear ion trap (cycle time approx. 1 s). An active exclusion window of 90 s was employed. Data were analyzed on a Sorcerer Solo system running Sorcerer-Sequest and by using the Mascot algorithm (V. 2.7 rev.11, Matrix Science, London, UK). Data were further analyzed and visualized using the Scaffold software package (v. 2.6, Proteome Software).

4.2.3.8: Growth Assay

A colony was picked for each *E. coli* strain and grown overnight with appropriate antibiotic. Cells were normalized to an OD₆₀₀ of 1 and diluted 1:50 in fresh media with antibiotic and unnatural amino acid (if applicable). For BL21 (DE3) strains, OD₆₀₀ was then measured every 30 minutes for 10 hours. For JX1.0 derived strains OD₆₀₀ was then measured every 60 minutes for 48 hours. Doubling times were then calculated from the exponential growth phase in each strain.

4.2.4: Results

4.2.4.1: Synthesis and Sequencing of the RF1-Deletion *E. coli* Strain JX1.0

To determine if RF1 is essential in *E. coli* we attempted to replace RF1 with a chloramphenicol resistance cassette (see methods and **Figure 4.2.1a**) in a variety of strains (**Figure 4.2.1b**) using an established method (Datsenko 2000). We included the BP5 α strain which harbors a glutamine suppressor tRNA recognizing UAG. The presence of this tRNA implies that this strain already has ambiguity on the UAG stop codon, and if successful would suggest that this is an important factor in removal of RF1. We also included the minimal genome MDS42 strain derived from K12 (Posfai 2006) that will determine if a reduction in genome size can assist with RF1 removal. We engineered three additional strains harboring mutations in the gene encoding RF2 in DH10 β and MDS42. The RF2 gene in K12 strains contains the Ala246Thr mutation lowering RF2 recognition of UAA 10-fold (Uno 1996). This mutation may severely impact the ability of RF2 to recognize all UAA stop codons upon removal of RF1, so it was reverted back to alanine in these two strains (The amino acid identity at this position for each strain is shown in **Figure 4.2.1b**). We also introduced a further engineered RF2 gene into MDS42 (named JX2.0) that will be described in detail in Chapter 4.3.

Knockout of RF1 was assayed by genomic PCR (**Figure 4.2.1b**) and proved successful in two of the strains used. RF1 deficient derivatives of BL21 (DE3) and JX2.0 were generated, named JX1.0 and JX3.0 respectively. JX3.0 was fully characterized and will be discussed in detail in Chapter 4.3. Surprisingly, knockout in all other strains, including MDS42, BP5 α , and two of our engineered strains was unsuccessful, suggesting that a minimal genome, a suppressor tRNA, and reversion of the A246T mutation are insufficient on their own to allow for RF1 removal. To further characterize JX1.0, full genomic sequencing was performed and compared to the parental BL21 (DE3) strain. RF1 deletion was verified, and a number of additional single nucleotide polymorphisms were found (**Table 4.2.1**), none of which correspond to known mutations that complement an RF1 deficiency (Zhang 1994; Dahlgren 2000; Kaczanowska 2004). Many of these are in the terminator region or are silent mutations in the *ycgX* gene that is of phage origin. One lies in the potential promoter region of a predicted protease, and another introduces a modest amino acid substitution in a sugar transporter. Overall, none of these mutations would explain or complement the deletion of RF1 and thus suggest that RF1 function has been removed.

4.2.4.2: RF1 Removal in JX1.0 Affords the Synthesis of EGFP Harboring Multiple Internal UAG Codons, Regardless of the Presence of a Suppressor System

If RF1 function has been removed from JX1.0, UAG stop codons should now be blank or nonsense codons. To ascertain if UAG is blank in JX1.0, a variety of EGFP mutants harboring 1, 2, 3, or 10 internal UAG codons were expressed with or without an unnatural suppression system. This suppression system inserts the unnatural amino acid *p*-Acetylphenylalanine (*pActF*) in response to UAG codons via an orthogonal

tRNA/synthetase pair (Wang 2003 p56). Normally this system competes with the endogenous release factors and is limited to efficient expression of proteins harboring only a single stop codon, and the addition of multiple stop codons dramatically reduces protein yields (Kaya 2009; Huang 2010). A single plasmid system (pAIO) harboring the orthogonal $tRNA_{CUA}^{Tyr}$, the LW1RS synthetase, and the EGFP gene (**Figure 4.2.2a**) was introduced into JX1.0 and the parental BL21 (DE3). Expression of full-length EGFP was then assayed by western blotting (**Figure 4.2.2b**) and in-cell fluorescence (**Figure 4.2.2d**). BL21 (DE3) showed efficient expression of the 1-TAG mutant when *pActF* was supplied that decreased with the addition of each TAG. The insertion of 3-TAG codons in EGFP essentially abolished expression of the protein. JX1.0 showed high expression of EGFP in all mutants, and protein levels were similar in 1, 2, or 3-TAG containing mutants regardless of the presence of *pActF* (**Figure 4.2.2b and d**). Levels of the 10-TAG mutant were also high in the presence of *pActF*, however very little protein expression was seen without the unnatural amino acid (**Figure 4.2.2b**).

To determine if the high level of EGFP expression in JX1.0 even without *pActF* added to the growth media was due to the pAIO plasmid, the 1 and 3-TAG EGFP mutants were cloned into the pBAD expression system. Any expression of EGFP from this plasmid in JX1.0 would represent endogenous machinery, and suggest that the background seen in the pAIO experiments was not due to the presence of the orthogonal tRNA or synthetase. These vectors were introduced into JX1.0 and the parental BL21 (DE3) strain, and expression was induced overnight with arabinose. Expression of EGFP was assayed via an in-cell fluorescence assay (methods). Not surprisingly, BL21 (DE3) showed no expression of full-length EGFP (**Figure 4.2.2e**). JX1.0 however had a high

level of expression in the 1-TAG mutant, and a small, but reproducible level of expression in the 3-TAG mutant (**Figure 4.2.2e**). The overall reduction in EGFP fluorescence is most likely due to differences in promoters, as the P_{Ara} promoter from pBAD is weaker than the T_5 promoter used in the pAIO plasmid system (Deuschle 1986). These data suggest that JX1.0 has the ability to read-through UAG codons without any exogenously supplied machinery.

4.2.4.3: Mass Spectrometry Shows that Anticodon-Codon Mismatches are Responsible for the Production of EGFP, and are Outcompeted by *pActF*

To further investigate the mechanism of UAG readthrough in JX1.0 tryptic mass spectrometry was performed on various EGFP mutants. Tryptic mass spectrometry of the 1-TAG EGFP protein from the pAIO and pBAD experiments revealed that tyrosine, glutamine, and tryptophan was inserted at UAG positions in the absence of *pActF* (**Figure 4.2.3a and b**). Monoisotopic masses for glutamine and tyrosine were within 0.02 Da of the expected masses (4437.17 and 4472.18 respectively), and tryptophan was within 1 Da of the expected mass (4495.19Da). No other peak corresponded to other amino acids in this analysis. To ensure that this is not from a suppressor mutation in the endogenous tyrosyl-tRNAs, these loci were resequenced and found to be wildtype in JX1.0. This suggests that the readthrough of UAGs in these proteins is due to anticodon-codon mismatching during translation. Interestingly, when *pActF* is added to the growth media, it appears to outcompete these misincorporations, as most peptides observed have *pActF* at UAG positions (**Figure 4.2.4 a, b, and c**). Top-down results for 3-TAG are consistent, with all three positions containing *pActF* and no other amino acid (**Figure 4.2.4a, b, and c**). Bottom-up results are mostly consistent with only *pActF*-containing

peptides in the 1 and 3-TAG EGFP samples (**Figure 4.2.4e**). However, tyrosine was found in the 2-TAG sample at position 39 (**Figure 4.2.4d**), suggesting that some level of tyrosine contamination is still present. A rough estimation of these two species suggests that tyrosine incorporation is around ~10% of *pActF*.

4.2.4.4: The Yields of EGFP Harboring Multiple *pActF* are Very Robust in JX1.0

Yields were determined for the 1, 2, 3 and 10-TAG samples grown with *pActF* in the growth media. Purification of these samples according to the described protocol was very efficient and pure, with little contamination (**Figure 4.2.2c**). Yields were very high and determined to be 8.5, 7.1, 9.7, and 1.2 mg/L in the 1, 2, 3, and 10-TAG EGFP samples respectively. There is a drop off in yield in the 10-TAG sample, suggesting that protein folding may be interfering with efficient expression of the 10-TAG EGFP sample. Regardless, these yields are very high (even compared to JX3.0, see Chapter 4.3), and combined with the mass spectrometry data suggest that the UAG codon is essentially reassigned to *pActF* in this strain.

4.2.4.5: JX1.0 has Severe Growth Defect Exacerbated by the Presence of *pActF*

We finally assessed the health of JX1.0 by a simple growth assay. JX1.0 is stable in culture, cloneable, and healthy with the exception of a general growth defect compared to the parental BL21 (DE3) (**Figure 4.2.5**). The doubling time for JX1.0 has more than tripled, taking 91 minutes compared to 25.5 minutes for the parental BL21 strain. Introduction of an unnatural suppression system increases this further, either from the general toxicity from the system, or from the presence of kanamycin (doubling times of 135 and 41 minutes for JX1.0 and BL21 (DE3) respectively). Finally, the addition of the unnatural amino acid to the media exacerbates this even further increasing the doubling

time to 253 minutes for JX1.0, while not affecting BL21 (DE3). This is interesting, as it suggests that the efficient incorporation of an unnatural amino acid at UAG positions throughout the proteome is more detrimental to cellular health than the endogenous, relatively weak, level of tyrosyl, glutamine, and tryptophan suppression.

4.2.5: Discussion

Overall, our data suggests that RF1 function has been removed in JX1.0, and UAG stop codons are essentially blank. The JX1.0 strain appears to compensate for these blank codons through codon:anticodon mismatches in the ribosome. This is most likely efficient enough to enable readthrough at the endogenous UAG stop codons, as expression of even a 10-TAG containing EGFP was achieved without an unnatural amino acid present (**Figure 4.2.2b**). However, some level of ribosomal drop off must also be occurring, as truncated products are seen in the EGFP western analyses particularly in the $-pActF$ lanes (**Figure 4.2.2b lane 3**). UAG appears to be primed for reassignment, and the introduction of an orthogonal system yields high protein expression, even in a 10-UAG containing mutant. However, this appears to put a burden on the strain, as growth is extremely perturbed under these conditions. This is internally consistent, as reassignment of UAG to $pActF$ is efficient and most likely nearly complete, lowering the amount of truncated products via ribosomal drop-off. This would presumably lead to increased toxicity via perturbed protein folding and function in the ~300 genes that end in a UAG stop codon.

The level of toxicity seen in JX1.0 with the addition of an unnatural suppressor system may mimic an intermediate state of code evolution. This strain is by no means healthy, and is most likely undergoing a large amount of pressure to remove or relieve

this toxicity. Further analysis of these adaptations can now be done in real time to access the response to such global changes to the proteome. JX1.0 thus represents a novel strain for the study of codon reassignment and adaptation. Unlike other ambiguous strains developed before (Pezo 2004; Silva 2007), JX1.0 reassignment appears complete with the addition of an unnatural suppressor system. Continued growth of this strain under such pressures will help to address many of the questions regarding the evolution of the genetic code that were previously untenable.

JX1.0 shows that the function of RF1 is not essential in the *E. coli* strain BL21 (DE3). Further work is required to pinpoint the differences between the K12 and B strain derivatives that allow for RF1 knockout. There are many known differences in these strains, including the insertion of the DE3 prophage (Jeong 2009). Numerous other SNPs exist between the two strains, and further experimentation will help to tease out what differences allow for removal of RF1. Detailed understanding of these differences will also help to address some questions regarding the continued evolution of the code itself. JX1.0 now affords us an *E. coli* strain to directly address such questions.

Finally, JX1.0 can be utilized immediately in synthetic biology, as it represents an unconditional knockout of RF1 not reliant on any exogenous factors of complementary mutations. Suppression of UAG by an orthogonal system is greatly enhanced, allowing for the efficient incorporation of multiple unnatural amino acids in a single protein. In affect, this strain represents the first strain harboring an expanded genetic code. The introduction of an orthogonal system expands the genetic code of *E. coli* to include a 21st amino acid. This helps to overcome a major hurdle in this field, and will prove extremely

useful in a variety of applications pertaining to structural, functional, and enzymatic biology.

Table 4.2.1: Single nucleotide polymorphisms between the derived JX1.0 and the parental BL21 (DE3)

Position	Reference ¹	Read	Number	Conf	Gene	Mutation
1200795	T	C	382	1	(2)	
1200805	C	T	385	1	(2)	
1200849	C	T	409	1	(2)	
1200862	C	T	397	1	<i>ycgX</i>	Silent
1200961	A	C	198	1	<i>ycgX</i>	Silent
1200982	T	C	217	1	<i>ycgX</i>	Silent
2530079	G	C	17	0.809	(3)	
3708564	T	C	422	1	ECD_03520 ⁴	Silent

(1) Reference genome for BL21 (DE3) is As# CP001509 (Jeong 2009).

(2) These mutations lie in the intergenic region between *ycgX* and *ycgW*.

(3) This mutation lies in the intergenic region between *suhB* and *yfhR*.

(4) ECD_03520 is a conserved gene involved in transport of hexuronate.

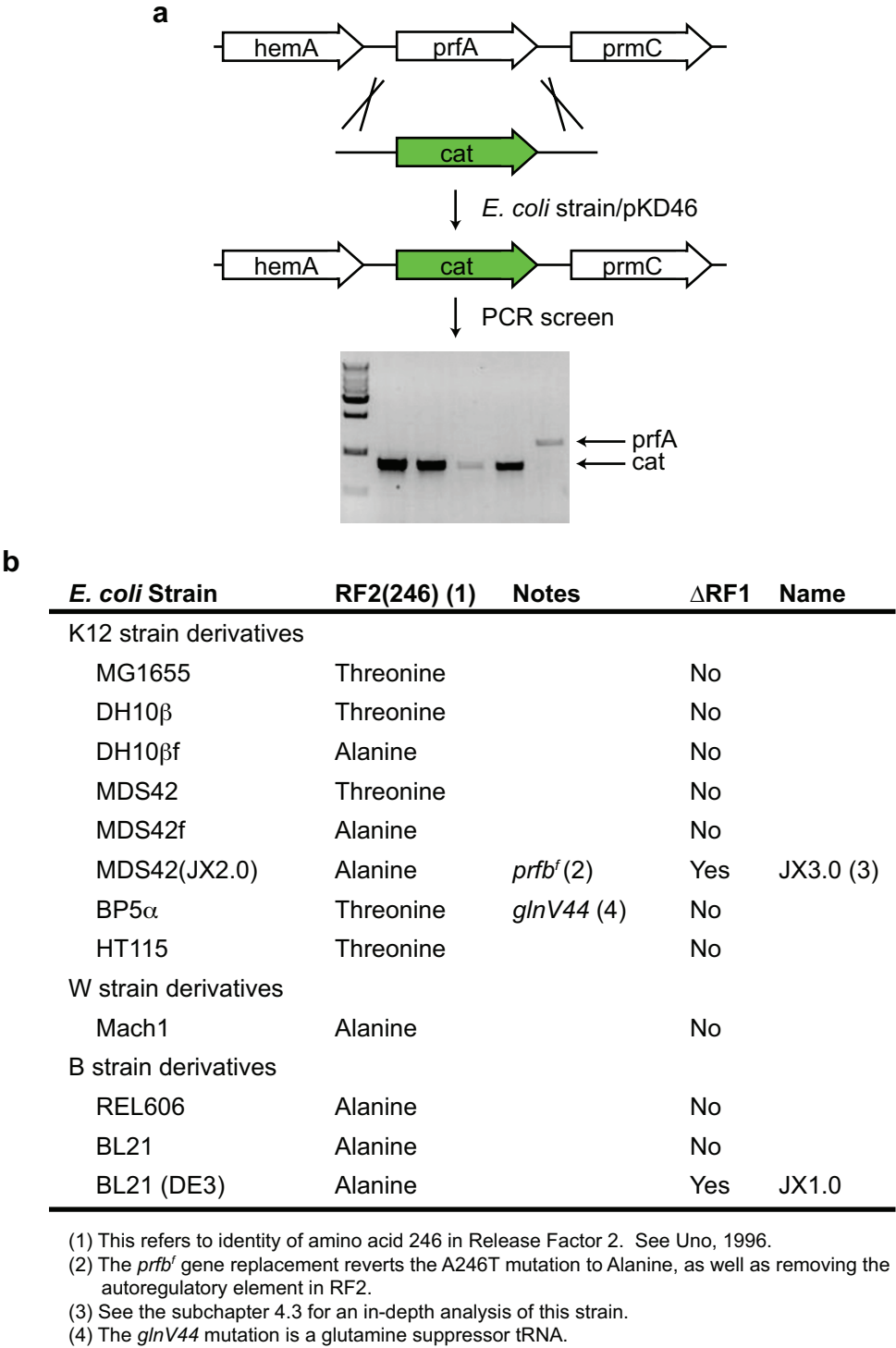


Figure 4.2.1: Knockout of RF1 in a variety of *E. coli* strains. **(a)** *prfA* knockout cassette and procedure used in these experiments. **(b)** A list of the strains used in this subchapter.

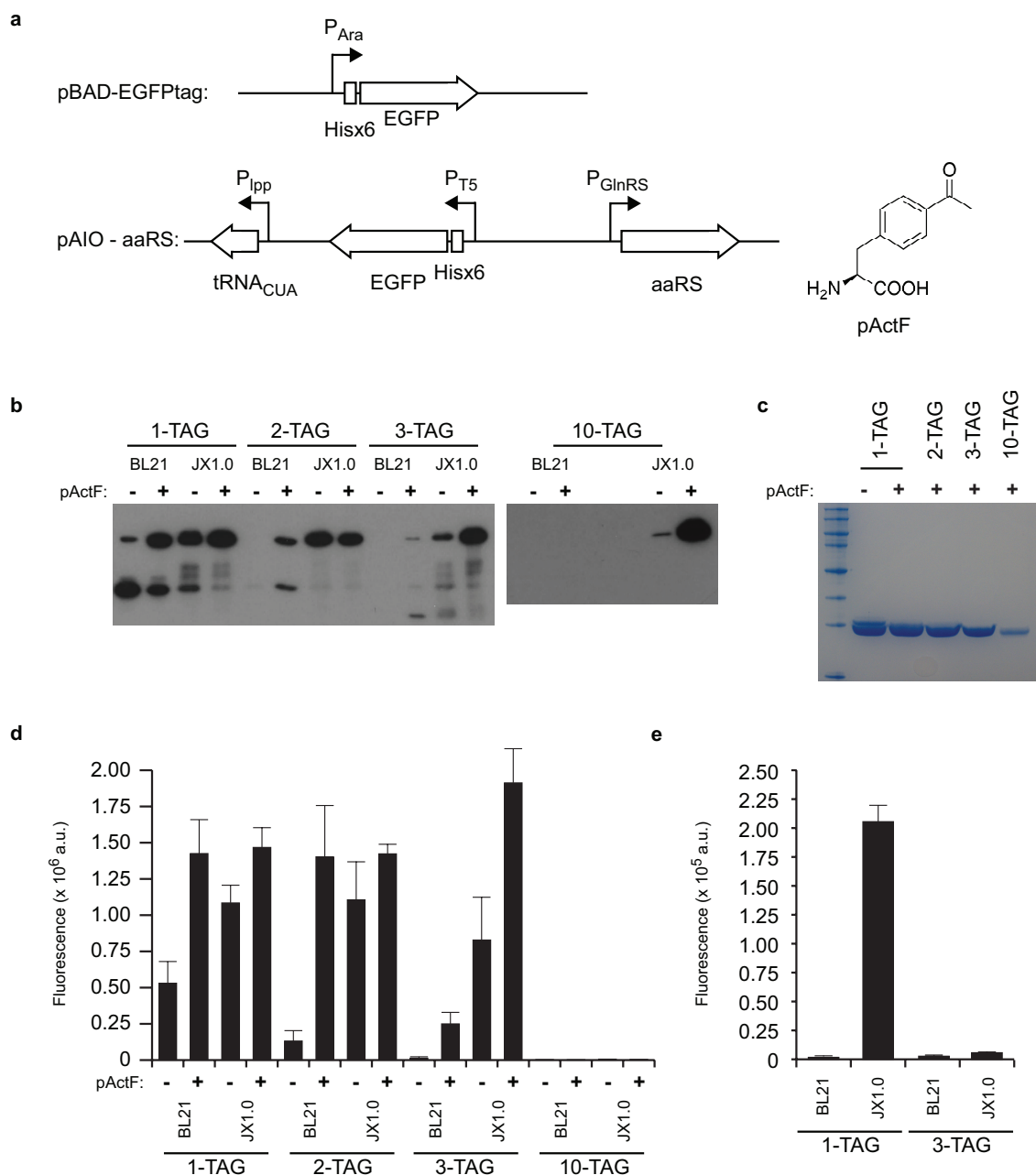


Figure 4.2.2: Removal of RF1 in JX1.0 enables the encoding of unnatural amino acids at multiple UAG sites. **(a)** The pBad and All-in-One plasmids. Structure of *pActF* is shown to the right. **(b)** Western analysis and **(d)** in-cell fluorescence of EGFP expression when UAG codons are decoded as *pActF* by *pAIO-LW1RS*. Measurement was performed on 3 independent batches of cells and error bars represent s.e.m.. **(c)** Commassie staining of various samples shows a high level of purity after purification. **(e)** in-cell fluorescence of EGFP expression in the pBAD vector with the 1 and 3-TAG containing EGFP genes. Conditions are the same in **(d)**.

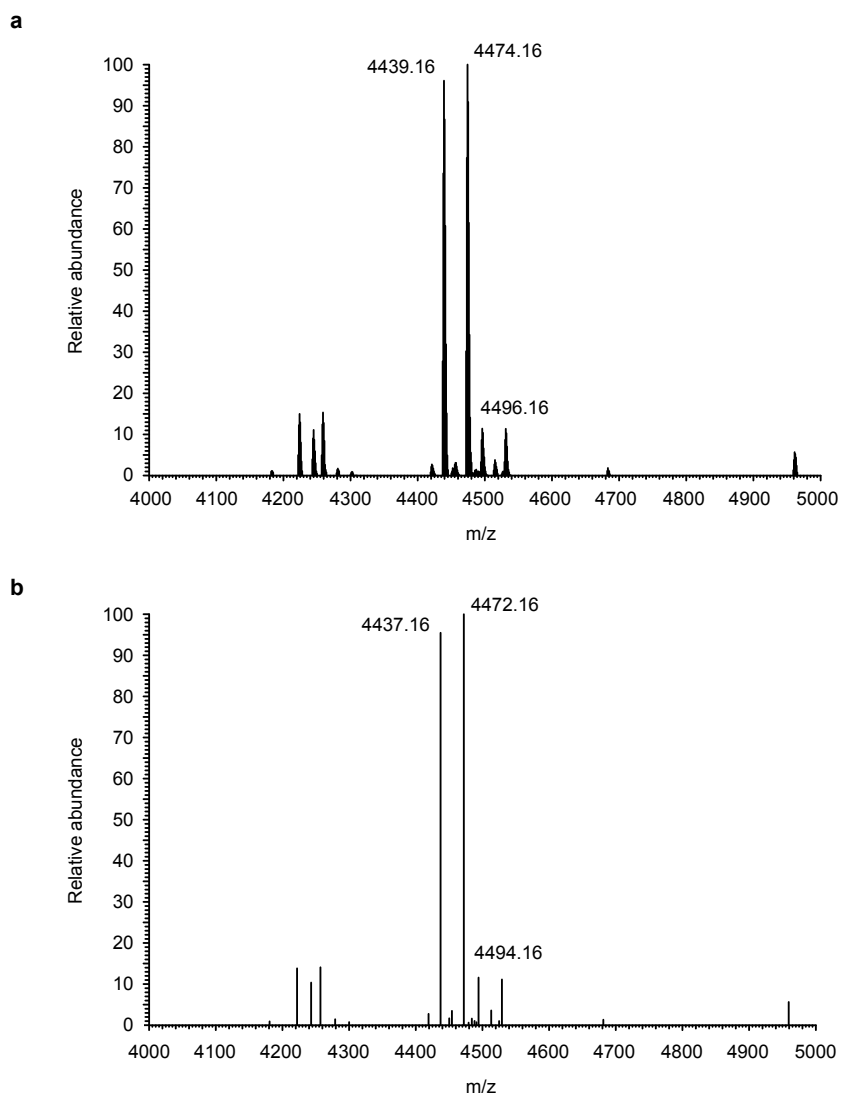


Figure 4.2.3: Top-down mass spectrometry of EGFP purified from pBAD without *pActF* in the growth media. **(a)** Deconvoluted mass spectrometry reveals three main species at position 182. **(b)** Monoisotopic masses from **a** shows that glutamine, tyrosine, and tryptophan are inserted at position 182 (ex. 4437.17, 4472.17, 4495.19 Da respectively).

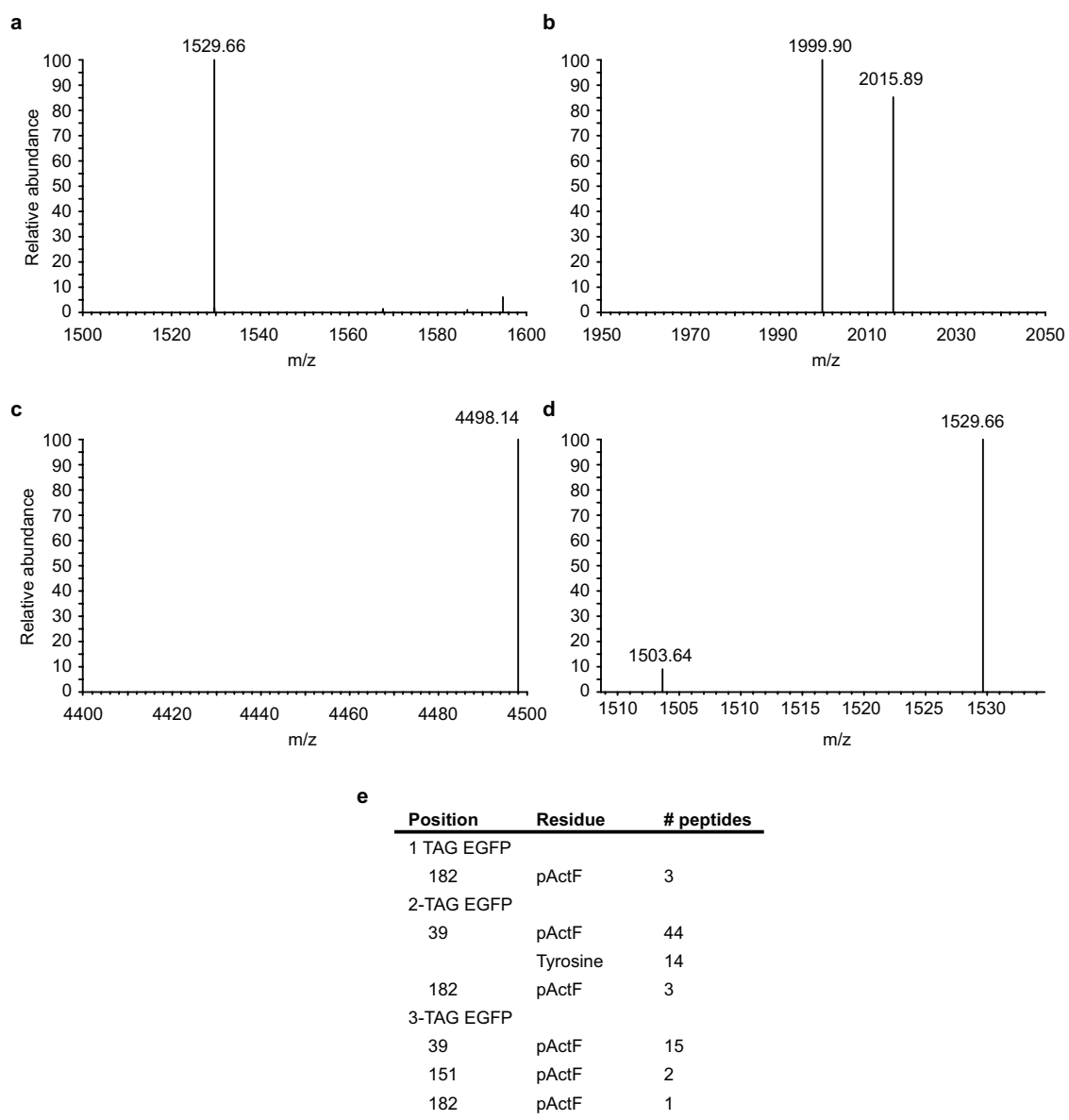


Figure 4.2.4: Mass spectrometry of various EGFP samples grown with *pActF* suggests that unnatural amino acid incorporation outcompetes tyrosine, glutamine, and tryptophan. Deconvoluted monoisotopic mass spectrometry results from the 3-TAG sample shows *pActF* at positions 39 (**a**), 151 (**b**), and 182 (**c**) (ex. 1529.65, 1999.89, and 4498.18 Da respectively). No other amino acids were found. The 2-TAG sample showed two species at position 39, one corresponding to *pActF* (expected mass: 1529.65) and the other corresponding to tyrosine (expected mass 1503.65). (**e**) Table showing compiled results from all three analyses.

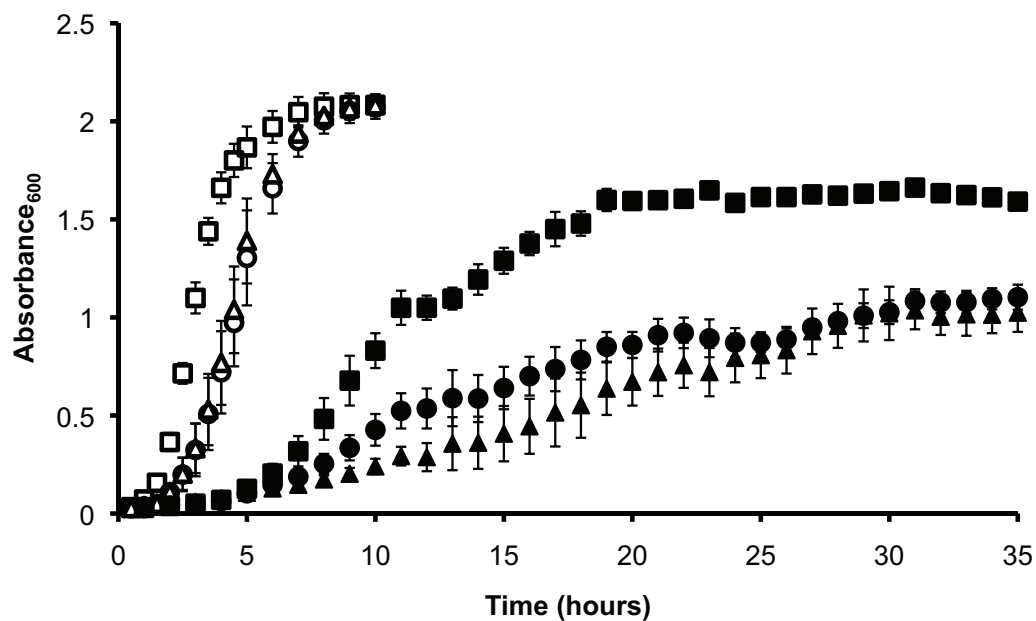


Figure 4.2.5: JX1.0 has a growth defect compared to the parental BL21 (DE3) exacerbated by the presence of an unnatural suppression system and unnatural amino acid. Cells grown without the pAIO plasmid are shown as squares, with the pAIO plasmid as circles, and with both the pAIO plasmid and *pActF* as triangles. BL21 (DE3) are open squares, circles and triangles, and JX1.0 strains are closed.

Chapter 4.3: Reassigning the Amber Stop Codon from Nonsense to Sense in *Escherichia coli*

4.3.1: Abstract

Deviations from the canonical genetic code suggest that the code is evolving and not frozen. Model organisms are lacking for laboratory investigation of the codon reassignment process and cellular adaptation to such challenge. Stop codons have been exploited for genetic incorporation of unnatural amino acids in live cells, but the efficiency is inherently low due to competition from release factors, limiting the power and scope of this technology. Here we knocked out the reportedly essential release factor 1 from *E. coli* by fixing release factor 2, successfully reassigning the amber codon UAG from nonsense to sense in the resultant JX3.0 strain. Various unnatural amino acids were efficiently incorporated into proteins at multiple UAG sites in JX3.0 but not in the parental *E. coli*. This work represents a synthetic reassignment of the UAG codon in a living organism. The resultant strain affords a previously unavailable model for investigating the code evolution process and cellular impact, and provides a unique host for synthesizing and evolving novel protein functions by encoding unnatural amino acids at multiple sites.

4.3.2: Introduction

The canonical genetic code specifies 61 sense codons for amino acids and 3 nonsense codons for stop signals in protein translation. It was once thought to be a “frozen accident”, because the code was universally preserved in all of the known organisms at the time and changes to it would affect all proteins simultaneously and be deleterious to the host (Crick 1968). Small deviations from the canonical code were later

discovered in the mitochondrial and nuclear codes of an increasing number of organisms (Osawa 1992; Knight 2001 p49). These include the reassignment of sense codons from one amino acid to another and the reassignment between nonsense and sense codons. These deviations in codon assignments suggest that the code is still evolving in extant lineages. The codon capture (Osawa 1989) and ambiguous intermediate (Schultz 1994) hypotheses are the two predominant explanations for code evolution. The codon capture hypothesis posits that shifts in genomic GC content can eliminate certain codons from the entire genome. When these codons reemerge, they can be captured by mutant tRNAs and reassigned (Osawa 1989). The ambiguous intermediate hypothesis proposes that a codon can have ambiguous meaning when decoding fidelity is impaired. The new assignment replaces the old one through adaptation (Schultz 1996; Schultz 1994). These hypotheses provide valuable insights into how the genetic code evolves on a theoretical level. An experimental system would be desirable to test these theories in extant organisms and to explore mechanisms of code change and adaptation in detail (Bacher 2005; Santos 2004). The ability to mimic the codon reassignment process in a live organism would enable the investigation of many outstanding questions that are otherwise intractable. Some major questions include: what constrains the code from less subtle changes, how do organisms survive and cope with codon reassignment, and how is a variant code fixed into an organism? Moreover, natural code evolution occurs over millions of years, while the development of a laboratory model can compact it into a practical time scale, potentially allowing the entire process to be studied in real time.

Stop codons have been exploited for the incorporation of natural and unnatural amino acids into proteins. Stop codons occur only once per gene, and this relative

scarcity may mitigate the damage caused by codon reassignment. Natural suppressor tRNAs decoding a stop codon as a common amino acid have been identified in *E. coli* and other organisms (Benzer 1962; Beier 2001). In addition, orthogonal tRNA-synthetase pairs have also been generated to incorporate unnatural amino acids into proteins in response to a stop codon (Wang 2001; Wang 2004). Using this approach, unnatural amino acids with a variety of functional groups have been genetically incorporated into proteins in both prokaryotic and eukaryotic cells (Wang 2009). However, the incorporation efficiency of either natural or unnatural amino acid is inherently limited, because the suppressor tRNA is in competition with endogenous release factors, whose native function is to recognize stop codons and terminate translation. The assignment of the stop codon is thus ambiguous in all cases, severely limiting the full exploitation and potential of this technology. Release factor competition results in truncated protein products, which may interfere with target protein function or be deleterious to the host cell. Low incorporation efficiency also prevents the synthesis of proteins containing unnatural amino acids at multiple sites. Protein yields drop precipitously with the addition of even a second stop codon. Therefore, it is currently infeasible to synthesize proteins with unnatural amino acid modifications at multiple sites and to explore novel protein and organism functions through experimental evolution.

To begin to address these questions, we aim to reassign the amber codon UAG from the stop signal to an amino acid in *E. coli*. This study will determine whether it is feasible to synthetically reassign a codon in *E. coli*, and if successful, afford a model system allowing for in-depth investigations into cellular adaption and evolution to codon reassignment. In prokaryotes, stop codons are recognized by two release factors, RF1 for

UAA/UAG and RF2 for UAA/UGA (Scolnick 1968). To achieve full reassignment of the amber codon UAG, RF1 must be removed from the system. However, the *prfA* gene encoding RF1 has been reported to be essential for *E. coli* (Rydén 1984; Gerdes 2003). Here we show that RF1 gene can be knocked out of the *E. coli* genome by fixing the expression of RF2. The RF1 knockout strain is stable and sustainable. This new strain enables the genetic incorporation of natural and a variety of unnatural amino acids into proteins at numerous UAG sites, essentially reassigning UAG to a sense codon. UAG codons in endogenous genes are also suppressed but result in no significant negative effect to overall cell fitness. Lastly, through whole genome sequencing we identified a novel mutation in RF2, which may also contribute to the unique phenotype of this new *E. coli* strain.

4.3.3: Methods

4.3.3.1: Strain Construction

All strains in this study were created using the lambda-red recombinase system (Datsenko 2000; Tischer 2006). JX2.0 and 3.0 were constructed as follows: first, a mutagenesis cassette was generated using overlapping PCRs. This cassette contained a mutated form of RF2 (*prfB^f*), a chloramphenicol resistance (Cm^{R}) cassette flanked by two I-SceI cut sites, and homologous regions on both the 5' and 3' end to facilitate recombination. The mutant *prfB^f* had the in-frame premature TGA element removed (Craigén 1985), a shine-delgarno like sequence mutated to a *Sac* II cut-site, and the A246T mutation reverted back to alanine (Uno 1996). This cassette was electroporated into MDS42 (Posfai 2006) cells harboring the pKD46 plasmid (Datsenko 2000). Cm^{R} colonies were screened using PCR to verify a correct knockin, and then by *Sac* II

digestion to verify that the mutant *prfB^f* was present. The resultant strain was transformed with the plasmid pATBSR, a derivative of pACBSR that has a tetracycline resistance (Tet^R) cassette in place of the original Cm^R gene (Herring 2003). Following induction with arabinose, cells were screened for removal of the Cm^R cassette using PCR and sequence verified. Curing of the pATBSR plasmid led to the final JX2.0 strain.

JX3.0 was created using JX2.0 as the parental strain. JX2.0 cells harboring the pKD46 plasmid were electroporated with a PCR cassette to knock out the endogenous *prfA* gene. This cassette contained a Cm^R gene flanked by 5' and 3' homologous overhangs to facilitate recombination. Cm^R colonies were again screened by PCR and sequence verified. The resultant strain, JX3.0, contains an exact replacement of *prfA* with the Cm^R gene. A Tet^R derivative of this strain was constructed from JX3.0 using a similar manner, and was used to express histone H3 using the two-plasmid system.

JX2.0 and 3.0 derivatives containing an N-terminal FLAG-tag in the *yfiA* and *sufA* genes were created as follows: A PCR cassette was synthesized using overlapping PCRs to yield a construct containing a 5' homologous region followed by an I-SceI flanked Kanamycin resistance (Kan^R) cassette and an N-terminal FLAG-tag appended onto the target gene, which itself serves as the 3' overhang. In addition, immediately 5' of the Kan^R cassette is 75bp of DNA that is perfectly homologous to 75bp on the 3' of the Kan^R cassette. This repeat element will help to leave a scarless insertion of the N-terminal FLAG tag upon excision of the Kan^R cassette. These constructs were electroporated into JX2.0 or 3.0 cells containing the pKD46 plasmid. Kanamycin resistant clones were screened for insertions using PCR and sequence verified for FLAG-tag insertion. Resultant strains were transformed with pATBSR, induced with arabinose, and screened

for removal of the Kan^R cassette. Sequence verified clones were then used for further studies.

To construct an MRA8 derivative harboring a *prfB*^f(A293E) gene identical to that in JX3.0, a construct was synthesized using overlapping PCRs harboring a full-length copy of *prfB*^f(A293E) from JX3.0 with a Kan^R cassette on the 3' end. This construct was electroporated into MRA8 cells harboring the pKD46 plasmid, and kanamycin resistant clones were screened for insertion using PCR. Resultant strains were sequence verified and used for further analyses.

4.3.3.2: Growth Assay

A colony was picked for each *E. coli* strain and grown overnight with appropriate antibiotic. Cells were normalized to an OD₆₀₀ of 1 and diluted 1:50 in fresh media without antibiotic. OD₆₀₀ was then measured every 20 minutes for 10 hours.

4.3.3.3: Plasmid Construction

All plasmids were assembled by standard cloning methods and confirmed by DNA sequencing. pAIO plasmids containing an EGFP gene with different TAG codons were synthesized as the following: EGFP cassettes with an N-terminal His6 tag, containing TAG's at various positions were created using overlapping PCRs. The following sites were used: Y182 for 1TAG; Y39 and Y182 for 2TAG; Y39, Y182 and Y151 for 3TAG; Y39, Y74, Y143, Y151, Y182 and Y200 for 6TAG; Y39, K101, D102, E132, D133, K140, E172, D173, D190 and V193 for 10TAG; a 10 tandem TAG codons in place of E172 and D173 for 10TAGtd. These cassettes were first cloned into pBP-Blunt (Biopioneer, San Diego, CA), and then digested and ligated into pBK-AIO vectors

containing the orthogonal $tRNA_{CUA}^{Tyr}$ and the *M. jannaschii* TyrRS (Wang 2001) or the LW1RS (Wang 2003 p56) using *Spe* I and *Bgl* II.

Histone H3 expression was accomplished using a two-plasmid system. The pTak plasmid contained the *M. barkeri* $tRNA_{CUA}^{Pyl}$ and various copies of human histone H3. This plasmid was synthesized by first shuttling the *Pst* I/*Dra* II fragment from the pRep plasmid (Santoro 2002) into pLei (Wang 2003 p56). Next, codon-optimization of human histone H3 was accomplished using Gene Design (Richardson 2006), and synthesized by overlapping PCRs using 40bp primers and shuttled into pBP-Blunt. Positive clones were verified by PCR and the resultant DNA was digested with *Spe* I and *Bgl* II and ligated into the pTak plasmid. Various mutant forms of histone H3 were then synthesized and also cloned into the pTak system. The following species of histone H3 were cloned: 1TAG – K9; 2TAG, K9 and K14; 3TAG – K9, K14, and K18; 4TAG – K9, K14, K18, and K23. The second plasmid, pBK-tAceK harbors the actK aminoacyl-tRNA synthetase (Neumann 2008). The six mutations characterized (D76G, L266V, L270I, Y271F, L273A, and C313F) were inserted into the wildtype pyrrolysyl synthetase from *M. barkeri* using overlapping PCR. This cassette was digested with *Nde* I and *Pst* I and ligated into the pBK-JYRS vector (Wang 2001). The *trc* promoter from pTrc (Invitrogen, Carlsbad, CA) was then used to replace the GlnRS promoter in pBK-JYRS using the *Bam* HI and *Nde* I cut sites.

4.3.3.4: Western Analyses

E. coli cells containing EGFP expression plasmids were grown at 37°C for 16 hours, harvested, washed 2 times with PBS and diluted to an OD₆₀₀ of 0.1 in PBS. One milliliter of cells was collected and resuspended in 100µl Blue Juice (Qbiogene,

Carlsbad, CA) incubated for 10 minutes at 95°C, and proteins were separated on 12% or 15% SDS polyacrylamide gel. After transfer, EGFP was detected using the HRP-conjugate penta-His antibody (Qiagen, Valencia, CA).

For Western analysis of YfiA, a modified version of an established protocol was used (Agafonov 2001). Briefly, one liter of *E. coli* cells harboring an N-terminal FLAG tagged *yfiA* gene were grown for 16 hours at 37°C. Cells were cold-shocked in ice-water for ten minutes followed by two hours of growth at 15°C. Cells were harvested by centrifugation and frozen at -80°C. SufA purification was also accomplished via an established procedure (Lee 2004). For *E. coli* cells harboring an N-terminal FLAG tagged *sufA* gene, a 50ml culture was grown for 16 hours at 37°C. Cells were diluted to an OD₆₀₀ of 0.02 in one liter of fresh media. Once the OD₆₀₀ reached 0.2, phenazine metholsulfate (Sigma, St. Louis, MO) was added to a final concentration of 0.1mM. Cells were harvested by centrifugation and pellets were frozen at -80°C after 90 minutes of growth at 37°C. Protein from both cell types was extracted using BPER reagent (Thermo Scientific, Rockford, IL) and then applied to an Anti-Flag agarose column from Sigma to remove the vast majority of contaminating protein. Purified YfiA and SufA were run on 15% or 20% SDS polyacrylamide gels, and detected by using a monoclonal FLAG M2 antibody (Sigma, St. Louis, MO).

All blots were developed using the pico chemiluminescence kit (Thermo Scientific, Rockford, IL) according to manufacturer's specifications.

4.3.3.5: Protein Purification

For EGFP preps, 100mL cultures were grown for 16 hours with or without unnatural amino acid. Cells were pelleted and lysed for 10 minutes at room temperature

in the appropriate volume of BPER. Cell lysate was collected after centrifugation at 12,000 x g for 5 minutes. Ni-NTA resin (Qiagen, Valencia, CA) was added directly to the lysate and shaken for 1 hour at 4°C. The lysate was added to a column, washed with 10 column volumes of wash buffer (50mM NaH₂PO₄, 300mM NaCl, 10mM Imidazole, pH7.5) and then eluted in six 1ml fractions of elution buffer (Same as wash, but 200mM Imidazole). Purified EGFP was buffer exchanged to 20mM Tris buffer (pH 8.0) using Microcon Ultracel YM-10 spin columns (Millipore, Billerica, MA), and further purified using 1 mL Resource Q ion exchange column on a UPC-900 FPLC (GE healthcare, Piscataway, NJ). The column was equilibrated with a low salt buffer (20 mM Tris·HCl, pH 8.0), and proteins were eluted with a linear gradient of 0 – 0.5 M NaCl. Peak fractions were analyzed by SDS-PAGE and pooled for further analysis.

For histone H3 preps, *E. coli* colonies transformed with the two-plasmid system were picked and grown 16 hours. Cells were diluted 1:100 into fresh media containing 5mM actK or 1mM *pActF*. For all ActK preps, nicotinamide was added to a final concentration of 5mM to minimize deacetylation. When the OD₆₀₀ reached 0.5, cells were induced using 0.4mM of IPTG and grown for 4 hours at 37°C. Cells were pelleted and frozen overnight to facilitate lysis. Pellets were thawed for ten minutes in a water bath, and then processed as described (Luger 1999). The final cell lysate was collected and applied to a Ni-NTA column pre-equilibrated with wash buffer (6M guanidine HCl in PBS pH 7.6, 25mM Imidazole). Lysate was applied to the column 2 times, followed by 20 times the column volume of wash buffer. Fractions containing Histone H3 were eluted with elution buffer (wash buffer plus 250mM Imidazole) and analyzed by SDS page. Histone H3 was further purified using 1 mL Resource Q ion exchange column on a

UPC-900 FPLC (GE healthcare, Piscataway, NJ). A gradient system with a low salt buffer (7M Urea, 10mM Tris-HCl, 1mM EDTA, 5mM β -Me, 100mM NaCl) and high salt (same as low salt but with 600mM NaCl) buffer was used. One-milliliter fractions were collected and those containing histone H3 were pooled for further analysis.

All protein concentrations and total yields were determined using the Bio-Rad protein assay kit (Hercules, CA) according to manufacturer's specifications.

4.3.3.6: Mass Spectrometry

Intact protein analysis by ESI-MS was performed with purified *pActF*-containing EGFP proteins dissolved in 1% formic acid and infused into a LTQ Velos mass spectrometer at 1 μ L/min by a syringe pump. MS scans were collected for 1 minute. About 1,600 MS spectra were collected for each sample. Spectra were averaged and the charge states were de-convoluted using a freeware MagTran (Zhang 1998 p225).

For Histone H3 intact protein mass spectrometry, the sample was analyzed by high resolution Fourier-transform mass spectrometry on a Thermo LTQ-Orbitrap XL. The sample was loaded onto a capillary column with integrated spray tip, which was filled with reversed phase material (Zorbax SB C-18, particle size 5 μ m). Protein was eluted with a gradient of 0.1% formic acid and an increasing proportion of acetonitrile at a flow rate of 300 nl/min. The eluate was electrosprayed directly into the mass spectrometer. Fourier-transform mass spectra were recorded a resolution of 60,000 for a scan range of $m/z = 400$ -1800. This was followed by a select ion scan of the most intense ion at resolution 60,000. Data were deconvoluted using the Thermo Qualbrowser 2.0 Xtract program. A four minute window (elution time 82-86 min) was found to contain a protein of mass $MH^+ = 16558.04$.

Tandem MS analysis was performed with purified *pActF*-containing EGFP, SufA and YfiA proteins solubilized in 50 mM Hepes (pH 7.2). The proteins were reduced and alkylated using 1 mM Tris(2-carboxyethyl)phosphine (Fisher, AC36383) at 95 °C for 5 minutes and 2.5 mM iodoacetamide (Fisher, AC12227) at 37 °C in dark for 30 minutes, respectively. *pActF*-containing EGFP was digested with 1:50 chymotrypsin (Roche, 11418467001). SufA was digested with 1:50 trypsin (Roche, 03708969001) and YfiA was digested by both trypsin and Lys-C (Roche, 11420429001) at 37°C overnight. Automated 2D nanoflow LC-MS/MS analysis was performed using LTQ tandem mass spectrometer (Thermo Electron Corporation, San Jose, CA) employing automated data-dependent acquisition. The detailed LC-MS/MS method can be found in previously published work (Tanner 2007; Castellana 2008). Briefly, the peptides were fractionated by the on-line SCX column using a series of 7 salt gradients (10 mM, 20 mM, 30 mM, 50 mM, 70 mM, 100 mM, and 1M ammonium acetate for 20 minutes), followed by high resolution reverse phase separation using an acetonitrile gradient of 0 to 80% for 120 minutes.

The full MS scan range of 400-2000 m/z was divided into 3 smaller scan ranges (400-800, 800-1050, 1050-2000) to improve the dynamic range. Both CID (Collision Induced Dissociation) and PQD (Pulsed-Q Dissociation) scans of the same parent ion were collected for protein identification and quantitation. Each MS scan was followed by 4 pairs of CID-PQD MS/MS scans of the most intense ions from the parent MS scan. A dynamic exclusion of 1 minute was used.

The raw data was extracted and searched using Spectrum Mill (Agilent, version A.03.02). The CID and PQD scans from the same parent ion were merged together.

MS/MS spectra with a sequence tag length of 1 or less were considered as poor spectra and discarded. The rest of the MS/MS spectra were searched against the NCBI (National Center for Biotechnology Information) RefSeq protein database (version 21, January 2007) limited to *E. coli* (16,324 sequences) plus the SufA and YfiA protein sequences with extended C-terminal sequences, as well as EGFP protein sequence. The enzyme parameter was limited to full chymotrypsin, tryptic or Lys-C peptides with a maximum miscleavage of 1. All other search parameters were set to SpectrumMill's default settings (carbamidomethylation of cysteines, +/- 2.5 Da for precursor ions, +/- 0.7 Da for fragment ions, and a minimum matched peak intensity of 50%). A variable modification of Gln to *pActF* (+61 Da) was used for *pActF*-containing EGFP database search. MS/MS spectra were validated using the criteria list in **Table 4.3.1**.

4.3.3.7: In-Cell Fluorescence Assay

In-cell fluorescence intensity was determined using a FluoroLog-3 (Horiba Jobin Yvon). *E. coli* colonies were picked and grown 16 hours with or without unnatural amino acids. Cells were washed two times in PBS buffer, and diluted in PBS to an OD₆₀₀ of 0.1. The emission spectrum of EGFP was measured using an excitation wavelength of 488 nm scanning from 503 to 560 nm. Fluorescence intensity of each sample was compared using the intensity at the maximal emission at 511 nm.

4.3.3.8: Genomic Sequencing of *E. coli* Strains

Genomic DNA from JX2.0 and JX3.0 was harvested and purified using a Qiagen DNeasy kit. One µg of genomic DNA was used to prepare DNA libraries for sequencing. Genomic DNA was fractionated using the Covaris S2 System (Applied Biosystems, Foster City, CA) using the following parameters: cycle number = 6, duty cycle = 20%,

Intensity = 5, cycles/burst = 200 and time = 60 seconds. Fractionated DNA was purified using a Qiagen PCR minielute purification kit. Purified DNA was repaired using the Epicentre End-It Repair Kit (Madison, WI) and purified using a Qiagen minielute column. Purified DNA was A-tailed using dATP and Klenow (3'-5' exo-) from New England Biolabs (NEB, Ipswich, MA) and then purified with a Qiagen minielute column. Purified DNA was then ligated overnight with Illumina genomic DNA adapters using T4 DNA Ligase from NEB and purified using a Qiagen minielute column. The ligated DNA was run on a 2% agarose gel and size selected to remove adapters. Gel extraction was performed on the gel slice using Qiagen minielute gel purification kit. Purified DNA was PCR amplified using 1 μ L of ligated DNA and Phusion Taq from NEB and size selected from a 2% agarose gel.

Genomic DNA libraries were sequenced using the Illumina Genome Analyzer II (Illumina, San Diego, CA) as per manufacturer's instructions. Sequencing of genomic DNA libraries was performed up to 82 cycles. Image analysis and base calling were performed with the standard Illumina pipeline (Firecrest v1.3.4 and Bustard v.1.3.4). Sequence alignments and SNP analysis were performed using the SHORE package (Ossowski 2008) according to the documentation provided with the software. In brief, the *E. coli* K-12 MG1655 reference genome was preprocessed into a SHORE acceptable format. Next, FASTQ files for each sample were converted to a SHORE flat file format. Reads were mapped using Genomemapper contained within the SHORE package using the following parameters $-n\ 4$, $-g\ 3$. Capitalizing on the large amount of coverage for this experiment we identified large deletions. We posited that any region of the reference genome where reads did not map were regions that were deleted in this strain. Therefore,

we subtracted all positions from the reference genome that were covered by at least one read. The set of positions left over were the ones we called deleted. From this analysis the only deletion different between the two strains was the *prfA* gene. FASTQ files for each sample have been deposited to the Short Read Archives (SRA Accession# SRA016379.1).

4.3.3.9: Statistics

Statistical analysis of the GCT-GAA knockin data was performed using Fisher's exact test on Prism software (GraphPad Software, La Jolla, CA).

4.3.4: Results

4.3.4.1: Generation of the RF1 Knockout Strain JX3.0

Knockout of RF1 is an enticing avenue for reassignment of the UAG codon given the overlap in recognition of UAA by RF2 and UAG being the least used stop codon in *E. coli*, with approximately 300 genes ending in UAG (Nakamura 2000). RF1 is considered essential in *E. coli* (Gerdes 2003), and only conditionally lethal knockouts have been described (Rydén 1984), suggesting that either accurate stoppage at UAG is an essential trait in *E. coli* or that stoppage of the numerous UAA stop codons is not possible with RF2 alone. Consistently, our initial attempts to directly knock out the corresponding *prfA* gene in the common DH10 β strain with a knockin cassette containing the chloramphenicol acetyltransferase gene (Cm^R knockin cassette) failed. We then tried a two-step strategy using a curable plasmid to express *prfA* in DH10 β cells. A gene cassette expressing *prfA* was cloned into pKD46, which encodes the λ -red recombination system and can be cured at 37 °C or higher temperature. In the presence of this plasmid expressing RF1 exogenously, the genomic *prfA* was knocked out by the Cm^R knockin

cassette, indicating a functional Cm^R knockin cassette and λ -red recombination system. However, these cells were unable to survive once transferred to higher temperatures to cure the *prfA* expressing plasmid. These experiments indicate that it is unfeasible to directly knock out the *prfA* gene in the genomic background of the *E. coli* strain DH10 β .

We next tried to understand the lethality caused by *prfA* knockout in order to possibly compensate for it. One possibility could be that endogenous genes ending with the UAG codon will not be terminated correctly during translation once RF1 is removed. Mutating UAG in these genes to UAA would eliminate this potentially lethal consequence. There are over 300 genes in *E. coli* that use UAG for termination and it would prove difficult to change all of them in a reasonable time scale. Only 9 (*pgpA*, *arcC*, *sucB*, *ycgX*, *yffB*, *ygaU*, *ygbA*, *yibA* and *priB*) of these genes are considered essential (Gerdes 2003), so repair of the UAG in these 9 genes was attempted. Nine corresponding knockin cassettes were synthesized that change UAG to UAA, and transformed into DH10 β harboring pKD46 together with the Cm^R knockin cassette for RF1. However, after attempting this experiment under different conditions for numerous times, no RF1 knockout clone was identified.

The tmRNA surveillance system for translational stalling could be another potential source of lethality. If the *prfA* gene is knocked out, endogenous mRNAs ending with UAG will stall the ribosome due to the absence of RF1-mediated peptide release. Stalled ribosomes will be recognized by the alanyl-tmRNA, which enters the stalled ribosome to disassemble and recycle ribosomal subunits. The nascent peptide is transferred to an alanine on the tmRNA, and translation switches from the original mRNA to a short open reading frame on the tmRNA encoding a degradation tag. After translation and

termination the tagged polypeptide is targeted for degradation (Moore 2007). This surveillance mechanism may result in the degradation of many corresponding proteins and cause cell lethality. We reasoned a tmRNA-deficient *E. coli* strain would avoid the degradation of these proteins. However, our attempts to knock out *prfA* in the tmRNA-deficient strain (X90 *ssrA1::cat*) (Keiler 1996) also failed.

RF2 expression in *E. coli* is tightly autoregulated by an in-frame UGA codon in its mRNA that requires a +1 frameshift to generate full-length RF2 (Craigen 1985). Interestingly, *E. coli* strains derived from K-12 contain a peculiar mutation (Ala246Thr) in RF2 that reduces its recognition of UAA 10-fold (Uno 1996). RF2 expressed at the endogenous level may therefore be unable to efficiently terminate all UAA codons once RF1 is removed in K-12 derivatives. To relieve the potential increased burden imposed on RF2, we removed the in frame UAG autoregulation element and mutated residue 246 back to Ala (**Figure 4.3.1a**). In addition, *E. coli* strain MDS42 with a reduced genome (Posfai 2006) was used as the parental strain, because the deletion of nearly 700 genes in this strain may further help to reduce the termination load imposed upon RF2. The gene encoding RF2, *prfB*, was first replaced by the “fixed” *prfB^f* gene coupled to a Cm^R gene cassette in MDS42 cells using the λ -red recombination system (Datsenko 2000) (**Figure 4.3.1b**). The Cm^R cassette was subsequently excised from the Cm resistant clones using the pACBSR plasmid system developed for markerless insertions (Tischer 2006). The resultant strain, JX2.0, has the *prfB* replaced by *prfB^f*. Knockout of *prfA* was then attempted in JX2.0 by electroporating the linear Cm^R knockin cassette with flanking sequence identical to those of *prfA* (**Figure 4.3.1c**). Genomic PCR screening of Cm resistant colonies showed that they contained the Cm^R cassette at the endogenous *prfA*

locus, indicating that knocking out of RF1 was achieved (**Figure 4.3.1c**). RF1 knockout was further confirmed with genomic sequencing (see the last section).

The initial RF1 knockout strain grew slowly, taking 24 hours to reach an OD₆₀₀ of 1.0. However, cells rapidly adapted within 3 passages and became a stable and sustainable strain, named JX3.0 (**Figure 4.3.1d**). In comparison to the parental JX2.0 strain, the final JX3.0 strain showed no significant change in growth under normal conditions and only a modest slowing of growth when an exogenous amber suppression tRNA-synthetase pair was introduced (**Figure 4.3.1e**, see sections below).

4.3.4.2: JX3.0 Enables the Incorporation of Tyrosine at Multiple Sites in EGFP Via the Amber Codon

Deletion of RF1 from *E. coli* presumably reassigns the UAG codon from the stop signal to a blank codon. Introduction of an orthogonal tRNA/synthetase pair which recognizes the UAG codon in JX3.0 would translate UAG with the amino acid for which the synthetase is specific, reassigning UAG to a sense codon. In the absence of RF1 competition, the incorporation efficiency at UAG should be significantly increased. This would also serve as a reporter allowing us to ascertain if RF1's function has been removed from JX3.0. A single All-in-One expression plasmid (pAIO) was used that contains an orthogonal amber suppressor tRNA, an orthogonal aminoacyl-tRNA synthetase, and an EGFP reporter with an N-terminal hexahistidine (His6) tag (**Figure 4.3.2a**) TAG mutations were introduced into tyrosyl sites in the EGFP gene to create 1, 2, 3, and 6 TAG-EGFP reporters (see methods). We first tested incorporation efficiency at UAG sites using the orthogonal $tRNA_{CUA}^{Tyr}$ /TyrRS derived from archaeobacterium *Methanococcus jannaschii* (Wang 2001), which inserts a tyrosine residue at UAG

codons. In parental JX2.0 cells, this system showed a reduction of EGFP protein yields with each additional UAG with no full-length EGFP in the 6TAG reporter as shown by Western blot (**Figure 4.3.2b**). In stark contrast, the RF1-deletion strain JX3.0 showed increased levels of protein and no reduction in EGFP protein yields across all TAG mutants. The overall protein yield for the 6TAG mutant was 6.8 mg/L, about half of the yield from wildtype EGFP expressed from the same plasmid system harboring the EGFP gene with no TAG mutations. In-cell fluorescence intensity was measured with fluorometry for each mutant in both the JX2.0 and JX3.0 strains. In JX2.0, the fluorescence intensity decreased upon the insertion of each additional TAG, while JX3.0 fluorescence was similar across all mutants and much higher than in JX2.0 (**Figure 4.3.2c and f**). These results demonstrate that the JX3.0 strain has much higher incorporation efficiency for tyrosine at UAG sites than the parental JX2.0 strain. The knock out of RF1 also allows multiple UAG sites to be efficiently suppressed with tyrosine, which is greatly diminished in the RF1-containing JX2.0 strain.

4.3.4.3: JX3.0 Enables the Incorporation of the Unnatural Amino Acid *p*-acetylphenylalanine at Multiple Sites in EGFP Via the Amber Codon

To determine if the RF1-deletion could allow an unnatural amino acid to be incorporated into a protein at multiple UAG sites, we performed the above experiments using the orthogonal $tRNA_{CUA}^{Tyr}$ /LW1RS pair (Wang 2003 p56), which is specific for the unnatural amino acid *p*-acetylphenylalanine (*pActF*). In JX2.0 cells, only the EGFP reporter containing a single TAG produced full-length EGFP protein as shown by Western blot. No full-length EGFP was detected in reporters containing 2, 3, or 6 TAGs (**Figure 4.3.2d**). In-cell fluorescence measurement also showed green fluorescence in

only the single TAG reporter (**Figure 4.3.2e and f**). The evolved LW1RS is known to be less active than the wt TyrRS in aminoacylation of their respective amino acids (Wang 2003 p56), which explains why small amounts of EGFP was detected in the 2 and 3 TAG reporters in JX2.0 when using the $tRNA_{CUA}^{Tyr}$ /TyrRS but not when using the $tRNA_{CUA}^{Tyr}$ /LW1RS. In JX3.0, however, large amounts of full-length EGFP was produced in the 1, 2, and 3 TAG reporters using the $tRNA_{CUA}^{Tyr}$ /LW1RS pair (**Figure 4.3.2d**). The protein yield did not decrease when the number of UAG codon increased from 1 to 3 (**Table 4.3.2**). In-cell fluorescence measurement showed a small decrease in the fluorescence intensity when *pActF* was incorporated at the second UAG site but no further decrease at the third UAG site. This is consistent with previous studies, which show that GFP fluorescence properties change with natural or unnatural amino acid mutations in a site-dependent manner (Tsien 1998; Wang 2003 p174). The 6TAG mutant also produced full-length protein in the JX3.0 strain (**Figure 4.3.2d**) with a lower yield than the 1, 2, or 3TAG mutants and exhibited no green fluorescence (**Figure 4.3.2e and f**). The introduction of 6 *pActF* into this protein may affect its folding and stability, reducing protein yields and abolishing fluorescence.

Yields for JX3.0 samples were determined to be 3.5 - 5.4 mg/L between the 1, 2, and 3 TAG mutants (**Table 4.3.2**). These yields represent about 25% of wildtype EGFP expressed without UAG codons, drastically higher than previously published reports (Huang 2010; Kaya 2009). The yield for 6TAG reporter with *pActF* incorporated was 0.5 mg/L. All yields were determined from proteins purified by Ni-NTA chromatography followed by FPLC using anion exchange. This is in contrast to some reports, which calculates yields right after the less stringent Ni-NTA purification (Huang

2010; Kaya 2009). When the His6 tag was appended at the N-terminus of the target protein, truncated protein products would be included in Ni-NTA purification and thus inflate the yield.

Mass spectrometry was used to confirm the incorporation of *pActF* at UAG sites. Fidelity for unnatural amino acid incorporation depends primarily on the substrate specificity of the orthogonal synthetase, and LW1RS has been established to be specific for *pActF* (Wang 2003 p56). Consistently, Electrospray ionization mass spectrometry (ESI-MS) of intact EGFP protein expressed by the 1-TAG reporter in JX3.0 showed two peaks (27801 and 27897 Da), which correspond to the fully mature *pActF*-containing EGFP missing the N-terminal methionine (theoretical mass 27799.2 Da) and the unfolded *pActF*-containing EGFP (theoretical mass 27899.2 Da), respectively. ESI-MS analysis of EGFP expressed by the 2- and 3-TAG reporters showed a single peak at 27898 and 27924 Da, respectively. These peaks lie within ± 2 Da of the theoretical masses of the 2 and 3 *pActF*-containing mature EGFP lacking an N-terminal methionine (27896.3 and 27922.3 Da, respectively). For all samples, no peaks corresponding to mutant EGFP with any natural amino acid at the UAG position were observed in the spectra. This corroborates our Western blot and in-cell fluorescence results showing that no significant amount of protein is expressed in the absence of *pActF* (**Figure 4.3.2d and e**). Liquid chromatography tandem mass spectrometry (LC-MS/MS) of chymotrypsin-digested protein samples was carried out to confirm the sequence of peptides containing UAG sites. LC-MS/MS allows characterization of mutant proteins with high sensitivity and dynamic range. The fragment ion masses were unambiguously assigned confirming the site-specific incorporation of *pActF* at the UAG site for all 1-, 2-, and 3-TAG EGFP

mutants (**Figure 4.3.3a**). Extracted ion chromatograms (EIC) clearly showed that the peptide with *pActF* incorporated at the UAG site was the dominant species, with only trace amount of Gln-containing peptide detectable (**Figure 4.3.3b and c**). Because both *pActF* and Gln have neutral side chains, we expect the two peptides containing *pActF* or Gln at the UAG site to be similar in ionization efficiency. This assumption has been proven valid when the target residue is not a basic amino acid (Chen 2007). Therefore, we used the peptide intensity calculated from peak area in EIC to determine the incorporation fidelity of *pActF* at all UAG sites (**Figure 4.3.3b and c**). Incorporation of *pActF* in JX3.0 was found >99.81% at all UAG sites (**Figure 4.3.3c**), consistent with the previously reported incorporation fidelity (>99.8%) of *pActF* at a single UAG site in the DH10 β strain (Wang 2003 p56). Overall, these results demonstrate that the JX3.0 strain efficiently and specifically incorporates the unnatural amino acid *pActF* into a protein at multiple sites encoded by UAG.

4.3.4.4: Ten UAG Sites can be Suppressed Simultaneously in JX3.0

To assess whether more UAG sites can be suppressed simultaneously for amino acid incorporation in JX3.0, we synthesized two 10-TAG containing EGFP reporters. One inserts 10 TAGs across various loop regions in EGFP (10-TAG) and the other has 10 TAGs inserted in tandem in one loop (10-TAGtd, **Figure 4.3.4a**). Loop sites were chosen to minimize potential mutational effects to protein folding and stability. Western blot results for these mutants showed that full-length EGFP was produced using the *tRNA^{Tyr}_{CUA}*/LW1RS to incorporate *pActF*, with a substantial increase in truncated products (**Figure 4.3.4b**). The expression of 10-TAGtd reporter was lower than the 10-TAG reporter, possibly because 10 consecutive *pActFs* would more negatively affect EGFP

folding and stability. To facilitate protein yield determination, the His6 tag was moved to the C-terminus of the 10 TAG reporters. Proteins were purified by Ni-NTA chromatography followed by FPLC. Incorporation of tyrosine using the *tRNA*_{CUA}^{Tyr}/TyrRS pair yielded full-length EGFP for both 10 TAG mutants, with a similar decrease in expression of the 10-TAGtd reporter (**Figure 4.3.4b**). For the 10-TAG mutant, protein yields were 0.4 mg/L for tyrosine incorporation and 0.5 mg/L for *pActF* incorporation. Fluorescence was abolished in all 10-TAG reporters regardless of identity of the amino acid incorporated, suggesting that fluorescence was affected by these mutations. The expression level of the 10-TAG reporter was reduced in comparison to the 1-, 2-, and 3-TAG reporters, presumably due to folding and stability issues caused by the large number of mutations. Nonetheless, the ability to produce any level of protein with 10 unnatural amino acids selectively incorporated is a novel property of the JX3.0 strain, and has never been accomplished before in any cells. If the protein folding and stability are not negatively affected by the amino acid incorporated, JX3.0 should allow suppression and unnatural amino acid incorporation at more than 10 TAG sites.

4.3.4.5: A Variety of Unnatural Amino Acids can be Incorporated at Multiple UAG Sites in JX3.0

To ascertain if incorporation at multiple TAG sites in JX3.0 was generally applicable to different unnatural amino acids, we assayed the expression of the 3-TAG EGFP reporter with a variety of other unnatural amino acids. EGFP was efficiently expressed with N ϵ -acetyl-L-lysine (ActK) (Neumann 2008), *p*-azido-L-phenylalanine (*pAzdF*), *p*-carboxymethyl-phenylalanine (*pCmF*) (Xie 2007), and *p*-iodo-phenylalanine (*pIodF*) (Xie 2004) incorporated, as shown by Western blotting (**Figure 4.3.5a**) and in-

cell fluorescence (**Figure 4.3.5b**). No full-length EGFP was detected by Western blot in the absence of the unnatural amino acid in the growth media. Protein yields varied with the unnatural amino acid, and were reduced in comparison to *pActF*. This reflects the relative activity of the evolved synthetases, suggesting that LW1RS specific for *pActF* was the most active among those tested. We obtained about 1 mg/L of purified protein containing ActK, *pAzdF*, and *pIodF* at 3 positions (**Table 4.3.2**). Overall, these results indicate that UAG codons in JX3.0 can be used to encode different natural or unnatural amino acids at multiple sites.

4.3.4.6: The Effect of RF1 Removal on Endogenous Genes Ending with a TAG

What would happen to over 300 endogenous genes that end with a TAG in *E. coli* genome upon removal of RF1? To address this question we divided these genes into two categories defined by their downstream context (**Figure 4.3.6a**). The majority of the TAG-ending genes have a secondary in-frame stop codon (TAA or TGA) downstream in the mRNA transcript before a transcriptional terminator, as represented by *sufA*. Upon RF1 deletion and introduction of an amber suppressor tRNA/synthetase pair, translation of these genes should be extended to the next stop codon in JX3.0. To facilitate sensitive detection, a scarless insertion of a FLAG-tag was appended to the N-terminus of the *sufA* gene in both the JX2.0 and 3.0 strains. As expected, SufA protein purified from JX3.0 harboring pAIO-TyrRS showed an increase in size on a Western blot corresponding to the expected molecular weight increase from extension to the next stop codon (**Figure 4.3.6b**). In contrast, there was no detectable extension of SufA protein in JX2.0 harboring pAIO-TyrRS, suggesting that suppression of these endogenous UAG codons is not efficient in the presence of RF1. Protein samples from JX3.0 harboring pAIO-TyrRS

were analyzed by LC-MS/MS (**Figure 4.3.6c**). Numerous peptide fragments were found that represent the extended peptide, confirming that translation was extended to the next stop codon and the UAG codon was suppressed with Tyr. The non-extended peptide fragment was also detected by mass spectrometry although it was not detectable on Western blot (**Figure 4.3.6c** in bold).

The second category of genes has a transcriptional terminator between the UAG and the next in frame UAA or UGA, as represented by *yfiA* (**Figure 4.3.6a**). Upon suppression of UAG, the ribosome will presumably stall at the 3' end of the mRNA as defined by the terminator hairpin. We appended a scarless N-terminal FLAG tag to the *yfiA* gene in the genomes of JX2.0 and JX3.0. A Western blot of YfiA expressed in the presence of pAIO-TyrRS showed a dramatic reduction of protein expression as well as the appearance of multiple bands in JX3.0 in comparison to JX2.0 (**Figure 4.3.6d**). The tmRNA trans-translation mechanism (Moore 2007) may explain this result. Suppression of the UAG in JX3.0 leads to the ribosome stalling at the end of the mRNA transcript. The stalled ribosome is recognized by tmRNA, which releases the ribosome and induces the degradation of the mRNA and extended polypeptide. In JX2.0, the presence of RF1 allows stoppage at the UAG codon to produce wildtype protein, and thus *yfiA* has much lower expression in JX3.0 than in JX2.0. To verify this, we analyzed YfiA purified from JX3.0 using LC-MS/MS to determine the identity of each band. Both Western blot and mass spectrometry did not find C-terminally extended peptides that indicate extension to the next in-frame UGA stop codon. This is presumably because the mRNA is processed back to the terminator hairpin, making the distal portion of the hairpin (red arrow in **Figure 4.3.6a**) the 3' end of the mRNA as previously reported (Abe 1999). No proteins

with the C-terminus extended to the end of mRNA were detected as well, indicating that these proteins were efficiently degraded via the tmRNA surveillance mechanism. The three bands resolved in lane 6 on the Western blot correspond to extensions of 0, 2, or 6 amino acids from the UAG site, respectively, with 0 and 2 amino acid extension being the dominant species in the Western blot (**Figure 4.3.6d**, 40, 44, and 16% of total intensity for 0, 2, and 6 amino acid extension, respectively). The release of the ribosome from the mRNA transcript early due to the presence of terminator hairpin structure and ribosome stalling at the mRNA end may yield this small amount of proteins containing these peptides, which are not tagged by tmRNA for degradation. Our results are consistent with and corroborate a previous report on tmRNA-induced degradation in non-stop mRNA (Ueda 2002), with the exception that we studied genes endogenously expressed in the genome whereas the other study expressed the target gene exogenously from a plasmid. Lastly, protein expression of SufA and YfiA in the absence of pAIO-TyrRS was similar in both strains, most likely due to the weak termination ability of mutant RF2 for the UAG codon (see section below).

4.3.4.7: Genomic Sequencing of JX3.0 Reveals a Novel Mutation in Release Factor 2

To further characterize JX3.0 and to understand how it can still terminate translation at endogenous amber stop codons in the absence of a suppressor tRNA/synthetase, full genomic sequencing was performed on JX2.0 and JX3.0 and compared to that of *E. coli* K-12 MG1655. Both JX2.0 and JX3.0 showed gene deletions identical to the parental MDS42 strain, a multiple-deletion derivative of MG1655 (Posfai 2006). The knockout of *prfA* in JX3.0 was confirmed, and this is the only deletion difference between JX2.0 and JX3.0. In addition, two single nucleotide polymorphisms

were found between the two strains (**Figure 4.3.7a**). One is a silent mutation in the coding region of the gene *ypdE* and the other results in an amino acid change (A293E) in RF2. Besides these, no other mutations were found in JX2.0 and JX3.0.

Although mutations have been identified in RF2 that allow it to recognize all three stop-codons (Ito 1998), the A293E has not been discovered in any previous complementation screens for RF1 deficiency (Ito 1998; Zhang 1994; Dahlgren 2000; Kaczanowska 2004). To characterize this novel mutation, we first determined if it was necessary for the survival of JX3.0. We attempted to repair this mutation in JX3.0 using a knockin cassette designed to revert Glu293 (GAG) of RF2 to alanine (GCT), but with a codon different from the one used in DH10 β (GCG) to facilitate screening and strain maintenance. This cassette was electroporated into JX3.0 harboring the pKD46 plasmid as well as DH10 β as a control. A total of 60% of the sequenced DH10 β clones contained the knocked-in GCT mutation, however no GCT mutants were identified in any of the sequenced JX3.0 transformants (**Figure 4.3.7b**). To control for any differences in knockin efficiency between JX3.0 and DH10 β , another knockin cassette was created that maintained the A293E mutation but with a silent single nucleotide change (GAG to GAA). The efficiency of this knockin was similar in both strains (**Figure 4.3.7b**), suggesting that knockin efficiency is not hindering the reversion of A293E in JX3.0. The knockin efficiency of these two cassettes (GCT vs. GAA) was significantly different ($P = 0.0004$) in JX3.0. We therefore conclude that JX3.0 requires the A293E mutation in RF2 for survival.

We next determined if the A293E mutation in RF2 is sufficient for rescuing RF1 function in *E. coli*. We replaced the endogenous RF2 gene of a tsRF1 strain (MRA8)

(Rydén 1984) with the RF2(A293E) gene from JX3.0 using established procedures (Datsenko 2000). If this mutation was sufficient to confer survival without RF1, it should complement the tsRF1 deficiency and rescue growth of this strain at 43°C. Surprisingly, no difference in growth phenotype was observed between the parental (MRA8) and mutant (MRA8 A293E) strains (**Figure 4.3.7c**), indicating that the A293E mutation is not able to rescue RF1 temperature sensitive phenotype in *E. coli*.

4.3.4.8: JX3.0 Enables the Efficient Expression of Human Histone H3 Containing Multiple actKs at Canonical Lysine Positions

To demonstrate the usefulness of this strain in the synthesis of proteins for structural, functional, and biological studies, we attempted to synthesize human histone H3 containing multiple acetylated lysine residues. Mimicking histone H3 modifications have thus far been limited to a single insertion of a methylated (Nguyen 2009) or acetylated (Neumann 2009) lysine residue, which does little to represent the wealth of modifications that can be made to the protein. Using an orthogonal tRNA/synthetase pair for the acetyl-lysine (ActK) (Neumann 2008), we expressed histone H3 with 1, 2, 3, and 4 ActKs inserted at known acetylation sites (**Figure 4.3.8a**, see methods). Expression of all mutants is shown in western blot for the full-length protein (**Figure 4.3.8b**). Yields were sufficient: 2.2, 0.8, 0.4, and 0.3 mg/L for the 1, 2, 3, and 4TAG H3 mutants respectively (**Figure 4.3.8c**). To ensure that ActK was introduced in these mutants, intact 1TAG Histone H3 protein was subjected to mass spectrometric analysis (**Figure 4.3.9a and b**). The large peak identified at 16558.01 Da corresponds to the expected monoisotopic mass of histone H3 harboring an ActK, a truncated N-terminal methionine, and a disulfide bond (expected mass of 16558.03) (**Figure 4.3.9a and b**). The oxidized version of this

peak is found at 16573.05 as well (expected 16574.02). Only one of the remaining small peaks corresponds to a natural amino acid at the ActK position, which is a lysine insertion (observed mass: 16514.95, expected: 16516.01) that most likely represent deacytlation of ActK (Neumann 2008). The drop off in yields with each subsequent TAG was not an ActK related phenomenon, as similar results were seen with *pActF* suppression of these proteins (**Figure 4.3.8b**). This may be due to the relative proximity of these sites, as all four actK positions lie within a 15 amino acid stretch in the N-terminus of the protein. This may invoke a similar rare-codon type of effect as seen with the 10TAG EGFP reporter. Regardless, 0.3 mg/L of the 4-actK containing protein is a sufficient yield and represents the first means of preparing such a protein synthetically.

4.3.5: Discussion

We describe here the generation of an RF1 knockout *E. coli* strain, JX3.0. When a tRNA/synthetase pair decoding the UAG codon is introduced into JX3.0, UAG codons are reassigned from a stop signal to a sense codon. RF1 is reported to be essential for *E. coli*. We think the success of the RF1 knockout in this study is a combined contribution of the reduced genome size, non-autoregulated RF2 expression, and the novel A293E mutation in RF2. A small genome reduces the termination load on RF2, and removing the negative autoregulation mechanism of RF2 would increase its expression level. Both factors help RF2 to cope with the increased termination load upon RF1 deletion. The role of the A293E mutation is intriguing, as it is necessary for JX3.0 survival, but not sufficient to rescue the RF1 temperature sensitive phenotype in MRA8 cells. The latter result suggests that RF2(A293E) is unable to substitute RF1 in the termination of UAG codons, as least in an *E. coli* strain with a normal size genome. These data also explain

why this mutation was never discovered in previous complementation assays performed in strains with normal size genomes (Ito 1998; Zhang 1994; Dahlgren 2000; Kaczanowska 2004). Based on the crystal structure of RF2 (Klaholz 2003; Vestergaard 2001), residue 293 is not in close proximity to either the conserved GGQ motif responsible for peptide termination (Zavialov 2002) or to the anticodon-like loop responsible for codon recognition (Wilson 2000), suggesting that the A293E mutation contributes to RF2 codon specificity, if any, in a subtle and indirect way. The exact contribution of A293E warrants further studies. Regardless, a delicate balance has been achieved in JX3.0, such that the termination of endogenous UAGs is sufficient for strain survival but outcompeted by the introduction of a tRNA/synthetase pair for amino acid incorporation.

Reassignment of a stop to a sense codon has been echoed in nature. Stop codons are common targets for alterations in organisms containing a nonstandard code (Osawa 1992; Knight 2001 p49). In mitochondria, the UGA codon has been reassigned from stop to Trp in many taxa, while UAG has been reassigned to Leu or Ala in plants. In the nucleus, the UAA and UAG codons have been reassigned to Gln in some diplomonads, ciliates and alga, and the UGA codon has been reassigned to Cys or Trp in some ciliates and mycoplasma. In addition, selenocysteine and pyrrolysine (Pyl), considered the 21st and 22nd amino acids respectively, are cotranslationally inserted at in-frame stop codons in methanogenic archaea and the bacterium *Desulfitobacterium hafniense* (Bock 1991; Srinivasan 2002). The incorporation of Pyl at the UAG codon is of particular interest. Similar to other 20 common amino acids and in contrast to selenocysteine, Pyl uses a specific tRNA/synthetase pair, the $tRNA_{CUA}^{Tyr}$ /PylRS pair for incorporation. However,

PylRS is able to aminoacylate tRNA^{Pyl} with Pyl even when the tRNA anticodon is mutated to CAA, AUA and CUU, and the *tRNA*_{CUA}^{Tyr}/PylRS pair can incorporate Pyl at UGA sites into proteins (Ambrogelly 2007; Yanagisawa 2008). This suggests that the tRNA^{Pyl}/PylRS had the flexibility of using many codons when this system emerged. The fact that UAG was used suggests that a stop signal is preferred for code expansion. Nonetheless, bioinformatic analyses of all UAG-containing genes in Pyl-utilizing organisms are inconclusive on whether the UAG is completely reassigned to Pyl or still in competition with the release factor (Zhang 2005).

Although natural variations in stop codon assignments are valuable in understanding the outcome of the reassignment, only limited information can be gained on the reassignment process and organismal adaptation. Our study demonstrates the feasibility of reassigning the UAG stop codon to a sense codon in the extant organism *E. coli*. This result provides laboratory evidence supporting such codon reassignment events during code evolution. The resultant JX3.0 strain affords a model system for studying the evolution of the genetic code experimentally. Genomic and proteomic profiling of the JX2.0 and JX3.0 strain will demonstrate the physiological changes in *E. coli* after RF1 removal making UAG a blank codon. When the orthogonal tRNA/synthetase pair is introduced into JX3.0, similar profiling will reveal how cells adapt to the resulting codon reassignment. Since we can reassign the UAG codon to various unnatural amino acids in JX3.0, for which natural systems cannot, it is of interest to determine whether encoding UAG with an unnatural amino acid affords any evolutionary advantage to the host and whether the UAG codon will be fixed as a sense codon. In addition, JX3.0 will also

enable the determination of whether the UAG codon, like other sense codons, will eventually spread throughout the genome upon further experimental evolution.

An immediate utility of the JX3.0 strain is in the selective incorporation of unnatural amino acids at multiple sites into recombinant proteins to enable new studies. JX3.0 is a stable strain without major growth or other deleterious defects. It was generated 3 years ago and no changes have been observed after many generations in culture. Incorporation of Tyr into EGFP at multiple UAG positions using the orthogonal $tRNA_{CUA}^{Tyr}$ /TyrRS pair, which is closest to an endogenous tRNA/synthetase pair in terms of activity among all orthogonal pairs tested, shows no decrease of EGFP yields from 1 to 6 UAG-containing EGFP reporter, a trait for UAG being a sense codon. This result indicates that the competition from release factors for the UAG codon is essentially removed in JX3.0. Protein yields in JX3.0 are striking with efficient incorporation of amino acids at even 10 UAG sites. More than 10 amber codons can be suppressed in JX3.0, if the protein folding and stability are not affected by the amino acid incorporated. Other systems that attempt to incorporate multiple unnatural amino acids into the same protein suffer dramatic reductions of yields when more than one UAG codon is introduced (Huang 2010; Kaya 2009). In addition to incorporating Tyr and *pActF*, for which an orthogonal tRNA/synthetase pair with high activity was generated, the JX3.0 strain also enables the multi-site incorporation of unnatural amino acids for which the evolved orthogonal tRNA/synthetase has relatively weak activity, such as ActK, *pAzdF*, and *pIodF*. For these unnatural amino acids, about 1 mg/L of protein containing each at 3 UAG sites was purified, which is sufficient for the requirements of most studies. The ability to selectively incorporate unnatural amino acid at multiple sites will prove useful

in many aspects of protein related research: multi-site incorporation of heavy-atom containing unnatural amino acids (e.g., *pIodF*) will enable phase determination of proteins with large molecular weight; chemical handles (e.g., *pActF*) will enable multi-site PEGylation or glycosylation to enhance protein in vivo stability and half-life; multi-site incorporation of fluorescent unnatural amino acid will facilitate single molecule imaging; and multi-site incorporation of posttranslational modification mimics (e.g., *ActK* and *pCmF*) will facilitate the study of epigenetics and signal transduction pathways. We have demonstrated the successful incorporation of these unnatural amino acids into proteins at multiple sites in JX3.0, including the multiple incorporation of *ActK* in human histone H3 (**Figure 4.3.5a, 4.3.8, 4.3.9, Table 4.3.2**)

The JX3.0 strain would also enable the laboratory evolution of bacteria in search for new protein property or organismal functions by exploiting the novel properties of unnatural amino acids. Such experiments were not effective or feasible before. Incorporation of an unnatural amino acid at one site in low efficiency is not likely to generate novel functions; novel functions may require an unnatural amino acid at multiple positions simultaneously. The presence of RF1 makes truncated protein products, which may interfere with the selection and evolution. The presence of RF1 also makes the meaning of UAG ambiguous, which hinders the cell to quickly fix the new assignment. All these problems do not exist in the JX3.0 strain. It has a “blank” UAG codon that can be reassigned to any amino acid for which an orthogonal tRNA/synthetase pair is introduced. This would greatly accelerate the directed evolution process in laboratory settings.

In summary, RF1 can be knocked out in *E. coli* after the RF2 gene is fixed. The resultant JX3.0 strain reassigns the amber UAG codon to a sense codon when an orthogonal tRNA/synthetase pair is introduced. Our results suggest that the genetic code of the extant organism *E. coli* is tolerant of codon reassignment and more flexible than expected in adapting to a new code. This work represents a fundamental re-structuring of the genetic code of *E. coli* with the first full reassignment of a stop codon in the laboratory, mimicking an evolutionary process that might have occurred in numerous taxa. The JX3.0 strain affords a previously unavailable model system for experimentally monitoring the physiological change and adaptation of a living organism to codon reassignment. This would give a glimpse into the past and reveal how organisms survived these genetic events through Darwinian selection, and shed light on the natural progression of the genetic code. Finally, the JX3.0 strain enables the efficient genetic incorporation of various unnatural amino acids into protein at multiple sites, and thus should be of great use in the study and evolution of proteins and protein-involved biological processes.

Table 4.3.1: Filtering criteria for database search results.

	1+ peptide	2+ peptide	3+ peptide
Cutoff Score	>9	>9	>12

Table 4.3.2: Protein yields of the various EGFP-TAG constructs expressed in JX2.0 and JX3.0.

Protein species	JX2.0 (mg/L)	JX3.0 (mg/L)
wildtype EGFP (no TAG)		14.9
6 - Tyr	N/D	6.8
10 - Tyr	N/D	0.4
1 - <i>pActF</i>	1.8	3.5
2 - <i>pActF</i>	N/D	3.5
3 - <i>pActF</i>	N/D	5.4
6 - <i>pActF</i>	N/D	0.5
10 - <i>pActF</i>	N/D	0.5
3 - <i>pAzdF</i>	N/D	0.8
3 - ActK	N/D	1.0
3 - <i>pCmF</i>	N/D	0.2
3 - <i>pIodF</i>	N/D	0.9

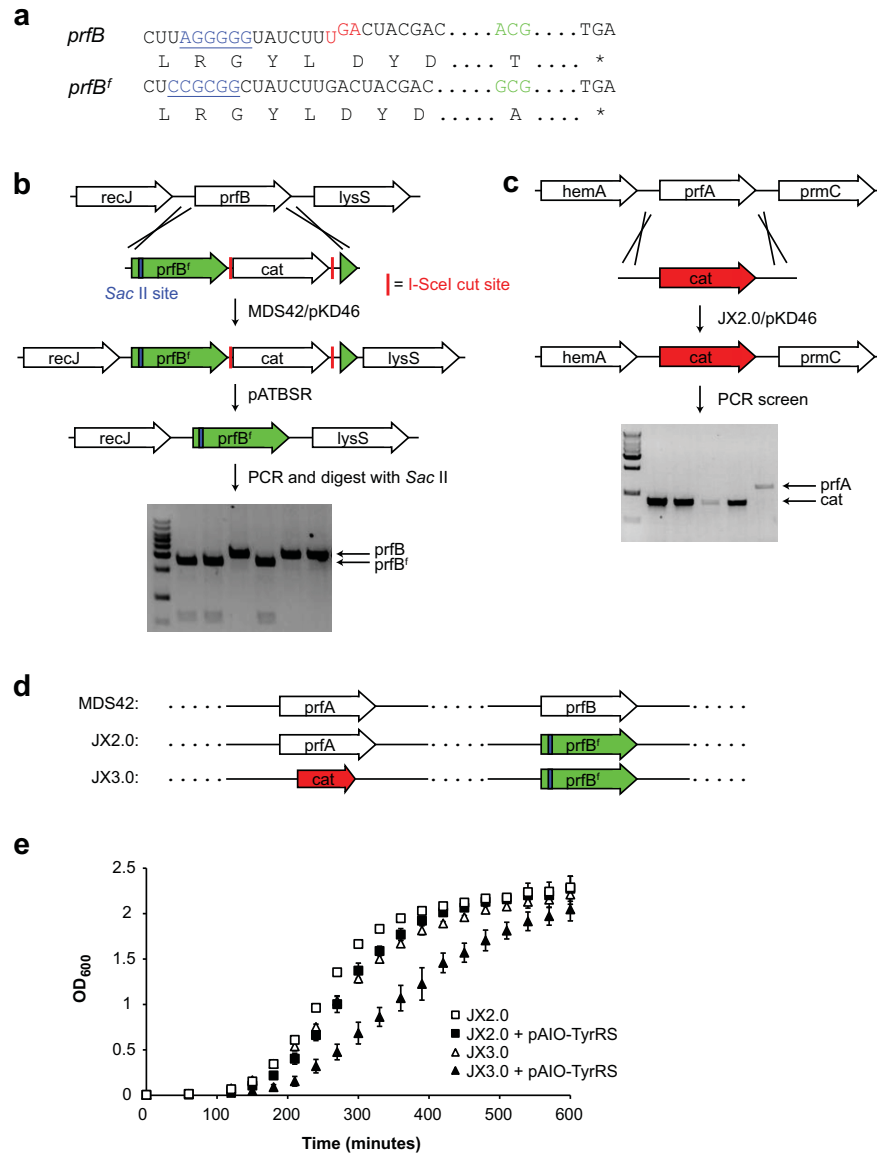


Figure 4.3.1: RF1 can be knocked out in a genome-reduced *E. coli* strain after RF2 is fixed. **(a)** Features of the *prfB* gene encoding the RF2 in K-12 *E. coli* strains: an in-frame UGA stop codon (red) for autoregulation of RF2 expression and the Ala246Thr mutation (green) impairing RF2's recognition of the TAA codon. In *prfB'*, we removed the UGA regulation and reverted residue 246 to Ala. A Shine-Dalgarno like sequence (blue) in *prfB* was silently mutated to a *Sac* II site (blue) in *prfB'* to facilitate the screening of *prfB'* knockin. **(b)** Generation of the JX2.0 strain. **(c)** Generation of the JX3.0 strain. **(d)** Illustration of main features of the three strains, MDS42, JX2.0 and JX3.0. **(e)** Growth rates of the JX2.0 and JX3.0 strains with or without the tyrosyl suppression system (pAIO-TyrRS, see **Figure 4.3.2**). Growth rates of JX2.0 (□) and JX3.0 (Δ) were assayed in the absence (filled) or presence (open) of the pAIO-TyrRS plasmid. Shown is the average from three independent measurements with error bars representing s.e.m..

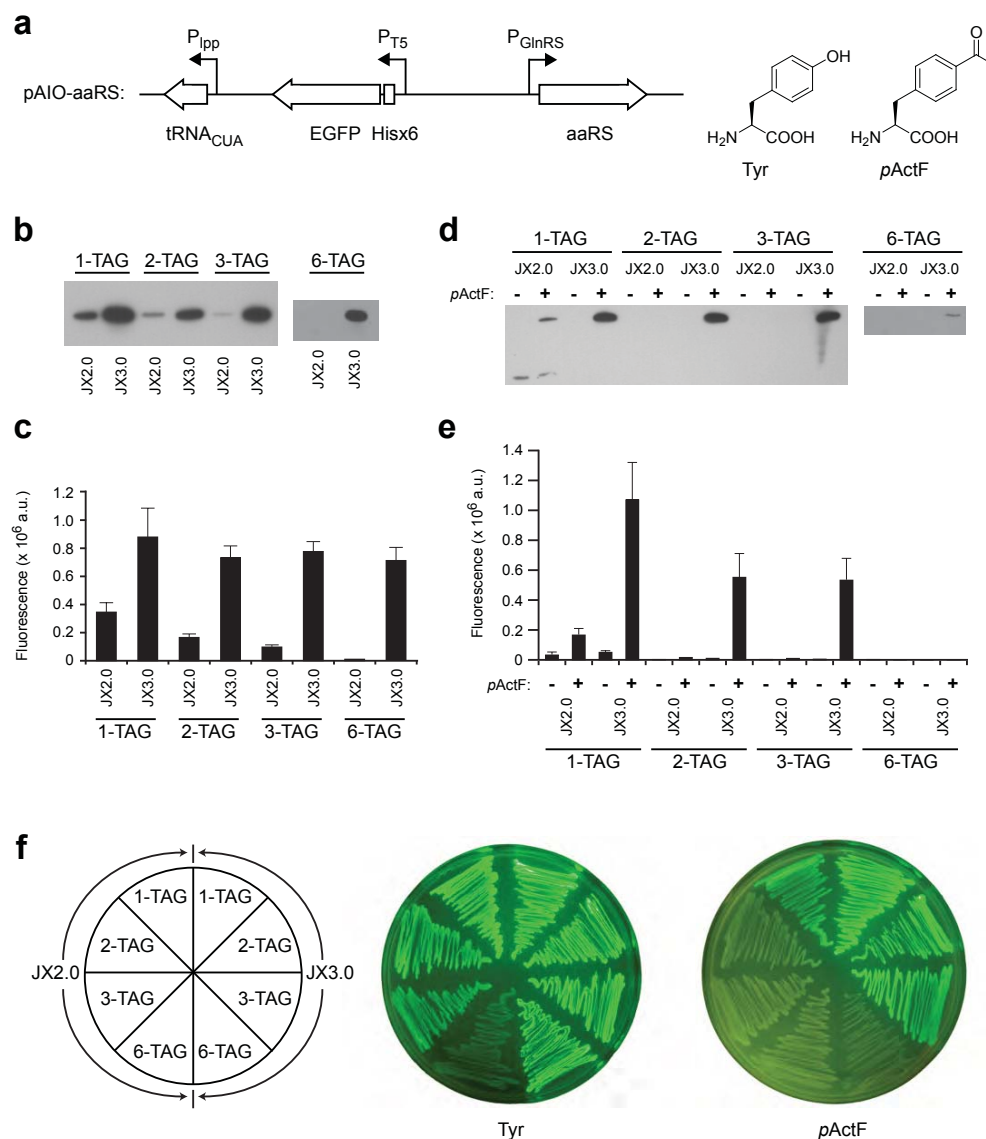


Figure 4.3.2: Removal of RF1 enables encoding natural or unnatural amino acids at multiple UAG sites in JX3.0. **(a)** The All-in-One plasmid. Structures of Tyr and *pActF* are shown to the right. **(b)** Western analysis and **(c)** in-cell fluorescence of EGFP expression when UAG codons are decoded as Tyr by pAIO-TyrRS. Measurement was performed on 3 independent batches of cells and error bars represent s.e.m.. **(d)** Western analysis and **(e)** in-cell fluorescence of EGFP expression when UAG codons are decoded as unnatural amino acid *pActF* by pAIO-LW1RS. Conditions are the same in **(b)**. For each sample, a duplicate of cultures were grown in the presence or absence of *pActF* in the growth media. Measurements were performed as in **(c)** with duplicate samples grown in the presence or absence of *pActF* in the growth media. $n = 3$, error bars represent s.e.m. **(f)** Fluorescence images of cells when UAG was decoded as Tyr (center) or *pActF* (right).

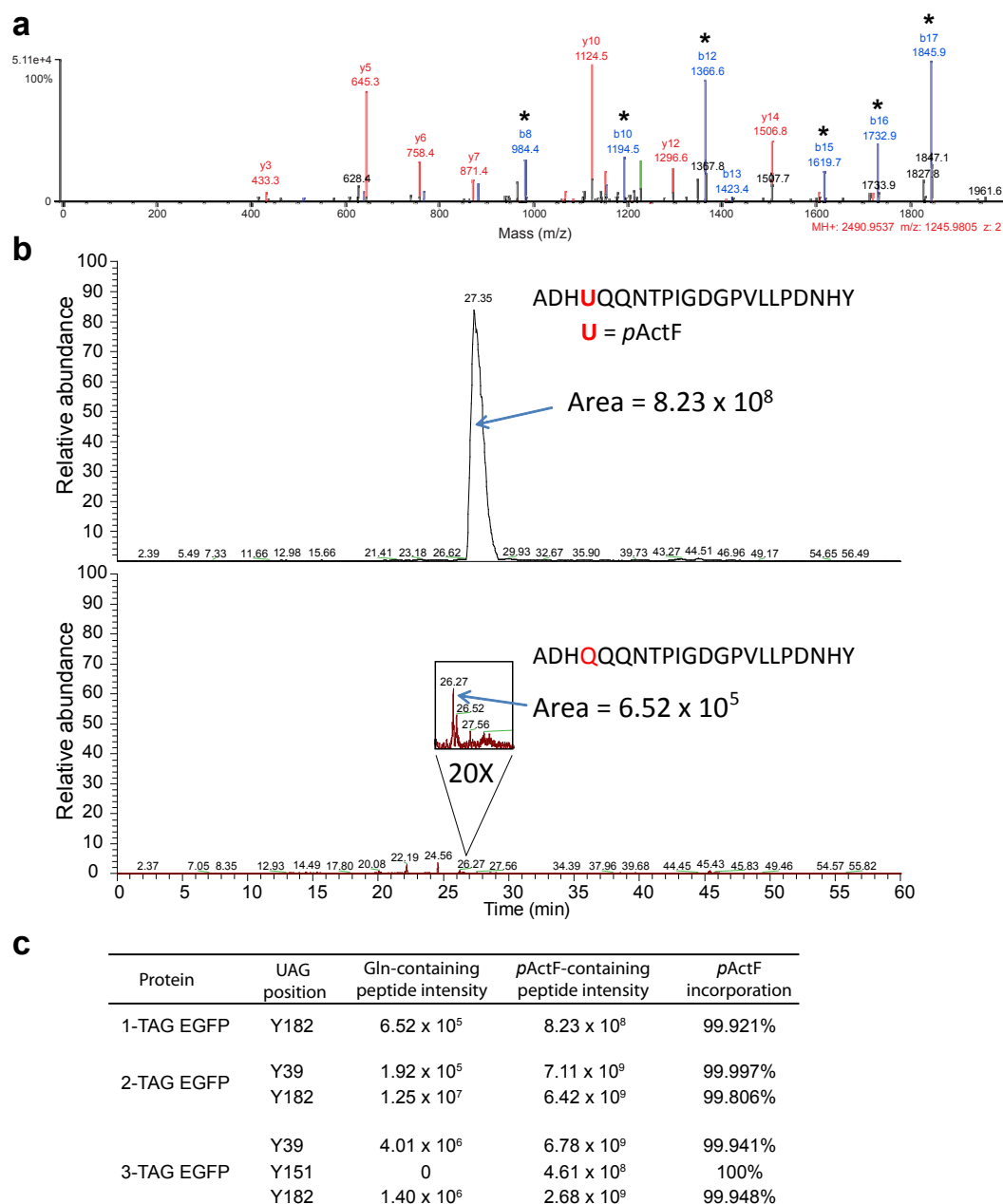


Figure 4.3.3: Mass spectrometric analyses of EGFP expressed in JX3.0 show that *pActF* was selectively incorporated at multiple UAG sites with high fidelity. **(a)** MS/MS spectrum of EGFP peptide ADHUQQNTPIGDGPVLLPDNHY. U represents the UAG codon at residue 182. Star (*) in the spectrum indicates peptide fragments containing *pActF*, which unambiguously indicate that *pActF* was incorporated at the UAG site. **(b)** Extracted ion chromatograms (EIC) of the above peptide containing *pActF* (top) or Gln (bottom) at the UAG 182 position. The peak areas are indicated. **(c)** Summary of all UAG site-containing peptides from the 1-, 2-, and 3-TAG EGFP mutants. Peptide intensities were determined by the area peak in the EIC and translated into the incorporation fidelity for *pActF*.

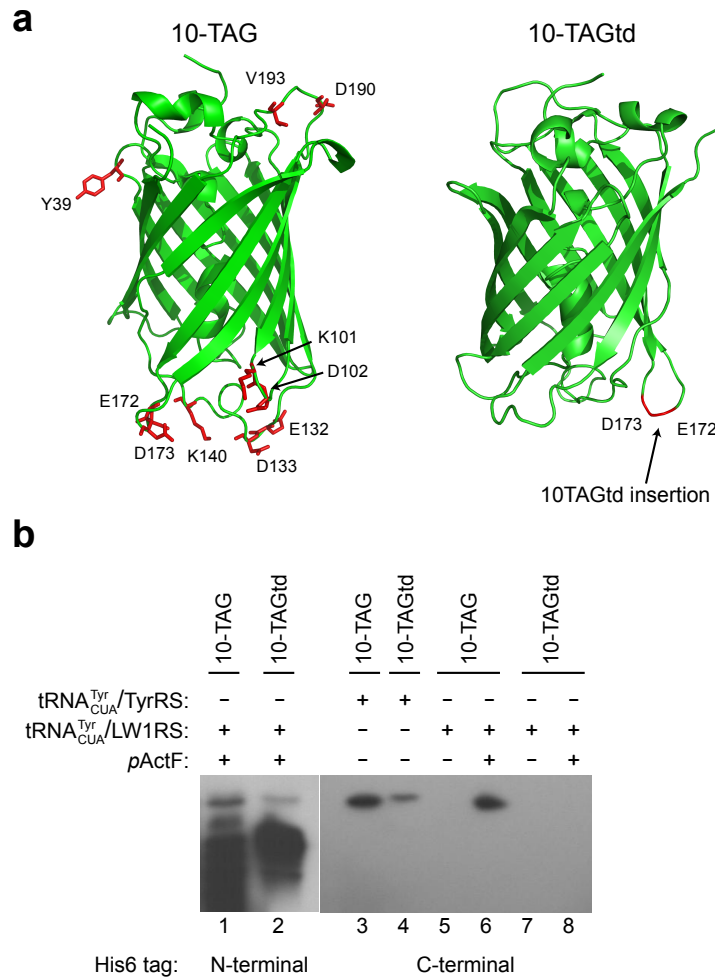


Figure 4.3.4: Ten UAG sites are simultaneously suppressed with natural or unnatural amino acids in JX3.0. **(a)** GFP structure (PDB 1GFL) illustrating the UAG sites introduced. In the 10-TAG reporter, codons for 10 residues highlighted in red were mutated to TAG. In the 10-TAGtd reporter, 10 consecutive UAG codons were inserted to replace E172 and D173. **(b)** Western blot analysis of the expression of 10-TAG and 10-TAGtd EGFP in JX3.0. The UAG codons were decoded as Tyr by $tRNA_{CUA}^{Tyr}/TyrRS$ or as $pActF$ by $tRNA_{CUA}^{Tyr}/LW1RS$. Cell lysates from same amount of cells were separated by SDS-PAGE and probe with a penta-His antibody.

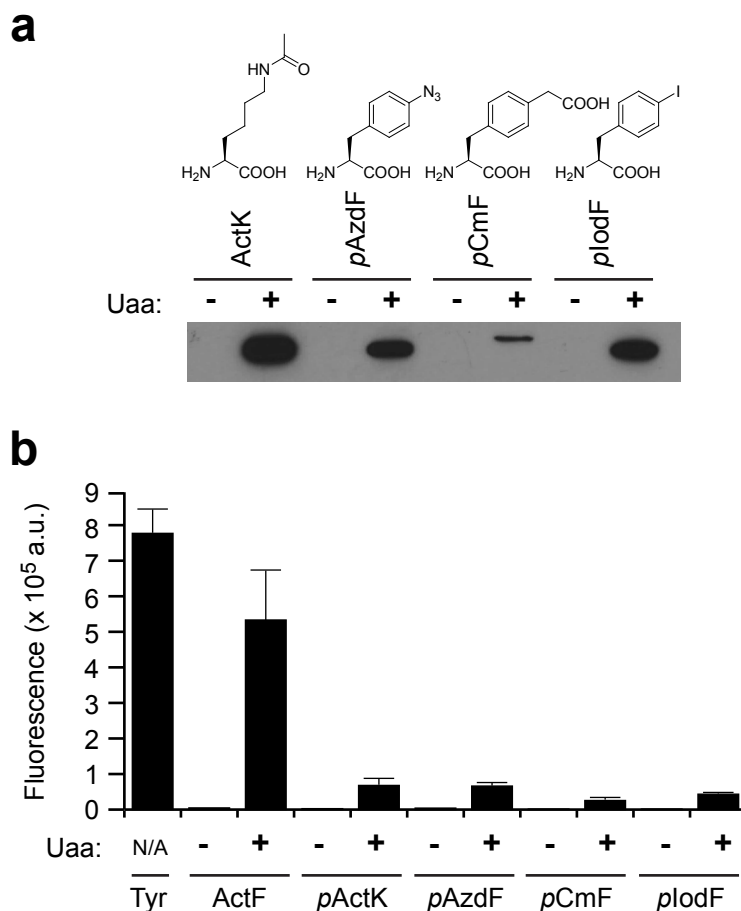


Figure 4.3.5: Various unnatural amino acids are incorporated at multiple UAG sites into EGFP in JX3.0. **(a)** Western analysis of the expression of 3TAG EGFP reporter in JX3.0 with the UAG codon decoded as different unnatural amino acids (Uaa) by different orthogonal tRNA/synthetase pairs. Densitometric analysis of Western blot bands and purified protein yields (**Table 4.3.2**) were consistent on incorporation efficiency. **(b)** In-cell fluorescence assay of these mutant EGFP proteins containing different unnatural amino acids at 3 UAG sites. Measurements were performed using 3 independent batches of cells. Error bars represent s.e.m..

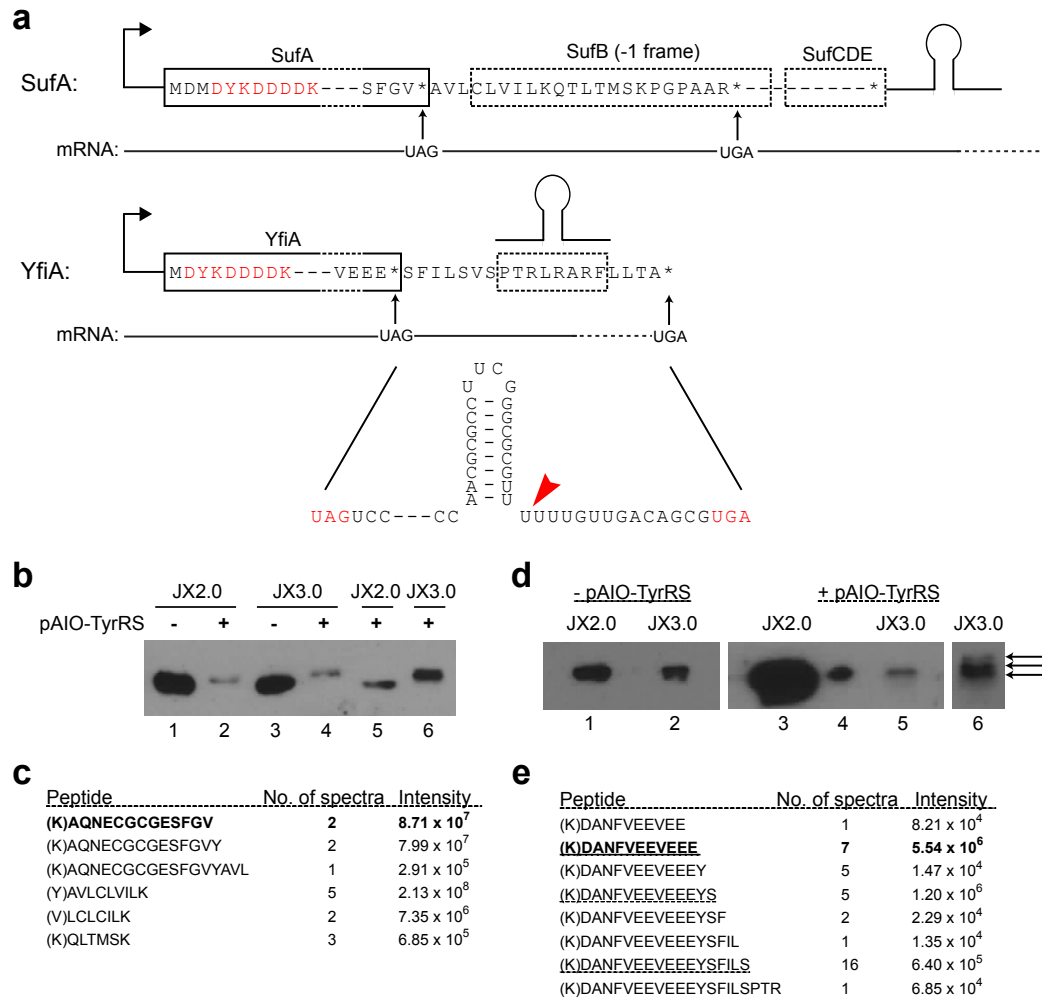


Figure 4.3.6: Endogenous genes ending in TAG extend their peptides to different positions in an mRNA context-dependent manner. **(a)** Genomic context of *sufA* and *yfiA*. **(b)** Western analysis of SufA purified from cells using the N-terminal FLAG tag. Extension of SufA to its next stop codon occurred only in the JX3.0 strain harboring the suppressor $tRNA_{CUA}^{Tyr}/TyrRS$ (lanes 4 and 6). All other strains showed a band that corresponds to the wildtype SufA molecular weight. **(c)** Tandem MS analysis of SufA protein purified from JX3.0 harboring $tRNA_{CUA}^{Tyr}/TyrRS$. A small amount of non-extended wildtype SufA C-terminal peptide (bolded) was also detected. **(d)** Western analysis of YfiA purified from cells using the N-terminal FLAG tag. Lane 6 is a re-run of lane 5 sample in a higher percentage gel (20% vs. 15%). YfiA in JX3.0 harboring $tRNA_{CUA}^{Tyr}/TyrRS$ (lane 5 and 6) showed 3 predominant bands, while other samples showed only one band. **(e)** Tandem MS analysis of YfiA purified from JX3.0 harboring $tRNA_{CUA}^{Tyr}/TyrRS$ showed that no peptide was extended to the next UGA codon. Various extensions before the terminator hairpin structure were identified. Peptides extended to the 0, 2nd and 6th amino acid after UAG were predominant and underlined.

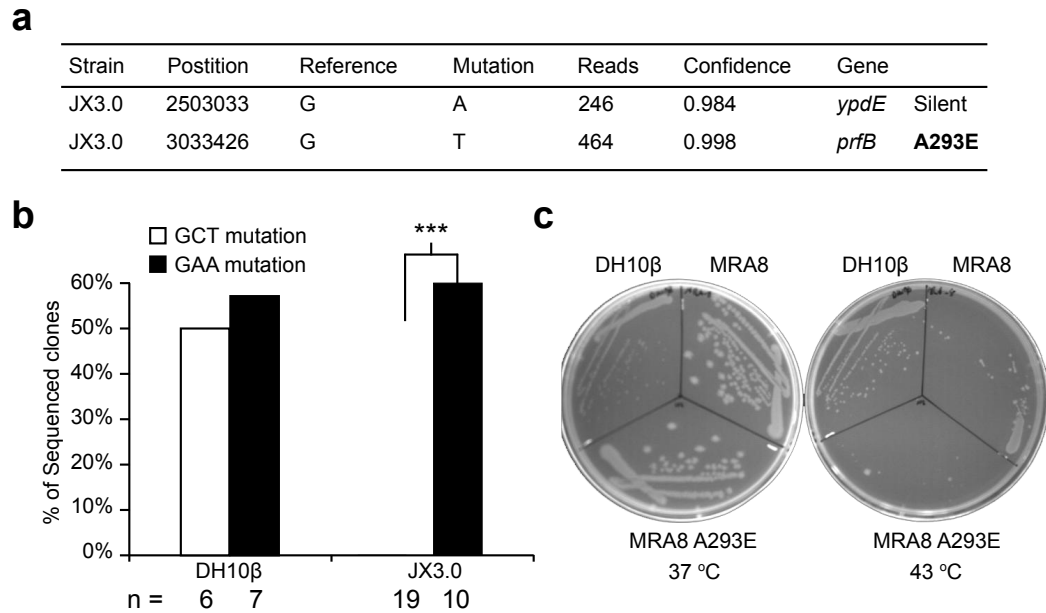


Figure 4.3.7: The A293E mutation in RF2 is necessary for survival of the JX3.0 strain but not sufficient for rescue of a tsRF1 strain. **(a)** Full-genome sequencing results for JX2.0 and JX3.0. Deletion of *prfA* in JX3.0 was confirmed and not listed here. **(b)** The A293E mutation cannot be reverted in the JX3.0 strain. Although GCT knockin in DH10β was efficient, no positive clones were obtained for JX3.0. In contrast, the control knockin GAA was efficiently introduced into a similar percentage of clones for both strains. The difference in efficiency between the two cassettes is significant ($P = 0.0004$) from a Fisher's exact test. **(c)** The A293E mutation cannot rescue the RF1 temperature sensitive phenotype of the MRA8 strain. The introduction of *prfB^f*(A293E) did not rescue the growth defect of MRA8 at 43 °C. The control strain DH10β had no such defect.

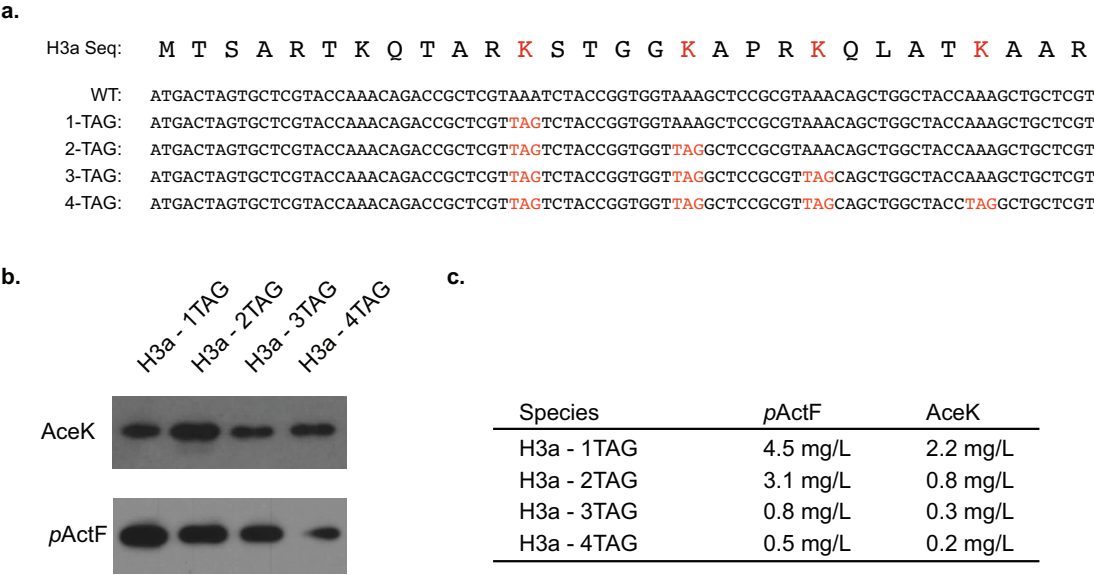


Figure 4.3.8: Human Histone H3 harboring multiple ActKs is efficiently expressed in JX3.0. (a) Amino acid and nucleotide sequence of the four mutants used in this study. (b) Western blot of histone H3 protein purified from JX3.0 harboring the four mutants coexpressed with either the ActK or *pActF* suppression systems. Protein was normalized in each sample. (c) Protein yields for the various histone H3 mutants coexpressed with either the ActK or *pActF* suppression systems.

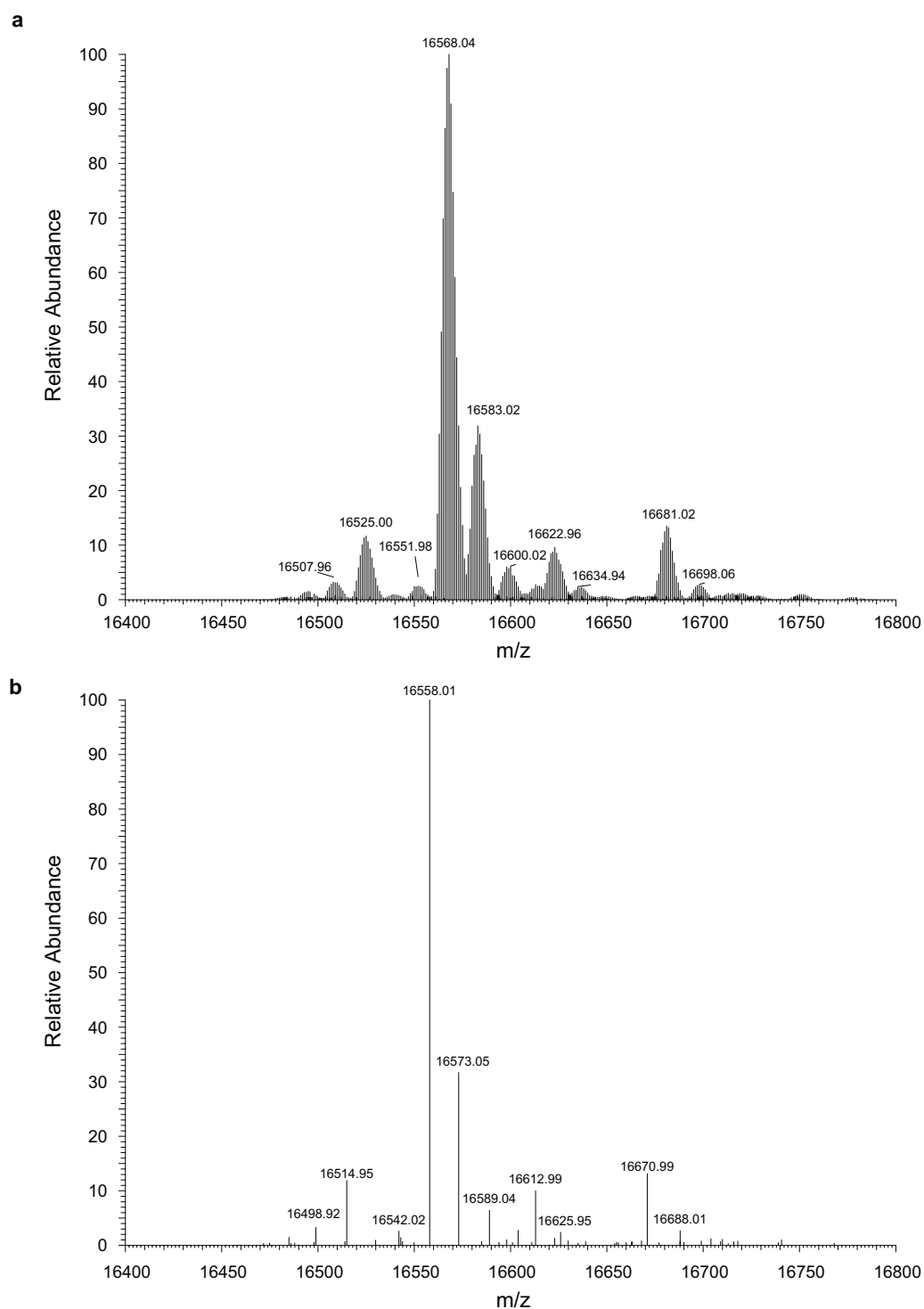


Figure 4.3.9: Intact protein mass spectrometry of 1-TAG histone H3 coexpressed with the ActK suppression system. **(a)** Spectra obtained from the histone H3-1TAG protein. Predominant peaks are labeled and discussed in text. **(b)** Monoisotopic results obtained from **(a)**. Predominant peaks are labeled and discussed in text.

Chapter 4.1 is a reprint in full of the material as it appears in the Wiley Encyclopedia of Chemical Biology, 2007. Johnson DBF, Takimoto JK, Xu J and Wang L (2007). Unnatural acids to investigate biological processes. Wiley Encyc Chem Bio. doi:10.1002/97804700486.72.webc585. The dissertation author, J.K. Takimoto, and J. Xu contributed equally to this manuscript. The dissertation author specifically contributed the introduction, methods, and future directions sections in this review.

Chapter 4.2 is a part of a manuscript currently being prepared for submission. Johnson DBF, Xu J, Schultz MD, Schmitz RJ, Ecker JR and Wang L Release Factor One is Non-Essential in *Escherichia coli*. The dissertation author is the principal author of the manuscript and performed all experiments except those described below. Jianfeng Xu was the creator of the JX1.0, JX2.0, and JX3.0 strains and designed the RF1 knockout cassette. Full genome sequencing was performed by Matthew Schultz and Robert Schmitz in Joseph Ecker's laboratory at the Salk Institute. All mass spectrometry experiments were performed by Jessica Read and Wolfgang Fischer in Vincent J Coates Foundation mass spectrometry center at the Salk Institute.

Chapter 4.3 is a reprint in full of a manuscript recently submitted for publication in Nature Chemical Biology. Johnson DBF, Xu J, Shen Z, Schultz MD, Schmitz RJ, Ecker JR, Briggs SP and Wang L. Reassigning the Amber Stop Codon from Nonsense to Sense in *Escherichia coli*. This paper was truly a collaborative effort. The dissertation author and Jianfeng Xu contributed equally to this work. Jianfeng Xu was responsible for the creation of the JX2.0 and JX3.0 strains, as well as the initial efforts to knock out RF1 from DH10 β . The dissertation author performed all other analyses and experiments excluding the mass spectrometry and full genome sequence. All EGFP, YfiA, and SufA

mass spectrometry was performed by Zhouxin Shen in the laboratory of Steven Briggs at UC-San Diego, and the histone H3 mass spectrometry was performed by Jessica Read and Wolfgang Fischer in Vincent J Coates Foundation mass spectrometry center at the Salk Institute. Full genome sequencing was performed by Matthew Schultz and Robert Schmitz in Joseph Ecker's laboratory at the Salk Institute. The dissertation author was the writer of the manuscript, excluding specific materials sections.

Chapter 5: Conclusions, Summary, and Perspectives

The work described here encompasses two distinct avenues of research, not only in their goals, but also in their methodology. They share a common thread, addressing the complex problem of code evolution and advancing our understanding of both the early and late evolution of the canonical genetic code. In addition, the methods used represent large departures from the common methodology in studies of code evolution, and will help to spur further advancements in each field.

Chapter 5.1: Early Evolution of the Genetic Code:

Previously, work in this field focused on theory and conjecture, with little empirical evidence. In fact, most of the empirical evidence presented was based off of an artificial *in vitro* system to select for amino acid-binding RNA molecules. The *in vivo* evidence for such interactions is limited, and overall much less compelling. Without application of these theories to an *in vivo* system, their precise roles in shaping the evolution of the genetic code remain mysterious.

We show evidence in both the ribosome data and the proteomic data from DJ.2 that stereochemical interactions may not only exist in a modern organism, but may also continue to influence cellular functions. The ribosome data is particularly compelling. The ribosome sits at the interface of the RNA/protein world, and as such may have been shaped and molded by these stereochemical interactions. When I first started this work, I expected very minimal, if any, interactions to be seen in the ribosome. We approached the existence of these interactions primarily as relics of a bygone era, and expected that the long evolutionary history of the ribosome may have erased many of these imprints. What we found however was an abundance of such interactions in all organisms studied. These interactions may not represent relics of a bygone era; instead, their abundance suggests that they still play a structural role in modern organisms. This speculation has support from the ribosome data itself. The amino acid residues participating in these interactions differ between organisms, and appear no more conserved than other non-interacting amino acids. If these interactions were just relics of a common ancestral ribosome, more overlap should be seen, especially in closely related organisms (the three bacterial species). If however, these interactions represented a fundamental structural

interaction important to ribosome structure and function, then they would be more fluid, conserving the local interaction rather than the residue/RNA itself. Evidence for this is currently anecdotal, and a complete analysis of such conservation has not yet been done. The noise in the current data set may be a major drawback to performing this experiment, and a more refined data set with more compiled structural data may be needed.

The ribosome data set has an interesting implication towards the early evolution of the genetic code. Our data, although robust statistically, does not encompass all amino acid residues. This strongly suggests a multivariable evolution of the genetic code; a full stereochemical code cannot be achieved with our data. The multistage evolution to the genetic code appears dependent on the order of amino acid additions, with newer amino acids being added through stereochemical interactions. This idea is antithetical to previous dogma (Knight 1999), which suggests that stereochemistry may have shaped the original primordial code, and played less of a role in the expansion of the code. There are some other explanations for our data outside of a multistage evolution of the code. Many of the “oldest” amino acids are smaller, and would appear much different in their free form compared to within a polypeptide (which is how all amino acids would appear in the ribosome). Therefore, we may not be able to access these amino acids within the boundaries of our system. This theory has some support from the DJ.2 data, which suggests that valine may interact with its anticodon. Another explanation stems from evidence of the potential shuffling of some of these older amino acids. Older amino acids would have presumably been shuffled more, masking their original positions. Lastly, and most interestingly, is that these older amino acids were not placed in the genetic code via

a stereochemical interaction. This speculates that there remains some other mysterious driving force for the original codon assignments of these amino acids.

The proteomic data generated from the ambiguous DJ.2 strain not only supports some of the conclusions regarding the existence of a stereochemical code, but also presents a novel and powerful means for teasing out such interactions. Currently, this work addresses two amino acids, isoleucine and valine, in the organism *E. coli*. This methodology is easily expanded to other amino acids (Zhai 2005; Fersht 1979; Roy 2004; Wong 2003; Dock-Bregeon 2000; Karkhanis 2007) and other organisms (Bacher 2007; Yao 2008). This work should provide a solid framework for future work and eventually help to unravel the complete stereochemical code. In addition, it may more accurately pinpoint those amino acids whose placement is independent of stereochemical interactions.

Chapter 5.2: Late Evolution of the Genetic Code

The RF1-deletion strains generated will have impacts to both evolutionary and synthetic biology. The JX1.0 strain represents the first organism in which a sense codon (the UAG stop codon) has been reverted to nonsense. The introduction of an exogenous suppressor system results in the efficient production of protein harboring multiple unnatural amino acids. This result suggests that the now nonsense UAG codon can easily be “captured” by the introduction of a new amino acid. This capture is not beneficial to the organism and severely impairs the already slow growth phenotype of JX1.0. JX1.0 maybe in an intermediate state of code evolution, potentially mimicking the natural processes that occurred during the natural expansion of the genetic code with selenocysteine and pyrrolysine. This presents a unique opportunity to study the processes of code evolution. It is obvious from its growth phenotype that this strain is under extreme pressures, and over continual growth will attempt to adapt to such pressures. JX1.0 allows monitoring such pressures in real time, potentially leading to further insights into adaptation. Many questions can be addressed using such a system: will the strain further reduce endogenous UAGs as stop codons, or will UAGs spread throughout the genome? Will the strain make novel complementary mutations to the translational machinery in an effort to make protein more efficiently? Finally, one can even imagine in the distant future that this strain, when coupled to an unnatural system, may evolve novel mechanisms associated with an expanded genetic code.

The differences in the various *E. coli* strains that allow for RF1 removal lend immediate insights into these processes. There are many differences between the strains we used in these studies, most notably in the gene encoding RF2. Interestingly, none of

the known mutations in RF2 were enough to confer RF1 removal on their own. In addition, the presence of an endogenous suppressor tRNA did not appear to influence RF1 removal. Every successful knockout relied on other factors. With JX1.0, these factors are still unknown. There are some notable differences between the BL21 (DE3) strain and the related BL21 and REL606 strains (Jeong 2009), and further exploration of these differences may lead to a specific gene or groups of genes whose function (or even removal) is responsible for allowing such massive changes to the genetic code. It is also important to address the sole SNP found in JX1.0 that may be important to these phenotypes. Specifically, the mutation found in the intergenic region between *suhB* and *yfhR* may be in the promoter of *yfhR*, which is a conserved peptidase. Proteases are intriguing targets to us, as they play a role in the degradation of misfolded proteins, including those present in a stalled ribosome (Moore 2007). It is obvious in our system that we have translational issues from ribosome stalling; including anticodon-codon mismatching and ribosome fall-off at UAG positions (see the truncated products in **Figure 4.2.2b** specifically lanes 3 and 4). Nothing is known about the function of this protease, but it is feasible that it plays a role in monitoring such issues, and its inactivity may be a prerequisite for removal of RF1. This would prove insightful, as it would imply that a fully functional degradation system is helping to prevent changes to the genetic code. The evolution of variant codes may therefore hinge on a balance between the natural defenses guarding against such changes and the competitive advantages conferred from the presence of such stress-related defenses (see Santos 1999).

It is important to note however that the type of ambiguity in JX1.0 is strikingly different than what occurred in *Candida*. Instead of ambiguous codons in which one

amino acid is replaced with another, potentially in the interior of a protein, JX1.0 has ambiguity on what was once stop codons. The “ambiguous” proteome of JX1.0 would be represented by extended C-termini of the 300 genes that normally stop via a UAG-codon. The growth defect of JX1.0, especially when grown in the presence of *pActF*, makes it clear that these changes still interfere with protein function or folding, however it may be to a lesser extent. This may speak to why stop codons are such a frequent target for natural variations. The ambiguity imposed would be less detrimental, allowing for codon minimalization to be enough for reassignment to begin; organisms would not have to rely on the complete removal of a codon for capture to occur.

The JX1.0 results are unique from previous reports in a number of ways. Our knockout is unconditional, and relies on no complementary mutations in the genome, or no additional exogenous factors. The strain appears to survive through a combination of ribosome fall-off and anticodon-codon mismatches to produce proteins whose genes end in UAG. Unlike other reports that rely on the exploitation of endogenous systems to force changes in the genetic code, our system does not lead to a replacement of an endogenous amino acid. In effect, the resulting strain harboring an exogenous and orthogonal unnatural amino acid system represents the first fully 21-amino acid organism to be constructed in a laboratory setting.

Studies of the later evolution of the genetic code are focused on the two major avenues of evolutionary and synthetic (technological) biology. Although JX1.0 is interesting to the former, the technological usefulness of JX1.0 is limited by its severe growth defect. Although large amounts of protein can be harvested (double the yields when compared to JX3.0), the general growth phenotype may not be conducive to

unnatural amino acid mutagenesis. JX3.0 in contrast, appears to contain a precise balance between incorporation efficiency and health. This is most likely due to a number of factors, including the reduced genome in combination with the novel A293E mutation. Although further work needs to be done to solidify the function of this mutation, it is enticing to speculate that it allows some small level of UAG recognition. The Western results of endogenous *E. coli* genes support this, as expression of wildtype SufA protein is high in JX3.0 without the suppression system present (**Figure 4.3.6b** lanes 3 and 5). A293E also lies in the same interface layer of RF2 as E167K, a known mutation that allows for RF2 recognition of all three stop codons (Ito 1998). However, the A293E mutation was unable to complement the MRA8 strain harboring a temperature sensitive mutation in RF1. This may be where the reduced genome of MDS42 plays a role, limiting the sheer volume of activity needed from RF2.

Regardless of the mechanisms underlying JX3.0's capabilities, it is clear that this strain allows for the efficient and rapid production of proteins harboring multiple unnatural amino acids. This overcomes a major technological hurdle for this field, and should prove extremely useful in the production of recombinant proteins. What JX3.0 lacks (conditional of the A293E RF2 mutation), JX1.0 makes up for, making these strains a very valuable combination for the continued research into the evolution of the genetic code. These strains should hopefully represent the beginning of a new onset of interest in the evolution of the code, both for purely biological and technological reasons. Overall, in combination with our work supporting a stereochemical influence on the establishment of the code, I believe that this work will make a lasting imprint on this field.

References:

- Abe H, Abo T and Aiba H (1999) Regulation of intrinsic terminator by translation in *Escherichia coli*: transcription termination at a distance downstream. *Genes Cells* 4:87-97
- Agafonov DE, Kolb VA and Spirin AS (2001) A novel stress-response protein that binds at the ribosomal subunit interface and arrests translation. *Cold Spring Harb Symp Quant Biol* 66:509-14
- Alff-Steinberger C (1969) The genetic code and error transmission. *Proc Natl Acad Sci USA* 64:584-91
- Ambrogelly A, Palioura S and Söll D (2006) Natural expansion of the genetic code. *Nat Chem Biol* 3:29-35
- Ambrogelly A, Gundllapalli S, Herring S, Polycarpo C, Frauer C and Söll D (2007) Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. *Proc Natl Acad Sci USA* 104:3141-6
- Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. *J Mol Evol* 44:473-6
- Antos JM and Francis MB (2004) Selective tryptophan modification with rhodium carbenoids in aqueous solution. *J Am Chem Soc* 126:10256-7
- Ardell DH (1998) On error minimization in a sequential origin of the standard genetic code. *J Mol Evol* 47:1-13
- Bacher JM and Ellington AD (2001) Selection and characterization of *Escherichia coli* variants capable of growth on an otherwise toxic tryptophan analogue. *J. Bacteriol.* 183:5414-25
- Bacher JM, de Crécy-Lagard V and Schimmel PR (2005) Inhibited cell growth and protein functional changes from an editing-defective tRNA synthetase. *Proc Natl Acad Sci USA* 102:1697-701
- Bacher JM and Schimmel P (2007) An editing-defective aminoacyl-tRNA synthetase is mutagenic in aging bacteria via the SOS response. *Proc Natl Acad Sci USA* 104:1907-12
- Bacher JM, Waas WF, Metzgar D, de Crécy-Lagard V and Schimmel P (2007) Genetic code ambiguity confers a selective advantage on *Acinetobacter baylyi*. *J Bacteriol* 189:6494-6
- Bae JH, Rubini M, Jung G, Wiegand G, Seifert MH, Azim MK, Kim JS, Zumbusch A, Holak TA, Moroder L, Huber R and Budisa N (2003) Expansion of the genetic

- code enables design of a novel "gold" class of green fluorescent proteins. *J Mol Biol* 328:1071-81
- Ban N, Nissen P, Hansen J, Moore PB and Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289:905-20
- Banghart M, Borges K, Isacoff E, Trauner D and Kramer RH (2004) Light-activated ion channels for remote control of neuronal firing. *Nat Neurosci* 7:1381-6
- Barrell BG, Bankier AT and Drouin J (1979) A different genetic code in human mitochondria. *Nature* 282:189-94
- Beier H and Grimm M (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res* 29:4767-82
- Benzer S and Champe SP (1962) A change from nonsense to sense in the genetic code. *Proc Natl Acad Sci USA* 48:1114-21
- Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, Harney JW and Larsen PR (1991) Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. *Nature* 353:273-6
- Bilgicer B, Xing X and Kumar K (2001) Programmed self-sorting of coiled coils with leucine and hexafluoroleucine cores. *J Am Chem Soc* 123:11815-6
- Bock A, Forchhammer K, Heider J, Leinfelder W, Sawers G, Veprek B and Zinoni F (1991) Selenocysteine: the 21st amino acid. *Mol Microbiol* 5:515-20
- Borovinskaya MA, Pai RD, Zhang W, Schuwirth BS, Holton JM, Hirokawa G, Kaji H, Kaji A and Cate JH (2007) Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. *Nat Struct Mol Biol* 14:727-32
- Bose M, Groff D, Xie J, Brustad E and Schultz PG (2006) The incorporation of a photoisomerizable amino acid into proteins in *E. coli*. *J Am Chem Soc* 128:388-9
- Bronskill PM and Wong JT (1988) Suppression of fluorescence of tryptophan residues in proteins by replacement with 4-fluorotryptophan. *Biochem J* 249:305-8
- Browne DR, Kenyon GL and Hegeman GD (1970) Incorporation of monofluorotryptophans into protein during the growth of *Escherichia coli*. *Biochem Biophys Res Commun* 39:13-19
- Buchan JR, Aucott LS and Stansfield I (2006) tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res* 34:1015-27
- Budisa N, Minks C, Medrano FJ, Lutz J, Huber R and Moroder L (1998) Residue-specific bioincorporation of non-natural, biologically active amino acids into

- proteins as possible drug carriers: structure and stability of the per-thiaproline mutant of annexin V. *Proc Natl Acad Sci USA* 95:455-9
- Budisa N (2004) Prolegomena to future experimental efforts on genetic code engineering by expanding its amino acid repertoire. *Angew Chem Int Ed Engl* 43:6426-63
- Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V and Briggs SP (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci USA* 105:21034-8
- Chapeville F, Lipmann F, Von Ehrenstein G, Weisblum B, Ray WJ and Benzer S (1962) On the role of soluble ribonucleic acid in coding for amino acids. *Proc Natl Acad Sci USA* 48:1086-92
- Chapman E, Thorson J and Schultz PG (1997) Mutational analysis if backbone hydrogen bonds in staphylococcal nuclease. *J Am Chem Soc* 119:7151-7152
- Chattoraj M, King BA, Bublitz GU and Boxer SG (1996) Ultra-fast excited state dynamics in green fluorescent protein: multiple states and proton transfer. *Proc Natl Acad Sci USA* 93:8362-7
- Chen S, Schultz PG and Brock A (2007) An improved system for the generation and analysis of mutant proteins containing unnatural amino acids in *Saccharomyces cerevisiae*. *J Mol Biol* 371:112-22
- Ciesiolka J, Illangasekare M, Majerfeld I, Nickles T, Welch M, Yarus M and Zinnen S (1996) Affinity selection-amplification from randomized ribooligonucleotide pools. *Methods Enzymol* 267:315-35
- Cohen GN and Cowie DB (1957) Total replacement of methionine by selenomethionine in the proteins of *Escherichia coli*. *C R Hebd Seances Acad Sci* 244:680-3
- Cohen GN and Munier R (1959) Effects of structural analogues of amino acids on growth, protein synthesis, and enzyme synthesis in *Escherichia coli*. *Biochim Biophys Acta* 31:347-56
- Cook SN, Jack WE, Xiong X, Danley LE, Ellman JA, Schultz PG and Noren CJ (1995) Photochemically initiated protein splicing. *Angew Chem Int Ed Engl* 34:1629-1630
- Cornish VW, Mendel D and Schultz PG (1995) Probing protein structure and function with an expanded genetic code. *Angew Chem Int Ed Engl* 34:621-633
- Craigen WJ, Cook RG, Tate WP and Caskey CT (1985) Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci USA* 82:3616-20

- Crick FH, Griffith JS and Orgel LE (1957) Codes without Commas. *Proc Natl Acad Sci USA* 43:416-21
- Crick FH, Barnett L, Brenner S and Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* 192:1227-32
- Crick FH (1967) An error in model building. *Nature* 213:798
- Crick FH (1968) The origin of the genetic code. *J Mol Biol* 38:367-79
- Crick FH, Brenner S, Klug A and Pieczenik G (1976) A speculation on the origin of protein synthesis. *Orig Life* 7:389-97
- Dahlgren A and Rydén-Aulin M (2000) A novel mutation in ribosomal protein S4 that affects the function of a mutated RF1. *Biochimie* 82:683-91
- Dale T, Sanderson LE and Uhlenbeck OC (2004) The affinity of elongation factor Tu for an aminoacyl-tRNA is modulated by the esterified amino acid. *Biochemistry* 43:6159-66
- Dale T, Fahlman RP, Olejniczak M and Uhlenbeck OC (2009) Specificity of the ribosomal A site for aminoacyl-tRNAs. *Nucleic Acids Res* 37:1202-1210
- Dang H, England PM, Farivar SS, Dougherty DA and Lester HA (2000) Probing the role of a conserved M1 proline residue in 5-hydroxytryptamine(3) receptor gating. *Mol Pharmacol* 57:1114-22
- Datsenko KA and Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 97:6640-5
- Dawson PE, Muir TW, Clark-Lewis I and Kent SB (1994) Synthesis of proteins by native chemical ligation. *Science* 266:776-9
- Dedkova LM, Fahmi NE, Golovine SY and Hecht SM (2003) Enhanced D-amino acid incorporation into protein by modified ribosomes. *J Am Chem Soc* 125:6616-7
- Deiters A, Groff D, Ryu Y, Xie J and Schultz PG (2006) A genetically encoded photocaged tyrosine. *Angew Chem Int Ed Engl* 45:2728-31
- Deuschle U, Kammerer W, Gentz R and Bujard H (1986) Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. *Embo J* 5:2987-94
- Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol* 29:288-93
- Di Giulio M (1999) The coevolution theory of the origin of the genetic code. *J Mol Evol* 48:253-5

- Di Giulio M (2000) Genetic code origin and the strength of natural selection. *J Theor Biol* 205:659-61
- Di Giulio M and Amato U (2009) The close relationship between the biosynthetic families of amino acids and the organisation of the genetic code. *Gene* 435:9-12
- Dock-Bregeon A, Sankaranarayanan R, Romby P, Caillet J, Springer M, Rees B, Francklyn CS, Ehresmann C and Moras D (2000) Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem. *Cell* 103:877-84
- Döring V and Marlière P (1998) Reassigning cysteine in the genetic code of *Escherichia coli*. *Genetics* 150:543-51
- Döring V, Mootz HD, Nangle LA, Hendrickson TL, de Crécy-Lagard V, Schimmel P and Marlière P (2001) Enlarging the amino acid set of *Escherichia coli* by infiltration of the valine coding pathway. *Science* 292:501-4
- Dougherty DA (2000) Unnatural amino acids as probes of protein structure and function. *Curr Opin Chem Biol* 4:645-52
- Dounce AL (1952) Duplicating mechanism for peptide chain and nucleic acid synthesis. *Enzymologia* 15:251-8
- Dunnill P (1966) Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids. *Nature* 210:1265-7
- Eckert DM, Malashkevich VN, Hong LH, Carr PA and Kim PS (1999) Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. *Cell* 99:103-15
- Eisenhauer BM and Hecht SM (2002) Site-specific incorporation of (aminooxy)acetic acid into proteins. *Biochemistry* 41:11472-8
- Ellman JA, Mendel D and Schultz PG (1992) Site-specific incorporation of novel backbone structures into proteins. *Science* 255:197-200
- England PM, Lester HA, Davidson N and Dougherty DA (1997) Site-specific, photochemical proteolysis applied to ion channels in vivo. *Proc Natl Acad Sci USA* 94:11025-30
- England PM, Zhang Y, Dougherty DA and Lester HA (1999) Backbone mutations in transmembrane domains of a ligand-gated ion channel: implications for the mechanism of gating. *Cell* 96:89-98

- Fersht AR and Dingwall C (1979) Evidence for the double-sieve editing mechanism in protein synthesis. Steric exclusion of isoleucine by valyl-tRNA synthetases. *Biochemistry* 18:2627-31
- Forster AC, Tan Z, Nalam MN, Lin H, Qu H, Cornish VW and Blacklow SC (2003) Programming peptidomimetic syntheses by translating genetic codes designed de novo. *Proc Natl Acad Sci USA* 100:6353-7
- Fox GE and Naik AK (2004) The Evolutionary History of the Translation Machinery. In: LR de Pouplana (ed) *The Genetic Code and the Origin of Life*. Landes Bioscience, New York, 92-105
- Freeland SJ and Hurst LD (1998) The genetic code is one in a million. *J Mol Evol* 47:238-48
- Freeland SJ, Knight RD and Landweber LF (2000) Measuring adaptation within the genetic code. *Trends Biochem Sci* 25:44-5
- Freeland SJ, Wu T and Keulmann N (2003) The case for an error minimizing standard genetic code. *Orig Life Evol Biosph* 33:457-77
- Gamow G (1954) Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature* 173:318
- Gandhi CS and Isacoff EY (2005) Shedding light on membrane proteins. *Trends Neurosci* 28:472-9
- Geoghegan KF and Stroh JG (1992) Site-directed conjugation of nonpeptide groups to peptides and proteins via periodate oxidation of a 2-amino alcohol. Application to modification at N-terminal serine. *Bioconjug Chem* 3:138-46
- Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN and Osterman AL (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* 185:5673-84
- Gilbert W (1986) The RNA World. *Nature* 319:618
- Giriati I and Muir TW (2003) Protein semi-synthesis in living cells. *J Am Chem Soc* 125:7180-1
- Gross D and Tarver H (1955) Studies on ethionine. IV. The incorporation of ethionine into the proteins of *Tetrahymena*. *J Biol Chem* 217:169-82

- Haig D and Hurst LD (1991) A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33:412-7
- Haig D and Hurst LD (1999) A quantitative measure of error minimization in the genetic code. *J Mol Evol* 49:708
- Hamano-Takaku F, Iwama T, Saito-Yano S, Takaku K, Monden Y, Kitabatake M, Soll D and Nishimura S (2000) A mutant *Escherichia coli* tyrosyl-tRNA synthetase utilizes the unnatural amino acid azatyrosine more efficiently than tyrosine. *J Biol Chem* 275:40324-8
- Hartman MC, Josephson K and Szostak JW (2006) Enzymatic aminoacylation of tRNA with unnatural amino acids. *Proc Natl Acad Sci USA* 103:4356-61
- Hellinga HW, Wynn R and Richards FM (1992) The hydrophobic core of *Escherichia coli* thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. *Biochemistry* 31:11203-9
- Hendrickson TL, Nomanbhoy TK and Schimmel P (2000) Errors from selective disruption of the editing center in a tRNA synthetase. *Biochemistry* 39:8180-6
- Hendrickson WA (2000) Synchrotron crystallography. *Trends Biochem Sci* 25:637-43
- Herrera F, Chen Q, Fischer WH, Maher P, Schubert DR (2009) Synaptojanin-1 plays a key role in astrogliogenesis: possible relevance for Down's syndrome. *Cell Death Differ* 16:910-920
- Herring CD, Glasner JD and Blattner FR (2003) Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli*. *Gene* 311:153-63
- Hohsaka T, Ashizuka Y, Taira H, Murakami H and Sisido M (2001) Incorporation of nonnatural amino acids into proteins by using various four-base codons in an *Escherichia coli* in vitro translation system. *Biochemistry* 40:11060-4
- Hohsaka T and Sisido M (2002) Incorporation of non-natural amino acids into proteins. *Current opinion in chemical biology* 6:809-15
- Huang Y, Russell WK, Wan W, Pai PJ, Russell DH and Liu W (2010) A convenient method for genetic incorporation of multiple noncanonical amino acids into one protein in *Escherichia coli*. *Mol Biosyst* 6:683-6
- Ibba M, Kast P and Hennecke H (1994) Substrate Specificity Is Determined by Amino Acid Binding Pocket Size in *Escherichia coli*. *Biochemistry* 33:7107-12
- Ibba M and Hennecke H (1995) Relaxing the substrate specificity of an aminoacyl-tRNA synthetase allows in vitro and in vivo synthesis of proteins containing unnatural amino acids. *FEBS Lett.* 364:272-5

- Ibba M and Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* 69:617-50
- Illangasekare M, Sanchez G, Nickles T and Yarus M (1995) Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* 267:643-7
- Illangasekare M, Kovalchuk O and Yarus M (1997) Essential structures of a self-aminoacylating RNA. *J Mol Biol* 274:519-29
- Illangasekare M and Yarus M (1999) Specific, rapid synthesis of Phe-RNA by RNA. *Proc Natl Acad Sci USA* 96:5470-5
- Illangasekare M and Yarus M (1999) A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA* 5:1482-9
- Illangasekare M and Yarus M (2002) Phenylalanine-binding RNAs and genetic code evolution. *J Mol Evol* 54:298-311
- Ito K, Uno M and Nakamura Y (1998) Single amino acid substitution in prokaryote polypeptide release factor 2 permits it to terminate translation at all three stop codons. *Proc Natl Acad Sci USA* 95:8165-9
- Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L, Hur CG, Park HS, Segurens B, Kim SC, Oh TK, Lenski RE, Studier FW, Daegelen P and Kim JF (2009) Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol* 394:644-52
- Johnston WK, Unrau PJ, Lawrence MS, Glasner ME and Bartel DP (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* 292:1319-25
- Judice JK, Gamble TR, Murphy EC, de Vos AM and Schultz PG (1993) Probing the mechanism of staphylococcal nuclease with unnatural amino acids: kinetic and structural studies. *Science* 261:1578-81
- Jukes TH (1973) Possibilities for the evolution of the genetic code from a preceding form. *Nature* 246:22-6
- Jukes TH, Osawa S and Muto A (1987) Divergence and directional mutation pressures. *Nature* 325:668
- Jungck JR (1978) The genetic code as a periodic table. *J Mol Evol* 11:211-24
- Kaczanowska M and Ryden-Aulin M (2004) Temperature sensitivity caused by mutant release factor 1 is suppressed by mutations that affect 16S rRNA maturation. *J Bacteriol* 186:3046-55

- Kano A, Ohama T, Abe R and Osawa S (1993) Unassigned or nonsense codons in *Micrococcus luteus*. *J Mol Biol* 230:51-6
- Karkhanis VA, Mascarenhas AP and Martinis SA (2007) Amino acid toxicities of *Escherichia coli* that are prevented by leucyl-tRNA synthetase amino acid editing. *J Bacteriol* 189:8765-8
- Kawaguchi Y, Honda H, Taniguchi-Morimura J and Iwasaki S (1989) The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. *Nature* 341:164-6
- Kaya E, Gutsmedl K, Vrabel M, Muller M, Thumbs P and Carell T (2009) Synthesis of Threefold Glycosylated Proteins using Click Chemistry and Genetically Encoded Unnatural Amino Acids. *Chembiochem* 10:2858-61
- Keeling PJ and Doolittle WF (1997) Widespread and ancient distribution of a noncanonical genetic code in diplomonads. *Mol Biol Evol* 14:895-901
- Keiler KC, Waller PR and Sauer RT (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science* 271:990-3
- Kent SB (1988) Chemical synthesis of peptides and proteins. *Annu Rev Biochem* 57:957-89
- Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, Peralta-Gil M, Santos-Zavaleta A, Shearer AG and Karp PD (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* 37:D464-70
- Kiick KL and Tirrell DA (2000) Protein engineering by in vivo incorporation of non-natural amino acids: control of incorporation of methionine analogs by methionyl t-RNA synthetase. *Tetrahedron* 56:9587-93
- Kirshenbaum K, Carrico IS and Tirrell DA (2002) Biosynthesis of proteins incorporating a versatile set of phenylalanine analogues. *Chembiochem* 3:235-7
- Klaholz BP, Pape T, Zavialov AV, Myasnikov AG, Orlova EV, Vestergaard B, Ehrenberg M and van Heel M (2003) Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. *Nature* 421:90-4
- Knight RD and Landweber LF (1998) Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem Biol* 5:R215-20
- Knight RD, Freeland SJ and Landweber LF (1999) Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci* 24:241-7

- Knight RD, Freeland SJ and Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet* 2:49-58
- Knight RD, Landweber LF and Yarus, M (2001) How mitochondria redefine the code. *J Mol Evol* 53:299-313
- Koh JT, Cornish VW and Schultz PG (1997) An experimental approach to evaluating the role of backbone interactions in proteins using unnatural amino acid mutagenesis. *Biochemistry* 36:11314-22
- Kohn T, Kohda D, Haruki M, Yokoyama S and Miyazawa T (1990) Nonprotein amino acid furanomycin, unlike isoleucine in chemical structure, is charged to isoleucine tRNA by isoleucyl-tRNA synthetase and incorporated into protein. *J Biol Chem* 265:6931-5
- Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127-36
- Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE and Cech TR (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* 31:147-57
- Kumar RK and Yarus M (2001) RNA-catalyzed amino acid activation. *Biochemistry* 40:6998-7004
- Kwon I, Kirshenbaum K and Tirrell DA (2003) Breaking the degeneracy of the genetic code. *J Am Chem Soc* 125:7512-3
- LaRiviere FJ, Wolfson AD and Uhlenbeck OC (2001) Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science* 294:165-8
- Lawrence MS and Bartel DP (2003) Processivity of ribozyme-catalyzed RNA polymerization. *Biochemistry* 42:8748-55
- Lawrence MS and Bartel DP (2005) New ligase-derived RNA polymerase ribozymes. *RNA* 11:1173-80
- Leder P and Nirenberg M (1964) RNA codewords and protein synthesis, II. Nucleotide sequence of a valine RNA codeword. *Proc Natl Acad Sci USA* 52:420-7
- Leder P and Nirenberg MW (1964) RNA codewords and protein synthesis, III. On the nucleotide sequence of a cysteine and a leucine RNA codeword. *Proc Natl Acad Sci USA* 52:1521-9

- Lee HY, Lee KH, Al-Hashimi HM and Marsh EN (2006) Modulating protein structure with fluorous amino acids: increased stability and native-like structure conferred on a 4-helix bundle protein by hexafluoroleucine. *J Am Chem Soc* 128:337-43
- Lee JH, Yeo WS and Roe JH (2004) Induction of the *sufA* operon encoding Fe-S assembly proteins by superoxide generators and hydrogen peroxide: involvement of OxyR, IHF and an unidentified oxidant-responsive factor. *Mol Microbiol* 51:1745-55
- Legiewicz M and Yarus M (2005) A more complex isoleucine aptamer with a cognate triplet. *J Biol Chem* 280:19815-22
- Lehmann U (1985) Chromatographic separation as selection process for prebiotic evolution and the origin of the genetic code. *Biosystems* 17:193-208
- Lemeignan B, Sonigo P and Marlière P (1993) Phenotypic suppression by incorporation of an alien amino acid. *J Mol Biol* 231:161-6
- Levine M and Tarver H (1951) Studies on ethionine. III. Incorporation of ethionine into rat proteins. *J Biol Chem* 192:835-50
- Lin L, Hale SP and Schimmel P (1996) Aminoacylation error correction. *Nature* 384:33-4
- Link AJ, Mock ML and Tirrell DA (2003) Non-canonical amino acids in protein engineering. *Curr Opin Biotechnol* 14:603-9
- Lohse PA and Szostak JW (1996) Ribozyme-catalysed amino-acid transfer reactions. *Nature* 381:442-4
- Lozupone CA, Knight RD and Landweber LF (2001) The molecular basis of nuclear genetic code change in ciliates. *Curr Biol* 11:65-74
- Lozupone C, Changayil S, Majerfeld I, Yarus M (2003) Selection of the simplest RNA that binds isoleucine. *RNA* 9:1315-1322
- Luger K, Rechsteiner TJ and Richmond TJ (1999) Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol* 304:3-19
- Lummis SC, Beene DL, Lee LW, Lester HA, Broadhurst RW and Dougherty DA (2005) Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* 438:248-52
- Lummis SC, D LB, Harrison NJ, Lester HA and Dougherty DA (2005) A cation- π binding interaction with a tyrosine in the binding site of the GABAC receptor. *Chem Biol* 12:993-7

- Maiti A and Roy S (2005) Switching DNA-binding specificity by unnatural amino acid substitution. *Nucleic Acids Res* 33:5896-903
- Majerfeld I and Yarus M (1998) Isoleucine:RNA sites with associated coding sequences. *RNA* 4:471-8
- Majerfeld I, Puthenvedu D and Yarus M (2005) RNA affinity for molecular L-histidine; genetic code origins. *J Mol Evol* 61:226-35
- Majerfeld I and Yarus M (2005) A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res.* 33:5482-93
- Mannironi C, Scerch C, Fruscoloni P and Tocchini-Valentini GP (2000) Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif. *RNA* 6:520-7
- McCutcheon JP, McDonald BR and Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet* 5:e1000565
- McGinness KE, Wright MC and Joyce GF (2002) Continuous in vitro evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem Biol* 9:585-96
- McGinness KE and Joyce GF (2003) In search of an RNA replicase ribozyme. *Chem Biol* 10:5-14
- Means GE and Feeney RE (1990) Chemical modifications of proteins: history and applications. *Bioconjug Chem* 1:2-12
- Mendel D, Ellman JA, Chang Z, Veenstra DL, Kollman PA and Schultz PG (1992) Probing protein stability with unnatural amino acids. *Science* 256:1798-802
- Mendel D, Cornish VW and Schultz PG (1995) Site-directed mutagenesis with an expanded genetic code. *Annual review of biophysics and biomolecular structure* 24:435-62
- Miller JC, Silverman SK, England PM, Dougherty DA and Lester HA (1998) Flash decaging of tyrosine sidechains in an ion channel. *Neuron* 20:619-24
- Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* 117:528-9
- Miller SL, Urey HC and Oró J (1976) Origin of organic compounds on the primitive earth and in meteorites. *J Mol Evol* 9:59-72

- Miller SL (1987) Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb Symp Quant Biol* 52:17-27
- Minks C, Alefelder S, Moroder L, Huber R and Budisa N (2000) Towards new protein engineering: in vivo building and folding of protein shuttles for drug delivery and targeting by the Selective Pressure Incorporation (SPI) method. *Tetrahedron* 56:9431-41
- Miranda I, Silva R and Santos MA (2006) Evolution of the genetic code in yeasts. *Yeast* 23:203-13
- Moore SD and Sauer RT (2007) The tmRNA system for translational surveillance and ribosome rescue. *Annu Rev Biochem* 76:101-24
- Moura GR, Lousado JP, Pinheiro M, Carreto L, Silva RM, Oliveira JL and Santos MA (2007) Codon-triplet context unveils unique features of the *Candida albicans* protein coding genome. *BMC Genomics* 8:444
- Muir TW, Sondhi D and Cole PA (1998) Expressed protein ligation: a general method for protein engineering. *Proc Natl Acad Sci USA* 95:6705-10
- Murakami H, Saito H and Suga H (2003) A versatile tRNA aminoacylation catalyst based on RNA. *Chem Biol* 10:655-62
- Nakamura Y, Gojobori T and Ikemura T (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* 28:292
- Neumann H, Peak-Chew SY and Chin JW (2008) Genetically encoding N(epsilon)-acetyllysine in recombinant proteins. *Nat Chem Biol* 4:232-4
- Neumann H, Hancock SM, Buning R, Routh A, Chapman L, Somers J, Owen-Hughes T, van Noort J, Rhodes D and Chin JW (2009) A method for genetically installing site-specific acetylation in recombinant histones defines the effects of H3 K56 acetylation. *Mol Cell* 36:153-63
- Nguyen DP, Garcia Alai MM, Kapadnis PB, Neumann H and Chin JW (2009) Genetically encoding N(epsilon)-methyl-L-lysine in recombinant histones. *J Am Chem Soc* 131:14194-5
- Nirenberg M and Leder P (1964) RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of SRNA to ribosomes. *Science* 145:1399-407
- Nirenberg MW and Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci USA* 47:1588-602

- Nissen P, Kjeldgaard M, Thirup S, Polekhina G, Reshetnikova L, Clark BF and Nyborg J (1995) Crystal structure of the ternary complex of Phe-tRNA^{Phe}, EF-Tu, and a GTP analog. *Science* 270:1464-72
- Nissen P, Hansen J, Ban N, Moore PB and Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science* 289:920-30
- Noren CJ, Anthony-Cahill SJ, Griffith MC and Schultz PG (1989) A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* 244:182-8
- Novozhilov AS, Wolf YI and Koonin EV (2007) Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct* 2:24
- Nowak MW, Kearney PC, Sampson JR, Saks ME, Labarca CG, Silverman SK, Zhong W, Thorson J, Abelson JN, Davidson N and et al. (1995) Nicotinic receptor binding site probed with unnatural amino acid incorporation in intact cells. *Science* 268:439-42
- Oba T, Andachi Y, Muto A and Osawa S (1991) CGG: an unassigned or nonsense codon in *Mycoplasma capricolum*. *Proc Natl Acad Sci USA* 88:921-5
- Orgel LE (1968) Evolution of the genetic apparatus. *J Mol Biol* 38:381-93
- Osawa S and Jukes TH (1989) Codon reassignment (codon capture) in evolution. *J Mol Evol* 28:271-8
- Osawa S, Jukes TH, Watanabe K and Muto A (1992) Recent evidence for evolution of the genetic code. *Microbiol Rev* 56:229-64
- Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N and Weigel D (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024-33
- Pelc SR and Welton MG (1966) Stereochemical relationship between coding triplets and amino-acids. *Nature* 209:868-70
- Pezo V, Metzgar D, Hendrickson TL, Waas WF, Hazebrouck S, Döring V, Marlière P, Schimmel P and De Crécy-Lagard V (2004) Artificially ambiguous genetic code confers growth yield advantage. *Proc Natl Acad Sci USA* 101:8593-7
- Polycarpo C, Ambrogelly A, Berube A, Winbush SM, McCloskey JA, Crain PF, Wood JL and Soll D (2004) An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. *Proc Natl Acad Sci USA* 101:12450-4

- Posfai G, Plunkett G, 3rd, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, Burland V, Harcum SW and Blattner FR (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science* 312:1044-6
- Pratt EA and Ho C (1975) Incorporation of fluorotryptophans into proteins of *Escherichia coli*. *Biochemistry* 14:3035-40
- Ralph RK (1968) A suggestion on the origin of the genetic code. *Biochem Biophys Res Commun* 33:213-8
- Rennert OM and Anker HS (1963) On the incorporation of 5',5',5'-Trifluoroleucine into proteins of *E. coli*. *Biochemistry* 2:471-6
- Richardson SM, Wheelan SJ, Yarrington RM and Boeke JD (2006) GeneDesign: rapid, automated design of multikilobase synthetic genes. *Genome Res* 16:550-6
- Richmond MH (1959) Incorporation of canavanine by *Staphylococcus aureus* 524 SC. *Biochem J* 73:261-4
- Richmond MH (1962) The effect of amino acid analogues on growth and protein synthesis in microorganisms. *Bacteriological reviews* 26:398-420
- Ronneberg TA, Landweber LF and Freeland SJ (2000) Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc Natl Acad Sci USA* 97:13690-5
- Root-Bernstein RS (1982) On the origin of the genetic code. *J Theor Biol* 94:895-904
- Root-Bernstein RS (1982) Amino acid pairing. *J Theor Biol* 94:885-94
- Roy H, Ling J, Irnov M and Ibba M (2004) Post-transfer editing in vitro and in vivo by the beta subunit of phenylalanyl-tRNA synthetase. *EMBO J* 23:4639-48
- Rydén SM and Isaksson LA (1984) A temperature-sensitive mutant of *Escherichia coli* that shows enhanced misreading of UAG/A and increased efficiency for some tRNA nonsense suppressors. *Mol Gen Genet* 193:38-45
- Sagle LB, Zimmermann J, Matsuda S, Dawson PE and Romesberg FE (2006) Redox-coupled dynamics and folding in cytochrome c. *J Am Chem Soc* 128:7909-15
- Santoro SW, Wang L, Herberich B, King DS and Schultz PG (2002) An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nat Biotechnol* 20:1044-8
- Santos MA, Cheesman C, Costa V, Moradas-Ferreira P and Tuite MF (1999) Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp. *Mol Microbiol* 31:937-47

- Schmeing TM, Huang KS, Kitchen DE, Strobel SA and Steitz TA (2005) Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol Cell* 20:437-48
- Schon A, Bock A, Ott G, Sprinzl M and Soll D (1989) The selenocysteine-inserting opal suppressor serine tRNA from *E. coli* is highly unusual in structure and modification. *Nucleic Acids Res* 17:7159-65
- Schubert D, Herrera F, Cumming R, Read J, Low W, Maher P, Fischer WH (2009) Neural cells secrete a unique repertoire of proteins. *J Neurochem* 109:427-435
- Schultz DW and Yarus M (1994) Transfer RNA mutation and the malleability of the genetic code. *J Mol Biol* 235:1377-80
- Schultz DW and Yarus M (1996) On malleability in the genetic code. *J Mol Evol* 42:597-601
- Schultz KC, Supekova L, Ryu Y, Xie J, Perera R and Schultz PG (2006) A genetically encoded infrared probe. *J Am Chem Soc* 128:13984-5
- Schumacher TN, Mayr LM, Minor DL, Jr., Milhollen MA, Burgess MW and Kim PS (1996) Identification of D-peptide ligands through mirror-image phage display. *Science* 271:1854-7
- Scolnick E, Tompkins R, Caskey T, Nirenberg M (1968) Release Factors differing in specificity for terminator codons. *Proc Natl Acad Sci USA* 61:768-774
- Selmer M, Dunham CM, Murphy FV, Weixlbaumer A, Petry S, Kelley AC, Weir JR and Ramakrishnan V (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* 313:1935-42
- Shimizu M (1982) Molecular basis for the genetic code. *J Mol Evol* 18:297-303
- Silva RM, Paredes JA, Moura GR, Manadas B, Lima-Costa T, Rocha R, Miranda I, Gomes AC, Koerkamp MJ, Perrot M, Holstege FC, Boucherie H and Santos MA (2007) Critical roles for a genetic code alteration in the evolution of the genus *Candida*. *EMBO J* 26:4555-65
- Sonneborn TM (1965) Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. In: VaV Bryson, H.J. (ed) *Evolving Genes and Proteins*. Academic Press, New York, 377-397
- Srinivasan G, James CM and Krzycki JA (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* 296:1459-62
- Starzyk RM, Webster TA and Schimmel P (1987) Evidence for dispensable sequences inserted into a nucleotide fold. *Science* 237:1614-8

- Suchanek M, Radzikowska A and Thiele C (2005) Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nat Methods* 2:261-7
- Sugita T and Nakase T (1999) Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus *Candida*. *Syst Appl Microbiol* 22:79-86
- Suzuki T, Ueda T and Watanabe K (1997) The 'polysemous' codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. *EMBO J* 16:1122-34
- Szathmary E (1995) The major evolutionary transitions. *Nature* 374:227-32
- Szathmary E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet* 15:223-9
- Tang Y and Tirrell DA (2001) Biosynthesis of a highly stable coiled-coil protein containing hexafluoroleucine in an engineered bacterial host. *J Am Chem Soc* 123:11089-90
- Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP and Bafna V (2007) Improving gene annotation using peptide mass spectrometry. *Genome Res* 17:231-9
- Thorson JS, Chapman E, Murphy EC, Schultz PG and Judice JK (1995) Linear free energy analysis of hydrogen bonding in proteins. *J Am Chem Soc* 117:1157-1158
- Tilley SD and Francis MB (2006) Tyrosine-selective protein alkylation using pi-allylpalladium complexes. *J Am Chem Soc* 128:1080-1
- Ting AY, Shin I, Lucero C and Schultz PG (1998) Energetic analysis of an engineered cation-pi interaction in staphylococcal nuclease. *J Am Chem Soc* 120:7135-7136
- Tischer BK, von Einem J, Kaufer B and Osterrieder N (2006) Two-step red-mediated recombination for versatile high-efficiency markerless DNA manipulation in *Escherichia coli*. *Biotechniques* 40:191-7
- Trifonov EN (2004) The triplet code from first principles. *J Biomol Struct Dyn* 22:1-11
- Tsien RY (1998) The green fluorescent protein. *Annu Rev Biochem* 67:509-44
- Ueda K, Yamamoto Y, Ogawa K, Abo T, Inokuchi H and Aiba H (2002) Bacterial SsrA system plays a role in coping with unwanted translational readthrough caused by suppressor tRNAs. *Genes Cells* 7:509-19
- Umbarger HE and Mueller JH (1951) Isoleucine and valine metabolism of *Escherichia coli*. I. Growth studies on amino acid-deficient mutants. *J Biol Chem* 189:277-85

- Uno M, Ito K and Nakamura Y (1996) Functional specificity of amino acid at position 246 in the tRNA mimicry domain of bacterial release factor 2. *Biochimie* 78:935-43
- Uy R and Wold F (1977) Posttranslational covalent modification of proteins. *Science* 198:890-6
- Valiyaveetil FI, Sekedat M, Mackinnon R and Muir TW (2004) Glycine as a D-amino acid surrogate in the K(+)-selectivity filter. *Proc Natl Acad Sci USA* 101:17045-9
- van Hest JC and Tirrell DA (1998) Efficient introduction of alkene functionality into proteins in vivo. *FEBS Lett.* 428:68-70
- Vestergaard B, Van LB, Andersen GR, Nyborg J, Buckingham RH and Kjeldgaard M (2001) Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. *Mol Cell* 8:1375-82
- Wallace CJ (1995) Peptide ligation and semisynthesis. *Curr Opin Biotechnol* 6:403-10
- Wang J, Xie J and Schultz PG (2006) A genetically encoded fluorescent amino acid. *J Am Chem Soc* 128:8738-9
- Wang L, Brock A, Herberich B and Schultz PG (2001) Expanding the genetic code of *Escherichia coli*. *Science* 292:498-500
- Wang L, Xie J, Deniz AA and Schultz PG (2003) Unnatural amino acid mutagenesis of green fluorescent protein. *J Org Chem* 68:174-6
- Wang L, Zhang Z, Brock A and Schultz PG (2003) Addition of the keto functional group to the genetic code of *Escherichia coli*. *Proc Natl Acad Sci USA* 100:56-61
- Wang L and Schultz PG (2004) Expanding the genetic code. *Angew Chem Int Ed Engl* 44:34-66
- Wang Q, Parrish AR and Wang L (2009) Expanding the genetic code for biological studies. *Chem Biol* 16:323-36
- Weber AL and Lacey JC, Jr. (1978) Genetic code correlations: amino acids and their anticodon nucleotides. *J Mol Evol* 11:199-210
- Wilson DN, Schlutzen F, Harms JM, Starosta AL, Connell SR and Fucini P (2008) The oxazolidinone antibiotics perturb the ribosomal peptidyl-transferase center and effect tRNA positioning. *Proc Natl Acad Sci USA* 105:13339-44
- Wilson KS, Ito K, Noller HF and Nakamura Y (2000) Functional sites of interaction between release factor RF1 and the ribosome. *Nat Struct Biol* 7:866-70

- Woese CR (1965) Order in the genetic code. *Proc Natl Acad Sci USA* 54:71-5
- Woese CR (1965) On the evolution of the genetic code. *Proc Natl Acad Sci USA* 54:1546-52
- Woese CR, Dugre DH, Dugre SA, Kondo M and Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. *Cold Spring Harb Symp Quant Biol* 31:723-36
- Woese CR, Dugre DH, Saxinger WC and Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci USA* 55:966-74
- Woese CR, Olsen GJ, Ibba M and Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* 64:202-36
- Woese CR (2001) Translation: in retrospect and prospect. *RNA* 7:1055-67
- Wohlgemuth I, Brenner S, Beringer M and Rodnina MV (2008) Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. *J Biol Chem* 283:32229-35
- Wolf YI and Koonin EV (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol Direct* 2:14
- Wong JT (1975) A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA* 72:1909-12
- Wong JT (1980) Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc Natl Acad Sci USA* 77:1083-6
- Wong JT (1983) Membership mutation of the genetic code: loss of fitness by tryptophan. *Proc Natl Acad Sci USA* 80:6303-6
- Wong FC, Beuning PJ, Silvers C and Musier-Forsyth K (2003) An isolated class II aminoacyl-tRNA synthetase insertion domain is functional in amino acid editing. *J Biol Chem* 278:52857-52864.
- Wu N, Deiters A, Cropp TA, King D and Schultz PG (2004) A genetically encoded photocaged amino acid. *J Am Chem Soc* 126:14306-7
- Xie J, Wang L, Wu N, Brock A, Spraggon G and Schultz PG (2004) The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. *Nat Biotechnol* 22:1297-301
- Xie J and Schultz PG (2005) Adding amino acids to the genetic repertoire. *Current opinion in chemical biology* 9:548-54

- Xie J (2006) Adding unnatural amino acids to the genetic repertoire. Ph.D. Thesis. The Scripps Research Institute 46-66
- Xie J, Supekova L and Schultz PG (2007) A genetically encoded metabolically stable analogue of phosphotyrosine in *Escherichia coli*. *ACS Chem Biol* 2:474-8
- Yanagisawa T, Ishii R, Fukunaga R, Kobayashi T, Sakamoto K and Yokoyama S (2008) Crystallographic studies on multiple conformational states of active-site loops in pyrrolysyl-tRNA synthetase. *J Mol Biol* 378:634-52
- Yao P, Zhou XL, He R, Xue MQ, Zheng YG, Wang YF and Wang ED (2008) Unique residues crucial for optimal editing in yeast cytoplasmic Leucyl-tRNA synthetase are revealed by using a novel knockout yeast strain. *J Biol Chem* 283:22591-600
- Yarus M (1988) A specific amino acid binding site composed of RNA. *Science* 240:1751-8
- Yarus M (1989) Specificity of arginine binding by the *Tetrahymena* intron. *Biochemistry* 28:980-8
- Yarus M (1991) An RNA-amino acid complex and the origin of the genetic code. *New Biol* 3:183-9
- Yarus M (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J Mol Evol* 47:109-17
- Yarus M, Caporaso JG and Knight R (2005) Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* 74:179-98
- Yarus M, Widmann JJ and Knight R (2009) RNA-amino acid binding: a stereochemical era for the genetic code. *J Mol Evol* 69:406-29
- Yokogawa T, Suzuki T, Ueda T, Mori M, Ohama T, Kuchino Y, Yoshinari S, Motoki I, Nishikawa K, Osawa S and et al. (1992) Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: evolutionary implications. *Proc Natl Acad Sci USA* 89:7408-11
- Yoshikawa E, Fournier MJ, Mason TL and Tirrell DA (1994) Genetically engineered fluoropolymers. Synthesis of repetitive polypeptides containing p-fluorophenylalanine residues. *Macromolecules* 27:5471-5
- Yusupova GZ, Yusupov MM, Cate JH and Noller HF (2001) The path of messenger RNA through the ribosome. *Cell* 106:233-41
- Zabin HB, Horvath MP and Terwilliger TC (1991) Approaches to predicting effects of single amino acid substitutions on the function of a protein. *Biochemistry* 30:6230-40

- Zavialov AV, Mora L, Buckingham RH and Ehrenberg M (2002) Release of peptide promoted by the GGQ motif of class 1 release factors regulates the GTPase activity of RF3. *Mol Cell* 10:789-98
- Zhai Y and Martinis SA (2005) Two conserved threonines collaborate in the *Escherichia coli* leucyl-tRNA synthetase amino acid editing mechanism. *Biochemistry* 44:15437-43
- Zhang B and Cech TR (1997) Peptide bond formation by in vitro selected ribozymes. *Nature* 390:96-100
- Zhang B and Cech TR (1998) Peptidyl-transferase ribozymes: trans reactions, structural characterization and ribosomal RNA-like features. *Chem Biol* 5:539-53
- Zhang S, Ryden-Aulin M, Kirsebom LA and Isaksson LA (1994) Genetic implication for an interaction between release factor one and ribosomal protein L7/L12 in vivo. *J Mol Biol* 242:614-8
- Zhang Y, Baranov PV, Atkins JF and Gladyshev VN (2005) Pyrrolysine and selenocysteine use dissimilar decoding strategies. *J Biol Chem* 280:20740-51
- Zhang Z and Marshall AG (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J Am Soc Mass Spectrom* 9:225-33
- Zhong W, Gallivan JP, Zhang Y, Li L, Lester HA and Dougherty DA (1998) From ab initio quantum mechanics to molecular neurobiology: a cation-pi binding site in the nicotinic receptor. *Proc Natl Acad Sci USA* 95:12088-93
- Zuckerlandl L. (1965) Evolutionary Divergence and Convergence in Proteins. In: VaV Bryson, H.J. (ed) *Evolving Genes and Proteins*. Academic Press, New York,