The Dissertation Committee for Drew Scott Tack Certifies that this is the approved version of the following dissertation:

Enabling evolution with expanded genetic codes

Committee:

Andrew D. Ellington, Supervisor

Hal Alper

Jeffrey Barrick

Bryan W Davies

Andreas Matouschek

Enabling evolution with expanded genetic codes

by

Drew Scott Tack, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2016

Dedication

For my wife, Allyson.

And for Lilly and Henry.

Acknowledgements

This work described in this dissertation was possible only with the help of many people in my life. First and foremost, Michelle Byrom and Arti Pothukuchy, who manage the Ellington lab and keep everything running, despite our best attempts to break things, and Ella Watkins and Terry Stewart, both worked to ensure needed materials were always available. Also, a special thanks to Barrett Morrow and Matthew Forster, two undergraduate researchers who pulled through on experiments when I needed them most.

I owe the army of post-docs in the lab most of my sanity. Randy Hughes set me up when I first got here, and taught me the basics of research. Ross Thyer, for continually throwing bad ideas back and forth with me, and for making any of my words sound good. Jared Ellefson, for introducing me to StarCraft, and helping in experimental design and troubleshooting. And Andre Maranhao, who, by the time this is finished, will technically be post-doc.

I must thank my wife Allyson, for putting up with my erratic schedule over the last five years, for understanding when I needed to go in to lab in the middle of the night, or right after the kids went to bed. She kept me straight, even when I wanted to give up and just move to Japan, she would push me forward. And I need to thank my kids. Lilly and Henry are my happiness when a day is only sad. They always pull through with a hug just at the right time.

I also want to thank Andrew Ellington, who took me into his lab, despite minimal experience. I already look back on my days in this lab with fond remembrance. The lab culture he incubates is unique, his enabling allows creativity to flourish.

Enabling evolution with expanded genetic codes

Drew Scott Tack, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Andrew D. Ellington

The natural genetic code is largely shared amongst all terrestrial life, though variations do exist that provide evidence of genetic code evolution. The field of synthetic biology has developed methods to augment and modify the genetic code, by reassigning codons and genetically incorporating noncanonical amino acids (NCAAs). Most commonly, amino-acyl tRNA synthetases and their cognate tRNAs are engineered to recognize (NCAAs), the tRNA charged with an NCAA is then used during translation to decode the amber stop codon. Using this method, over 200 unique NCAAs have been added to the genetic code, allowing for the site-specific incorporation of useful chemistries, including covalent bond formation and fluorescence. NCAAs have been implemented in biological research and protein engineering. The second chapter of this dissertation explores the biophysical parameters that allow efficient crosslinking with a NCAA. We investigated several environmental factors which could impact crosslinking between two interacting proteins. We define a set of parameters that permit efficient crosslinking.

Noncanonical amino acids are a potentially powerful tool in understanding genetic code evolution. Evolutionary experiments using NCAAs have already been used to explore proposed theories on codon evolution, including codon capture and ambiguous intermediate theories. Though several studies have explored the implications of NCAAs

in evolution, significant obstacles have prevented the long-term evolution of wild-type organisms with expanded genetic codes. In chapter three of this dissertation we demonstrate that a single essential enzyme can be engineered to be structurally dependent on a genetically encoded NCAA. This confers a functional addiction to the NCAA and the addiction can be transferred to new bacterial hosts using a DNA plasmid. Chapter four expands on this idea, attempting to use the chemistries bestowed by the NCAA for active site chemistry. Finally, in chapter five we demonstrate that a protein-conferred cellular NCAA addiction is sufficient for the long-term evolution of bacteria with a recoded amber codon. We investigated the genetic and phenotypic impact of evolution with an expanded genetic code.

Table of Contents

List of Tables

List of Figures

Chapter 1: The Genetic Code: Structure, Plasticity, and Evolution

The genetic code is nearly universal, with all known lifeforms relying upon the same five nucleotides and 22 proteinogenic amino acids. Yet natural variations exist which provide evidence of the evolutionary history of the genetic code. While several theories have been proposed on the evolution of the genetic code and the arrangement of the codon table, experimental exploration of these theories has largely been limited to computational modeling and other *in silico* approaches. Over the last few decades, advances in synthetic biology have allowed for the augmentation of the genetic code, through the reassignment of codons and the addition of noncanonical amino acids into the proteome. These advances should provide insights into genetic code evolution, as well as enable the evolution of organisms that utilize expanded genetic codes throughout their proteome. While progress has been made towards these goals, barriers still exist which inhibit the full implementation of expanded genetic codes in biology.

The genetic encoding of the components necessary to maintain life are held in polymeric nucleic acids and stably stored in genomic DNA, the biomolecule responsible for heritability (Avery et al., 1944). The nucleotide sequence of genomic DNA is used as a template to transcribe a copy using the chemically similar RNA backbone, which in turn is translated by the ribosome though the RNA-sequence defined assembly of amino acids, coded for by triplet nucleotide codons. This general idea was hypothesized more than half a century ago, (Crick, 1958) and the then-imminent elucidation of the genetic code and decoding of triplet codons verified this hypothesis (Matthaei et al., 1962).

The triplet nature of codons paired with the four available genetic nucleotides allows for 64 possible codons (4x4x4=64) in the genetic code (**Figure 1-1** and **Figure 1-2**). Typically, 61 codons code for 20 standard amino acids which are decoded using

transfer RNA (tRNA), and the remaining three used to signal the termination of protein synthesis. Before being used in translation, tRNAs are charged with the appropriate amino acid by their amino-acyl tRNA synthetase (aaRS), which ligates the correct amino acid onto its corresponding tRNA. The charged tRNA is then recognized by an elongation factor (EF-Tu in bacteria) and is delivered to the ribosome where it participate in the translation of an mRNA codon. The remaining three codons, amber (UAG), ochre (UAA), and opal (UGA), code for the termination of protein synthesis, accomplished in the ribosomal A-site by protein release factors which recognize these stop codons and release the translated peptide from the ribosome (Scolnick et al., 1968). This arrangement of the codons was thought to be universal, and famously hypothesized to be a "frozen accident" (Crick, 1968). This conjures images of a primordia ooze, with a milieu of diverse organisms containing the foundation of the genetic code; DNA, RNA, and proteins, but with different arrangements of codons, or an alternate set of amino acids in each organism. Yet our single arrangement of 20 amino acids somehow persisted, by chance or merit, and life is now constrained to operate within any limitations of the code.

But how frozen is the genetic code? Is life forever confined to this, perhaps suboptimal, code? Or is it evolving even now, and it always has been? Evolution of something so fundamental as the genetic code would likely require enormous time periods, making observation and experimentation difficult. In this era of biotech, can we augment the foundation of life, and peak into mechanisms of genetic code evolution? Can we evolve new genetic codes, either through rational engineering or directed evolution of whole organisms? This dissertation focuses on using stop codon reassignment, and adding additional noncanonical amino acids to the genome, as a method to explore the evolutionary impact of expanded genetic codes.

1.1 Expanding the Genetic Code

Since the elucidation of the genetic code, the dogma that it is a frozen artifact of evolution; ridged, unalterable, and inflexible, has slowly thawed. In the era of bioinformatics, natural variations to the code have been discovered with additional proteinogenic amino acids, demonstrating natural variations among extant life. While simultaneously, biotechnological advances have made it possible to engineer the genetic code, altering it to fit specific needs. Together, the once untouchable central dogma of biology is proving to be flexible and engineerable.

1.1.1 Natural Expansion of the Code

For nearly 30 years the frozen accident hypothesis seemed to fully describe life. The 20 standard proteinogenic amino acids were considered the entirety of the genetic code, and contained all information that was necessary for survival of Earth's diverse biome. Recently, however, examples have been identified that demonstrate that life has evolved beyond the 20 standard proteinogenic amino acids (**Figure 1-1**). Nature has devised mechanisms to use two additional amino acids which provide unique chemistries beyond the standard genetic code. The noncanonical amino acids selenocysteine and pyrrolysine are evidence that the genetic code is not entirely frozen, and nature itself has evolved to use new amino acids. Further demonstrating the flexibility of the code, both of these amino acids use stop codons as signals for incorporation, defying the canonical translation scheme. Importantly, these examples of natural variation to the standard genetic code demonstrate that nature can use new amino acids to accomplish new functions, and thus can reach higher fitness peaks with additional amino acid chemistries.

1.1.1.1 Selenocysteine

In the late 1980s, the discover of selenocysteine, and is often called the 21st amino acid, was the first example of natural expansion of the genetic code. Selenocysteine is biosynthesized by modifying a serine-charged opal (UGA) stop codon suppressor tRNA, and is encoded site-selectively with signaling from an mRNA structural element and a specialized elongation factor (Berry et al., 1991). Selenocysteine is used across all domains of life; it is essential in some, and nonessential in others, and is an ancient component of the genetic code (Allmang and Krol, 2006; Atkinson et al., 2011; Böck et al., 1991; Kaiser et al., 2005; Kryukov et al., 1999). Selenocysteine brings unique chemistry to the code that cannot be achieved with canonical amino acids. It is one of two proteinogenic amino acids with side chains readily capable of forming covalent bonds, the other being cysteine, though no structural diselenide bonds have been discovered in nature. The discovery of selenocysteine was irrefutable evidence that the universal genetic code was not so universal, and that variations could arise.

1.1.1.2 Pyrrolysine

More recently, pyrrolysine was identified as the 22nd proteinogenic amino acid. Unlike selenocysteine, pyrrolysine has only been identified in methanogenic archaea and a single eubacterium, *Desulfitobacterium hafniense* (Herring et al., 2007), likely indicating an evolutionary appearance after selenocysteine. Like selenocysteine, pyrrolysine is inserted at a stop codon, though a different codon, using amber (UAG) suppression instead of opal suppression, as selenocysteine does. Pyrrolysine is biosynthesized in the cell, then charged on a tRNA using a pyrrolysine specific aaRS (Atkins and Gesteland, 2002; Gaston et al., 2011; Hao et al., 2002). Like selenocysteine, pyrrolysine performs a chemical function which cannot be accomplished by any of the 20 standard proteinogenic amino acids, and is essential for enzymatic function of

methyltransferases in some archea (Hao et al., 2002), again demonstrating nature can evolve new chemistries to reach higher fitness peaks.

1.1.1.3 Implications of selenocysteine and pyrrolysine

For nearly 30 years, the universal genetic code was considered all that was necessary for life. Since then, two exceptions to the universal code have been discovered in nature. Each of these expanded amino acids provide unique functions that cannot be accomplished with the universal code. It's likely that selenocysteine and pyrrolysine evolved at different periods in history, as selenocysteine is used among all domains of life, while pyrrolysine is limited to a specific branch of archaea. Additionally, both amino acids use ambiguous codons for incorporation, each utilizing a stop codon, selenocysteine using the UGA codon, and pyrrolysine utilizing UAG codon. We are now in an era where genome sequencing is relatively easy, and bioinformatics is a major focus. It's possible that other variations to the code exist, and have yet to be identified. But certainly, these two natural expansions demonstrate that the genetic code is far from frozen, and is, like all things, amenable to evolution.

1.1.2 Demonstrations of code plasticity *in vitro*

1.1.2.1 In vitro *incorporation of NCAAs at cognate codons*

After the elucidation of the genetic code and the underlying mechanisms of protein translation, it was apparent that aaRSs played an important role in maintaining = translation fidelity. Each aaRS required specificity toward both cognate tRNA and appropriate amino acid. Other translational machinery, including the ribosome and elongation factors, is generally shared amongst amino acids and tRNAs. Since the aaRS was largely responsible for amino acid recognition, circumventing the aaRS might allow incorporation of NCAAs by the ribosome and elongation factors. This was first

demonstrated when synthetically acylated tRNAs charged with L-4'-[3-(trifluoromethyl)-3*H*-diazirin-3-yl]phenylalanine were readily used by the ribosome in an *in vitro* assay using rabbit reticulocyte lysate (Baldini et al., 1988). The tRNA was read as a phenylalanine, and encoded into translating proteins at phenylalanine codons.

1.1.2.2 In vitro *incorporation of NCAAs at stop codons*

With NCAAs readily incorporated by cellular translation machinery, a codon was needed to signal for site specific NCAA incorporation. While 61 of the 64 possible codons are already occupied by natural amino acids, it was well known that some organisms used suppressor tRNAs complementary to stop codons as alternatives for typical release factor mediated translational termination (Murgola, 1985), and that some aaRSs would recognize their tRNA even if mutated to complement a stop codon (Bruce et al., 1982). As such, stop codons were identified as possible sites for NCAA incorporation. Using this approach, two groups reported that the UAG codon could be used as a signal for NCAA incorporation *in vitro*, one using a biologically produced tRNA which was chemically charged with a NCAA (Noren et al., 1989), and another which used a synthetic tRNA (Bain et al., 1989) charged with an NCAA. This demonstrated that the core genetic code was amenable not only to NCAA incorporation, but to the conversion of a stop codon to a cognate NCAA incorporation site. While laborious, using synthetically charged tRNAs for *in vitro* NCAA incorporation had demonstrated uses in exploring protein structure, elucidating pathways, and understanding protein mechanisms, as was demonstrated in elucidating the role of a glutamate thought to be responsible for enzymatic activity in staphylococcal nuclease (D Mendel et al., 1995).

While *in vitro* methods for NCAA incorporation have largely been superseded by efforts towards *in vivo* methods (**Chapter 1.1.3**), the flexibility of *in vitro* systems allows for NCAA incorporation at multiple codons, or incorporation of NCAAs that are largely rejected by incorporation machinery *in vivo*. A benefit to *in vitro* translation is that each component of translation can be individually added or removed for the translation mixture, including release factors, tRNAs, aaRSs, and even canonical amino acids (Forster et al., 2003; Goto et al., 2011; Shimizu et al., 2001). In fact, *in vitro* translational systems have been developed which function in an entirely different fashion than natural translational properties, with *de novo* ribozymes that can charge tRNAs with NCAAs that are difficult to incorporate using *in vivo* methods described below, including D-amino acids (Goto et al., 2011; Ohuchi et al., 2007), without engineering aaRS specificity for each desired NCAA.

1.1.3 Methods for *in vivo* Incorporation of NCAAs

While expansion of the code had successfully been demonstrated *in vitro*, the process was laborious, and synthetically charged tRNAs could only be used for a single codon translation event, severely limiting production of proteins. Ideally, aaRSs could be used to charge tRNAs with an NCAA *in vivo*, allowing for unlimited turnover and continuous production of protein. As such, several approaches have been developed to incorporate NCAAs into proteins *in vivo*.

1.1.3.1 Using aaRS Promiscuity for NCAA incorporation in vivo

AaRS fidelity is crucial to translation and thus biological function. Despite this, it has long been known that certain amino acid analogs could be recognized by aaRSs and would be charged to the corresponding tRNA. The tRNA could then be used by the ribosome for the translation of its natural codon complement. This was first demonstrated

when methionine was replaced with the ethionine in rat proteins (Levine and Tarver, 1951), and subsequent studies showed similar aaRS flexibility (Cowie and Cohen, 1957; Richmond, 1962). An exploration of the promiscuity of the *E. coli* methionyl-synthetase demonstrated that a number of methionine analogs, including chemically reactive, terminally unsaturated amino acids could be charged to the methionyl-tRNA and would be used during translation at ATG codons by the ribosome (van Hest et al., 2000). Analysis of translated proteins revealed that ATG codons were read ambiguously during translation and contained either methionine, the natural metRS substrate, or the methionine analog supplemented to the host.

Using the promiscuity of aaRSs has allowed for the introduction of chemically reactive amino acids, including several capable of forming covalent bonds, such as azidohomoalanine (Dieterich et al., 2007). This methodology requires no changes to the genome, providing a facile method to monitor *in vivo* protein synthesis (Ngo and Tirrell, 2011) and detailed proteomic analysis of individual cells in complex multicellular organisms (Yuet et al., 2015). This same general process has also been used to force the tolerance of toxic amino acids into the genome of growing cells (Lemeignan et al., 1993).

1.1.3.2 Engineering host aaRSs for relaxed substrate specificity

While aaRS promiscuity allows for *in vivo* incorporation of NCAAs, nature has evolved aaRSs for specificity, and the range of NCAAs that can be used as a substrate is limited. Expanding the number of NCAAs which can be encoded required engineering aaRSs for relaxed or changed amino acid specificity. As recombinant DNA technology and protein structural information became available, engineering aaRSs became feasible. The first demonstration of relaxed aaRS fidelity used a single amino acid substitution in *E. coli* phenylalanyl-synthetase (A294G) (Ibba et al., 1994). This single mutation allowed

for tRNA charging with the NCAA *p*-chlorophenylalanine (PCFA). Follow up work demonstrated that the tRNA charged with PCFA was used in protein synthesis *in vivo* by the host organism (Ibba and Hennecke, 1995). This same mutation, as well as this mutation paired with a secondary mutation (T251G), later showed the ability to incorporate a number of tyrosine analogs, including chemically reactive amino acids *p*-cyano-, *p*-azido-, and *p*-keto- phenylalanine (Datta et al., 2002; Kirshenbaum et al., 2002), and photoreactive amino acids including benzofuranylalanine (Bentin et al., 2004).

1.1.3.3 Assigning a Codon for NCAAs

While successfully demonstrating aaRSs could be engineered to incorporate NCAAs *in vivo*, there were detrimental fitness effects to the host when incorporating the NCAA. Fitness effects were later identified as a consequence of global NCAA incorporation at cognate codons throughout the genome (Ibba and Hennecke, 1995). Beyond fitness cost, commandeering a cognate codon for NCAA incorporation results in a mixture of amino acids at a given codon. For efficient incorporation, and to maintain cellular viability, a designated codon was needed to signal NCAA incorporation. Preferably, the NCAA would be the sole amino acid coded by the codon, which significantly reduced the codon options to stop codons. The UAG stop codon had been used for NCAA incorporation *in vitro* (Noren et al., 1989), and was followed by a demonstration of *in vivo* suppression, when an *E. coli* UAG suppressor tRNA and corresponding tyrRS were heterologously expressed and shown to function in *S. cerevisiae* (Edwards and Schimmel, 1990). UAG is the rarest codon throughout the genome, and signals for termination, and so assignment to an NCAA meant only translational competition between the NCAA and termination (**Figure 1-3**)

1.1.3.4 Engineering orthogonal translational machinery

The semi-rational engineering of an aaRS for the improved mis-incorporation of NCAAs at cognate codons throughout a genome demonstrated the plasticity of the genetic code translational machinery (**Chapter 1.1.3.2**). As high resolution structural information became available, and methods to generate large libraries of protein variants were perfected, it became possible to generate aaRSs which were NCAA specific as opposed to amino acid generalists. This requires the genetic encoding a new aaRS into an organism, and engineering the aaRS for NCAA specificity, as well as the tRNA. For high fidelity incorporation of an NCAA into the genome, the supplemented aaRS and tRNA must function in the new host, while simultaneously not interacting with host tRNAs or aaRSs. A number of these orthogonal translation systems, comprised of an aaRS and tRNA from a genetically distinct lineage from the host, have been developed for use in a variety of model organisms.

The original, and still most common, OTS engineered for NCAA incorporation is the *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase and corresponding tRNA. The *Mj* tyrRS/tRNA pair is optimal because the cognate codons for tyrosine (TAT/TAC) are similar to the amber codon (TAG) that is often used for NCAA incorporation, and the *Mj* tyrRS has a minimal anticodon recognition domain (Steer and Schimmel, 1999). This meant that the *Mj*-tyrRS would still recognize and charge the tRNA even if the anticodon of the tRNA was mutated to complement the UAG stop codon. Furthermore, an ideal OTS would function without interfering with, or interference from the host translational machinery (**Figure 1-4**). Since most NCAA incorporation experiments are performed in *E. coli*, it was essential to use an aaRS/tRNA pair which was orthogonal to the *E. coli* translational machinery. The archaeon *M. jannaschii* is a genetically distinct and far

removed from *E. coli*, and the tyrRS/tRNA pair has been shown to be orthogonal in *E. coli*.

The structure of the wild type *Mj* tyrRS:tRNA complex has been solved (Kobayashi et al., 2003) (**Figure 1-5**), providing insight into the orthogonality of the pair in *E. coli*, but also informing attempts to engineer the *Mj*OTS. In fact, many of the *Mj*OTSs engineered for NCAA incorporation have been selected by randomizing the same residues, those responsible for amino acid binding and recognition, including Tyr32, His70, Asp158, Ile159, and Leu162. Many of these engineered versions of the *Mj*OTS that are specific for NCAAs have been crystallized and the structures solved. These structures have further shed insight into the functionality of the synthetase in amino acid and tRNA recognition (Cooley et al., 2014; Sakamoto et al., 2009) (**Figure 1-5**).

While the *Mj* tyrRS/tRNA pair has been successfully engineered to incorporate a range of diverse NCAAs (see **Chapter 1.2.2**), the ability to successfully evolve an aaRS for NCAA recognition depends on the structural limitations of the aaRS active site. As such, several other aaRSs, based on many of the existing natural amino acids, have been engineered to orthogonally incorporate NCAAs in model organisms. Like the *Mj*tyr OTS, a number of OTSs have been derived from archaea, including the pyrrolysine synthetase from the methanogenic archaea *Methanosarcinaceae*, which naturally recognized the amber stop codon, has been demonstrated to introduce a number of NCAAs with minimal engineering (Tuley et al., 2014). Other archaeal OTSs include the Glutamyl-RS from *Methanosarcina mazei*, and the Lysyl-RS from *Pyrococcus horkoshii*, both of which have been further engineered to recognized quadruplet codons (Anderson and Schultz, 2003; Anderson et al., 2004). The *Saccharomyces cerevisiae* tryptophanyl-tRNA synthetase has been engineered for orthogonal NCAA incorporation in *E. coli* (Ellefson et al., 2014;

Hughes and Ellington, 2010). Other *S. cerevisiae* OTSs have been used for NCAA incorporation, including the aspartyl-RS, glutamyl-RS, tyrosyl-RS (Furter, 1998; Liu et al., 1997; Ohno et al., 1998; Pastrnak et al., 2000). And the *E. coli* leucyl- and tyrosyl-synthetases has been engineered for NCAA incorporation into yeast and mammalian cells (Brustad et al., 2008).

While the aaRS is often the focus of engineering, as the gate responsible for both amino acid and tRNA recognition, the corresponding orthogonal tRNA can also exert significant influence on translation. Foreign tRNAs can have interplay with host aaRS, resulting in charging with undesired natural amino acids, as was seen when an *E. coli* tRNA was charged with lysine when expressed in *S. cerevisiae* (Furter, 1998). Harboring an OTS carries a significant burden on the host, and work toward engineering the *Mj*-tyrtRNA showed that it was possible to decrease misacylation by native aaRSs while still retaining NCAA charging by the *Mj*-tyrRS (Wang and Schultz, 2001). Somewhat surprisingly, later work revealed that the tRNA was important in NCAA recognition, and could be evolved to increase NCAA specificity (Guo et al., 2009; Maranhao and Ellington, 2016) and decrease fitness effects typically associated with harboring an OTS (Maranhao and Ellington, 2016).

1.1.5 Conclusions

The universal genetic code was, for several decades, considered frozen and immovable. Since then, a series of natural discovers as well as engineering breakthroughs have demonstrated that the genetic code is not entirely ridge. Nature itself provides examples of expanded codes, having evolved pathways for the addition of at least two noncanonical amino acids. Beyond that, aaRSs have some amino acid promiscuity, and can incorporate chemically reactive NCAAs that are structurally similar to the natural

amino acid. Furthermore, aaRSs can be engineered at the amino acid binding pocket for NCAA recognition and specificity, and can be introduced from distant organisms into host cells with minimal reactivity with host translational machinery. The tRNA is also important in NCAA incorporation, and can be used to further tune NCAA specificity and reduce the fitness cost of harboring an OTS.

1.2 Expanding the Chemistries of Proteins

Chemical modification of proteins has been a powerful tool in elucidating biological interactions and metabolic pathways, as well as in biochemical engineering. The development of amino-acid specific crosslinkers has been a boon in interactome studies, the most prominent include protein surface labeling with *N*-hydroxysuccinimide (NHS) or Sulfo-NHS, targeting proteinogenic amines or carboxylic acids (Staros et al., 1986), and succinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate (SMCC), which reacts selectively with nucleophilic amines and thiols (Brinkley, 1992). Other forms of protein modification include enzymatic modification of proteins (Hwang et al., 2004), and the encoding of biochemically unique proteins or protein domains to the protein of interest, including GFP (Hanson et al., 2004; Wilson et al., 2004) and SNAP tag (Keppler et al., 2003), which provides a unique physiochemical environment for the generalized labeling of proteins (Gautier et al., 2008; Juillerat et al., 2003). Nature has even evolved sequence specific modifications, which can be used to label proteins (Dierks et al., 1999; Rush and Bertozzi, 2008). Yet, the development of site specific NCAA incorporation has superseded these technologies, allowing the incorporation of amino acids which can accomplish each of these tasks, and more, with the replacement of a single amino acid.

1.2.1 Noncanonical amino acids for protein modification

The addition of noncanonical amino acids (NCAAs) to the genetic code allows for the incorporation of novel and useful chemistries not available through the standard proteinogenic amino acids. While a series of approaches have been used to modulate protein structure and function, NCAAs allow for site specific modification of proteins with minimal changes to the target protein sequence and overall structure. NCAAs provide site specificity, orthogonal reactivities, and scarless insertion. The utility of an augmented code was recognized over 50 years ago as a way to study protein function by inserting amino acid analogs with chemical synthesis of proteins (Hofmann and Bohn, 1966), and was later recognized as a tool for use in peptide and protein design and engineering (Balaram, 1992; Link et al., 2003). Since then, a series of progressive advancements have led to the *in vivo* site specific incorporation of an NCAAs with unique and useful chemistries.

1.2.2.1 Crosslinking

The chemical labeling of proteins has been a powerful tool for understanding and engineering biology. Historically, proteins have been chemically modified after translation at specific reactive groups present in the natural genetic code, or with the addition of large protein tags. Historical methods are nonspecific – they react with all surface exposed amino acids containing the target chemistry (i.e. every cysteine, every lysine, or every glutamate *and* aspartate, *etc*.). Site specific crosslinking is a highly desirable characteristic, but is difficult to accomplish in most proteins using these traditional labeling methods. Several systems to incorporate NCAAs with crosslinking chemistry have been engineered. Crosslinking NCAAs have distinct advantages over traditional methods; many are bioorthogonal, a number have been incorporated that can be crosslinked at physiologically relevant pH and temperatures, preserving protein

structure and function, and they can be incorporated *in vivo*, and several can be activated for crosslinking *in vivo.*

Ketones and aldehydes are chemically reactive functional groups that are often used as reaction sites in organic chemistry. The reactivity of carbonyls is largely outside the realm of biomolecules, allowing covalent bond formation in biological conditions with minimal interference from proteins and nucleic acids. Additionally, their small size has made them amenable to incorporation using NCAAs. *p*-acetyl-L-phenylalanine has been incorporated into proteins using engineered OTSs, and has facilitated protein complex assembly using alkoxyamine- or hydrazide- crosslinking functional groups, in slightly acidic conditions (Kim et al., 2012). These requirements ensure crosslinking does not occur *in vivo*, but when initiated by protein treatments and the presence of the associated crosslinkers.

Possibly one of the most common crosslinking systems used with NCAAs is azide- and alkyne- reactions. At least two sets of OTSs have been developed for each of these functional groups, using the *M. jannaschii* OTS, and the pyrrolysine OTS (Chin et al., 2002a; Deiters and Schultz, 2005; Tuley et al., 2014; Wang et al., 2012, 2013). Azide- and alkyne- NCAAs are attractive because they can interact with each other in a bioorthogonal manner, through 3+2 cycloadditions.

The NCAA *p*-benzophenone (pBpa) has been incorporated using an engineered OTS and is chemically inert in low light conditions. Upon UV irradiation, pBpa crosslinks preferentially with inactivated C-H bonds, which are common in the proteome (Chin et al., 2002b; Galardy et al., 1973). Because UV irradiation can induce crosslinking in living cells, pBpa has been used to map protein-protein interactions *in vivo* (Pham et al., 2013; Shiota et al., 2013). In addition to amino acid crosslinking, pBpa is also capable

of crosslinking with DNA, expanding the potential interaction mapping capabilities beyond protein interactions (Lee et al., 2009a).

Alternatively, the NCAA L-DOPA can be oxidized and crosslinks with reactive nucleophiles, including the amino acids cysteine, lysine, and histidine (Liu et al., 2006). The crosslinking of L-DOPA to canonical amino acids not only provides protein interaction information, but also sequence information as well, as it reacts with a limited set of natural amino acids. L-DOPA has been used to determine the binding site of a peptide-protein complex *in vitro*, using oxidation as the crosslinking trigger (Burdine et al., 2004).

1.2.2.2 Post-translational modifications

Nature uses chemical modifications through post-translational modification (PTM) of proteins for the activation or deactivation of enzymes and cellular signaling pathways, and for controlling transcription, and gene expression (Wold, 1981). Studying and understanding PTMs is essential for drug discovery, understanding disease states, and a general understanding of protein structure and function. PTMs are tightly controlled in cellular environments, and controlling the PTM state of proteins can be difficult, as reproducing the conditions that induce PTMs can be impossible, or purification of the protein in its natural modified condition can be prohibitive. The utility of NCAAs in studying PTMs has been realized for quite some time (Hortin and Boime, 1983), and in this vein, NCAAs provide a facile method to introduce PTMs into proteins using bacteria, simplifying protein expression and purification.

OTSs have been engineered to incorporate PTM or PTM analogs during translation, which have helped researchers understand the role a PTM has on protein activity or interactions. A number of different demonstrations have shown phosphoserine,

as well as analogs of phosphoserine and phosphotyrosine can be incorporated *in vivo* (Lee et al., 2013; Park et al., 2011; Rogerson et al., 2015; Xie et al., 2007)). Additional engineering of translational machinery, specifically engineering EF-Tu to recognize a tRNA charged with phosphoserine and improve phosphoserine utilization in the ribosome, substantially increased NCAA incorporation (Rogerson et al., 2015).

A number of other PTMs have been genetically encoded using NCAAs, including N^{ε}-acetyllysine (Neumann et al., 2008a) which was used for histone regulation (Struhl, 1998), and nitrotyrosine, an oxidation product of tyrosine, which may be a biomarker, but has been proposed as a PTM to alter function (Cooley et al., 2014; Neumann et al., 2008b; Ng et al., 2013; Siles et al., 2005). Additionally, complex PTMs, such as glycosylation or ubiquitination, can be added using crosslinking NCAAs (van Kasteren et al., 2007).

1.2.2.3 Fluorescence

The development of biologically relevant fluorescent trackers, largely based on GFP has revolutionized *in vivo* protein monitoring, allowing for protein localization analysis (Feilmeier et al., 2000), protein-protein interaction mapping (Cabantous et al., 2013; Kamiyama et al., 2016), and chemical environment indicators (Cannon and Remington, 2006; Hanson et al., 2004). GFP requires the encoding of a ~25 kDa protein onto the proteins being studied. To remove the large protein tag necessary for fluorescence, a series of fluorescent amino acids have been developed of *in vitro* and *in vivo* incorporation. Early attempts at genetic encoding of fluorescence used host translational machinery, but required chemical acylation of the tRNA (Pantoja et al., 2009). Other methods have used crosslinking amino acids to conjugate with fluorophores, this requires chemical modification after translation, but has the benefit of the NCAA

sidechain and the pre-fluorophore being nonfluorescent, but becoming fluorescent upon conjugation (Song et al., 2008). Since these, *in vivo* incorporation of several different fluorescent amino acids has been accomplished using engineered OTSs, and been used in protein localization studies and NCAA environmental determination (Charbon et al., 2011; Chatterjee et al., 2013; Summerer et al., 2006).

1.2.2.4 Additional NCAA properties

Other useful chemistries have been engineered into the genetic code using NCAAs, including the incorporation of heavy atom containing amino acids, which have been used to solve the phase problem of proteins structure elucidation (Lee et al., 2009b; Sakamoto et al., 2009), and metal chelation (Lee et al., 2009b; Luo et al., 2016), which has been used to stabilize protein motifs.

1.2.3 Conclusions

Noncanonical amino acids allow for the incorporation of many useful chemistries into proteins as they are synthesize *in vivo*. While many NCAAs accomplish the same tasks as tradition protein modification techniques, NCAAs provide the combined benefits of being site specific, small, and efficient. As protein design and engineering becomes easier, the number of chemistries that can be engineered into the proteome will likely continue to expand. It may be possible to incorporate glycosylation without downstream chemistry, or FAD or NADH analogs, or even nucleobase amino acids. The limits of translational machinery are just beginning to be understood, and it's possible even the current limitations can be overcome as computer modeling and additional OTS structural information data is collected from the many diverse life forms on earth.

1.3 Evolution of the Genetic Code

Once deciphered, it was quickly realized that the genetic code is largely shared among all life (Hinegardner and Engelberg, 1963; Woese, 1964), though examples of minor modifications exists, including eukaryotic organelles (Barrell et al., 1979), and some fungi and bacteria (Fitzpatrick et al., 2006; Santos and Tuite, 1995; Yamao et al., 1985). The universal genetic code is nonrandom, with spatially adjacent codons typically coding for the same or chemically similar amino acids. This has been hypothesized as an evolutionary mechanism to reduce the effects of translational errors or genomic mutations, and computational analysis has supported this hypothesis (Freeland and Hurst; Haig and Hurst, 1991). Despite the nonrandom nature of the code, computational modeling has identified further optimized table arrangements that could theoretically be achieved. Despite this, the code remains as it is, likely because mechanisms of codon reassignment make the evolutionary transition to these optimized codes prohibitory (Koonin and Novozhilov, 2009; Santos and Monteagudo, 2011).

While further optimized codes have been proposed, optimality is condition dependent, as can be seen through diverse phenotypes throughout biology. The conditions that existed and were responsible for shaping genetic code evolution are far different than those that exist today. Yet, the code suffices to maintain life in today's environments. Understanding the level of code optimization in its current form requires a set of parameters that objectively define optimality. Probably the most important feature is adoption of mutations. While mutation is the driving force behind evolution and required for changing organismal characteristics, most mutations are detrimental or lethal to the host. Adopting a code tolerant to mutation was likely a driving force in the arrangement of the code. A second feature to consider is translation errors, which is closely related to genomic mutations. Translation has evolved to be somewhat

ambiguous, with wobble pairing (Crick, 1966; Varani and McClain, 2000), and frameshift codons allowing alternate translations of the same mRNA sequence. These features appear counterintuitive or costly to the natural code yet they have persisted during evolution and are likely beneficial. Tolerance of these translational anomalies must play some role in the formation of the code.

1.3.1 Amino acid evolution

Understanding the evolution of the genetic code is simplified by considering the 20 standard proteinogenic amino acids, as well as two sets of base-pairing nucleotides (adenine-thymine, and guanine-cytosine), and that codons are composed of triplet nucleotide sequences. Though a number of exceptions to these rules exist, and are addressed in this work, they can be viewed as evidence for evolutionary processes instead of complications of code evolution analysis. Early discussion on the evolution of the genetic code largely concluded that the code was frozen, and incapable of further changes. The freezing of the code happened after settling on the current 20 amino acid system and corresponding assignment of codons. Three theories have been proposed to describe the evolutionary processes that led to the current set of amino acids.

1.3.1.1 Stereochemical theory

One of the earliest hypotheses used to explain the evolution of genetic code was the stereochemical theory. The stereochemical theory proposes that the triplet codons have an inherent affinity for the amino acids they code for in the codon table. This is a satisfying theory when considering evolution in a ‘primordial ooze’, where a preliminary propensity for binding would allow the evolution of the two parts, anticodon and amino acid, in close physical proximity to each other. Significant efforts were made to explore this possibility, even with some potentially encouraging results showing anticodons were

attracted to their associated amino acid (Pelc, 1965; Pelc and Welton, 1966; Saxinger et al., 1971). The discovery that DNA and RNA sequences could have structurally based specific binding to ligands (Ellington and Szostak, 1990, 1992) led credence to the idea that anticodons could have been early aptamers for amino acids, and some work showed aptamers against amino acids did enrich for sequences containing the anticodon coding for the amino acid (Knight and Landweber, 1998). Preliminary statistical analysis showed it unlikely that these occurred at random, through a more comprehensive analysis, considering features of the code as it exists, showed that these enrichments were likely just coincidence. The stereochemical theory, as proposed, has little evidence supporting it today.

1.3.1.2 Adaptive theory

The vast increases in computational power over the last few decades have allowed for analysis of the genetic code in comparison to other codes which could have evolved. Quantitative features are required to compare the fitness levels of hypothetical codes against the standard code, and a number of approaches have been taken to examine the fitness of the code. Typically, physiochemical similarity of amino acids is used as a parameter, although the exact definition of physiochemical properties is unclear. While phylogenetic analysis, looking at which amino acids replace generally conserved residues is a protein, has been used to determine which amino acids are physiochemically similar. The utility of this method for characterizing amino acid similarity is dubious at best, as these mutations and substitutions are a product of the code, and thus are a poor way to define physiochemical similarity. Other approaches, including hydrophobicity index, and polar requirement scale have been used to characterize amino acids, though a precise, objective description of amino acid relatively has yet to be accepted.

There is no doubt that the code is special. Compared to other randomized codes, the standard code has been estimated to have a one in a million odds of randomly settling in its current layout (Freeland and Hurst). This is largely because, despite the lack of a strict definition of physiochemical properties, generally similar amino acids share similar codons. Interestingly, the most important indicator of the physiochemical property of the coded amino acid is the central nucleotide of a codon, which has also been shown to be translated with the highest fidelity. Mistranslations of both the third and first nucleotides are more likely to mis-pair during translation (Kramer and Farabaugh, 2007; Parker, 1989). This reduces the number of 'crucial' nucleotides in the codon from three to one, and would favor amino acids of similar properties to flow 'down' the codon table.

1.3.1.3 Coevolutionary theory

Finally, the coevolutionary theory postulates that a prerequisite to amino acids adoption into the genetic code was the biosynthesis of the amino acid. It is probable that amino acids appeared in the genetic code in some temporal sequence, and did not all appear simultaneously. A number of amino acids are known to occur naturally under conditions that could have existed early in terrestrial history (Kobayashi et al.). While several amino acids could have been chemically synthesized in a prebiotic environment, many proteinogenic amino acids seen today likely required biosynthesis for existence at concentrations high enough for utility in biology. While it's difficult to imagine that both biosynthesis of an amino acid and its use in the genetic code would occur simultaneously there are a few possibilities that could describe the evolutionary path.

First, today there exist a number of nonproteinogenic amino acids that are biosynthesized and used as neurotransmitters (Misu et al., 2002), pigments (Roffler-Tarlov et al., 2013), pheromones (Yambe et al., 2006), and defense mechanisms (Yeung

et al., 2002). The possibility that amino acids perhaps served alternative functions in pre-code life, and were then coopted during evolution of the code is a clear possibility. In this instance, biosynthesis of an amino acid, (e.g. cysteine) would evolve, and free cysteine might be useful in some function in the proto-cell or cell (reducing agent). A second possibility would be the evolution of amino acid biosynthesis and the hijacking of a promiscuous aaRS would allow immediate incorporation into the code, and over time, differentiation between the two amino acids, and codons which code for each, could evolve. A third possibility postulates that new amino acids could be synthesized by modified, pre-charged tRNAs. This is seen extensively in nature (Sheppard et al., 2008), and some evidence suggests it as a mechanism during genetic code evolution (Francklyn, 2003; Roy et al., 2003).

1.3.2 Codon assignment evolution

Once the extant amino acids were established as part of the code, rearrangement of the codon table itself could be used to optimize the code through mutational tolerance and misincorporation during translation. It has been postulated that codon reassignment with a new or existing amino acid would result in global replacement of an amino acid throughout a host genome and would be expected to severely compromise fitness or likely be lethal to the host organism (Crick, 1968; Hinegardner and Engelberg, 1963). While the basic structure of the genetic code is shared among all life, examples of alternate codes provide evidence of previous evolution, and even the current transitioning of codon assignment. Three theories attempt to explain codon assignment and evolution.

1.3.2.1 Codon capture, or the disappearing codon

A prominent theory on codon reassignment is codon capture, which proposes that during evolution a codon can disappear entirely from a genome (Osawa and Jukes)

(**Figure 1-7**). Once the codon is extinct, the complementary tRNA is no longer necessary and may also be removed from the hosts genome. The codon can then reappear, along with the near simultaneous mutation of a tRNA, resulting in a complementary anticodon to the renewed codon, and the codon now codes for a different amino acid. This process is facilitated at small genome sizes, which makes extinction of a codon statistically more favorable. Evidence for codon capture includes the alternate assignment of UGA codons to signal tryptophan incorporation in the bacterial genus *Mycoplasma*, which have small genomes. The capture of UGA is largely facilitated by the existence of two alternative stop codons (UAG and UAA) which provide mutational destinations for any genomic UGA codons used for translational termination. In fact, *Mycoplasma* have no ORFs which terminate in UGA (Yamao et al., 1985). The UGA to UAA mutation requires a single, transitional mutation, which is a relatively favorable (Collins and Jukes, 1994). The result would be UGA to UAA, and would terminate protein synthesis in the same fashion as UGA codons, having little to no effect on translation or translated products. Once UGA was removed from the genome, a duplicated tryptophanyl-tRNA could acquire an anticodon mutation complementary to the UGA codon, which could still translate UGG tryptophan codons with wobble pairing. Any mutations which then introduced UGA codons to the genome would result in tryptophan incorporation, and evolution would occur with this redefined codon. Besides *Mycoplasma*, a number of other examples exist to provide evidence for the codon capture hypothesis, including mitochondrial genomes of animals and yeast (Inagaki et al., 1998; Macino et al., 1979).

1.3.2.2 Ambiguous intermediate hypothesis

Another proposed mechanism to explain genetic code evolution is the ambiguous intermediate hypothesis. This explanation proposes that a tRNA mutation would provide

an alternate translational solution to a codon, in addition to its natural translation (**Figure 1-8**). The codon would be read as either the traditional or the new amino acid each time it reached in the ribosome. Over time, the new tRNA could become the prominent translator, or the primary tRNA could be deactivated or removed from the genome, giving an alternative reading of the codon exclusively. Again, this example could occur with stop codons, which are not read by tRNAs, but with release factors. In this instance, an encountered stop codon could be read by the tRNA, or code for termination by a release factor, similar to NCAA incorporation methodologies. Evidence for the ambiguous intermediate hypothesis includes suppressor tRNAs. In *E. coli*, a number of UAG, UGA, and UAA complementary tRNAs are known to incorporate amino acids at canonical stop codons (Murgola, 1985). Additionally, in some *Candida*, CUG codons are read as leucine (the canonical translation) at low levels, but most often codes for serine, an alternative decoding of CUG (Suzuki et al., 1997). It's possible that *Candida* could evolve to a strict serine translation over time. An ambiguous intermediate could also be used to explain the conversion of UGA codons in *Mycoplasma*.

1.3.2.3 Genome streamlining

A third hypothesis in evolution of the arrangement of the genetic code is genome streamlining, which suggests there that genomes are under a pressure to reduce to a minimal necessary set of tools for survival (Giovannoni et al., 2005; Lynch, 2006; Massey and Garey, 2007). With the current code, reducing the number of degenerate codons and tRNAs is a simple, straightforward approach to genome streamlining. The clearest examples of genome streamlining include small genome bacteria and eukaryotic organelles including mitochondria and chloroplasts, which have greatly reduced the genetic code material in their genome.

1.3.3 Conclusions

The genetic code evolved into its current form and likely went through transitional codes that contained a reduced number of amino acids and varied codon assignments. Multiple mechanisms have been theorized for the evolution of the genetic code, but evidence for them is limited to snapshots provided by the extant variations to the genetic code, and actual code evolution has been difficult monitor, either *in situ* or in the laboratory. It is likely that all proposed theories described above played roles in evolution. The extent that each affected, and currently affects genetic code evolution is unknown. Exploration of these theories is largely limited to computer modeling, as recoding the genetic code is onerous. Though, advancements in genetic code expansion, as well as evolutionary approaches are beginning to open the door to thorough investigations of genetic code evolution.

1.4 Evolution and Expanded Codes

Nature has evolved at least two different expansions to the standard genetic code, with stop codon incorporation of selenocysteine and pyrrolysine. Selenocysteine incorporation has evolved with a complex regulatory pathway, establishing two unique conditions for the possible decoding of UGA codons: selenocysteine incorporation of termination. Selenocysteine is used throughout all domains of life, and has found uses in a number of biological functions. Alternatively no regulatory elements have been discovered for pyrrolysine, and so UAG codons are read ambiguously (Théobald-Dietrich et al., 2004). Despite this, pyrrolysine has a single known purpose, and UAG codons have not appeared throughout the genome of pyrrolysine encoding archaea or bacteria, in fact, most in-frame UAG codons in archaea are followed closely by UGA or UAA codons (Zhang et al., 2005). So, with two available translations of UAG, pyrrolysine

incorporating archaea appear to have eliminated UAG suppression necessity, if not actual termination.

Evolution with expanded or altered genetic codes using NCAAs has not been greatly explored, with most examples having been accomplished very recently. Experimental evaluation of the proposed evolutionary theories is the ideal tool to explore methods of evolution. Expanded genetic codes, and NCAAs in general, provide a toolkit for the exploration of genetic code evolution. Two general methodologies have been proposed to explore evolution using an expanded genetic code as defined by (Bacher et al., 2004): a top-down method, where an organism is required to evolve in order to survive in an altered or expanded amino acid environment, and the bottom-up method, using rational engineering of the genetic code.

1.4.1 Top-down genetic code evolution

Biology has evolved exclusively using the standard genetic code, generally limited to the 20 standard proteinogenic amino acids, with exceptions to reach 22. These amino acids are sufficient for the rapid adaption to a range of unfavorable conditions. Short and long term evolutionary experiments have found that bacteria quickly evolve when they encounter non-ideal growth conditions. Examples include the evolution of ethanol tolerance (Goodarzi et al., 2010), butanol tolerance (Smith and Liao, 2011), thermal tolerance (Rudolph et al., 2010), and antibiotic resistance (Petrosino et al., 1998). When faced with challenges nature evolves, and thus evolutionary pressure with an altered amino acid pool is a reasonable approach to study the adoption of altered genetic codes.

1.4.1.1 Evolution with an altered set of amino acids

While the genetic code was considered frozen, an early test of genetic code flexibility demonstrated that the eubacteria *Bacillus subtilis* rapidly adapted to an alternate genetic code (Wong, 1983). By simply excluding tryptophan and including of 4-fluoro-tryptophan (4FW) in growth media, and providing a means to increase the rate of genomic mutation, cells coopted the new amino acid into their genome. Cells not only tolerated the new amino acid, but became dependent on 4FW for growth and survival, eventually preferring 4FW to tryptophan. To some extent, the adoption of 4FW does not seem surprising, as 4FW and tryptophan vary only by the replacement of a hydrogen with a tryptophan, with minimal effect on structure, shape, or size. Yet, the polarity is altered significantly, and 4FW is insufficient for sustaining growth of wild-type *B. subtilis*, and reduces some enzymatic activity by greater than 90% (Pratt and Ho, 1975). The 4FW dependent *B. subtilis* has recently been genome sequenced and characterized, in an attempt to understand mechanisms for the adoption of augmented genetic codes. Surprisingly, few mutations were needed to replace genomic tryptophan with 4FW (Yu et al., 2014). Hence, the 14600 predicted tryptophan incorporation sites were generally amenable to 4FW replacement, demonstrating the malleability of the genetic code in a natural organism; the entire code rapidly adapted to a new amino acid environment.

A similar undertaking using *E. coli* seems to demonstrate that the code is not . Rounds of selection on 4FW resulted in a strain that could largely tolerate 4FW throughout the genome, but still grew significantly better with tryptophan. Despite a lack of complete acceptance, the strain still began adopting to the new code, with genomic modifications in tryptophan transporters and the tryptophanyl-synthetase contained mutations presumably to build a tolerance for the new amino acid. The resistance to adopt the new amino acid in *E. coli* was surprising, given the rapid acceptance in the

genome of *B. subtilis*. Evolutionary adoption of new amino acids is perhaps dependent on the overall growth environment (Bacher et al., 2004), selection conditions, and the organism itself.

A third attempt at adopting new amino acids into the genome by evolution using the bacteriophage Qβ showed that the phage rapidly adapted to 6-fluorotryptophan in place of tryptophan (Bacher et al., 2003a). A number of mutations that were found did not alter the ratio of tryptophan codons throughout the genome, but were adaptations to the replace of tryptophan with the analog 6-FW throughout the genome. Mapped mutations were largely involved the metabolism of tryptophan. Characterization of the proteome revealed 6FW incorporation, along with tryptophan, throughout the genome. These results again supported the hypothesis of ambiguously read intermediate codons during evolution.

1.4.2 Conclusions

While only a limited data exists, the experimental evidence discussed above supports the ambiguous intermediate hypothesis of codon evolution. *B. subtilis* and the bacteriophage Qβ rapidly adapted to new amino acids, altering the genome to tolerate these NCAAs were possible over relatively short evolutionary timeframes. These examples required no engineering of translational machinery, but instead demonstrate the ability of nature to adapt to new genetic codes without any rational engineering.

1.4.2 Bottom-Up genetic code design and evolution

Biotechnological advances in genetic engineering have led to site specific incorporation of unique chemistries *in vivo* using NCAAs. The OTSs used for NCAA incorporation are engineered for minimal interference with the host genome and are specific for the desired NCAA. While OTSs allow the incorporation of additional amino

acids, usually at amber stop codons, further attempts to engineer the genetic code involve the removal of release factor mediated termination, as well as the removal of entire codons throughout a genome. These efforts mirror the theory of the codon capture hypothesis of evolution.

1.4.2.1 OTS engineering

The simplest form of bottom-up genetic code evolution is engineering OTSs. The recoding of UAG codons for NCAA incorporation has been discussed in detail above. But, briefly, as the top-down approaches represent ambiguous intermediate hypothesis, OTS incorporation of NCAAs represent the codon capture hypothesis, using a codon unassigned for translation (UAG), and repurposing it for NCAA incorporation. While technically still ambiguous, as the codon can be read as 'stop' or NCAA, further genomic engineering made the full transition to codon capture possible.

1.4.2.2 Removing codon competition at UAG

In *E. coli*, the competition between release factor one and suppressor tRNAs results in reduced translational efficiency of UAG codons by complementary tRNAs. The simplest solution is deletion of *prfA*, the gene encoding release factor one. Release factor one is responsible for recognition of UAG codons in the ribosome, and completing translational termination. Temperature sensitive versions of *prfA* demonstrated that stop codon suppression could be substantially increased by release factor one inactivation (Rydén and Isaksson, 1984), but failed attempts to fully remove *prfA* from the genome suggested that *prfA* was essential in wild-type *E.* (Baba et al., 2006; Gerdes et al., 2003). Since these initial attempts, several elegant solutions have been found to circumvent this essentiality of *prfA*, allowing genomic knockouts and near cognate translation of UAG codons in the ribosome.

The first successful *prfA* knockout required the encoding of seven essential genes which terminate with amber codons in genome to be given the alternative termination codon TAA on a bacterial artificial chromosome (Mukai et al., 2010). While the remainder of the genomic TAG codons did not require recoding or supplementation, the strain required UAG suppression to survive. It is hypothesized that secondary stop codons, often found within 3 to 10 triplets of genomic TAG codons terminate most translating genes, allowing them to retain protein function. Subsequent investigation showed that only one of the seven BAC supplemented genes was necessary for cellular viability (Mukai et al., 2010). Removal of *prfA* allowed for greatly improved NCAA incorporation at UAG codons (**Figure 1-6**).

Later attempts to knockout *prfA* used a particular feature of release factor two (*prfB*) (Johnson et al., 2011). In *E. coli*, release factor one is responsible for recognition and termination and amber (UAG) and ochre (UAA) codons, while release factor two recognizes and terminates at ochre (UAA) and opal (UGA) stop codons. Using a *E. coli* strain with a minimalized genome (Pósfai et al., 2006), and "fixing" *prfB*, the *prfA* gene could be excised successfully. Fixing of *prfB* required only the removal of an in-frame UAG codon, used as an autoregulatory element, and introducing a T246A mutation, which together increased *prfB* expression and allowing enhanced recognition of ochre codon and likely prevented amber codon stalling in the ribosome. This demonstration was largely recapitulated in full genome of *E. coli* by simply using a B strain *E. coli*, which code for A246 naturally (Johnson et al., 2012). Even though these strains were capable of survival and reproduction, the fitness of these release factor one knockouts was severely compromised.

More recently, radical engineering of an *E. coli* BL21 genome allowed for the deletion of release factor one with the reassignment of 95 genomic amber codons to

alternative stop codons, while leaving the remaining 178 TAG codons in the genome. This allowed for the simple knockout of *prfB* with minimal growth effects (Mukai et al., 2015).

1.4.2.3 Engineering genomes for unassigned codons

While near cognate decoding of the amber codon can be accomplished with *prfA* knockout strains, the remaining genomic UAG codons must be suppressed with NCAAs, reducing the site specificity and possibly impacting general fitness of the organism. The solution to these issues is the recoding of an entire genome to remove all UAG codons. This was largely made possible by the degeneracy of the codon table, the presence of alternative stop codons, and multiplexed automated genome engineering (MAGE), which streamlines the editing of multiple targets across a genome (Wang et al., 2009). By systematically removing UAG codons from the genome, *prfA* could be removed without the requirement for an externally supplemented OTS or UAG suppressor tRNAs (Lajoie et al., 2013). This strain (C321.ΔA) is largely unaffected by encoded OTSs or NCAA incorporation.

The generation of the genetically recoded genome of *E. coli* C321.Δ has demonstrated that redundant codons can be removed from the genome with minimal effect. While 64 codons are available in the codon table, they code only for 20 standard amino acids. The degeneracy of the codon table suggests that other codons could be recoded to synonymous codons with little effect on the host. In theory, the total number of codons could be reduced to as few as 21, 20 coding for the standard amino acids (one for each amino acid), and one signaling termination, although this is complicated with wobble pairing. This would allow multiple NCAAs to be incorporated at unique sites throughout the genome, providing a platform for in the incorporation of interacting

NCAAs (i.e. alkyl/alkyne NCAAs to perform click chemistry *in vitro*) while being bioorthogonal to all natural host chemistries. While a number of methods have been demonstrated to incorporate multiple unique NCAAs *in vivo*, they often require inefficient methodologies including quadruplet codon and rare codon suppression (Neumann et al., 2010). As the number of available NCAAs increases, the need for more codons for NCAA incorporation increases. More recently, progress has been made towards entirely removing seven rare codons from the genome of *E. coli*, opening translational space for the site specific incorporation of multiple amino acids in a single host (Ostrov et al., 2016).

1.4.2.4 Evolutionary studies with unassigned codons

The first real demonstration of genetic code evolution with an unassigned codon use a *prfA* knockout strain of *E. coli* and the T7 bacteriophage phage. Despite the removal of *prfA* from the genome, and the requirement for UAG suppression for survival, incorporated OTSs lose activity in hosts relatively quickly. A simply method to ensure an OTS remains functional over evolutionary timepoints is to 'refresh' the OTS. While most cellular life is not amenable to genetic 'refreshing,' phage function with the genetic code of the host, and switch to new hosts often. By harboring the OTS in the phage host, (*E. coli*), the phage has a continuously active OTS after each transfection, and thus always has a functioning OTS. Using this, T7 bacteriophage utilized an expanded genetic code to reach new fitness peaks (Hammerling et al., 2014). This was conducted with the relatively inert amino acid 3-iodo-tyrosine, but was still found to provide distinct advantages that was not be accomplished with the standard amino acids.

In another recent study using C321.Δ, it was demonstrated that expanded codes can provide new methods to rifampicin resistance, a phenotypic trait which is

representative of a potential beneficial trait in nature. By evolving *E. coli* with an expanded genetic code, using the UAG codon as a cognate codon, the code landscape was altered, which provided a new mutational path rifampicin resistance (Hammerling et al., 2016).

The full implications of expanded genetic codes, both of biological function and exploration of genetic code evolution, have yet to be realized. Few examples exist, all of them reviewed above, that use NCAAs to explore evolution, and more information is needed to fully understand the mechanisms nature uses to evolve new codes, or how it adopts new codes once they exist.

1.4.3 Center-out evolution

A significant portion of this dissertation (**Chapter 3**, **Chapter 4**, and **Chapter 5**) focuses on an intermediate between top-down and bottom-up evolutionary approaches to expanded genetic codes. The approach described here, which I've labeled as 'center-out,' uses an engineered genetic element to enforce and OTS into the genome of a host cell indefinitely. Engineering the OTS and tool to enforce the OTS falls squarely in the 'bottom-up' camp, yet these tools provide a means to explore evolution from the 'top-down' by evolving the organism with its newly expanded genetic code, providing a means to monitor evolutionary processes as they occur.

1.5 FIGURES

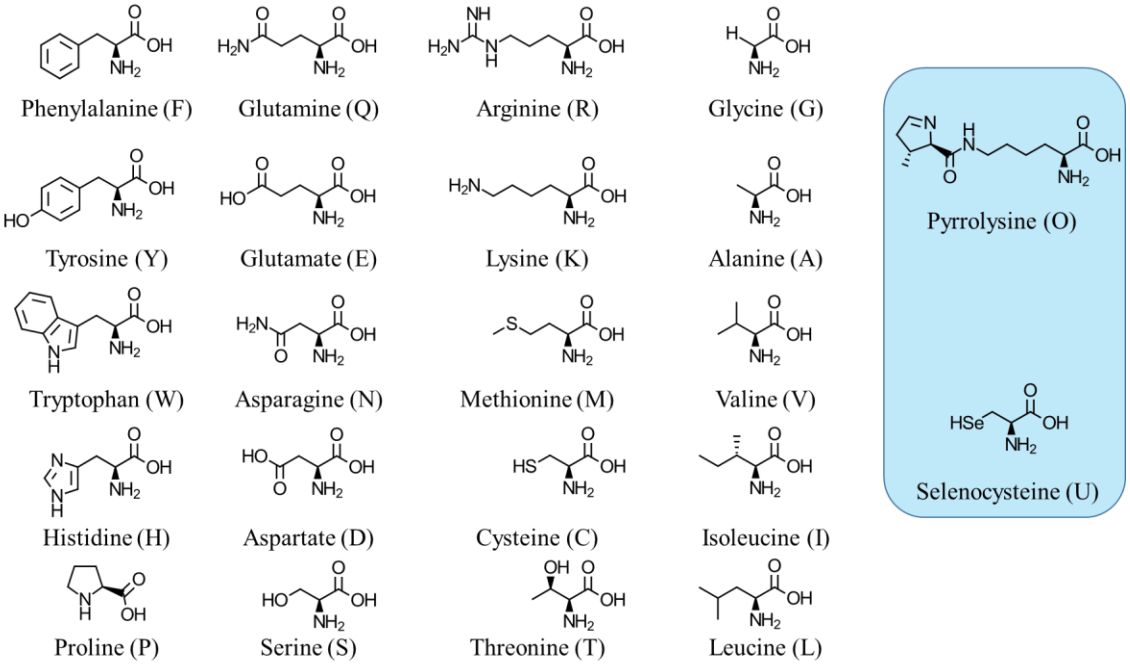

Figure 1-1: Canonical proteinogenic amino acids

The 20 canonical proteinogenic amino acids shared among all life, as well as the two rare proteinogenic amino acids selenocysteine and pyrrolysine, highlighted in blue.

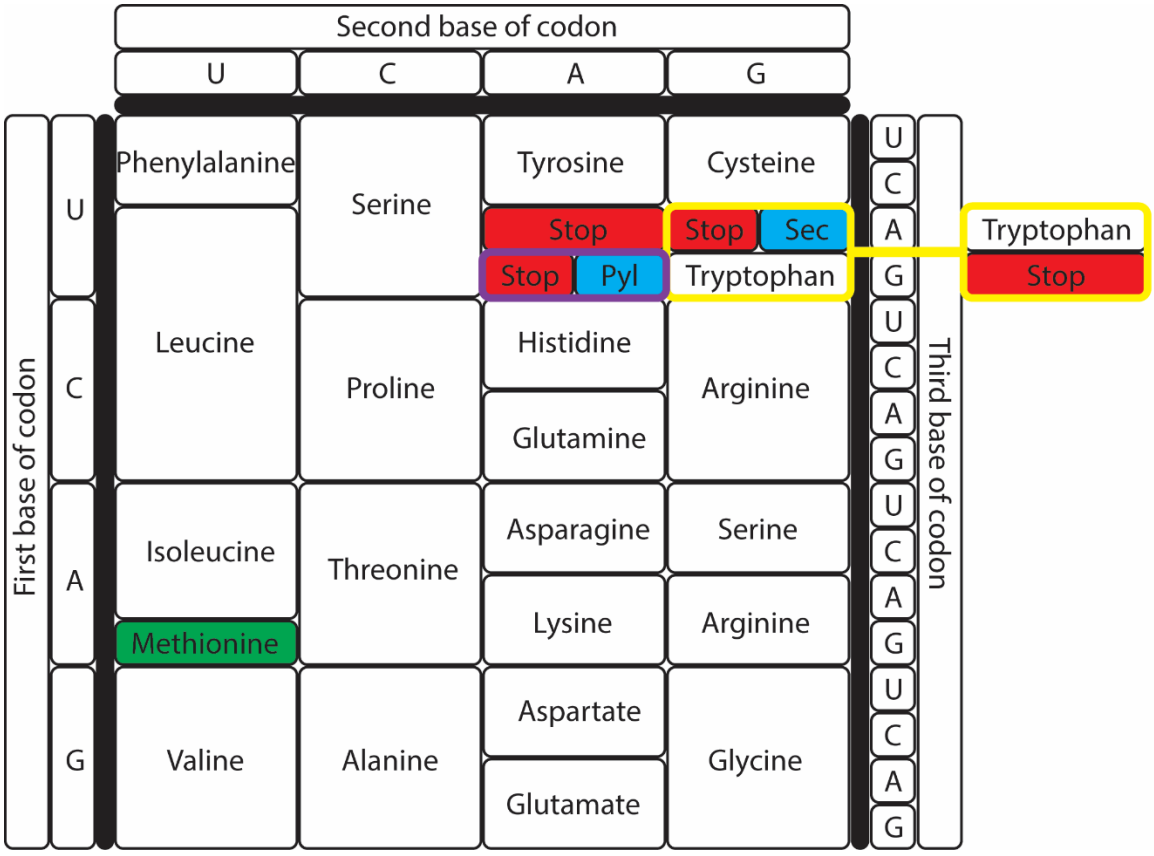

Figure 1-2: The standard codon table with variations

The standard codon table, with start codons marked in green, and stop codons marked in red. Pyrrolysine (Pyl) and selenocysteine (Sec) codons are marked in blue, and are incorporated at stop codons. *Mycoplasma* have inverted codon assignments in the yellow box, with tryptophan and UGA termination codon swapped.

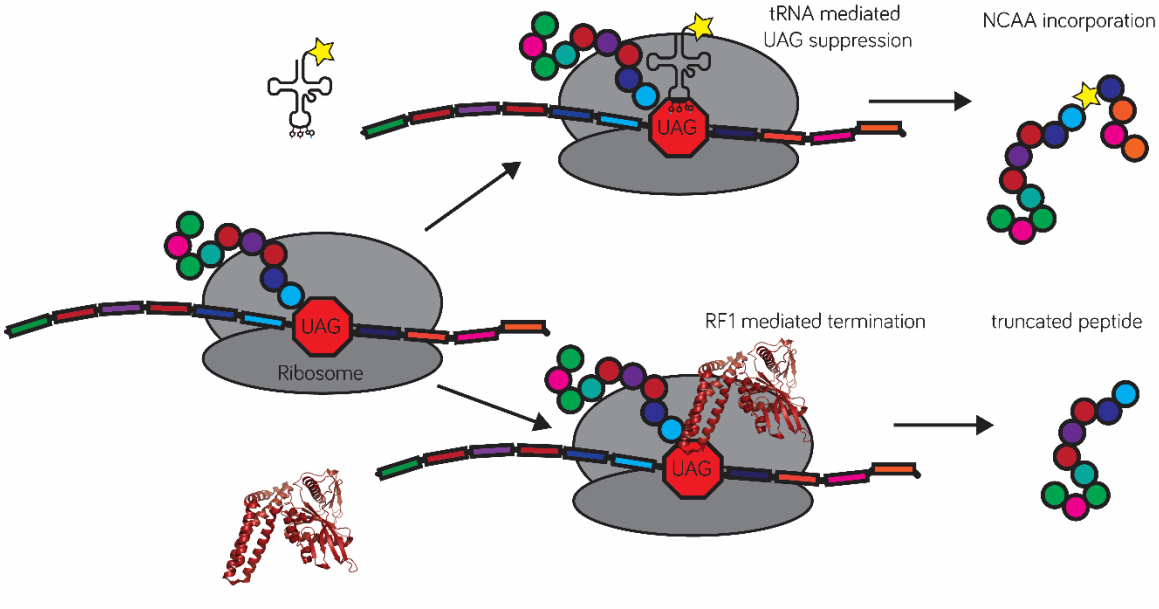

Figure 1-3: NCAA incorporation by the ribosome

The simplest method to expand the genetic code *in vivo* involves the expressions of heterologous aaRS/tRNA pairs in a host cell. The NCAA is charged to a tRNA that complements the amber (UAG) stop codon by the engineered aaRS (**Figure 1-4**). When the ribosome encounters the UAG codon, two different translational events can occur: a) (top pathway) the NCAA charged tRNA is used to translate UAG and the NCAA is inserted into the peptide and translation continues, or b) the traditional reading of the UAG codon (bottom pathway), release factor 1 (RF1) terminates translation, and the truncated peptide is release from the ribosome.

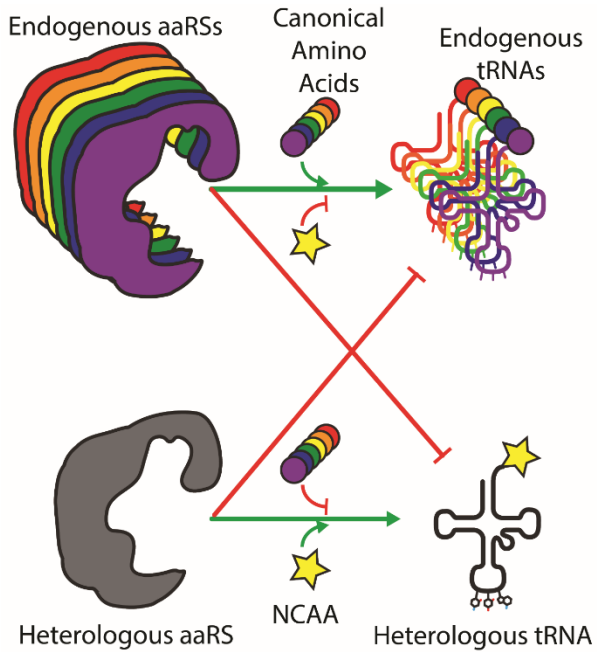

Figure 1-4: Engineering orthogonal translation systems

A heterologous aaRS expressed in a host cell is engineered for specificity to its corresponding tRNA, as well as the desired NCAA. The aaRS should not interact with host tRNAs. Endogenous aaRSs do not charge the heterologous tRNA, or the supplemented NCAA. A lack of interaction between the translation systems (heterologous translational machinery and endogenous translational machinery), are described as orthogonal.

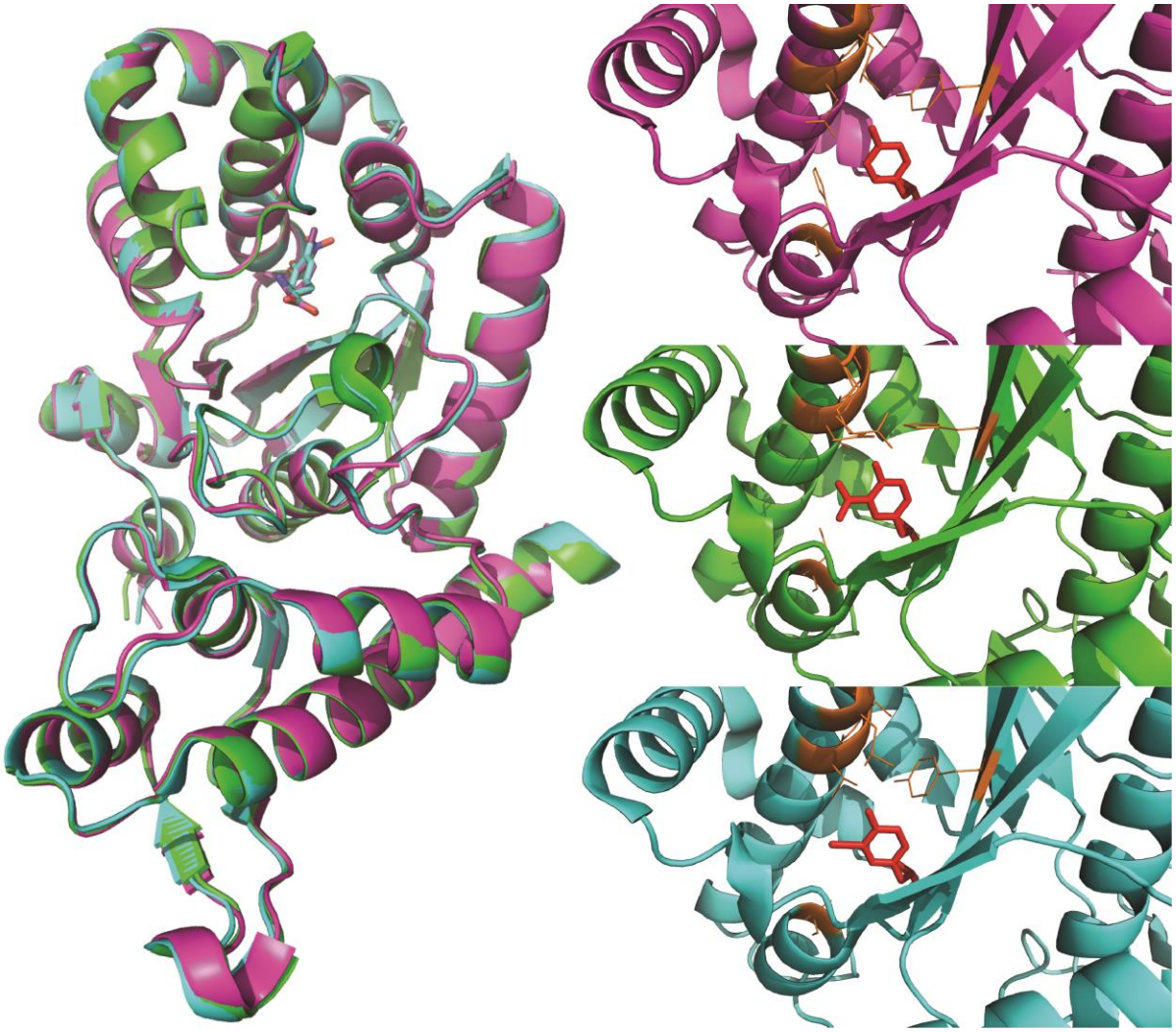

Figure 1-5: Structure of *M. jannaschii* tyrosyl-synthetases

The crystal structure of the *M. jannaschii* tyrosyl-tRNA synthetase (pink), aligned with the same synthetase engineered for specific recognition of the NCAAs 3-nitro-L-tyrosine (green), and 3-iodo-L-tyrosine (teal). The amino acid binding pockets of the synthetases (right images) are specific to the NCAA, and exclude the natural amino acid tyrosine.

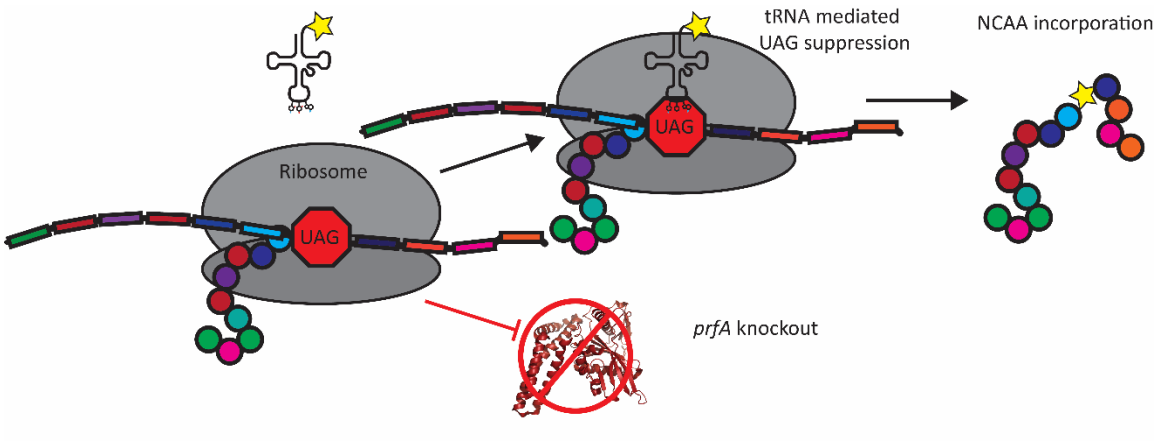

Figure 1-6: Release factor 1 knockouts improve NCAA incorporation efficiency

Over the last 10 years, methods to knockout *prfA* from the *E. coli* genome have greatly increase NCAA incorporation efficiency. With *prfA* removed, UAG codons are only recognized by OTS suppressor tRNAs. This results in ~100% incorporation efficiency of NCAAs at UAG codons.

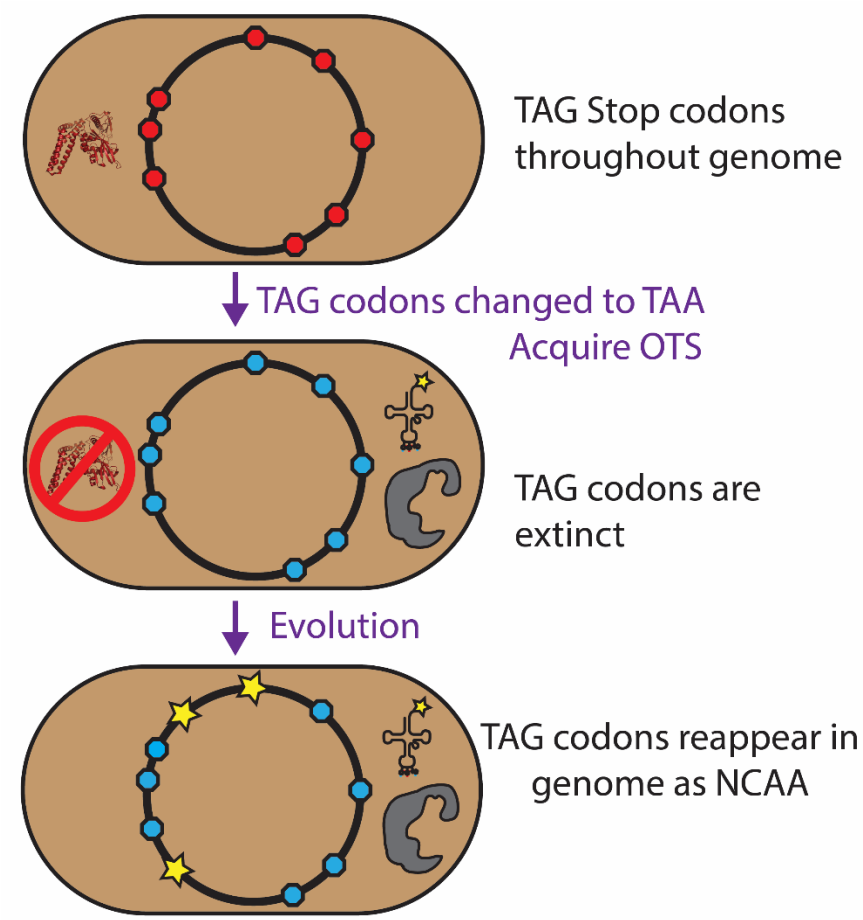

Figure 1-7: Codon capture of amber codons

Codon capture has been theorized as a method for genetic code evolution. Codon capture has been synthetically accomplished with genetically recoded organisms, specifically C321.Δ. During codon capture, a codon, the TAG stop codon in this example, becomes removed from the genome during evolution or through genome engineering. Its translational machinery is no longer necessary and is removed from the genome, release factor one in this example (a tRNA could be removed if a cognate codon went extinct). A new translator for the codon, here an OTS, appears and evolution TAG codons appear in the genome, resulting the evolution of the genetic code.

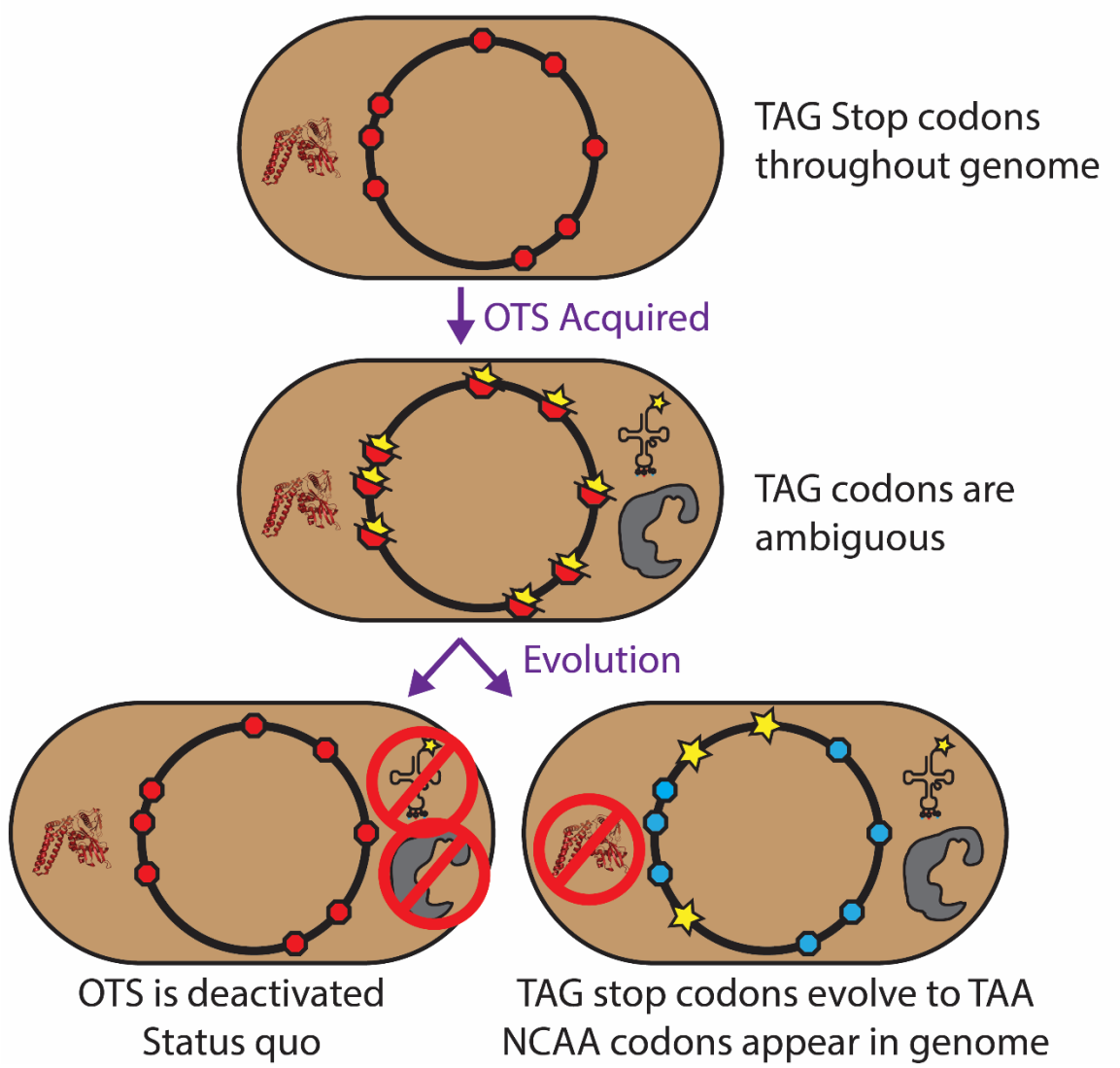

Figure 1-8: Ambiguous intermediate transition of amber codons

Ambiguous intermediate codons have been proposed as a mechanism in genetic code evolution. In this example a cell acquires an OTS, and UAG codons are translated ambiguously as termination codons, recognized by release factor one, or as cognate codons for NCAA incorporation. Over evolutionary time periods, either the OTS or release factor one are removed from the genome, and UAG codons are exclusively cognate codons.

CHAPTER 2: Structure-based noncanonical amino acid design to covalently crosslink an antibody–antigen complex[1]

Engineering antibodies to utilize non-canonical amino acids (NCAAs) should greatly expand the utility of an already important biological reagent. In particular, introducing crosslinking reagents into antibody complementarity determining regions (CDRs) should provide a means to covalently crosslink residues at the antibody–antigen interface. Unfortunately, finding the optimum position for crosslinking two proteins is often a matter of iterative guessing, even when the interface is known in atomic detail. Computer-aided antibody design can potentially restrict the number of variants that must be explored in order to identify successful crosslinking sites. We have therefore used Rosetta to guide the introduction of an oxidizable crosslinking NCAA, L-3,4-dihydroxyphenylalanine (L-DOPA), into the CDRs of the anti-protective antigen scFv antibody M18, and have measured crosslinking to its cognate antigen, domain 4 of the anthrax protective antigen. Computed crosslinking distance, solvent accessibility, and interface energetics were three factors considered that could impact the efficiency of DOPA-mediated crosslinking. In the end, 10 variants were synthesized, and crosslinking efficiencies were generally 10% or higher, with the best variant crosslinking to 52% of the available antigen. The results suggest that computational analysis can be used in a pipeline for engineering crosslinking antibodies. the rules learned from DOPA crosslinking of antibodies may also be generalizable to the formation of other crosslinked interfaces and complexes.

[1] Xu, J., Tack, D., Hughes, R.A., Ellington, A.D., Gray, J.J. (2014). Structure-based non-canonical amino acid design to covalently crosslink an antibody-antigen complex. J Struct Biol. 185(2): 215-222. I, Drew Tack, was responsible for experimental design and execution in this chapter, I also contributed to writing the manuscript.

2.1 INTRODUCTION

2.1.1 Background and Rationale

2.1.1.1 Computational improvement of antibodies

Antibodies are key components of the immune system with broad diversity to recognize a variety of antigens. Antibody-based therapeutic, diagnostic, and industrial applications frequently require antibodies having high stability and strong binding affinity. With the development of computational techniques and a number of successful experiences in protein modeling and design (Lippow and Tidor, 2007; Mandell and Kortemme, 2009), computational antibody design has begun to play an important role in predicting improvements to antibody function. Computational design of antibodies has been used to enhance binding affinity (Barderas et al., 2008; Clark et al., 2006; Lippow et al., 2007), to improve stability by improvement of thermal tolerance or aggregation resistance (Chennamsetty et al., 2009; Miklos et al., 2012), to alter binding specificity (Farady et al., 2009), and others (A. Caravella et al., 2010; Kuroda et al., 2012; Midelfort et al., 2004; Pantazes and Maranas, 2010).

To date, most computational design methods have focused on manipulating the twenty natural proteinogenic amino acids to modify molecular forces such as electrostatics (Lippow et al., 2007), hydrophobic interactions (Chennamsetty et al., 2009), hydrogen bonds (Clark et al., 2006), and salt bridges (Miklos et al., 2012). However, recent advances in engineering the translation system have now allowed for the site-specific insertion of non-canonical amino acids (NCAAs) with a variety of functionalities into proteins with good efficiency (Wang et al., 2006b; Xie and Schultz, 2006). Such NCAAs can be used to improve the stability and pharmacokinetics of therapeutic proteins (Cho et al., 2011), to augment binding (Liu et al., 2009), and to provide a myriad of

chemical handles to study protein structure and function (Jones et al., 2009; Tsao et al., 2006; Zhang et al., 2002).

2.1.1.2 The Potential of NCAAs in antibody design

The generation of protein-protein crosslinks by inserting NCAAs into proteins could prove useful for a variety of applications. To this end, a number of crosslinking-capable NCAAs have been incorporated into proteins in a site-specific manner utilizing an array of functionalized amino acids. These crosslinking functionalities include photo-crosslinkable aryl-azides (Chin et al., 2002a), benzophenones (Chin et al., 2002b) and diazirines (Ai et al., 2011) as well as the oxidizable crosslinker, L-DOPA (Alfonta et al., 2003). While any of the crosslinkers might benefit from a quantitative placement methodology, we chose L-DOPA because the periodate induced oxidation would allow for more control over the crosslinking conditions compared to the photo-inducible crosslinkers which could spuriously crosslink during sample handling (Chin et al., 2002b). In addition, the nucleophile-driven cross-linking mechanism of L-DOPA has been extensively characterized with every available proteinaceous nucleophile (Liu et al., 2006).

2.1.1.3 Utility of L-DOPA crosslinking

L-DOPA has previously been used to successfully crosslink the monomeric domains of a dimeric sortase A for structural studies (Umeda et al., 2009), to enhance the affinity of low-affinity peptide probes for a kinase SH3 bioassay (Umeda et al., 2010), and to site-specifically label proteins with polysaccharides (Ayyadurai et al., 2011). While previously reported use of L-DOPA as a site-specific crosslinker has yielded examples of effective crosslinking (as shown on SDS-PAGE gels or Western blots), the quantization of the efficiencies of these crosslinking events have not been reported

(Burdine et al., 2004; Umeda et al., 2009, 2010). These previous reports indicated that it is possible to place L-DOPA by intuition, but the placement parameters as they relate to crosslinking efficiency are unknown (Umeda et al., 2009, 2010).

L-DOPA crosslinking to map the antibody-antigen interface

In this paper, we explore using the Rosetta suite of computational protein design tools to predict the site-specific incorporation of L-DOPA into an antibody-antigen complex. A better understanding of where and how to insert crosslinking moieties into an antibody CDRs could lead to the development of tools for validating antibody-antigen structural models and reagents capable of binding analytes with extremely high affinities and specificities. As a proof-of principle demonstration, we chose a complex with a known structure, the anti-anthrax antibody M18 bound to anthrax protective antigen (PA) (Leysath et al., 2009). PA is a component of the tripartite toxin secreted by *Bacillus anthracis*. It binds to cellular receptors and assists host cellular targeting and transport of the lethal factor (LF) and edema factor (EF) into cytoplasm (Moayeri and Leppla, 2004; Young and Collier, 2007). M18 is an affinity-improved neutralizing antibody (Leysath et al., 2009) engineered from monoclonal antibody 14B7 (Harvey et al., 2004; Little et al., 1988), which binds to the fourth domain of PA (PAD4) and effectively blocks PA binding to cellular receptors such as CMG2 to mitigate anthrax toxicity. The success of the computational predictions for the L-DOPA mediated crosslinking of M18-PAD4 is not only useful for the design of better antibody therapeutics, but also helpful to improve antibody-based biosensors for anthrax detection (Kim and Yoon, 2010) and beneficial to the development of better antibody epitope mapping techniques via chemical crosslinking and mass spectrometry (Pimenova et al., 2008).

2.2 APPROACH

We wished to test whether we could use computational design for the precise placement of crosslinks between antibodies and antigens. Our general approach was to use simple criteria and the Rosetta suite of software (Kaufmann et al., 2010; Leaver-Fay et al., 2011) to identify where we might place the NCAA L-DOPA, which is activated when oxidized by periodate to form the ortho-quinone. (Burdine et al., 2004; Liu et al., 2006) The ortho-quinone is susceptible to nucleophilic attack, and can crosslink to a variety of nucleophilic, proteinogenic amino acids, in particular the histidine imidazole and the ϵ-amine of lysine. Thus, we first used Rosetta to build structural models of the antibody-antigen complex with candidate L-DOPA mutations, using energy parameters and rotamer libraries recently developed for NCAAs (Renfrew et al., 2012). For each L-DOPA position on the antibody, we evaluated various biophysical measurements that might affect crosslinking to an adjacent histidine or lysine on the bound antigen. The incorporation of L-DOPA into proteins was accomplished using an evolved tRNA synthetase from *Methanacoccus jannaschii* that when expressed in *E. coli* can specifically charge a suppressor tRNA with L-DOPA (Alfonta et al., 2003) . In the following sections we detail each of these steps.

2.2.1 Biophysical Criteria for Rational Design

2.2.1.1 Initial Choice of Residues for L-DOPA Substitution

The X-ray structure of the M18-PAD4 complex, solved at 3.8 Å resolution (PDB ID 3ETB), reveals a binding interface of ~1700 $Å^2$, with roughly equal contributions from the M18 light and heavy chains. All M18 antibody CDR loops except L3 contact the antigen, and CDR loops H3 and L2 appear to have the strongest interaction with PAD4 (Leysath et al., 2009). Histidine and lysine residues are the only nucleophiles present on

the antigen surface, however only lysine residues are near the epitope (**Figure 2-1**). To identify candidate positions, we selected all pairs of antigen nucleophilic residues and antibody residues that have atoms within 10 Å. Thus, we considered three nucleophilic lysines, PAD4 residues K684, K679 and K653, as crosslinking targets for sixteen possible L-DOPA locations on the antibody (thirteen on the heavy chain and three on the light chain).

Several biophysical parameters can be suggested as criteria for the evaluation and prediction of sites for L-DOPA insertion. First, the distance between the L-DOPA mutant and the reactive antigen nucleophilic residue should be close enough for them to react. Second, L-DOPA should be introduced to the antibody position where it is exposed to the solvent, so that periodate anions can access the reaction pair and trigger the reaction. Third, the original binding of M18/PAD4 should not be affected by the introduction of L-DOPA, meaning the interface must be compatible with it, or the change of interface energy should be moderate. Each of these criteria will be examined in turn as the basis for design.

2.2.1.2 Crosslinking Distance between L-DOPA and Nucleophile

The most obvious structural criterion for L-DOPA crosslinking is that the residue must be close enough to a nucleophile to crosslink. To assess this criterion, we took those L-DOPA atoms that were already within 10 Å of a lysine, as measured from C_β to C_β, and further calculated the distance between C2, C5, and C6 on the ring and the lysine N_ϵ atom. In preliminary calculations, Rosetta tended to extend both L-DOPA and lysine residues into solution, and thus the calculated distances would exclude the possibility of crosslinking. To bring the reactive atoms closer, we optimized side-chain conformations using an energy function augmented with a constraint potential (see Methods). The

crosslinking distance thus obtained, dXL, varied from 2.9 to 9.8 Å (**Table 2-1**). For comparison, we also analyzed one L-DOPA substitution at position far from the interface (L_D17, dXL = 32.1 Å).

2.2.1.3 Solvent Accessibility of L-DOPA

A second hypothesis is that a good crosslinking site needs access to the solvent so that the L-DOPA can be easily oxidized by periodate anions. In the presence of periodate, the hydroxyl groups on the phenolic ring of the encoded L-DOPA on M18 are oxidized to form the quinone intermediate, which activates the ring for potential nucleophilic attack by a neighboring lysine nucleophile on PAD4. We chose two methods to quantify the degree of solvent exposure. First, a simple metric was to count of the number of neighboring antibody and antigen atoms within 3 Å of the L-DOPA side chain. Mutants predicted to have fewer neighboring atoms were expected to have better chances for effective crosslinking. Second, the solvent-accessible surface area was calculated using a 1.4 Å probe. Both calculations were performed on the structure of antibody-antigen complex in which the L-DOPA and neighboring residue side chains were optimized by packing from a rotamer library using standard Rosetta "fixbb" protocol to find an energetically favorable conformation (see Computational Methods). We ultimately sought to test whether there was a minimal solvent accessibility or whether crosslinking would improve as solvent accessibility increased (**Table 2-1**).

2.2.1.4 Mutant Interface Compatibility

While it is difficult to computationally determine the impact of L-DOPA substitution on antigen binding, we tested whether Rosetta's estimated energetics of the substituted interface would be helpful. Two energy measures were tested. First, we hypothesized that a successful crosslink would require that the introduced L-DOPA

would not disrupt the favorable energy of interaction between the antigen and antibody. We therefore calculated the interface score, Isc, a measure used in docking calculations to approximate a binding energy (see Methods) (Chaudhury and Gray, 2008; Chaudhury et al., 2011). Data from docking 116 complexes (Chaudhury et al., 2011) indicate that most protein-protein complexes have negative interface scores in the range of -4 to -12 Rosetta Energy Units (REU), and comparisons to experimental alanine mutants suggest that changes in interface scores above +1 REU are likely to characterize mutations that will significantly reduce binding affinity (Kortemme and Baker, 2002). One complication is that NCAAs necessitate use of a Rosetta force field variation. In the case of the M18-PAD4 complex, the predicted interface score was -21.6 with the Rosetta's (standard) score12 and +2.7 with the NCAA score function, mm_std. The mm_std interface scores of the L-DOPA substituted variants ranged from -1.4 to 4.8 REU, with most interface scores being slightly positive, indicating that the L-DOPA mutated M18-PAD4 complex may be slightly disfavored.

A second energetic hypothesis is that an activated conformation with proximity between L-DOPA and the target lysine should be energetically accessible. Thus, we evaluated the interface score for the structure we generated with the crosslinking constraint that brought the reacting atoms together (I_sc^constr; see Methods). **Table 2-1** shows that these values vary from -0.6 to over 150 kcal/mol, showing that some cases exhibit significant energetic perturbations (e.g. H_D56) and even atomic clashes (H_R99 or H_D54) when the L-DOPA/Lys pair is constrained.

2.2.2 Crosslinking Studies

As the generation of a large series of point mutants for crosslinking studies is impractical, we selectively tested candidate mutations based on their predicted placement

within the protein-protein interface. We anticipated that the distance between the L-DOPA and a lysine nucleophile was likely to be the most critical parameter for successful crosslinking. Therefore, we made the closest eight substitutions within M18 (**Table 2-1**) (crosslinking distance of between 2.9 and 8.3Å), excluding only H_R99 because the high constrained interface energy (I_sc^constr) suggested that it would bind antigen poorly. As controls for the crosslinking distance, two additional mutants were made by placing L-DOPA at two distance-extended locations at H_I51 (9.8 Å, solvent excluded) and L_D17 (32.1 Å, solvent exposed) within the antibody structure.

The recombinant M18 antibodies were expressed in the periplasm of *E. coli* as a 28.9kDa single-chain variable fragment (scFv) with a N-terminal FLAG epitope tag and a C-terminal His tag to enable purification via immobilized metal affinity chromatography (IMAC). The FLAG tag allows ready detection of the antibody via Western blot analysis. Similarly, the PAD4 antigen was expressed as a 17.5kDa protein with a C-terminal His tag to enable purification of the antigen via IMAC. L-DOPA was introduced site-specifically into the chosen sites in the CDRs of M18 by changing the codon for the naturally occurring amino acid at these positions to an Amber (TAG) stop codon. The M18 amber variants were then co-expressed with an orthogonal tRNA suppressor and tRNA synthetase pair that had previously been evolved to be specific for L-DOPA (Alfonta et al., 2003). Each M18 variant was incubated with PAD4, and crosslinking was catalyzed by the addition of sodium periodate. Crosslinking between predicted L-DOPA-containing M18 variants and PAD4 was probed and quantitated using a Western blot assay specific to the FLAG epitope tag (**Figure 2-2**). A covalent crosslink between antibody and antigen was apparent by a shift in the molecular mass of the antibody. The efficiency of crosslinking was determined by the percentage of antibody that underwent a mass shift. Each variant was tested a minimum of two times. The long-distance

crosslinking controls (H_I51 and L_D17) showed no appreciable crosslinking (< 3%), and eight of the substitutions exhibited crosslinking values ranging from 5% to 55% (**Table 2-1**).

2.2.3 Correlations between Rosetta criteria and crosslinking

We can inspect the biophysical criteria and determine whether and how they impact crosslinking efficiency. **Figure 2-3** shows the plots of each of the five biophysical criteria described above versus the experimentally observed crosslinking extent. Crosslinking efficiency drops off appreciably (< 3%) at distances greater than 7 Å [e.g. variants H_G55 (8.3Å), H_I51 (9.8Å), and L_D17 (32.1 Å)] relative to L-DOPA substitutions that are predicted to lie closer to lysines (**Figure 2-3a**). There is no correlation between solvent accessibility measures and crosslink efficiency when all data are considered (**Figure 3b** and **Figure 2-3c**). However, when the points representing mutants with high crosslinking distance are excluded (black in **Figure 2-3b** and **Figure 2-3c**), both the number of residues that surround an L-DOPA substitution and solvent accessible surface area (SASA) are show a relationship to crosslinking efficiency (**Figure 2-3c**). In particular, variants H_Y52 and H_W33, with SASA of only 25.1 and 14.1 Å respectively, have the lowest crosslinking efficiencies (19.7% and 13.9%, respectively) among all variants with dXL under 7 Å.

The energetic parameters do not provide clear trends for the analysis (or prediction) of crosslinking efficiency even after filtering for distance and solvent accessibility (**Figure 3d** and **Figure 3e**). However, it is notable that the two most efficient crosslinking sites (H_S31 and H_G53) have low energies (I_sc and I_sc^constr of 1 REU or less in both cases). At best, we can say that by also choosing positions that pass the energy filter, we, in general, identify variants that exhibit efficient crosslinking.

Overall, these plots emphasize the key importance of distance, followed by solvent accessibility, in identifying positions capable of significant crosslinking.

2.2.4 Interface Structure of a successful crosslinking case

Figure 2-4 shows a detail of the structural model of the best crosslinking case with the L-DOPA mutant at position H_S31 targeting the lysine at position 679. This mutant exhibits 52% crosslinking (standard deviation 12%). The crosslinking distance is a moderate 5.7 Å, and the SASA of 132 $Å^2$ shows good solvent accessibility. The calculated interface energies of Isc = 0.7 REU and I_sc^constr = -0.6 REU suggest that both the unconstrained and constrained conformations are energetically favorable in the context of the bound antibody-antigen complex. As seen in **Figure 2-4**, the aromatic carbon atoms C_2, C_5, and C_6 are all positioned for nucleophilic attack by the amine.

2.3 DISCUSSION

The insertion of new, functional amino acids into proteins, even proteins of known structure, remains an enterprise that is fraught with uncertainty. While there have been several successful demonstrations of the rational placement of crosslinking amino acids into proteins (Alfonta et al., 2003; Forné et al., 2012; Sato et al., 2011; Umeda et al., 2009, 2010), to date there has not been a thorough analysis of how a detailed structural modeling might be used to guide the placement of amino acids that would crosslink efficiently. Currently, placement of an NCAA into a protein is based on intuition and no computational tools exist to guide the design of crosslinks. A set of computational tools that can streamline the design process while preserving the target protein function and enhancing the functionality of the introduced NCAA would be extremely desirable for the engineering of proteins and protein based materials.

As a first step towards the development of tools for the engineering of proteins containing abiotic functionalities, we sought to develop a rational rules-based approach for the placement of a novel NCAA within a protein-protein complex. To do this, the parameters affecting the placement of a functionalized NCAA (L-DOPA) and its ability to effectively perform its function (i.e. crosslinking) within a protein-protein interface were explored. While this approach entailed the making and testing of a number of different mutants that positioned L-DOPA in various locations of the antibody CDRs, it allowed us to elucidate the biophysical parameters which govern productive placement of an NCAA into a protein-protein interface.

In the case of antibody M18 and PAD, one may think that positions H_D54, H_D56, or H_Y52 on the antibody are the best locations for L-DOPA since they appear close to a lysine in the complex structure. While L-DOPA mutants at all three of these positions showed crosslinking (19.7-36.2%), the most efficient crosslinking (40.3% and 51.9%) occurred at H_S31 and H_R99, positions that scored well by all three biophysical criteria.

Typical crosslinking reactions observed for rationally placed variants in the literature have been reported to suffer from a lack of specificity, low yield, and uncharacterized side reactions (Fancy and Kodadek, 1999; Fancy et al., 1996; Sinz, 2006). Additionally, several commercially available ‘random’ crosslinkers have been reported as inefficient and hardly detectable by MS analysis (Dihazi and Sinz, 2003). In our study, we found that adjacency is an excellent primary criterion for success, and additionally, that greater solvent accessibility increases crosslinking efficiency for proximal residues. Although energetic calculations are noisy, the best two positions for crosslinking L-DOPA moieties showed a low interface energy, both when free and when constrained to be near the nucleophile. In the end, we were able to obtain an antibody that

could crosslink with very high efficiency (55%) with little to no observable side reactions.

Direct comparison of the crosslinking efficiencies to previously published use of L-DOPA as a site-specific crosslinker is difficult because the crosslinking efficiency was not considered or reported in these previous works (Burdine et al., 2004; Umeda et al., 2009, 2010). Instead simple (gel or blot based) demonstrations of conditional L-DOPA mediated crosslinking were reported to demonstrate the crosslinking capabilities of periodate oxidized L-DOPA. Data reported in our study was collected using a fixed set of reaction conditions to effectively evaluate the contributions of various placement parameters on overall crosslinking efficiency. Additional optimization of crosslinking conditions independent of placement parameters such as buffer conditions and periodate concentrations may lead to further enhancements of crosslinking efficiency.

As this study used a single antibody-antigen target and a known crystal structure of the complex, one must interpret the conclusions with caution. Nonetheless, exploration of the M18-PAD4 case suggests a general strategy for identifying high-probability antibody-antigen crosslinking positions that consists of a three-part filter that requires: (i) crosslinking distance under 7 Å; (ii) sufficient solvent-accessibility (SASA over 90 $Å^2$ or fewer than ten neighboring atoms); and (iii) compatible interface energies (Isc under 2.5 REU and I_sc^constr under 10 REU). Application of these filters in this case led to the quick and accurate prediction of a small set of substitutions from which generally excellent crosslinking antibodies were derived. This differs from previous attempts where iterations of insertion and crosslinking conditions had to be considered in order to identify the same levels of crosslinking.

This study employed a crystal structure of the antibody-antigen complex as a starting point, but the effort to determine the complex structure can outweigh the effort of

a blind scan of potential crosslinking sites. Further studies are needed to test the utility of these approaches with antibody homology models (e.g. (Marcatili et al., 2008; Pantazes and Maranas, 2010; Sivasubramanian et al., 2009)) or docked structures (e.g. (Sircar and Gray, 2010)); certainly accurate complex models will be required. It is our hope that the application of a parametric analysis will lead to the computational prediction of a one-hit prediction for the functional placement of NCAAs into a protein scaffold, given only basic structural information or even structural models.

2.4 Experimental Methods

2.4.1 Plasmid Construction

The pMB1 origin of replication in the dual aminoacyl tRNA synthetase/tRNA expression vector, pRST.11B (Hughes and Ellington, 2010), was replaced by the p15 origin of replication from vector pACYC184 by amplification of the p15 origin from pACYC184 using primers p15AA.1, and p15A0.2 (Supplementary Material). The pRST.11B vector was amplified with primers VSP.2 and VSP.3 to generate a ~6kbp fragment that lacks the original origin of replication and is flanked by the unique SpeI and XmaI restriction endonuclease sites. The 1.2kbp PCR product containing the p15A orgin and the 6kbp vector fragment was digested with SpeI and XmaI and ligated together to yield vector pRST.11C. The Nap1 mutant Methanococcus jannaschii tyrosyl amber suppressor tRNA (Guo et al., 2009) was assembled via PCR from the four overlapping oligonucleotides Nap.1, Nap.2, Nap.3 and Nap.4 (Supplemental Material). The assembled tRNA gene was digested with KpnI and BsrGI restriction enzymes and ligated into a similarly digested pRST.11C vector to yield vector pRST.11C-Nap1. A redundant XbaI site in pRST.11C-NapI was removed by quick change mutagenesis using primers, Qcxbaprstc.1 and Qcxbaprstc.2. The gene for the evolved L-DOPA utilizing

aminoacyl tRNA synthetase (Alfonta et al., 2003) was assembled from overlapping oligonucleotides in-house using automated protein fabrication gene assembly process (Cox et al., 2007). The assembled gene was digested with XbaI and XhoI and ligated into a similarly digested pRST.11C vector to yield the L-DOPA incorporating tRNA synthetase/tRNA vector, pDopa.

The pAK400-M18 scFv antibody and pAK400-pAD4 expression vectors (Leysath et al., 2009) were obtained from George Georgiou's group at the University of Texas at Austin. Amber (TAG) codons were introduced into the coding sequence of the M18 antibody via quick change mutagenesis or Gibson Assembly PCR.

2.4.2 Expression and Purification of M18 variants with NCAAs

The M18 antibody variants were expressed using a condensed culture labeling method (Liu et al., 2010) in the presence (or absence) of supplemented L-DOPA. Briefly, the M18 antibody and variants were expressed by inoculating 900 mls of 2xYT media containing 35 μg ml^{-1} chloramphenicol and 100 μg ml^{-1} ampicillin with 1 ml of a saturated overnight culture. Expression cultures were grown at 37°C to OD_{600} ~0.8. Cultures were centrifuged at 3500*g* for 10 minutes, and resuspended in 100 ml 2xYT containing 5 mM DTT, 1.5 mM L-DOPA, and 1.5 mM IPTG. Condensed cultures were grown at 26°C for 12 hours. The PAD4 antigen was grown in Terrific Broth, induced at OD_{600} = 1.0 with 1 mM IPTG, and allowed to grow at 30°C for 12 hours.

Expression cultures were centrifuged at 3500*g* for 15 minutes, and resuspended in PBS with 1 μg ml^{-1} lysozyme and 0.25 U ml^{-1} benzonase, and incubated on ice for 30 minutes. Cells were then sonicated for 4 minutes to further lyse the cells. Lysates were centrifuged at 35,000*g* for 45 minutes, after which the liquid fraction was poured over a 1.5 ml Ni-NTA agarose column. The resin was washed with 45 ml wash buffer (60 mM

imidazole, 200 mM NaCl, 50 mM Phosphate at pH 8), and the protein was eluted with elution buffer (400 mM imidazole, 200 mM NaCl, 50 mM phosphate at pH 8). The Proteins were concentrated using an Amicon cellulose based centrifugal concentrators. Concentrations of the proteins were determined using Abs280 measurements. DOPA incorporation was verified via blue tetrazolium staining of nitrocellulose transferred SDS-page purified protein samples (Gieseg et al., 1993).

2.4.3 Crosslinking Assays

81 pmol (2.7μM) PAD4 and 81 pmol (2.7μM) of the M18 scFv variant were mixed in assay buffer (200 mM NaCl, 50 mM Phosphate at pH 8.5). Crosslinking was accomplished by adding sodium periodate to 3.3 mM. Samples were incubated at room temperature for 30 minutes. Periodate was quenched with addition of DTT to 100 mM and 5x SDS loading dye. Samples were denatured by heating to 98 °C for ten minutes.

Samples were then run on Novex 4-12% Bis-tris SDS gels using MES-SDS running buffer at 200 V for 35 minutes. Proteins were transferred to nitrocellulose at 25 V for 1 hour, using Invitrogen XCell II Blot Module. Nitrocellulose was blocked at room temperature for 30 minutes using Superblock in PBS (Thermo) (for α-His antibody) or Superblock in TBS (Thermo) (for α-Flag antibody). Nitrocellulose was rinsed three times with PBS or TBS. Nitrocellulose was immersed in 50 ml of PBS or TBS with 10 μL of α-His or α-Flag for 1 hour at room temperature. Nitrocellulose was rinsed with PBS or TBS three times, and resolved with Promega Western Blue Stabilized AP substrate for α-His-AP conjugate, or with Promega ECL western blotting substrate for luminescent detection using the α-Flag-HRP conjugate.

M18 variants were resolved at ~30 kDa, PAD4 resolves at 14 kDa. Crosslinked product appears at ~45 kDa. Luminescence of western bands were quantified using ImageJ (Schneider et al., 2012).

2.5 COMPUTATIONAL METHODS

Rosetta computational modeling and calculations were a contribution from the Dr. Jeffrey J. Gray lab, from the Johns Hopkins University Department of Chemical and Biomolecular Engineering. Dr. Jianqing Xu was the primary collaborator from the Gray lab.

2.5.1 L-DOPA surface-interface models

Models for various mutants of the antibody–antigen complex were created using Rosetta (Leaver-Fay et al., 2011) with l-DOPA placed in various positions within the antibody paratope. Coordinates for the wild-type M18-PAD4 complex were downloaded from the Protein Data Bank (Berman et al., 2000) (PDB ID 3ETB). To remove crystal packing effects and obtain a Rosetta-minimized reference structure, fixed-backbone side-chain packing and minimization (1000 decoys) on the entire protein complex was performed using Rosetta's score12. The lowest-scoring structure was used for the calculations for the predictive introduction of l-DOPA into the CDRs of M18. The Rosetta 3.4 (revision 51671, available at www.rosettacommons.org) command line used to run the “fix_bb” protocol was:

```
fixbb.linuxgccrelease -s Crystal.pdb -nstruct 1000
 -use_input_sc
 -minimize_sidechains
 -run:multiple_processes_writing_to_one_directory
 -packing:repack_only -ex1 -ex2aro
```

For each interface Lys on the antigen, neighboring antibody residues within 10 Å (C_β–C_β distance) were selected as potential mutation sites. Each antibody residue within

the cutoff distance was substituted to L-DOPA individually, followed by fixed-backbone side-chain packing (20 decoys) of the nearby residues (<10 Å) to accommodate the local changes. For these and any further calculations where L-DOPA is present, Rosetta uses the molecular mechanics-based scoring function (mm_std) and associated NCAA rotamer library (Renfrew et al., 2012).

To carry out these calculations, the position of the l-DOPA mutation and the positions of the neighboring residues were specified in a "resfile", and the same "fixbb" protocol read the "resfile" and substituted the target residue to l-DOPA, followed by side chain repacking including all the neighboring residues. A Rosetta 3.4 (revision 51671) command line example is:

```
fixbb.linuxgccrelease -s Best_Prepacked.pdb
 -nstruct 20 -use_input_sc
 -resfile 315_LYS_J_679-139_SER_H_31.resfile
 -score:weights mm_std
 -minimize_sidechains -ex1 -ex2
```

2.5.2 Model relaxation with crosslink constraint

Some measures are performed on a structure where the L-DOPA is artificially constrained to be proximal to the target lysine residue. For these calculations, we employed an empirically-determined linear constraint potential, E_{constr}=-100+400*|dXL-3.5|, where d_{XL} is the distance in Ångstroms between C_γ atom on the l-DOPA ring (the atom bound to the $C_\beta$atom) and the Lys N_ε and E_{constr} is the constraint energy in Rosetta Energy Units (REU). The constraint weights were chosen to bring the l-DOPA and Lys into proximity, in order to evaluate interface compatibility. This constraint energy was not included in the final calculated interface score. All the neighboring residues within 10 Å (C_β–C_β distance) of the L-DOPA/Lys pair were repacked to accommodate the constraint. The constrained conformation was generated using a custom PyRosetta script

with PyRosetta 2.012, revision 51671 (PyRosetta available at www.rosettacommons.org, script available upon request).

2.5.3 Crosslinking distance

After selecting the L-DOPA position and the target Lys, all the distances of potential crosslinking atom pairs (lysine N_ε atom and L-DOPA atoms C_2, C_5, and C_6) were evaluated, and the one with the minimum value represented the crosslinking distance (d_{XL}).

2.5.4 Solvent-accessibility measures

As a proxy for solvent accessibility, the number of atoms within 3 Å of the L-DOPA side-chain atoms was computed (Simons et al., 1999). All non-l-DOPA atoms were considered, both from the antibody and antigen, and including hydrogens (as placed by Rosetta). A second solvent accessibility measure is the solvent accessible surface area (SASA) (Le Grand and Merz, 1993), calculated using an embedded function in PyRosetta 2.012 using a probe radius of 1.4 Å.

2.5.5 Rosetta energy calculations

I_{SC} is defined as $I_{sc}=S_{bound}-S_{unbound}$ where S_{bound} and $S_{unbound}$ represent the total scores of the antibody-antigen complex in bound and unbound form, respectively. The unbound form was scored after separating the antibody and antigen to a very large distance without additional side chain repacking. Scores are given in Rosetta Energy Units (REU), which approximate kcal mol^{-1}.

After repacking side chains under a constraint on the L-DOPA/Lys pair, the new pose of the M18/PAD4 complex was saved in a separate file. The constrained interface score, $I_{constrsc}$, was then obtained by rescoring the new conformations both as a complex and as unbound components without retaining the constraint potential,

$I^{constr}_{sc}=S^{const}_{rbound}-S^{constr}_{unbound}$, where the superscript indicates that the constrained conformations are used.

2.6 TABLES AND FIGURES

Antigen Residue	Antibody Position	Crosslinking Distance, d_{XL} (Å)	Number of Neighbors	SASA (Å^2)	Interface Score, I_{sc} (REU)	Interface Score I_{sc}^{constr} (REU)	Experimental Crosslinking
J_K684	H_D54	2.9	4	124.9	0.8	52.0	30.4% (15.8%)
J_K684	H_D56	3.0	2	120.3	1.1	28.9	36.2% (5.0%)
J_K684	H_Y52	3.0	14	25.1	1.3	15.6	19.7% (4.4%)
J_K684	H_W33	3.4	9	14.1	4.8	8.4	13.9% (8.4%)
J_K684	H_N58	5.5	6	111.3	3.7	6.1	24.4% (3.3%)
J_K679	H_S31	5.7	1	132.3	0.7	-0.6	51.9% (11.5%)
J_K679	H_G53	6.1	8	124.8	1.0	0.2	40.3% (13.8%)
J_K684	H_R99	6.6	8	28.2	0.2	151.1	not tested
J_K684	H_G55	8.3	1	161.6	2.8	1.8	5.1% (1.3%)
J_K653	L_N31	8.6	9	75.3	1.2	-1.1	not tested
J_K684	H_S57	8.8	2	122.7	2.7	12.7	not tested
J_K684	H_S32	8.9	18	42.9	1.3	12.2	not tested
J_K653	L_Y50	9.0	11	10.5	2.8	0.1	not tested
J_K684	H_G96	9.3	11	56.8	-1.4	11.5	not tested
J_K684	H_I51	9.8	22	0.5	2.8	13.4	2.7% (0.14%)
J_K653	L_D17	32.1	4	139.9	2.7	0.2	1.5% (0.49%)

Table 2-1: Crosslinking rates and biophysical parameters of selected residues

Candidate crosslink residue pairs, calculated biophysical properties, and experimental crosslinking of the M18/PAD4 interface with an L-DOPA substitution. Antigen residue nucleophile residues paired with antibody positions for placement of L-DOPA. The calculated values of factors (accessible distance, number of neighbors, interface score) after L-DOPA mutation that can affect L-DOPA-Lys crosslinking. Mutants are sorted by accessible distance. Grey shading indicates poor conditions that may reduce or prevent crosslinking (dXL > 7 Å, over ten neighbors, SASA < 90 Å2, Isc > 2.5 REU, or Iscconstr > 10 REU). For the two cases where antibody positions have multiple candidate crosslink targets, only one is shown in the table (Position H_G53's second target is J_K684 at dXL = 4.1 Å with Isc = 1.0 REU and Iscconstr = 11.5 REU, and position H_S31's second target is J_K684 at dXL = 8.6 Å with Isc = 0.7 REU and I_sc^constr = 4.7 REU.)

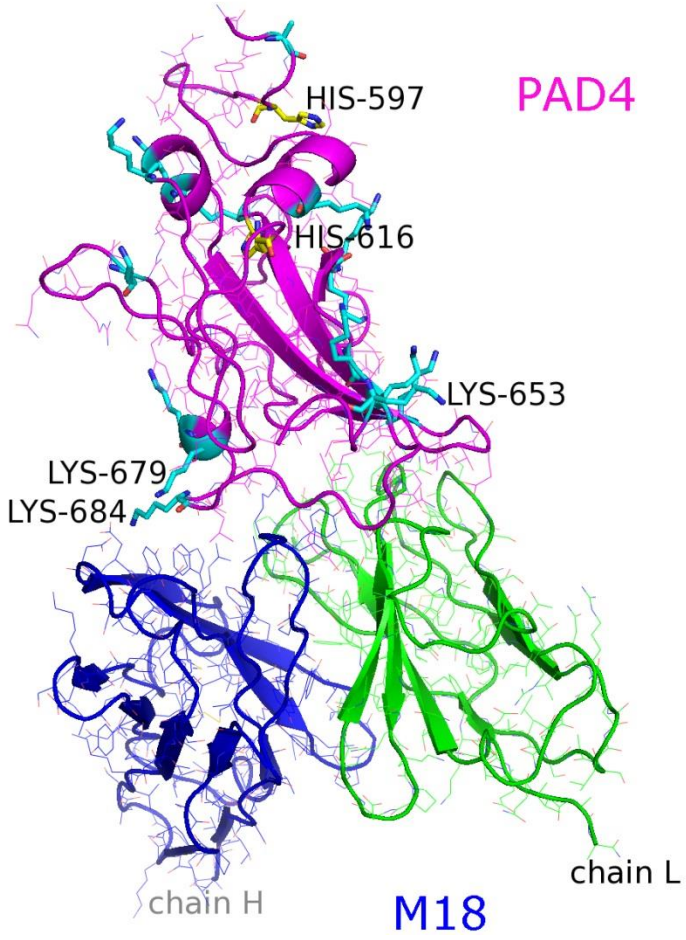

Figure 2-1: The M18/PAD4 antibody-antigen complex

The M18/PAD4 antibody-antigen complex structure from PDB 3ETB highlighting the antigen's candidate crosslinking residues, lysines (cyan sticks) and histidines (yellow sticks).

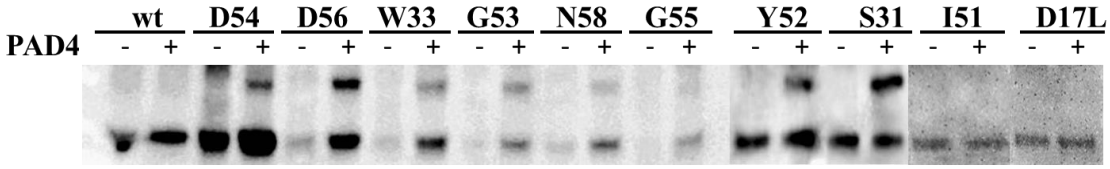

Figure 2-2: Western blot gel shift assay.

A western blot of M18 and variants treated with periodate in the absence and presence of antigen PAD4. α-Flag-HRP was used to detect a Flag-tag on the M18 variants. The shift in mass in the presence of PAD4 corresponds to the mass of PAD4 antigen (~14 kDa), signaling covalent crosslinking between antibody and antigen. Bands were quantified using ImageJ to determine crosslinking efficiencies, which were listed in Table 1. D17L is the only tested position on M18 light chain. The rest of variants were all on heavy chain.

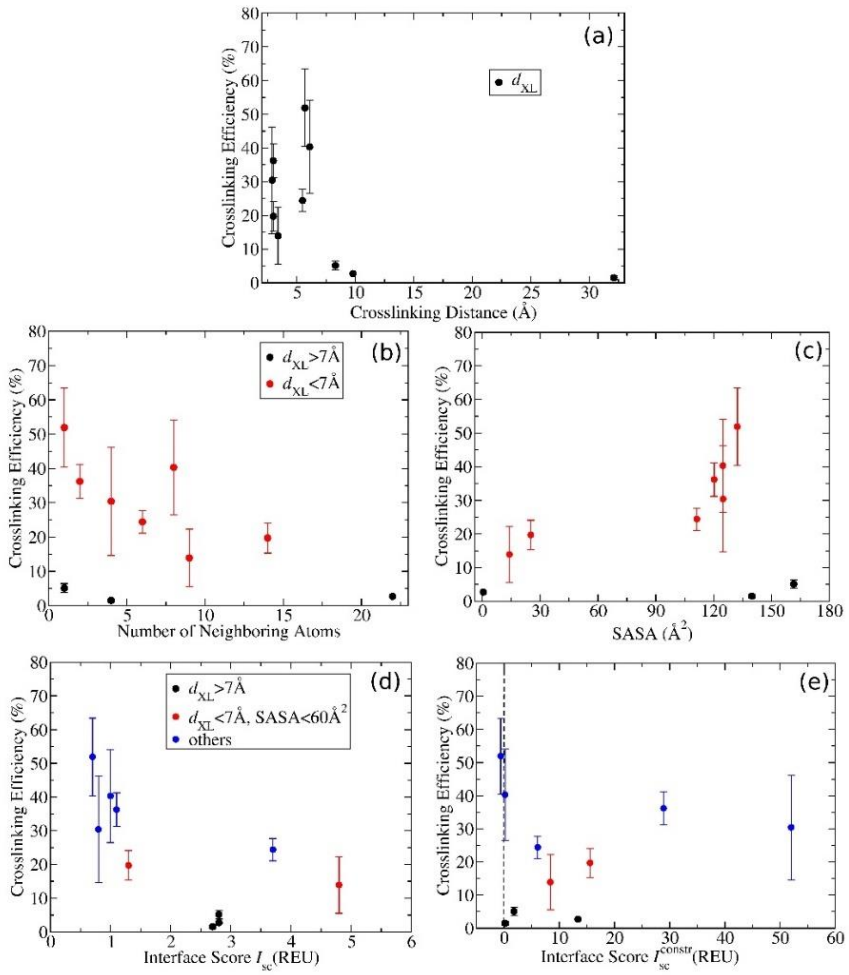

Figure 2-3: Cross efficiency against biophysical criteria.

Plots relating the biophysical criteria to experimental crosslinking efficiency. Error bars display the standard deviations obtained from multiple trials in experiments. (a). Effect of crosslinking distance. (b,c). Effect of solvent accessibility by (b) L-DOPA neighboring atoms or (c) solvent-accessible surface area. Black circles represent positions with dXL over 7 Å (H_G55, H_I51, and L_D17). (d,e). Effect of interface energetics by (d) interface energy or (e) crosslink fluctuation energy. Again black circles represent positions with d_XL over 7 Å, and additionally red circles represent positions with solvent-accessible surface area under 60 Å (H_Y52 and H_W33).

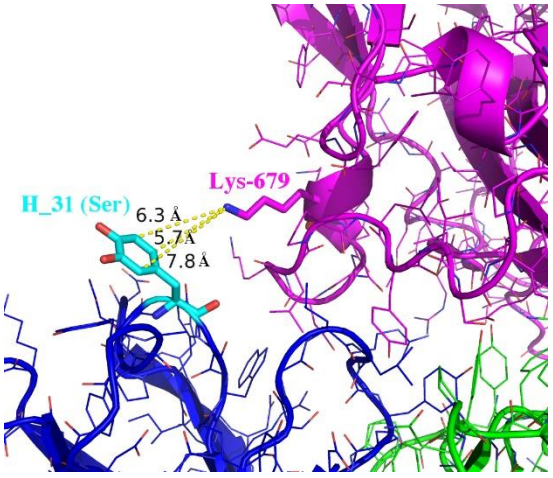

Figure 2-4: Structural modeling of crosslinking residue K679.

Structural detail of the successfully crosslinking L-DOPA mutant at position H_S31 with its targeted Lys at position 679. Crosslinking distance of 5.7 Å, 132 Å^2 of solvent-accessible surface area, and calculated interface energies of I_{sc} = 0.7 REU and I_{sc}^{constr} = -0.6 REU. Aromatic carbon atoms C_2, C_5, and C_6 are all positioned for nucleophilic attack by the amine. The crosslinking distances for these three sites are 7.8 Å, 6.3 Å, and 5.7 Å, respectively.

Chapter 3: Addicting Diverse Bacteria to a Noncanonical Amino Acid[2]

Engineered orthogonal translation systems (OTSs) comprised of tRNA and aminoacyl-tRNA synthetase (aaRS) pairs have greatly enabled the expansion of the genetic code using non-canonical amino acids (NCAAs), typically added across from the amber (UAG) stop codon (Liu and Schultz, 2010). However, the biological, biochemical, and biophysical impact of NCAAs on organismal evolution remains unclear, in part because it is difficult to force the adoption of new genetic codes in organisms; orthogonal translation machinery is often lost or inactivated, even when UAG suppression is required (Wang et al., 2014). To overcome this limitation, we have reengineered the structure of TEM-1 β-lactamase to generate an antibiotic resistance marker that depends on the incorporation of the tyrosine analogues 3-nitro- and 3-iodo- tyrosine. Using the engineered NCAA dependent beta-lactamase variant, we maintained a functional OTS and concomitant NCAA dependence in *E. coli* cells for more than 250 generations without detectable escape. In addition, our engineered β-lactamase allowed the rapid introduction and retention of the OTS across a range of different bacterial species (*Shigella flexneri*, *Salmonella enterica*, *Yersinia ruckeri*, and *Acinetobacter baylyi*).

3.1 Introduction

3.1.1 Previous NCAA evolutionary studies

Life on Earth has been shaped by the universal and almost static genetic code set early in evolutionary history. Life's proteome has been confined to 20 canonical amino acids, as well as selenocysteine and pyrrolysine, two rare, naturally occurring non-

[2] Tack, D.S., Ellefson, J.W., Thyer, R., Wang, B., Gollihar, J., Forster, M.T., Ellington, A.D. (2016). Addicting diverse bacteria to a noncanonical amino acid. *Nature Chemical Biology*. **12**, 138-140. I, Drew Tack, was responsible for all components of this paper, including experimental design and execution, as well as manuscript preparation and submission.

canonical amino acids (NCAAs). Attempts to modify or expand this code have required significant reworking of the genome, either randomly (Wong, 1983) or rationally (Mandell, 2015; Rovner, 2015). Examples include generation of a *Bacillus subtilis* strain via untargeted mutagenesis which preferentially incorporated 4-fluorotryptophan (Wong, 1983), and an *Escherichia coli* strain which required azaleucine incorporation to maintain thymidylate synthase activity via rational replacement of an active site arginine (Lemeignan et al., 1993). In recent years, the development of orthogonal translation systems (OTSs), comprised of aminoacyl-tRNA synthetase (aaRS)/tRNA pairs orthogonal to the host translation machinery, has expanded the genetic code to add new amino acid chemistries and created the potential for new biological functionality including crosslinking (Chin et al., 2002a), fluorescence (Wang et al., 2006a) and mimicking post-translational modifications (Xie et al., 2007). Reengineering the binding pocket of aaRSs to accommodate a wider variety of NCAAs has allowed the limited insertion of new amino acids, typically across from the UAG amber stop codon, resulting in 21 letter amino acid alphabets in a range of model organisms (Liu and Schultz, 2010). However, the ability to expand the use of NCAAs across a proteome will require additional engineering and more likely directed evolution that requires or utilizes the functionality of the new amino acid. For example, we have previously evolved both Qβ and T7 bacteriophages with novel genetic codes (Bacher et al., 2003b; Hammerling et al., 2014). Directed evolution of the genetic code to even larger genomes has been attempted (Bacher and Ellington, 2001), but widespread adoption of a 21st amino acid across the proteome of a whole bacteria has not yet been observed. We believe that this is in part due to the fact that the proteome is not functionally ‘addicted’ to the new amino acid, and thus there are many paths to fitness that do not rely upon it, even in engineered NCAA systems (Kato, 2015; Wang et al., 2014).

3.1.2 Enforcing genetic codes

One way to enforce the maintenance of an OTS has been to engineer a number of essential genes with UAG codons, either at permissive locations or through computational design (Mandell, 2015; Rovner, 2015). This approach has only been applied to *E. coli*, since it requires some foreknowledge of the genome of an organism and the relative orthogonality of its extant translation system. In order to extend this approach to a more diverse set of organisms, we have instead chosen to engineer an antibiotic resistance marker that is dependent on 3-nitro-tyrosine (3nY) or 3-iodo-tyrosine (3iY), NCAAs for which highly active and selective OTSs were available, and which have been shown to be useful in conferring unique and beneficial phenotypes in engineered and evolved proteins (Cooley et al., 2014; Ohtake, 2015; Sakamoto et al., 2009). The resistance marker thus makes the NCAA part of its functional genetic code and, because it is essential for survival, the host is required to maintain an active OTS.

3.2 APPROACH

3.2.1 Selection of NCAA insertion sites

To engineer NCAA dependence we chose to addict TEM-1 β-lactamase, which is known to confer high level β-lactam resistance across many Gram-negative bacteria (Bradford, 2001), and for which high resolution structural data is available (Fonze, 1995). To facilitate the production of an addicted protein, we first needed to identify inner-core amino acids in TEM-1 β-lactamase that abolish β-lactamase activity when replaced with 3nY and 3iY. Potential substitutions were identified by inspecting the structure of TEM1 β-lactamase (PDB: 1XPB) (Fonze, 1995), focusing on solvent shielded amino acids with side chains located in or pointing towards the interior of the protein. Residues involved in catalysis or substrate binding were excluded. We identified three residues that we

hypothesized would be good candidates for NCAA substitution: F66, L162, and T189. TEM-1 variants with single UAG codons were constructed, and introduced into *E. coli* cells containing tyrosyl-tRNA synthetase:tRNA (MjYRS/$tRNA_{CUA}$) pairs derived from *Methanocaldococcus jannaschii* that were capable of incorporating either the natural amino acid tyrosine, or the 3-substituted tyrosine analogues 3nY and 3iY (Cooley et al., 2014; Sakamoto et al., 2009) (**Figure 3-1**). Substitutions at positions L162 and T189 abolished β-lactamase activity with all three amino acids, while F66 retained some activity with tyrosine (**Table 3-1**).

3.2.2 Library Design and Selection

3.2.2.1 Initial selection

To evolve TEM-1 variants dependent on NCAA incorporation for activity, we designed random codon (NNS) libraries that spanned the six residues that had side chains in closest proximity to the suppressed position (**Table 3-2**, **Figure 3-2**). Libraries were selected in the presence of 3nY or 3iY on solid media supplemented with carbenicillin. Selection plates revealed a survival rate of ~0.1%, with a total surviving population of 10^5 cells for each library. One hundred colonies from each selection were assayed in the presence of carbenicillin with or without the corresponding NCAA. Most clones (>99%) from the F66 and T189 libraries did not require the NCAA for survival on carbenicillin, and sequencing indicated a high rate of mutation of the original UAG codon. However, in the L162 library greater than 95% of clones required the NCAA for survival and the UAG codon was retained.

3.2.2.2 Counter-selection with tyrosine

In order to limit pathways for reversion to a natural genetic code, we attempted to eliminate 3nY- and 3iY-dependent β-lactamase variants that would also be active with

the structurally similar canonical amino acid tyrosine at position 162, which is available in the genetic code via a single mutation of the amber codon. To accomplish this, 96 NCAA-dependent clones from each of the L162 libraries (3nY and 3iY) were transformed with a plasmid encoding an OTS for the incorporation of tyrosine at UAG codons. This preliminary screen revealed a total of 24 clones (14 from the 3nY library and 10 from the 3iY library) with minimal carbenicillin resistance with tyrosine incorporation, suggesting that replacing the NCAA with tyrosine strongly impaired β-lactamase activity. These 24 clones were sequenced and were found to comprise 23 unique sequences, with a single duplicate (**Table 3-3**, **Table 3-4**). In these clones, significant enrichment was observed for M68F (17/24) and F72G (11/24) mutations, while sequencing of NCAA active β-lactamases before the tyrosine counter-screen saw no convergence at these positions. The F72G mutation appears to make space for the larger aromatic side chains of the NCAA residues (**Figure 3-3**). M68F was frequently found in combination with F72G (10 out of 24). There were no obvious differences between variants selected on 3nY and those selected on 3iY.

3.2.3 Characterization of selected β-lactamase variants

3.2.3.1 Enzyme activity with probable canonical amino acid replacements

It was also possible that reversion might occur via the wild-type amino acid at position 162, leucine, which is also available via a single mutational step from the amber codon. We chose to evaluate the activity of the 23 unique β-lactamases with leucine and tyrosine codons replacing the UAG codon in each clone. These cognate codons are read at higher efficacy by *E. coli* translational machinery than OTS incorporation of amino acids, and by eliminating leucine and tyrosine active clones, we reduced the number of clones dependent on NCAAs to four variants (**Table 3-5**). Two of these variants showed

reduced enzymatic activity with either L or Y (MIC <1000 μg ml^{-1} carbenicillin; Fig. 1a, TEM-1.B9 and TEM-1.F2), one variant showed reduced activity with L, and one showed reduced activity with Y (**Figure 3-4**, TEM-1.G3 and TEM-1.D1 respectively).

Interestingly, all four variants were active with both 3nY and 3iY, as might have been expected given the similarity of sequenced variants selected on these two tyrosine analogues. All four variants also had higher activity with 3nY than with 3iY (**Figure 3-4**). Additionally, during the course of this experiment we observed higher carbenicillin resistance for colonies containing the 3iY OTS on plates containing 3nY than those containing the 3nY OTS, suggesting the synthetase originally evolved for 3iY incorporation was in fact better at incorporating 3nY than the synthetase evolved for 3nY incorporation. This was subsequently confirmed by UAG suppression assays using a GFP reporter (**Figure 3-5**), and the 3iY OTS was used for both 3nY and 3iY incorporation in all subsequent experiments.

3.2.3.2 Selecting and testing enzyme function with other canonical amino acids

Having eliminated tyrosine and leucine from the pool of amino acids that could replace the NCAA in the β-lactamase, we wanted to determine if any other canonical amino acids could substitute. Thus, NNS codon libraries at position 162 were constructed for the four selected TEM-1 variants. Libraries were selected on carbenicillin and TEM-1 genes from 15 random colonies were sequenced from each library. Canonical amino acids that could support activity were identified for all TEM-1 variants (**Table 3-6**), with phenylalanine being the most abundant. Variant TEM-1.B9 was observed to have the fewest canonical solutions, with all 15 clones returning TTC codons corresponding to phenylalanine. To confirm this reduced set of canonical solutions for TEM-1.B9, 20 individual variants were constructed corresponding to all 20 canonical amino acids at

position 162, and the MICs determined for each (**Figure 3-6**). This experiment confirmed that phenylalanine was the only canonical amino acid capable of rescuing NCAA dependence. We also explored stability and activity of TEM-1.B9, and control variants.

3.2.3.3 Long Term OTS Addiction

Since the codons for phenylalanine (TTT and TTC) cannot be accessed by a single point mutation from a TAG codon, and cannot be suppressed by a single mutation in either of the genes encoding *E. coli* $tRNA_{Phe}$ (*pheU* and *pheV*), we hypothesized that NCAA dependence and concomitant OTS functionality might be maintained during prolonged serial culture experiments with this β-lactamase. We therefore serially cultured *E. coli* cells containing either TEM-1.B9 or a control TEM-1 variant with an amber codon replacing residue D254 (TEM-1.D254tag) and the 3iY OTS, in triplicate. TEM-1.D254tag is located on a solvent exposed surface loop (**Figure 3-7**) and substitutions at this location with either canonical or NCAAs retain wild-type activity. Cells were cultured for over 250 generations, and despite screening $>5*10^{11}$ cells for carbenicillin resistance in the absence of 3nY in both solid and liquid media, no escaped cells were detected for cultures containing TEM-1.B9 (**Figure 3-8, Figure 3-9**). Control cultures containing the TEM-1.D254tag variant escaped NCAA dependence within 10 generations (one passage to confluence), and occurrences of NCAA independent β-lactam resistance increased in subsequent generations. The lack of escape mutants observed for cells containing TEM-1.B9 indicates that the combination of reducing the pool of canonical solutions, and eliminating solutions accessible by single point mutations is an effective strategy for enforcing NCAA dependence and OTS functionality.

We also investigated the frequency of escape for the other variants that, based on our substitution analyses, should have more paths to the use of a canonical code for

survival. Each of the other three variants was cultured in triplicate for 50 generations and $3*10^9$ cells screened from each line every 10 generations to determine the escape frequency. These variants had significantly fewer escape events than the TEM-1.D254tag control, but all had at least one escape event over the 50 generations (**Figure 3-8**). Screening cells in liquid cultures after 50 generations produced a similar result (**Figure 3-9**), confirming that eliminating canonical solutions at the NCAA position is important for preventing escape events. Additional serial culturing experiments at lower carbenicillin concentrations (0.25 ml^{-1} and 1 mg ml^{-1}; **Figure 3-10**) or in mutation-prone *E. coli* strains (Baba et al., 2006) (**Figure 3-11**) further confirmed the stability of the TEM-1.B9 allele. The lack of escape mutants observed for cells containing TEM-1.B9 indicates that both reducing the pool of canonical solutions and eliminating solutions accessible by single point mutations is an effective strategy for enforcing NCAA dependence and OTS functionality.

3.2.3.4 Transferring addiction to different bacterial hosts

Given β-lactams have a broad spectrum of activity, and that the tRNA and aaRS of the 3iY OTS might be orthogonal in species other than *E. coli*, we hypothesized that other bacterial species could be addicted to an NCAA using our evolved TEM-1.B9 β-lactamase. We used a single plasmid containing both OTS and TEM-1.B9 to test this hypothesis in *E. coli* strains TOP10, XL1-Blue, BL21 and RTΔA (derived from C321.ΔA) (Lajoie et al., 2013; Thyer et al., 2015) the *Enterobacteriaceae* species; *Shigella flexneri*; *Salmonella enterica*; *Yersinia ruckeri*; and *Acinetobacter baylyi*, which is from the order *Pseudomonadales* and is genetically distant from the *Enterobacteriales*. All strains and species became dependent upon 3nY for growth in the presence of ampicillin (**Figure 3-12**, **Figure 3-13**). To our knowledge, this was the first

demonstration of NCAA incorporation with M. jannaschii tyrosyl-tRNA synthetase in any *Shigella*, *Salmonella*, *Yersinia* or *Acinetobacter* and suggests broad compatibility of the archaeal translation machinery in bacteria. Serial culturing of all tested species showed continuous dependence on 3nY for ampicillin resistance. (**Figure 3-14** and **Figure 3-15**).

3.3 Discussion

By evolving TEM-1 β-lactamase to function selectively upon incorporation of the NCAAs 3nY or 3iY at the amber stop codon, we could maintain an OTS in diverse bacterial species for a hundred generations without any detected escape of NCAA dependence. This strategy should prove successful with other enzymes and other amino acids. In general, it should be possible to reconfigure internal pockets within proteins to accommodate an unnatural amino acid selectively, with little chance of accessing canonical solutions. More importantly, the ability to addict a range of bacterial species to NCAAs through engineered β-lactamases should provide a much more facile means of introducing new amino acids into the genetic codes of many organisms, including industrially relevant bioprocess strains.

3.4 Materials and Methods

3.4.1 Strains and reagents

TOP10 *E. coli* (Life Technologies) was used for routine cloning, screening of the libraries, and MIC measurements. *Salmonella enterica* (ATCC #9270), *Shigella flexneri* (ATCC #12022), and *Yersinia ruckeri* (ATCC #29473) were obtained through VWR international. Amberless *E. coli* strain RTΔA (derived from C321.ΔA) was described previously (Thyer et al., 2015), *Acinetobacter baylyi* ADP1 was a gift from Dr. Jeffery Barrick, University of Texas at Austin.

Restriction enzymes XhoI, HindIII, and DpnI, and Antarctic Phosphatase were obtained from New England Biolabs (NEB). All PCRs were performed using Phusion polymerase (NEB), 200 μM dNTPs and 0.5 μM of each oligonucleotide primer (IDT). Gibson assembly was used for site specific mutagenesis and assembly of the single plasmid systems (Gibson, 2009). Antibiotics required for plasmid maintenance were used at the concentrations listed unless otherwise indicated; chloramphenicol (34 μg ml^{-1}), gentamycin (10 μg ml^{-1}), kanamycin (50 μg ml^{-1}) and spectinomycin (100 μg ml^{-1}).

3.4.2 Plasmid construction

The OTSs were used in pNCAA (Genbank KU055485) and pAA (KU055484). Briefly, these plasmids contain an orthogonal translation cassette consisting of the Methanocaldococcus jannaschii tyrosyl-tRNA synthetase (*Mj*YRS) expressed from the *E. coli tyrS* promoter and the Nap3 UAG suppressor tRNA under the control of the *E. coli lpp* promoter. The pNCAA-3nY variant was constructed by replacing the 3iY-RS gene with a codon optimized *Mj*YRS evolved for incorporation of 3nY (Cooley et al., 2014), pNCAA contained gentamycin resistance cassette, and pAA contained spectinomycin resistance cassette.

The TEM-1 expression plasmid (pBla, Genbank KU055483) was derived from plasmid pBR322 (GenBank J01749.1). An XhoI restriction site was added immediately downstream of the $bla_{\text{TEM-1}}$ coding sequence, and a HindIII site added immediately 5′ of the start codon. The tetA(C) gene was replaced with the cat gene encoding chloramphenicol acetyl-transferase, which conferred chloramphenicol resistance from the plasmid pACYC184 (GenBank X06403.1). The $bla_{\text{TEM-1}}$ variant TEM-1.D254tag control was generated using Gibson assembly.

The single plasmid expression system was derived from plasmid pTH18kr (GenBank AB019603.1) which contains the low copy, medium host range SC101 origin (Hashimoto-Gotoh, 2000). The region containing P_{lac}, the MCS, and *lacZα* was replaced with a bicistronic insert encoding a codon optimized (for *E. coli*) 3iY synthetase and TEM-1 variants under the control of the constitutive EM7 promoter (Genbank KU055486). Oligos DT.15 and DT.16 (**Table 3-7**) were used to amplify pTH18kr backbone, for Gibson assembly. The Nap3 tRNA cassette was cloned from the pNCAA-3iY plasmid located 3' of synthetase and TEM-1 variants. TEM-1 variants were swapped into plasmid through amplification from pBla plasmid with oligos DT.17 and DT.18, with backbone application using DT.19, and DT.20. The plasmid was maintained in *E. coli* strains in LB media (Fisher BioReagents) supplemented with 25 μg ml^{-1} kanamycin, 50 μg ml^{-1} carbenicillin, and 1 mM 3nY (Sigma). The IncQ plasmid was created by replacing the TEM-1 β-lactamase on plasmid pMMB66EH (GenBank X15234) with TEM-1 variants, tRNA, and 3iY synthetase, using primers DT.21 and DT.22 to amplify pMMB66EH, and primers DT.23 and DT.24 to amplify machinery from the SC101 based plasmids.

3.4.3 Residue selection and characterization

TEM-1 β-lactamase libraries were designed by manual inspection of the TEM-1 structure (PDB: 1XPB) (Fonze, 1995) using the molecular visualization program pymol (v1.3). Individual variants with amber codons replacing F66, L162, or T189 were generated by Gibson assembly. The MICs of these variants with ampicillin were determined by plating $5*10^7$ cells (TOP10 *E. coli*) containing TEM-1 variants, and either pAA-Y, pNCAA-3nY, or pNCAA-3iY on solid media supplemented with 1 mM 3nY or 3iY as appropriate. MIC E-test strips (Biomerieux) were added to each plate, followed by

incubation for 16 hours at 37°C. MICs were recorded as the lowest concentration of ampicillin with no cellular growth following the manufacturer's instructions.

3.4.4 Library design and construction

For the L162 library, the six residues with side chains in closest proximity to the leucine side chain were selected for site-saturation (NNS) mutagenesis. These corresponded to residues M68, F72, L139, L148, L152, and L169 (**Figure 3-2)**. Degenerate oligonucleotides (IDT) were used to amplify the *bla*$_{\text{TEM-1}}$ gene in three segments. The N-terminal fragment was amplified using primers DT.1 and DT.3, the C-terminal fragment with DT.2 and DT.6, and the central fragment with DT.4 and DT.5. Library fragments were purified by gel extraction (Promega) and were assembled by overlap PCR using an equimolar ratio with primers DT.1 and DT.2. The assembled product was digested with XhoI and HindIII and ligated into a watermarked pBla backbone (E166 and P167 codons replaced with TAA stop codons). The *bla*$_{\text{TEM-1}}$ library and backbone were mixed in a 2:1 molar ratio and ligated with T4 DNA ligase (NEB) at 16°C for 16 hours. Ligated DNA was desalted by column purification (Zymo) and transformed into MegaX DH10B T1R Electrocompetent cells (Invitrogen, C6400-03). Library diversity was calculated by dilution plating and was determined to exceed $3*10^8$. Supercoiled library DNA was recovered by plasmid minipreps (Qiagen).

3.4.5 Primary β-lactamase selections and tyrosine counter-screening

Electrocompetent cells (*E. coli* TOP10 containing either pNCAA-3nY or pNCAA-3iY) were prepared by culture in SOB medium at 37°C to mid log phase (~ OD600=0.6), incubation on ice for 20 minutes, centrifugation at 2500*g* for 10 minutes, and two washes with equal volumes of chilled 10% glycerol. Cells were then resuspended in half culture volumes of chilled 10% glycerol along with 5 µg of TEM-1 library DNA.

Cells were electroporated using a Biorad micropulser and library diversity measured by dilution plating. Each library was plated on LB-agar supplemented with 250 μg ml^{-1} carbenicillin and either 1mM of 3nY or 3iY, as appropriate, and incubated at 37°C for 16 hours. Both libraries had ~10^5 colonies after selection, with calculated transformation efficiencies exceeding 3*10^8. Two-hundred colonies from each library were picked from the selection plates and streaked on three plates: the first containing 1 mM of NCAA and carbenicillin (250 μg ml^{-1}), the second without the NCAA, and a third without NCAA or carbenicillin.

Following incubation, clones which did not survive on the second plate (lacking NCAA) were picked and cultured in LB at 37°C for 16 hours, and then used to inoculate fresh 2 ml LB cultures. After incubation for 3.5 hours, cultures were chilled on ice for 20 minutes, centrifuged at 2500*g* for 10 minutes and resuspended in 200 μL of TSS buffer (Lajoie et al., 2013) with plasmid pAA-Y, and incubated for one hour on ice. LB medium was added to a final volume of 2 ml and cultures were recovered at 37°C for 2 hours. An aliquot of 200 μL was plated on LB-agar supplemented with chloramphenicol and spectinomycin, and incubated at 37°C for 16 hours. Colonies from these plates were used to determine MICs of individual clones as described above. Clones which appeared to be inactive with tyrosine incorporation were analyzed by Sanger sequencing (**Table 3-3**, **Table 3-4**).

3.4.5 Rephenotyping β-lactamase variants

The TAG codon at position 162 in all unique *bla*$_{\text{TEM-1}}$ genes isolated from the primary selection and screening was replaced with either leucine (TTG) or tyrosine (TAT) codons using Gibson assembly (oligo DT.9 with DT.10 and DT.11 respectively). Leucine and tyrosine variants were sequence verified, and MICs were determined by

plating 3 μL of mid-log cultures on LB-agar supplemented with chloramphenicol and a range of carbenicillin concentrations (0, 25, 50, 100, 250, 375, 500, 750, and 1000 μg ml^{-1}). Clones selected for further characterization (TEM-1.B9, TEM-1.D1, TEM-1.F2, TEM-1.G3) were transformed into cells containing either pNCAA-3nY or pNCAA-3iY, to determine NCAA dependent MICs. Each was plated on three series of plates covering a range of carbenicillin concentrations; one lacking NCAAs, one with 1 mM 3nY, and one with 1 mM 3iY. Carbenicillin concentrations used were 0, 50, 100, 250, 375, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 5000, 6000, 8000 and 12000 μg ml^{-1}.

3.4.6 Characterization of TEM-1.B9

TEM-1.B9 and TEM-1.D253tag were transformed into RTΔA containing the 3iY-OTS plasmid, and grown in 1000 μg ml^{-1} carbenicillin with 10 mM 3nY at 30°C for 20 hours. Wild-type TEM-1 was transformed into TOP10 *E. coli*, and grown in LB at 30°C for 20 hours. Samples were purified via periplasmic osmotic release (Stec et al., 2005). Cells were rinsed twice in 30 mM tris at pH 7.2, then resuspended in 20% sucrose with 5 mM EDTA on ice for 30 minutes. Cells were centrifuged and resuspended in ice cold water. Samples were concentrated with cellulose based centrifugation concentrators with 10 kDa MWCO. Samples were further purified with FPLC HiTrap Q column (GE) with 10 mM tris buffer at pH 7, and eluted with a 0 to 100 mM NaCl gradient. Kinetic analysis was conducted using 1 nM β-lactamase while monitoring the conversion of nitrocefin (EMD chemicals) at a range of concentrations (0, 10, 20, 30, 50, 75, 100, 150, 200, 250, 300, 350 μM) by measurement of absorbance at 486 nm in 100 mM phosphate buffer at pH 7 using a Tecan Infinite M200 Pro, technical duplicates were averaged, and error bars represent standard deviation (**Figure 3-16**). Kinetic values were calculated with Michaelis-Menten kinetics. Melting curves were calculated by monitoring fluorescence

with Texas Red, using excitation and emission wavelengths of 577 and 620 nm respectively, (**Figure 3-17**) with technical duplicates, using a Roche LightCycler 96, and data analysis was performed with LightCycler 96 SW 1.1 software.

3.4.7 Identifying additional canonical solutions

The TAG codon at position 162 of selected clones was randomized using a degenerate codon library (NNS), representing all 20 canonical amino acids. Gibson assembly was used for library construction with primers DT.9 and DT.14. Libraries were prepared as described above, and 12 naïve clones from each were sequenced to ensure a good codon distribution, and acceptable bias. Libraries were transformed into TOP10 cells and plated on LB-agar supplemented with 2000 μg ml^{-1} carbenicillin, above normal lab culture condition, but below the NCAA dependent MICs of the selected clones, which allowed us to identify amino acids that resulted in high enzyme activity. Fifteen random colonies from each library selection were recovered in liquid media and bla_{TEM-1} genes analyzed by Sanger sequencing. Additionally, a sample of library was grown in liquid media at several carbenicillin concentrations, plasmids from each was purified and sequenced with Sanger sequencing. Chromatogram analysis shows a convergence at higher carbenicillin concentrations to phenylalanine codons (**Figure 3-18**). To further characterize the TEM-1.B9 variant, 16 individual variants were pulled from NNS library corresponding to all 20 canonical amino acids at position 162 (tyrosine and leucine were cloned previously at this location, methionine and tryptophan were constructed with DT.9 paired with DT.12 and DT.13 respectively). Each was sequence verified and MICs determined as described previously. Biological triplicates of each codon were measured to ensure reproducibility, with all triplicates in close agreement.

3.4.8 Serial Culture Experiments

pBla containing our identified TEM-1 variants, including TEM-1.D254tag, were transformed into *E. coli* cells containing the pNCAA-3iY plasmid and plated on LB-agar supplemented with gentamycin and chloramphenicol. Colonies were picked in triplicate and cultured in liquid medium overnight. Overnight cultures were diluted 1:1,000 to start new 5-ml cultures of LB medium supplemented with 1 mM 3nY and carbenicillin (2,500 μg ml−1 for all clones except TEM-1.G3, which was grown at 1,500 μg ml−1). Cultures containing TEM-1.B9 and TEM-1.D254tag were serially grown to confluence (~10 doublings) for 25 d. TEM-1.B9 and TEM-1.D254tag were additionally cultured in TOP10 and JW0105-1 *E. coli* knockout strains (Baba et al., 2006) at 1,000 μg ml^{-1} for 10 d, and TOP10 cells were additionally cultured at 250 μg ml^{-1} for 10 d using the same protocol. Cells containing other bla_{TEM-1} variants were cultured to confluence for 5 d.

3.4.9 Characterization of Single Plasmid System in Bacterial Species

All bacterial species and strains described above were made electrocompetent through serial glycerol washes during mid-log phase growth, as described above. *Yersinia ruckeri* was grown and maintained at 26°C. Transformed cells were grown on LB-agar plates with kanamycin and 10 μg ml^{-1} ampicillin with 1 mM 3nY. Colonies were picked from plates in triplicate or greater to ensure reproducibility and grown in LB medium. At mid-log, an aliquot of $5*10^7$ cells was plated on LB-agar with and without 1 mM 3nY. MIC Etest strips were then added to each plate. MICs were recorded after 16 h of incubation (**Figure 3-12**). Data reported as mean of biological replicates ± s.e.m. Sample sizes (n) for each species reported were as follows: TOP10, 12; XL1-Blue, 3; BL21, 3; RTΔA, 6; *S. enterica*, 6; *S. flexneri*, 6; *Y. ruckeri*, 3; *A. baylyi*, 3. Strains with n>3 were used as internal controls to ensure consistency over multiple Etest strip packages.

For serial growth experiments, strains were cultured in LB medium with 1 mM 3nY and carbenicillin for 10 d, with daily 1,000-fold dilutions. After the tenth day, cultures were tested for reversions by plating on LB-agar with carbenicillin in the absence of 3nY.

3.4.10 Reversion Assays

Escape frequencies were determined for cultures containing TEM-1 variants following each of the first five passages to confluence, and additionally following the 10th and 25th passages for TEM-1.B9 and TEM-1.D254tag cultures. Aliquots of ~$3*10^9$ cells from overnight cultures were centrifuged at 2500*g* for 10 minutes and resuspended in 5 ml of LB medium supplemented with chloramphenicol and gentamycin, and incubated at 37°C for five hours to minimize transfer of 3nY. Following incubation, cultures were centrifuged and resuspended in 5 ml of fresh LB and ~$3*10^9$ cells were plated on solid media containing 2500 μg ml^{-1} carbenicillin (1500 μg ml^{-1} carbenicillin for TEM-1.G3). Colony counts on plates were used to determine escape frequencies.

All cultures were challenged by liquid reversion assays at day 5, and TEM-1.B9 again at day 25. Approximately $3*10^{10}$ cells were recovered by centrifugation and resuspended in 20 ml fresh LB with chloramphenicol and gentamycin, and incubated at 37°C for five hours. Cells were then centrifuged and resuspended again, and used to inoculate 200 ml of LB supplemented with 2500 μg ml^{-1} carbenicillin (1500 μg ml^{-1} carbenicillin for TEM-1.G3). Cell growth was monitored by OD600 readings taken at 0, 1, 2, 3, 4, 8, 20, 32, and 48 hours post dilution.

Single plasmid reversion assays were conducted in triplicate by passaging *E. coli* strains and other bacterial species in LB supplemented with 1 mM 3nY and 50 μg ml^{-1} ampicillin for all strains except RTΔA (100 μg ml^{-1}) and *Y. ruckeri* (conditioned to grow

at 50 μg ml^{-1} after 3 days, *Y. ruckeri* was passaged at 26°C). A 1:1000 dilution was used daily to allow for ~10 doublings per passages. After 10 passages, $3*10^9$ cells were centrifuged and resuspended in fresh LB without NCAA, and grown for 4 hours, then centrifuged and resuspended, and plated on LB-agar plates with the ampicillin concentrations as used during serial culture. Colony counts at a range of dilutions were used to determine reversion rates.

3.4.11 Molecular Modeling of TEM1 Library Variants

The wild-type structure of TEM1 (PDB: 1XPB) (Fonze, 1995) was prepared for mutational analyses using the Molecular Operating Environment (MOE.09.2014) software package from Chemical Computing Group. The structure was inspected for anomalies and protonated/charged with the Protonate3D subroutine (310K, pH 7.4, 0.1 M salt)21. The protonated structure was then lightly tethered to reduce significant deviation from the empirically determined coordinates and minimized using the Amber10:EHT forcefield with R-field treatment of electrostatics to an RMS gradient of 0.1 kcal mol-1 Å-1. Next, we created rotamer libraries for each non-canonical amino acid using a low-mode molecular dynamics (LowModeMD) methodology. Library positions of representative variants from the selection were mutated and repacked within the local environment. Conformational analysis of the loop was evaluated with LowModeMD22. Mutated structures were then solvated with water and counterions in a 6 Å sphere and minimized to an RMS gradient of 0.001 kcal mol-1 Å-1. 2-D contact maps of the non-canonical solutions were inspected with Ligand Interactions23 after arbitrarily setting the non-canonical amino acid to the ligand position. Measurements of stability and potential energy were scored within MOE.

3.5 DEPOSITED ACCESSION CODES

Sequencing data generated from this work are deposited under accession codes KU055483, KU055484, KU055485 and KU055486.

3.6 TABLES AND FIGURES

	Tyr	3nY	3iY
F66tag	256	4	4
L162tag	24	4	4
T189tag	8	4	3

Table 3-1: MICs of UAG TEM-1 variants

MICs of TEM-1.tag variants with tyrosine, 3-nitrotyrosine, or 3-iodotyrosine incorporated at UAG codons with engineered OTSs from *Methanocaldococcus jannaschii*.

Library	F66tag	L162tag	T189tag
Randomized Codons (NNS)	Arg43	Met68	Leu75
	Arg61	Phe72	Leu152
	Glu64	Leu139	Met155
	Pro67	Leu148	Asp157
	Tyr264	Leu152	Thr160
	Thr266	Leu169	Met186

Table 3-2: Library positions

Positions randomized in each TEM-1.tag library.

TEM-1	Met68	Phe72	Leu139	Leu148	Leu152	Leu162	Leu169
A5	Ser	Gly	Met	Met	Leu	Amb	Leu
B8	Phe	Gly	Thr	Leu	Met	Amb	Leu
B9	Phe	Gly	Thr	Leu	Met	Amb	Leu
C3	Tyr	Gly	Ile	Ile	Leu	Amb	Leu
C8	Met	Trp	Ser	Leu	Leu	Amb	Ser
C10	Leu	Phe	Ile	Leu	Leu	Amb	Cys
D1	Phe	Gly	Leu	Leu	Ala	Amb	Leu
D2	Phe	Gly	Thr	Met	Leu	Amb	Leu
D3	Phe	Gly	Phe	Leu	Val	Amb	Leu
E1	Phe	Ile	Met	Val	Leu	Amb	Ile
E3	Phe	Gly	Ala	Leu	Ala	Amb	Leu
E6	Val	Arg	Leu	Met	Leu	Amb	Cys
E9	Met	Phe	Ala	Ile	Ile	Amb	Cys
E10	Thr	Gly	Met	Leu	Leu	Amb	Leu
F2	Phe	Gly	Leu	Leu	Thr	Amb	Leu
F3	Met	Gly	Leu	Ile	Met	Amb	Leu
F4	Leu	Gly	Met	Leu	Ala	Amb	Leu
F9	Phe	Gly	Ala	Ile	Leu	Amb	Leu
G3	Met	Ala	Ala	Leu	Val	Amb	Cys
G7	Ile	Gly	Phe	Met	Met	Amb	Leu
G9	Phe	Gly	Ile	Val	Ala	Amb	Leu
H1	Met	Gly	Thr	Leu	Leu	Amb	Val
H3	Met	Ala	His	His	Leu	Amb	Cys
H8	Phe	Gly	Ile	Leu	Leu	Amb	Val

Table 3-3: Amino acid sequences of TEM-1 variants

Sequences of the 24 clones isolated after tyrosine OTS characterization. Gray indicates selection on 3nY, blue indicates selection on 3iY. Bold identifiers indicate variants characterized extensively.

TEM-1	Met68	Phe72	Leu139	Leu148	Leu152	Leu162	Leu169
A5	AGC	GGC	ATG	ATG	CTG	TAG	TTG
B8	TTC	GGG	ACC	CTC	ATG	TAG	TTG
B9	TTC	GGG	ACC	CTC	ATG	TAG	TTG
C3	TAC	GGG	ATC	ATC	TTG	TAG	TTG
C8	ATG	TGG	TCG	CTC	TTG	TAG	TCG
C10	TTG	TTC	ATC	CTG	TTG	TAG	TGC
D1	TTC	GGG	CTG	TTG	GCG	TAG	TTG
D2	TTC	GGC	ACC	ATG	TTG	TAG	CTC
D3	TTC	GGC	TTC	CTC	GTG	TAG	CTG
E1	TTC	ATC	ATG	GTC	TTG	TAG	ATC
E3	TTC	GGG	GCC	TTG	GCC	TAG	CTG
E6	CTG	CGG	CTG	ATG	CTG	TAG	TGC
E9	ATG	TTC	GCG	ATC	ATC	TAG	TGC
E10	ACG	GGC	ATG	TTG	CTC	TAG	CTG
F2	TTC	GGC	CTC	CTC	ACC	TAG	CTG
F3	ATG	GGC	CTG	ATC	ATG	TAG	TTG
F4	TTG	GGC	ATG	CTC	GCC	TAG	CTG
F9	TTC	GGG	GCC	ATC	CTG	TAG	TTG
G3	ATG	GCC	GCG	CTC	GTC	TAG	TGC
G7	ATC	GGG	TTC	ATG	ATG	TAG	TTG
G9	TTC	GGC	ATC	GTC	GCG	TAG	TTG
H1	ATG	GGG	ACC	CTC	TTG	TAG	GTC
H3	ATG	GCG	CAC	TTG	CTC	TAG	TGC
H8	TTC	GGC	ATC	TTG	CTC	TAG	GTC

Table 3-4: Codon sequences of TEM-1 variants

Sequences of the 24 clones isolated after tyrosine OTS characterization. Gray indicates selection on 3nY, blue indicates selection on 3iY. Bolded and underlined identifiers indicate the variants extensively characterized throughout this manuscript.

TEM-1	Met68	Phe72	Leu139	Leu148	Leu152	Leu162	Leu169
TEM-1.B9	Phe	Gly	Thr	Leu	Met	Amb	Leu
TEM-1.D2	Phe	Gly	Thr	Met	Leu	Amb	Leu
TEM-1.F2	Phe	Gly	Leu	Leu	Thr	Amb	Leu
TEM-1.G3	Met	Ala	Ala	Leu	Val	Amb	Cys

Table 3-5: Sequences of four potential addicted TEM-1 variants

Mutations of the β-lactamases identified with reduced canonical functionality, and reduced, or eliminated rates of NCAA independence.

TEM1.Variant		
Codon	*N*	Amino Acid
TEM1.B9		
TTC	15	Phe
TEM1.D1		
TTC	9	Phe
TGG	4	Trp
TAC	2	Tyr
TEM1.F2		
TTC	11	Phe
TGG	4	Trp
TEM1.G3		
TTC	6	Phe
ATC	5	Ile
ATG	1	Met
GTG	1	Val
GTC	1	Val
CTG	1	Leu

Table 3-6: Canonical amino acid solutions

Sequencing results from Amb162NNS libraries of the four TEM-1 variants fully characterized. Results indicate that each selected clone has at least one canonical amino acid which can substitute 3nY.

Name	Primer Sequence
DT.1	AATGCTTCAATAATATTGAAAAAGGAATAAAGCTTATGAGTATTCAA
DT.2	ATCAATCTAAAGTATATATGAGTAAACTTGGTCCTCGAGTTA
DT.3	TGGAAAACGTTCTTCGGGGCG
DT.4	CGCCCCGAAGAACGTTTTCCANNSATGAGCACTNNSAAAGTTCTGCTATGTGGCGCGGTA
DT.5	CCTAGCGAGTTACATGATCCCCCATGTTGTGSNNAAAAGCGGTSNNCTCCTTCGGTCCTCC GATCGTTGTSNNAAGTAAGTTGGCCGCAGTGTTATCAC
DT.6	TGGGGGATCATGTAACTCGCTAGGATCGTTGGGAACCGGAGNNSAATGAAGCCATACCAA ACGACGAGC
DT.7	CTGGGGCCATAGGGTAAGCCCTCCCGTATCG
DT.8	CTTACCCTATGGCCCCAGTGCTGCAATG
DT.9	GCGAGTTACATGATCCCCCATGTTGTG
DT.10	GGGGATCATGTAACTCGCTTGGATCGTTGGGAACCGGAG
DT.11	GGGGATCATGTAACTCGCTATGATCGTTGGGAACCGGAG
DT.12	GGGGATCATGTAACTCGCATGGATCGTTGGGAACCGGAG
DT.13	GGGGATCATGTAACTCGCTGGGATCGTTGGGAACCGGAG
DT.14	GGGGATCATGTAACTCGCNNSGATCGTTGGGAACCGGAG
DT.15	CAGTCAGATTTTGAGACGATTAGAAAAACTCATCGAGCATCAAATGAAACTGCA
DT.16	GATATACTATGCCGATGATTCCTGGGGTGCCTAATGAGTGAGCTAAC
DT.17	TTTGCTCACCCAGAAACGCTGGT
DT.18	TCGCTGAGATAGGTGCCTCACTGATT
DT.19	AATCAGTGAGGCACCTATCTCAGCGA
DT.20	ACCAGCGTTTCTGGGTGAGCAAA
DT.21	CCCAGTCAGCTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTCTG
DT.22	CCTAATGAGTGAACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGG
DT.23	TGAAAAAGGAAGAGTTCACTCATTAGGCACCCCAATCATCG
DT.24	GGTCTGACAGCTGACTGGGTTGAAGGCTCTCAAGG

Table 3-7: Oligonucleotides used in this study

	$k_{cat}(s^{-1})$	k_{cat} st. dev.	K_M (μM)	K_M st. dev.
TEM-1	195.9	16.5	23.3	9.1
TEM-1.D254tag	189.1	13.8	32.7	9.6
TEM-1.B9	154.2	5.2	33.2	4.6

Table 3-8: TEM-1 β-lactamase kinetics

Kinetic parameters for wild-type TEM-1, TEM-1.B9, and TEM-1.D254tag, as determined with the β-lactam substrate nitrocefin.

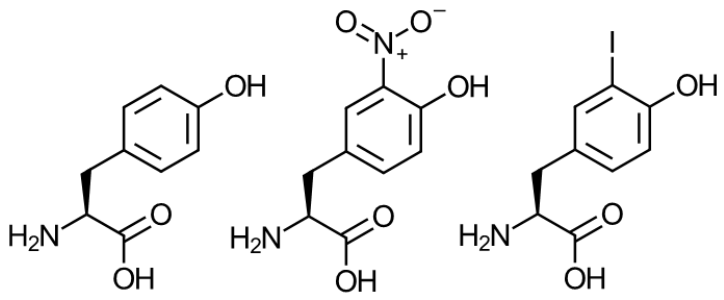

Figure 3-1: NCAA structures

The structure of the amino acids used in UAG-suppression OTSs throughout this study; the canonical amino acid L-tyrosine (Tyr, left), and the NCAAs 3-nitro-L-tyrosine (3nY, center) and 3-iodo-L-tyrosine (3iY, right).

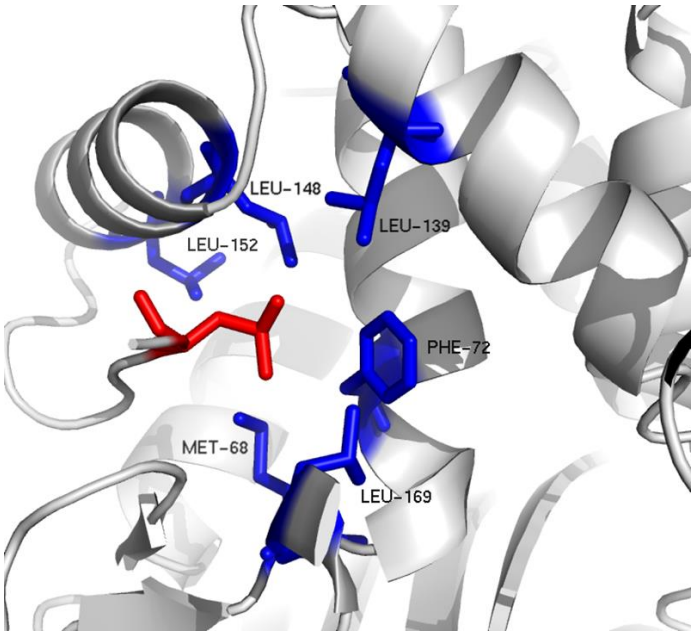

Figure 3-2: TEM-1 L162 Structure

Structure of wild-type TEM-1 β-lactamase (PDBID: 1XPB) at position L162 (red) with the amino acids randomized in the L162tag library (blue).

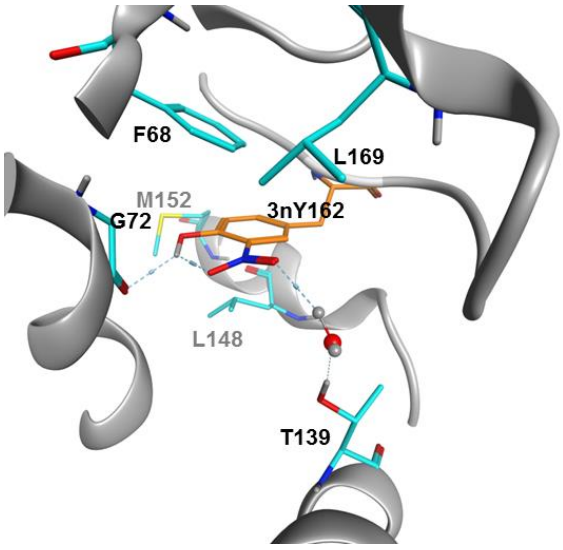

Figure 3-3: MOE modeled image of TEM-1.B9

Computational modeling of TEM-1.B9 focusing on the 3nY162 position and accommodating mutations.

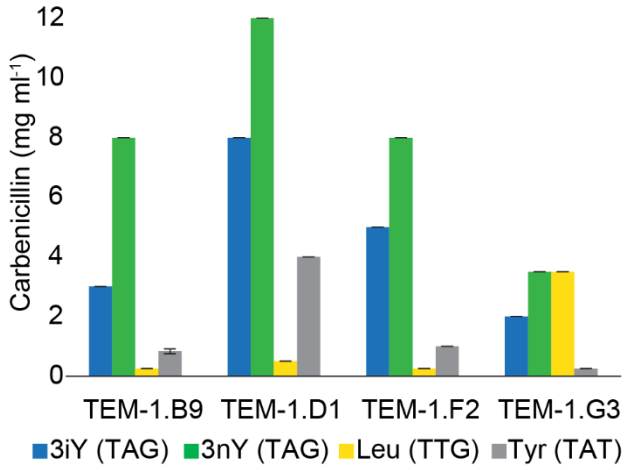

Figure 3-4: MIC characterization

Carbenicillin MIC values of four selected TEM-1 β-lactamase variants with 3iY, 3nY, leucine and tyrosine at position 162. 3nY and 3iY were incorporated at an encoded TAG codon. Tyrosine and leucine were coded for with cognate *E. coli* codons TAT and TTG respectively. Data represent mean ± s.e.m. of biological triplicates.

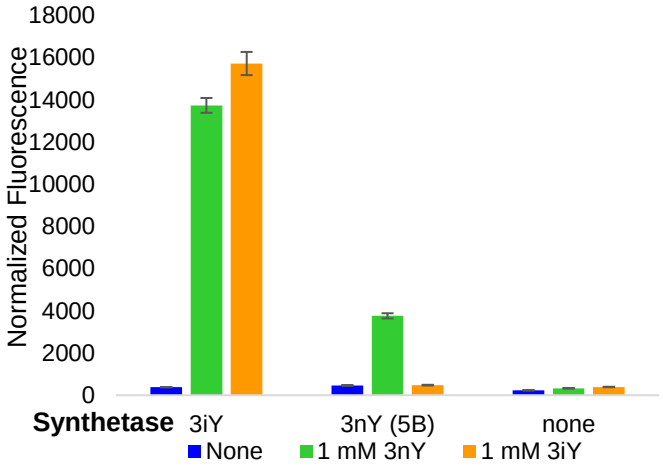

Figure 3-5: 3nY and 3iY incorporation efficiencies

Activities of engineered MjYRSs previously evolved to incorporate 3nY or 3iY. Each was characterized using 3iY and 3nY OTS in the presence of 1mM 3nY or 1mM 3iY. NCAA incorporation was quantified by measuring fluorescence of GFP-Y39tag. Values are mean ± s.d. of biological triplicates, technical duplicates of each triplicate.

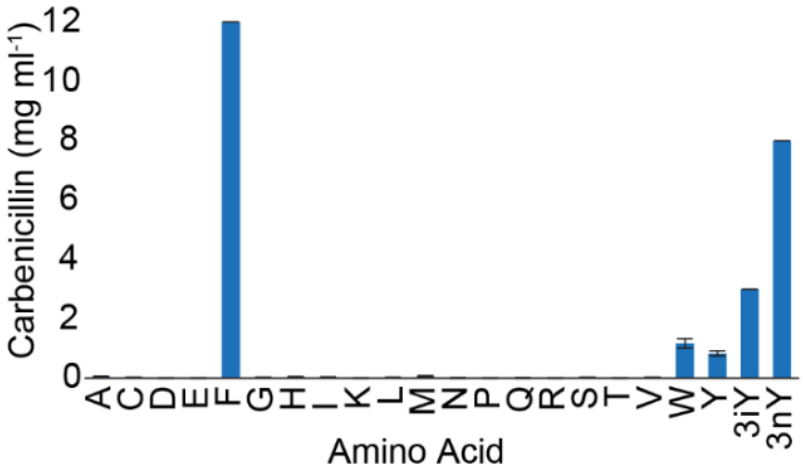

Figure 3-6: MICs of TEM-1.B9

MICs of TEM-1.B9 with each canonical amino acid and the NCAAs 3iY and 3nY at position 162. Data represent mean ± s.e.m. of biological triplicates.

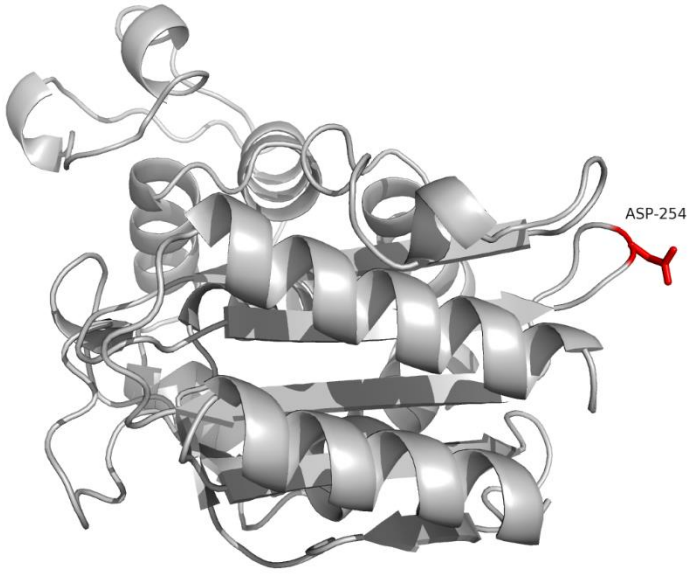

Figure 3-7: Structure of TEM-1.D254X

TEM-1 β-lactamase with solvent exposed loop amino acid D254 (red), used as a control throughout chapter three.

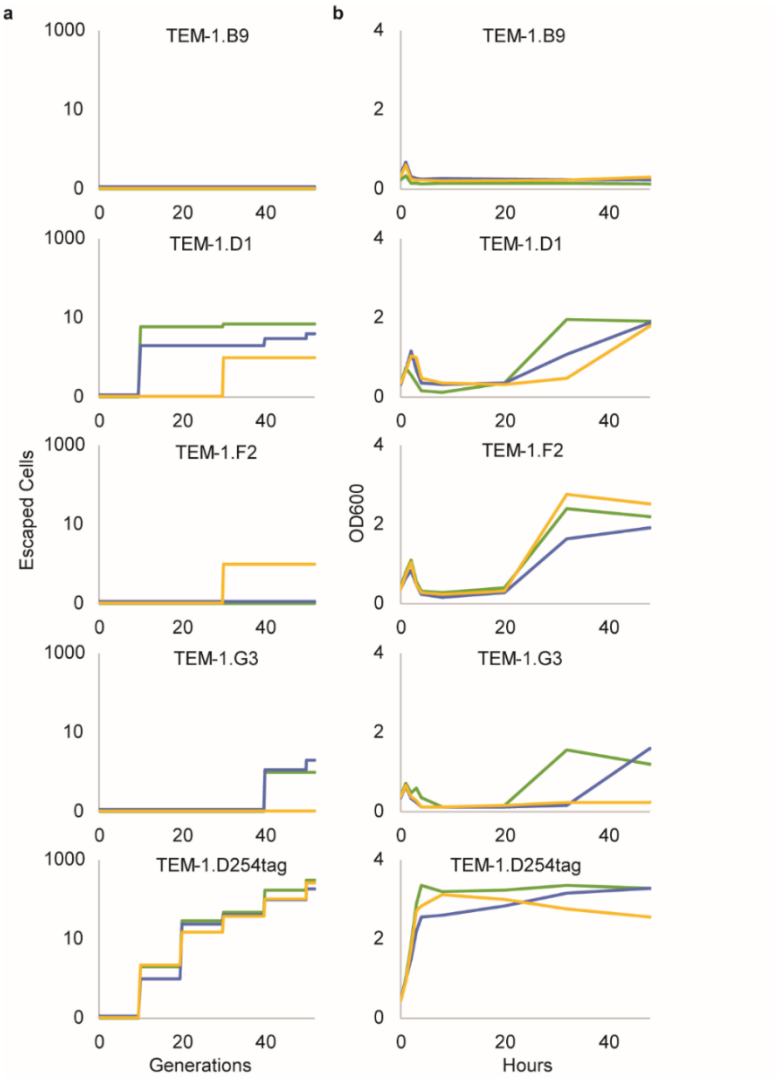

Figure 3-8: Monitoring NCAA dependence over time

(Left) Escape events of four selected clones and TEM-1.D254tag monitored through 50 generations. A sample of cells were screened for escape events every 10 generations, in triplicate.

Figure 3-9: Liquid culture assays for loss of addiction after 50 generations

(Right) After 50 generations of growth, a liquid media assay was used to screen large a larger *E. coli* population. Cultures containing carbenicillin were inoculated and growth in the absence of 3nY was monitored by OD_{600} measurements over 48 hours.

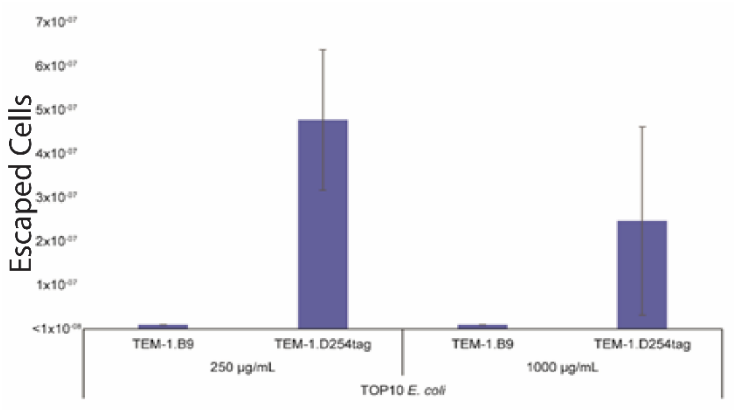

Figure 3-10: Bacterial escape at low antibiotic concentrations

Escaped cells of TOP10 *E. coli* at lower carbenicillin concentrations. Samples were passaged in triplicate for 100 doublings in 250 µg ml-1 and 1000 µg ml-1, and plated on solid media without 3nY to determine escape occurrences. TEM-1.B9 had no escape events. Data reported as mean of biological triplicates ± s.d.

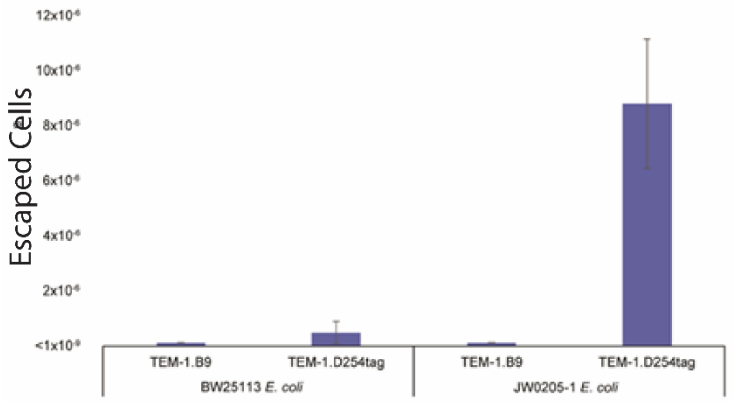

Figure 3-11: Escape occurances of mutation-prone *E. coli*

Strains from the Keio knockout collection (Baba et al., 2006) were passaged in triplicate for 100 doublings, and plated on solid media without 3nY to determine escape occurrences. Data reported as mean of biological triplicates ± s.d. The parental strain (BW25113), and a knockout of dnaQ, responsible for 3' to 5' proofreading of DNA polymerase III (Echols et al., 1983).

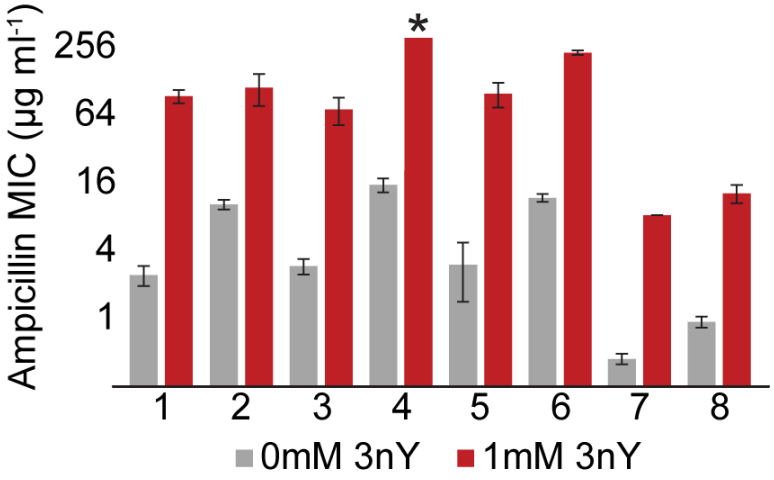

Figure 3-12: MICs of different bacterial species

MIC values of *E. coli* strains and different bacterial species in the presence and absence of 1 mM 3nY. (1) TOP10 *E. coli*; (2) XL1-Blue *E. coli*; (3) BL21 *E. coli*; (4) RTΔA *E. coli*; (5) *S. enterica*; (6) *S. flexneri*; (7) *Y. ruckeri*; (8) *A. baylyi*. Asterisk denotes that the value for RTΔA does not represent the MIC as the strain grew at the maximal ampicillin concentration tested; this is expected as 3nY incorporation does not compete with RF1-mediated termination in RTΔA.

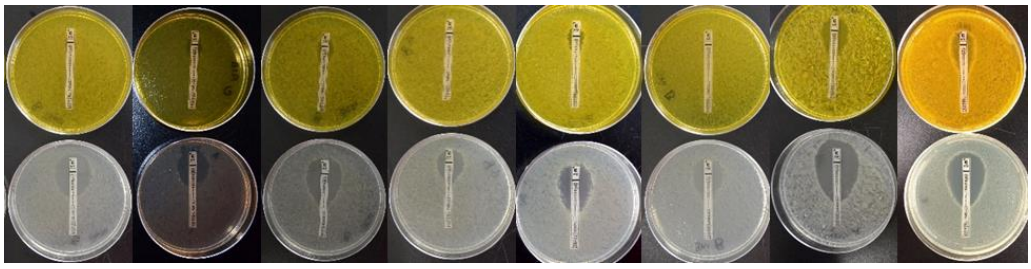

Figure 3-13: MIC plates of different bacterial species

MIC results for TEM-1.B9 in bacteria using MIC strip testing on LB-agar with 1 mM 3nY (top row) and without NCAA (bottom row). From left to right: Top10 *E. coli*, XL1-Blue *E. coli*, BL21 *E. coli*, RTΔA *E. coli*, *S. enterica*, *S. flexneri*, *Y. ruckerii*, and *A. baylyi*.

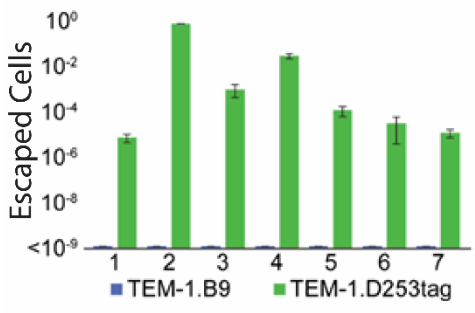

Figure 3-14: Escaped cells in different bacterial species

Escape frequency of *E. coli* strains and bacterial species after 100 generations. (1) TOP10 *E. coli*; (2) XL1-Blue *E. coli*; (3) BL21 *E. coli*; (4) RTΔA *E. coli*; (5) *S. enterica*; (6) S. *flexneri*; (7) *Y. ruckeri*. Notably, XL1-Blue cells contain the *supE* suppressor tRNA and are intrinsically ampicillin resistant with TEM-1.D254tag. Data represent mean ± s.e.m. of three or more biological triplicates.

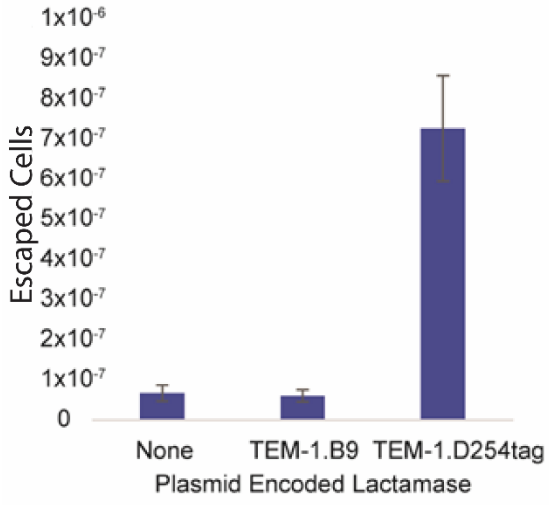

Figure 3-15: *Acinetobactor baylyi* escaped cells

Rates of 3nY independent ampicillin resistance development in *A. baylyi* ADP1 with and without plasmid encoded TEM-1 variants. TEM-1.B9 and TEM-1.D254tag were serially passaged to 50 doublings, and then a sample of 10^8 cells were plated onto LB-agar supplemented with 25 µg ml-1 ampicillin without 3nY. To determine the background rate of ampicillin resistance development in *A. baylyi* ADP1, wild type bacteria with no plasmids were grown to confluence and plated on 25 µg ml-1 ampicillin. Data reported as mean of biological triplicates ± s.d.

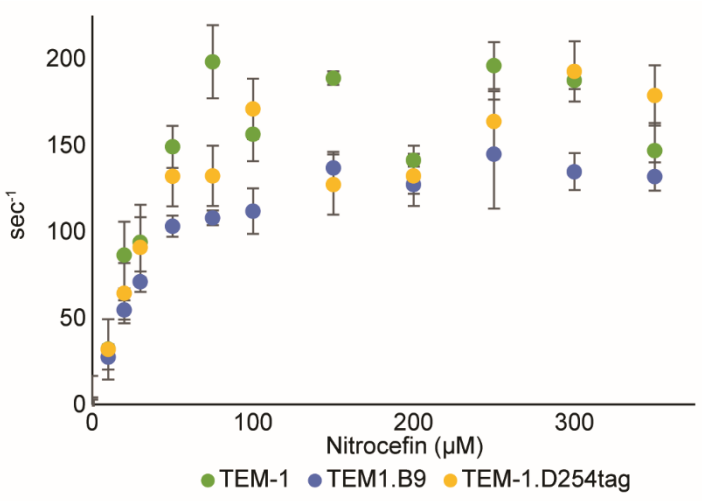

Figure 3-16: β-lactamase kinetics

Kinetics data from wild-type TEM-1, TEM-1.B9, and TEM-1.D254tag. Data points are the mean of two technical replicates ± s.d. Kinetic parameters are described in **Table 3-8**.

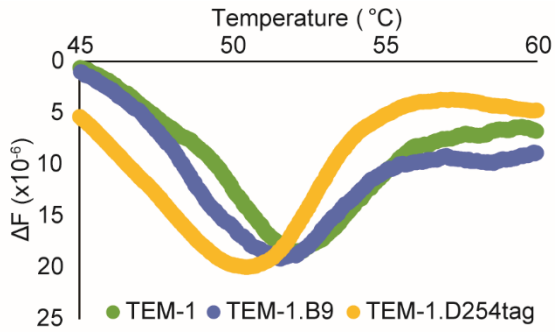

Figure 3-17: Melting curves of TEM-1 variants

Melting curves of wild-type TEM-1, TEM-1.B9, and TEM-1.D254tag. Melting temperatures were determined to be 52.1±1°C for wild-type TEM-1, 51.5±2°C for TEM-1.B9, and 50.4±1°C for D254tag.

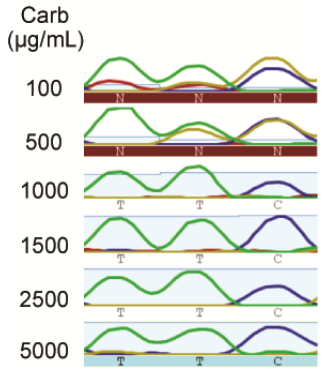

Figure 3-18: Sequencing reads of the TEM-1.B9 tag162NNS library

The single codon library was grown in liquid media at a gradient of carbenicillin concentrations. Sequencing results indicated phenylalanine was the only amino acid that provided enzyme activity at carbenicillin concentrations greater than 500 µg ml-1. This was reaffirmed through solid media selections, and a comprehensive analysis of TEM-1.B9 activity with all 20 amino acids replacing tag162.

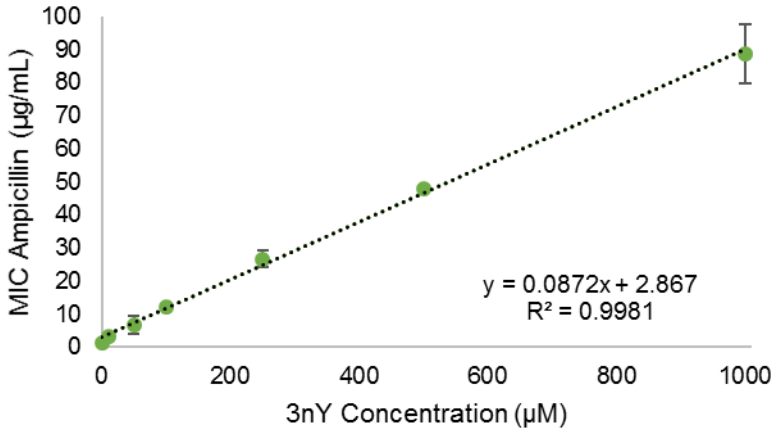

Figure 3-19: NCAA concentration dependent ampicillin resistance

Ampicillin MIC in Top10 *E. coli* as correlated to NCAA concentration. The MIC is directly correlated to 3nY concentration in the solid media. Data represent mean of biological triplicates ± s.d.

Chapter 4: Progress Towards NCAA-Chemistry Dependence

4.1 Introduction

4.1.1 Previous work on generating NCAA dependent enzymes

By evolving TEM-1 β-lactamase to function selectively upon site specific incorporation of the NCAAs 3nY or 3iY at the amber stop codon, we could maintain OTS activity in several bacterial species for at least one hundred generations without escape of NCAA dependence. The selected TEM-1 variant, TEM-1.B9, was structurally dependent on the physiochemical properties of 3nY or 3iY, and could not be compensated for with any amino acid within a single nucleotide polymorphism of the UAG stop codon. Despite this, the unique chemistry of the NCAAs was not used in any enzymatic role, but in a passive role for structural integrity of TEM-1. Additionally, while TEM-1.B9 successfully addicted several bacterial species, β-lactam antibiotics are entirely limited in use to prokaryotes, and largely Gram-negative eubacteria. NCAAs have been added to the genome of many organisms besides bacteria, including yeast and mammalian cells, and a more generalizable approach to NCAA addiction would have greater utility. Additionally, a lack of 3nY or 3iY was not inherently detrimental to a bacterium addicted to 3nY with TEM-1.B9, but required a supplemented β-lactam to be effective. An addiction system based on essential cellular function, which would become inactive in the absence of an NCAA could be used for OTS rentention, as well as biocontainment. Previous efforts to this end required significant reworking of a large number of genes throughout a genome (Mandell, 2015; Rovner, 2015). A more succinct effort, based on a single engineered enzyme would greatly facilitate trans-species NCAA addiction.

4.1.2 T7 RNA Polymerase

The RNA polymerase from T7 bacteriophage (T7 RNAP) is single protein capable of transcription without any additional proteins or cofactors (Schenborn and Mierendorf, 1985). T7 RNAP recognizes a small 17 base-pair promoter that is orthogonal to and functions well in prokaryotic and eukaryotic hosts (Conrad et al.; Lieber et al., 1989; Peretz et al., 2007; Studier and Moffatt, 1986). T7 RNAP has been used for *in vitro* RNA synthesis (Schenborn and Mierendorf, 1985), and *in vivo* RNA synthesis and protein overexpression (Studier and Moffatt, 1986). It is well-studied, with a comprehensive understanding of promoter recognition, and high resolution structural data in various conformations available (Cheetham et al., 1999; Temiakov et al., 2004; Yin and Steitz, 2004). T7 RNAP is amenable to engineering, and variants have been selected which recognize new promoter sequences while losing specificity for the native promoter (Meyer et al., 2015; Temme et al., 2012). These qualities make T7 RNAP an attractive target for engineering NCAA dependence.

Having previously demonstrated that a protein can be engineered to be structurally dependent on NCAA incorporation, while having no role in enzymatic function, we wished to use the unique physiochemical properties bestowed by NCAAs for an active site chemistry. Several previous studies have shown that T7 RNAP could be engineered to recognize a panel of orthogonal promoter sequences, and that these could be used simultaneously *in vivo* without cross reactivity (Meyer et al., 2015; Temme et al., 2012). While the T7 RNAP specificity loop is responsible for promoter recognition, the primary specificity-determining contacts are made through DNA promoter interactions with residues R746, N748, R756, and Q758 (**Figure 4-1**), and mutations in these amino acids often results in complete loss of T7 RNAP functionality (Imburgio et al., 2000). Indeed, the promoter specificity of T7 RNAP can be shifted for preference towards

unfavorable promoter sequences with just single amino acid substitutions in the promoter recognition loop (Imburgio et al., 2000; Raskin et al., 1993; Rong et al., 1998). As such, we desired to find promoters that were specifically recognized by NCAAs encoded at promoter interacting residues within T7 RNAP.

4.2 APPROACH

Since it is possible to change the promoter specificity of T7 RNAP with single amino acid substitutions at the promoter-interacting residues R746, N748, R756, and Q758 (Imburgio et al., 2000; Raskin et al., 1993; Rong et al., 1998), we chose to use these site for NCAA incorporation. We generated T7 RNAP variants with a UAG codon replacing one of each of the selected amino acids, and thus created four T7 RNAP variants; R746 (T7RNAP-R746X), N748 (T7RNAP-N748X), R756 (T7RNAP-R756X), or Q758X (T7RNAP-Q758X). Additionally, we created a control T7 RNAP with a UAG codon replacing the codon coding for D10 (T7RNAP-D10X), which is near the N-terminus of T7 RNAP, is not structurally important, and does not significantly affect T7 RNAP functionality (**Figure 4-2**). We used these constructs to search for promoters which were sequence specifically recognized by 3nY or 3iY at these positions.

4.2.1 T7 Promoter Recognition with NCAAs

We first needed to determine the activity of each T7 RNAP variant with the wild-type T7 promoter (P_{T7}). T7 RNAP expression was placed under the tetracycline expression system (Lutz and Bujard, 1997), which is tightly regulated, with minimal background, on the plasmid pRNAP. We used the *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase and corresponding tRNA that had previously been engineered to incoported 3iY, and has been shown to incorporate 3nY. We used a GFP reporter with the T7 promoter (P_{T7}) under control of the lacO operator system as an indicator of T7

RNAP transcriptional activity. Since mutations with canonical amino acids at positions R746, N748, R756, or Q758 disrupt promoter recognition and largely deactivate T7 based transcription, we expected that 3iY or 3nY replacing these positions would result in significantly reduced transcription, which in turn could be monitored with the GFP reporter under P_{T7}. We used *E. coli* strain RTΔA (Thyer et al., 2015) to test T7 RNAP variant activities, and determine background expression of GFP under different expression conditions; in the presence and absence of anhydrous tetracycline (aTc), to induce T7 RNAP expression, and the presence and absence of IPTG to induce GFP expression. We tested each in media lacking NCAA, and with .5 mM of NCAA (either 3nY or 3iY). The results (**Figure 4-3**) demonstrate that GFP expression, and thus T7 transcription, is dependent on IPTG, aTc, and NCAA supplementation. Additionally, these results confirmed that T7RNAP-D10X is functional when incorporating either 3nY or 3iY, and thus makes an appropriate positive control. Finally, and most importantly, these results confirm that T7 RNAP is largely deactivated with 3nY and 3iY substitutions at every tested position used for promoter recognition. These results confirm that R746, N748, R756, and Q758 are acceptable T7 RNAP variants to select promoters to complement, as T7 RNAP with NCAAs at any of these positions is unable to recognize its natural promoter.

4.2.2 GFP Selection

We designed a simple screen to select for promoters which would be recognized by T7 RNAP variants with 3nY or 3iY at each position. Since the four tested positions recognize a small region of the promoter consisting of five nucleotides, (positions -7 to -11 in P_{T7}) (**Figure 4-2**), it was reasonable to create a promoter library randomizing all five of these positions that could be selected with all four T7 RNAP variants. There exist

1024 possible promoters in this randomized space (4 nucleotide possibilities at each of the 5 randomized nucleotides, 4^5=1024). The library was generated in front of GFP, and exceeded 10-fold coverage of the theoretical library size. A subsampling of 10 randomly selected colonies showed good sequence variation at the randomized positions. The library was transformed into RTΔA harboring each of the T7 RNAP variants and pSC101-NCAA, and plated on LB-agar with aTc and IPTG and either 1 mM 3nY, or 1 mM 3iY. The following day, green colonies were selected for sequencing under blue light, (**Figure 4-4**), a similar ratio of fluorescent colonies was seen. To sample promoter diversity, we selected colonies with a range of fluorescent intensity from each T7 RNAP variant and each NCAA. A total of 96 were picked, 12 from each selection condition. Plasmids from each were subsequently purified and sequenced. Sequencing revealed an artifact in the promoter region of all 96 selected clones (**Figure 4-5** and **Figure 4-6**) which resulted in a predicted strong prokaryotic promoter, as determined by Neural Network Promoter Prediction (Reese, 2001).

4.2.3 CPR Selection of Active Promoters

The GFP selection failed to identify T7 promoters that were selectively recognized by T7 RNAP variants in an NCAA dependent matter, instead enriching artifacts that encoded cryptic consensus sequence *E. coli* promoters. As such, we determined the best way to proceed would be a methodology that allowed for sequential positive and negative selections. The positive selection would enrich for promoters that are active with a T7 RNAP variant, while the negative selection would eliminate any consensus sequence *E. coli* promoters that might arise, from artifacts or the N5 sequence. T7 RNAP is an excellent enzyme for selections, as any gene (protein or RNA) can be expressed with a simple T7 promoter. As such, we elected to go with compartmentalized

partnered replication (CPR), an emulsion based positive selection system which has been successfully used to engineer T7 RNAP previously (Ellefson et al., 2014; Meyer et al., 2015).

4.2.3.1 Positive Selection

CPR (**Figure 3-7**) uses the cellular production of *Taq* DNA polymerase as a selection criteria. Cellularly produced *Taq* DNA polymerase is then used for PCR amplification of the DNA sequence under selection. Cells are emulsified during PCR to ensure that *Taq* only amplifies the DNA from the host cell, and sequential rounds enrich for more active *Taq* producing sequences. We generated a T7 promoter library identical to that of the GFP screen, to drive the expression of *Taq* DNA polymerase. We conducted two rounds of CPR positive selection on each of the T7 RNAP variants with both 3nY and 3iY. We used a negative selection between the two positive rounds to eliminate any cryptic promoters which may have arisen during library assembly.

4.2.3.2 Negative Selection

For a negative selection, we used a plasmid encoded *E. coli* phenylalanyl-tRNA synthetase (*pheS*) containing the mutation A294G (m*pheS*). This mutation relaxes amino acid specificity, allowing for the incorporation of several aromatic NCAAs into the genome at phenylalanine codons (see **Chapter 1**). When growth media is supplemented with *p*-chloro-L-phenylalanine (PCPA) and m*pheS* is expressed, PCPA is incorporated throughout the genome and host fitness is affected, or the host become nonviable (Ibba and Hennecke, 1995). This negative selection has been used as a negative selection with CPR previously (Maranhao and Ellington, 2016). After the first round of positive selection with CPR, the library was cloned into the m*pheS* negative selection plasmid backbone, and transformed into TOP10 *E. coli*. Transformant cultures were grown

overnight in 5 mM PCPA and 100 μM IPTG. After overnight growth, the plasmid was purified from the culture and the promoter sequences were amplified using PCR, and cloned into the *Taq* expression vector for a second round of CPR.

4.2.3.3 GFP Characterization

After a second round of CPR positive selection, the selected promoters were cloned into in front of a GFP expression cassette, and transformed into RTΔA *E. coli* containing pSC101-NCAA and the proper T7 RNAP variant. Transformants were plated and grown overnight and then individual colonies from each condition were picked and tested for GFP expression. We tested each selected promoter with the T7 RNAP variant and NCAA with which it had been selected. We tested 16 from each condition, for a total of 160 promoters (D10X, R746X, N748X, R756X, and Q758X). We identified two potential promoters from this characterization, one which showed a 10-fold increase over background fluorescence (promoter D3, P_{D3}) and one which showed a 2-fold increase (promoter A2, P_{A2}). Both of these promoters were selected for activation with 3nY incorporated into the promoter recognizing region, P_{A2} was selected with T7RNAP-R746X, and P_{D3} was selected with T7RNAP-N748X. We sequenced both promoters, P_{D3} was a single nucleotide away from the wild-type cognate T7 promoter, (GACTC to GCCTC), and P_{A2} had three mutations from the wild-type promoter (GACTC to GGTTG).

We cloned the two identified promoters, P_{D3} and P_{A2} into another GFP construct and retested to verify promoter activity with the corresponding T7 RNAP variants. We also wanted to explore possible cross-reactivity with T7RNAP-D10X, to determine orthogonality to wild-type T7 RNAP (**Figure 4-8**). Our results verified P_{D3} activity with 3nY incorporated into T7RNAP-748X, but also showed that P_{D3} was recognized by

T7RNAP-D10X with near cognate efficiency, suggesting that our promoter was not specific for 3nY. Upon recharacterization, we saw no increase in GFP expression with P_{A2}. Additional characterizations of P_{A2} showed no increase in GFP expression with any T7 RNAP variants.

4.3 DISCUSSION

T7 RNAP is an attractive target for NCAA addiction. It is functional in a wide range of organisms, can be used to drive positive and negative selections, and can be incorporated into genomes to drive essential pathways. Utilization of an NCAA in promoter recognition is possible, as was demonstrated here with P_{D3}. While P_{D3} was not orthogonal to wild-type T7 RNAP recognition, this work demonstrates that promoters can be made to accommodate NCAAs in the promoter recognition domain of T7RNAP. The theoretical library size of the promoter selected here (~1000 variations) is relatively small, and while R746, N748, R756, and Q758 only recognize these nucleotides, upstream or downstream effects in the promoter, or other T7 RNAP promoter recognition domain may play some role in recognition, as seems apparent in previous studies (Meyer et al., 2015), and a larger library could increase the chances of finding NCAA specific promoters. Alternatively, accommodation of 3nY or 3iY in the positions selected could have significant structural effects on T7RNAP, which could inactivate some T7 RNAP variants entirely. Additionally, selection with NCAAs capable of unique interactions would likely be beneficial, amino acids such as DOPA, which can readily form multiple hydrogen bonds, or metal chelating amino acids might provide unique interactions which could aid in nucleic acid recognition in a way that may not be compensated with the canonical genetic code.

4.4 Materials and Methods

4.4.1 Strains and Reagents

TOP10 *E. coli* (Life Technologies) was used for routine cloning of plasmids and the m*pheS* negative selection. Amberless *E. coli* strain RTΔA (derived from C321.ΔA) was described previously (Thyer et al., 2015) was used for positive selection with CPR. NCAAs 3-nitro-L-tyrosine, 3-iodo-L-tyrosine, as well as PCPA were purchased from Sigma-Aldrich. LB media (Fisher BioReagents) was used for culturing all *E. coli* in all conditions, and LB-agar (Fisher BioReagents) was used for plating of cultures and cells for screens and cloning. Antibiotics were used at concentrations of 50 μg ml^{-1} kanamycin, 100 μg ml^{-1} ampicillin or carbenicillin, and 150 μg ml^{-1} spectinomycin. Libraries and plasmids were constructed with oligos purchase from Integrated DNA Technology. Plasmids were sequence verified with Sanger sequencing at the University of Texas at Austin Core Facilities.

4.4.2 Plasmid Construction

Plasmid pSC101-NCAA was constructed using the pNCAA plasmid (Ellefson et al., 2016) as a template for OTS, and inserted into the pTH18kr backbone (GenBank AB019603.1) which contains the low copy SC101 origin (Hashimoto-Gotoh, 2000) and kanamycin resistance marker, using Gibson cloning (Gibson, 2009). T7 RNAP plasmid was provided by Shaunak Kar at the University of Texas at Austin. Briefly, it was built upon the TetR/O regulatory system (Lutz and Bujard, 1997) with the P15A origin of replication, and a weak RBS (BioBricks Part Number BBa_J61101). In-frame UAG codons were added using the Gibson method for site directed mutagenesis (Gibson, 2009). GFP, *Taq* DNA polymerase, and m*pheS* were encoded in the pCDF-duet backbone (Novagen), with the CloDF origin and spectinomycin resistance marker. DNA

sequencing surrounding encoded T7 promoter were reduced to a minimal sequence to eliminate cryptic prokaryotic promoter formation.

4.4.3 Library Construction

DNA oligos with 5 randomized positions were used to amplify pCDF-GFP or pCDF-*Taq*. PCR products were purified with agarose gel electrophoresis, and reassembled with Gibson method (Gibson, 2009). Individual clones, as well as library mixture, were sequenced to verify random promoter sequences with no detected biases.

4.4.4 GFP Screen

Plasmids pSC101-NCAA and pTET-T7 for each T7 RNAP variant were transformed into RTΔA *E. coli* and plated on plates supplemented with kan and amp. A single colony from each was selected and grown to midlog phase (OD_{600} approximately 0.6), and made electro-competent using serial glycerol washes, followed by electroporation as described above with purified promoter library plasmid. Transformants were recovered in LB media for 1 hour at 37°C and plated at various dilutions on agar plates supplemented with spec, kan, amp, 100 μM IPTG and 100 ng ml^{-1} aTc, as well as .5 mM 3nY or 3iY. Plates were incubated overnight at 37°C, and imaged under blue light. Green colonies of various fluorescent intensities were selected for sequencing. Sequences were analyzed for prokaryotic promoters using Neural Network Promoter Prediction (Reese, 2001).

4.4.5 CPR Positive Selection

pSC101-NCAA plasmid and pTET-T7 plasmid, containing each T7 RNAP variant, were transformed into RTΔA *E. coli* and plated on plates supplemented with kan and spec. A single colony from each transformation was selected and made electrocompotent as described above, and transformed with p*Taq* plasmid with promoter

library. Cultures were recovered at 37°C in LB for 1 hour, then established overnight in LB supplemented with spec, kan, and amp.

A 100 µL aliquot from each overnight culture was used to inoculate 5 ml of fresh LB with .5 mM 3ny or 3iY, and kan, spec, and amp. Cultures were grown at 37°C for 1 hr then induced with 50 ng ml^{-1} of aTc and 100 µM IPTG, and incubated for 3 more hours at 37°C. After the 3 h expression, cells (300 µl) were spun down (15 min, 1500*g*, 4 °C) and washed twice with *Taq* buffer (50 mM KCl, 10 mM Tris-HCl pH 8, 1.5 mM $MgCl_2$). Washed cells were resuspended in 300 µl CPR buffer (50 mM KCl, 10 mM Tris-HCl pH 8, 1.5 mM $MgCl_2$, 200 µM dNTP, 0.3 nM each CPR1 primer, 100 µM).

Resuspended cells were added to a 2 ml microtube containing: 876 µl Tegosoft (Evonik), 240 µl mineral oil (Sigma), 84 µl AbilEW09 (Evonik), and the rubber tip from a 1 ml syringe. This mixture was chilled at 4°C while rotating for 5 minutes. The mixture was emulsified using a TissueLyser LT (Qiagen), which was operated at 42 Hz for 4 min at 4°C. Emulsified cells were dispersed in 100 µL aliquots and thermal cycled [95°C:3 min, 20 cycles of (95°C:30s, 59°C:30s, 72°C:10s), 72°C:5 min]. Post PCR emulsions were pooled and centrifuged (10 min at 12000*g*), and the top layer of oil was removed. The remaining emulsion was broken by applying an equal volume of phenol:chloroform mixture (1:1) (Sigma) and vortexing. The broken emulsion was then applied to a PhaseLock tube (5 Prime) and centrifuged for 2 minutes at 16000*g*. The top aqueous fraction (upper layer) was transferred to a microtube and an equal volume of chloroform (Sigma) was added, vortexed and placed in a new PhaseLock tube and centrifuged for 2 minutes at 16000*g*. The aqueous layer from this second extraction was collected and column purified (Zymo Research Corp). The elution from column purifications was used as template for the secondary recovery PCR using Pfusion polymerase (Invitrogen) and CPR2 primers specific to the overhang regions of CPR1 primers. These secondary PCR

products were run on a 1.5% agarose gel. The band corresponding to the desired PCR amplicon was excised from the agarose gel and extracted (Promega). This purified secondary amplicon was used as template for the third PCR, which generated sufficient product for Gibson cloning into the pCDF-m*pheS* plasmid and transformed into TOP10 *E. coli*.

4.4.6 PCPA Negative Selection

Transformation efficiencies for the Gibson cloning after CPR positive selection exceeded 10^3 on each selection condition, as determined by dilution plating. The promoter libraries in pCDF-m*pheS* were recovered in LB at 37°C for 1 hour, then added to LB supplemented with spec, 1 mM IPTG, and 5 mM PCPA and incubated overnight for no more than 12 hours. After 12 hours, plasmids from cultures were purified and used as a template for PCR with CPR3 primers. PCR product was run on 1.5% agarose gel and band extracted as described above. Purified product was used to Gibson clone into p*Taq* backbone, which were then used for a second round of positive CPR selection.

4.4.7 GFP Characterization

After the second round of CPR positive selection, gel purified PCR product containing the selected promoter library was cloned into pCDF-GFP plasmid. The purified promoter library was transformed into RTΔA containing pSC101-NCAA and appropriate T7 RNAP variant as described above, and plated on kan, amp, and spec supplemented LB-agar and incubated at 37°C for 12 hours. Individual colonies were selected and grown overnight in LB with kan, amp, and spec, and incubated for 16 hours overnight. Saturated overnight cultures were diluted 1:1000 to inoculate 1 ml of fresh LB supplemented with .5 mM 3nY or 3iY as appropriate, kan, amp, and spec at half normal concentration, after 1.5 hours, 50 ng ml^{-1} of aTc and 100 μM IPTG were used to induce

cultures, and incubator an additional 5 hours. After induction period, cultures were serially centrifuged at 2000*g* for 10 minutes and washed with PBS. Fluorescence was measured in Tecan M200 pro, using 475 nm excitation and reading at 525 nm wavelength for emission, and normalized to OD_{600} measurement. D10 with P_{T7} and three samples from P_{LIB} were used as controls on each plate. The same protocol was followed for recharacterization of P_{D3} and P_{A2}.

4.5 TABLES AND FIGURES

Name	Sequence
10Xf	TCGCTAAGAACTAGTTCTCTGACATCGAACTGGC
10Xr	AGAACTAGTTCTTAGCGATGTTAATCGTGTTGGATCC
746Xf	ACAAGAAGCCTATTCAGACGTAGTTGAACCTGATGTTCCTCGGTCAGTTCCGC
746Xr	CGTCTGAATAGGCTTCTTGTATTCCTGCCACACAGG
748Xf	GAATACAAGAAGCCTATTCAGACGCGCTTGTAGCTGATGTTCCTCGGTCAGTTCCGCTTACAGC
748Xr	CGTCTGAATAGGCTTCTTGTATTCCTGCCACACAGG
756Xf	GAACTGACCGAGGAACATCAGGTTCAAGCGC
756Xr	GATGTTCCTCGGTCAGTTCTAGTTACAGCCTACCATTAACACCAACAAAGATAGCGAGATTGATGC
758Xf	GAACTGACCGAGGAACATCAGGTTCAAGCGC
758Xr	GATGTTCCTCGGTCAGTTCCGCTTATAGCCTACCATTAACACCAACAAAGATAGCGAGATTGATGC
CPR1 f	GCAAAACCGATCAGGCGCCCTTCCCGGCTCTCCCTTATGCGACTCCTGCATTAGGAAATTAATAC
CPR1 r	CGGATCTCAGCACTGCAGCGGTGGTCTCCCCATGGTATATCTCCTTAACAATTGTTATCCGCTCACAATTCCCCTATAGT
CPR2 f	GCAAAACCGATCAGGCGCCCTTCCCGG
CPR2 r	CGGATCTCAGCACTGCAGCGGTGGTCTCCC
CPR3 f	CTCTCCCTTATGCGACTCCTGCATTAGGAAATTAATAC
CPR3 r	CATGGTATATCTCCTTAACAATTGTTATCCGCTCACAATTCCCCTATAGT
CDFbbf	ACTATAGGGGAATTGTGAGCGGATAACAATTGTTAAGGAGATATACCATG
CDFbbr	GTATTAATTTCCTAATGCAGGAGTCGCATAAGGGAGAG
CDFTLibf	GCGACTCCTGCATTAGGAAATTAATACNNNNNACTATAGGGGAATTGTGAGCGGATAACAATTG
CDFTLibr	GTATTAATTTCCTAATGCAGGAGTCGCATAAGGG
pTHf	CAAGCTTATCGATGTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTC
pTHr	GAGAGCCTTCGCCTGGGGTGCCTAATGAGTGAG
OTSf	ACCCCAGGCGAAGGCTCTCAAGGGCATCGGTC
OTSr	GATGAGTTTTTCTAACATCGATAAGCTTGATGCGTGGAAGATTGATCG

Table 4-1: Primer sequences

Sequences of the primers used in **Chapter 4**

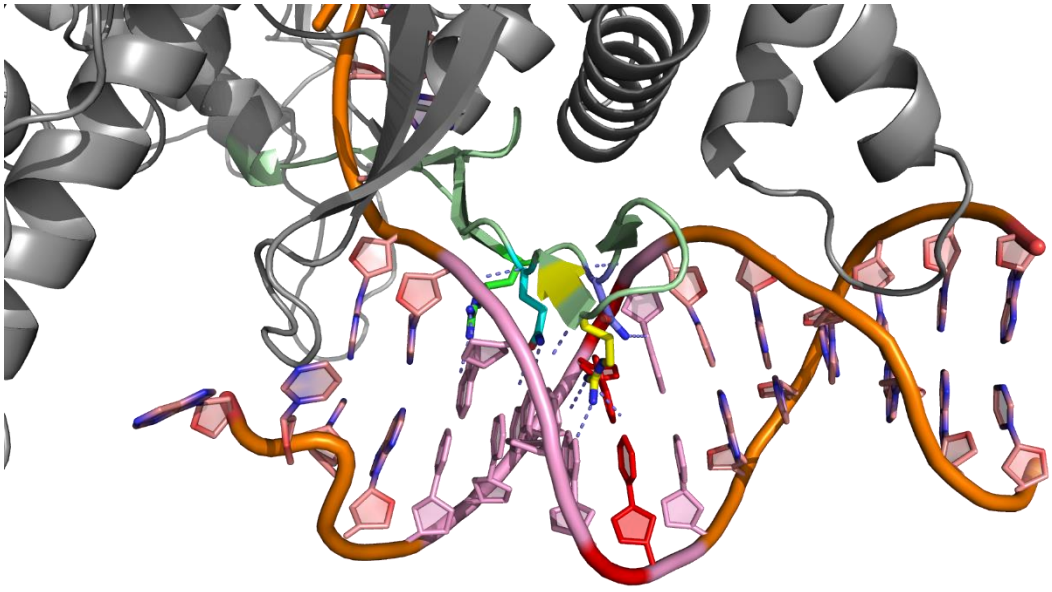

Figure 4-1: T7 RNAP promoter recognition

Structural basis of T7 RNAP promoter recognition. T7 RNAP specificity loop (pale green) interacts with the DNA promoter sequence. Residues R746 (green), N748 (blue), R756 (yellow), and Q758 (teal) recognize nucleotide positions -7 to -11 within the T7 promoter seqeunce (pink/red nucleotides). Pink and red nucleotides were randomized in the T7 promoter library, and screened against T7 RNAP variants with 3nY or 3iY replacing four colored amino acides in T7 RNAP specificity loop.

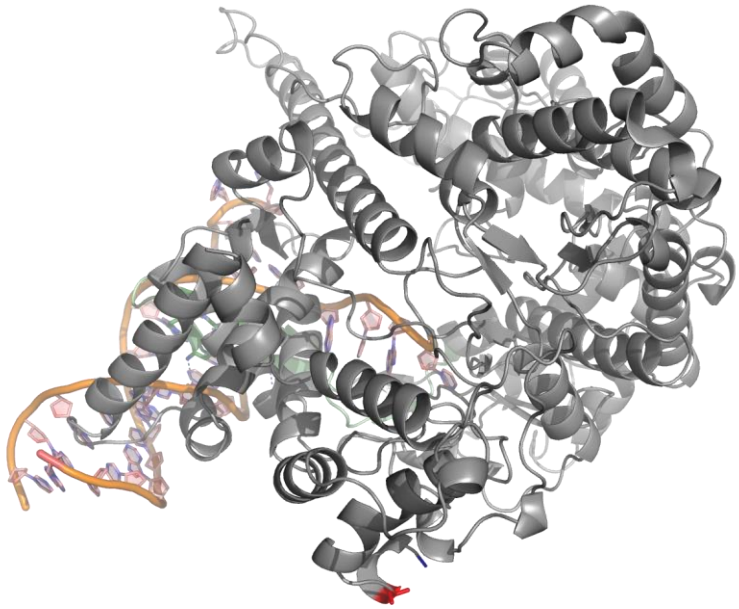

Figure 4-2: T7 D10X Control

Structure of the D10X positive control used throughout this work. D10 (bottom, red), does not affect T7 RNAP activity, but requires UAG suppression to function (**Figure 4-3**).

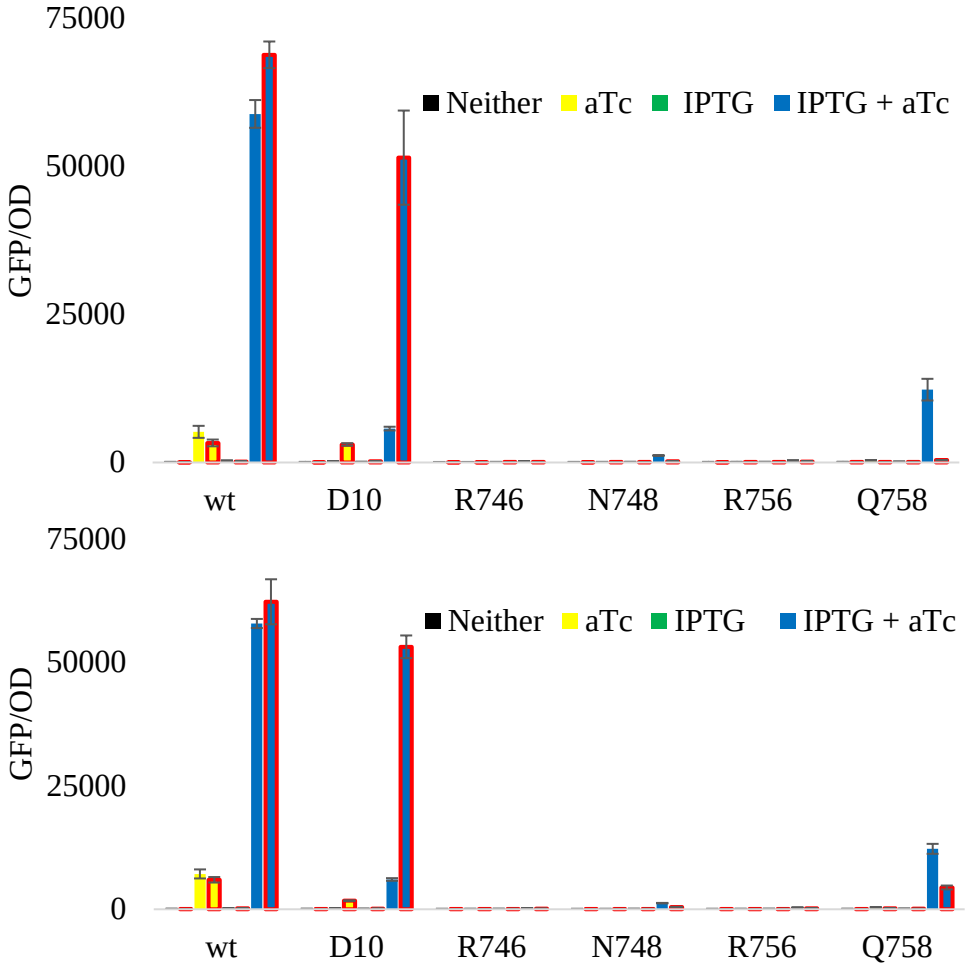

Figure 4-3: T7 RNAP functionality assays

T7 RNAP functionality assays as determined by a GFP reporter. T7 RNAP activity was tested in media without aTc or IPTG (black bars), in media with 100 ng ml^{-1} aTc (yellow bars), in media with 500 µM IPTG (green bars), and media supplemented with aTc and IPTG (blue bars). All were tested without NCAA supplementation (no border), and with 500 µM of NCAA (red borders), using both 3nY (top), 3iY (bottom). Wildtype T7 RNAP was active with aTc and IPTG, while control D10X required NCAA supplementation to be fully active. R746X, N748X, R756X, and Q758X were largely inactive under all conditions, except for Q758X, which has some activity without the NCAA added, suggesting background UAG suppression of some natural amino acid results in active T7 RNAP.

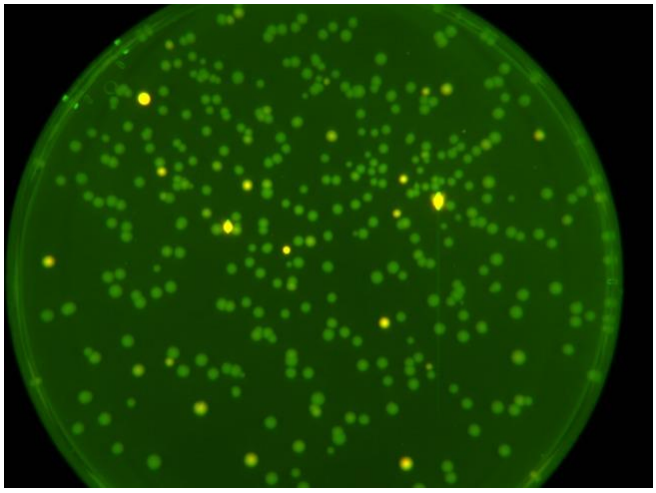

Figure 4-4: GFP screen result

Imaging of promotor library driving GFP from a library selection plate. Green colonies were selected for sequencing (**Figure 4-5** and **Figure 4-6**)

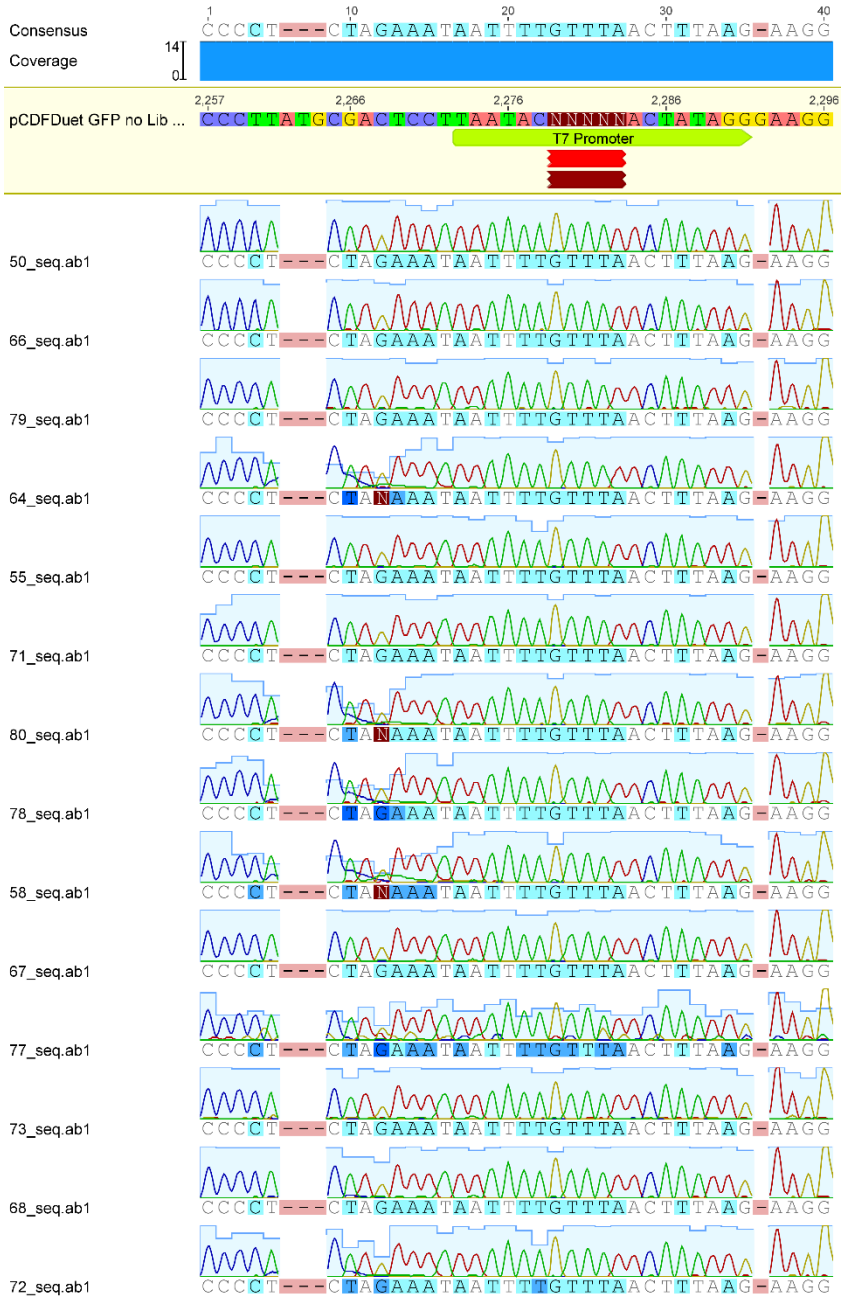

Figure 4-5: GFP screen T7 promoter sequences

Sequencing of promoters selected from P_{lib} GFP selection plates. Green colonies were selected for sequencing and all selected clones revealed a promoter artifact that expressed GFP in a T7 independent manner using a predicted *E. coli* promoter.

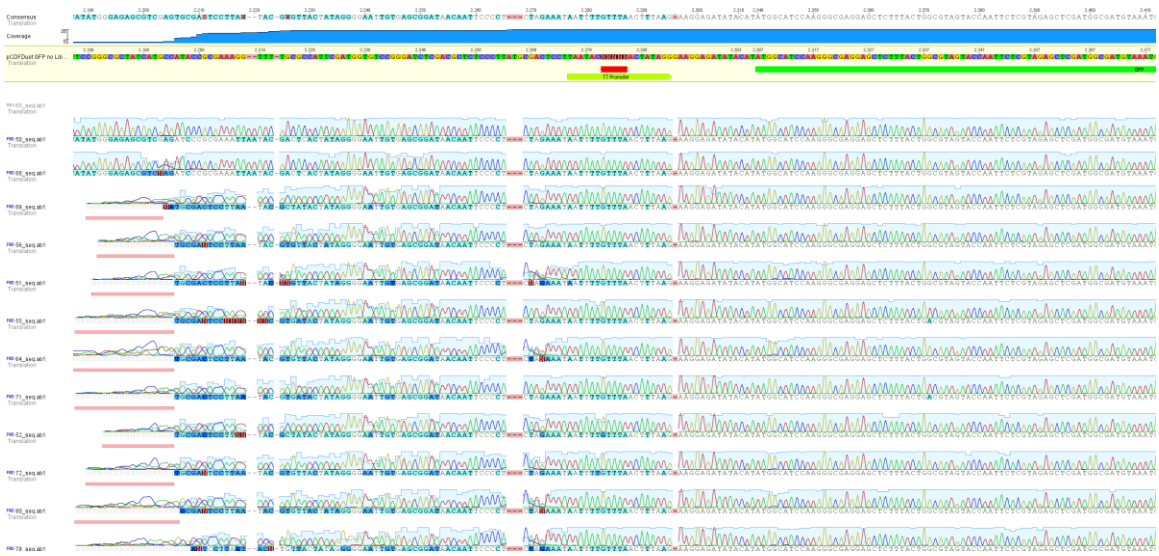

Figure 4-6: GFP cryptic promoter sequences

Imaging of promotor library GFP selection plate. Green colonies were picked for sequencing, and revealed a promoter artifact that expressed GFP in a T7 independent manner.

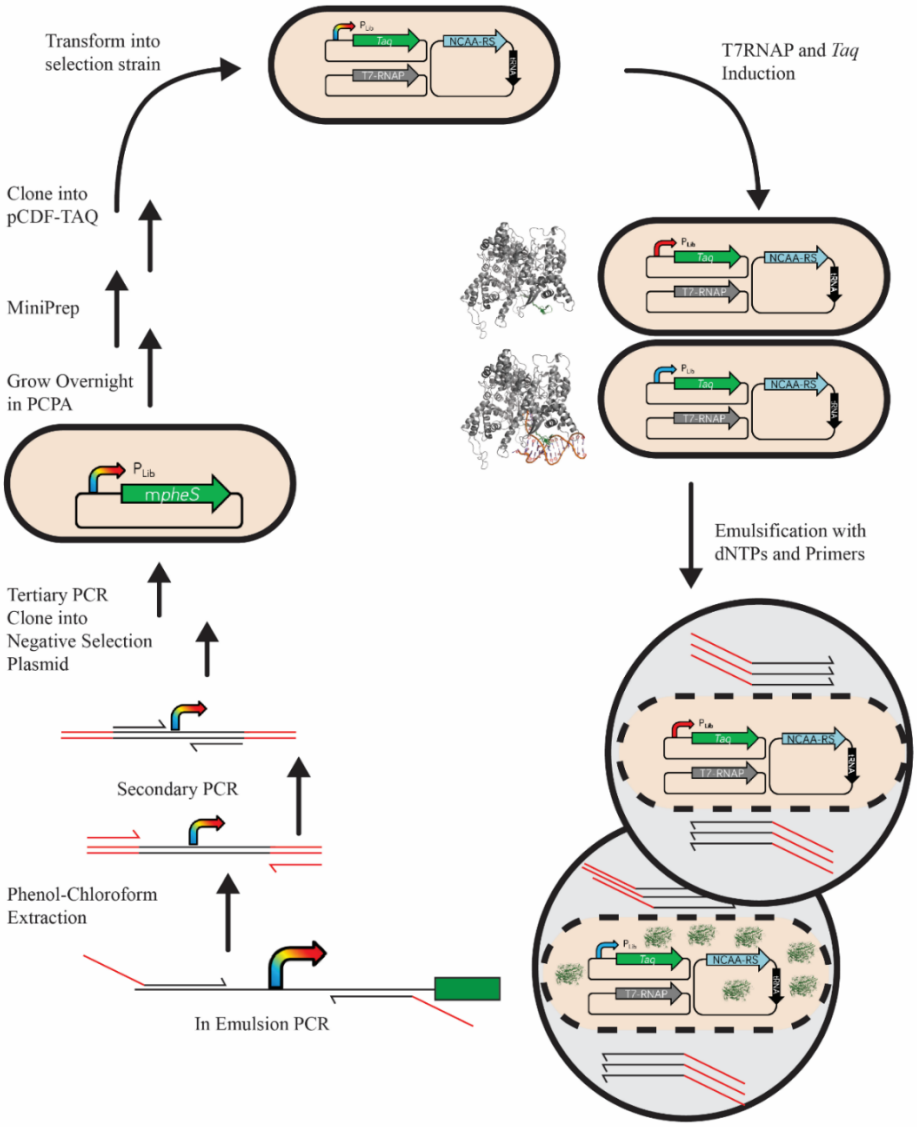

Figure 4-7: CPR Selection Scheme

The P_{Lib} library was built to drive the expression of *Taq* DNA polymerase. Cells containing promoters which were read by the T7 RNAP variant they contained would express and build up *Taq* DNA polymerase. Cells were emulsified with primers to amplify the promoter sequence, and if *Taq* was present, PCR would successfully amplify the active promoters. Active promoters were then cloned into a negative selection vector, to remove cryptic promoters, as seen in the GFP selection attempt. A second round of CPR positive selection was performed, followed by GFP characterization of selected promoters.

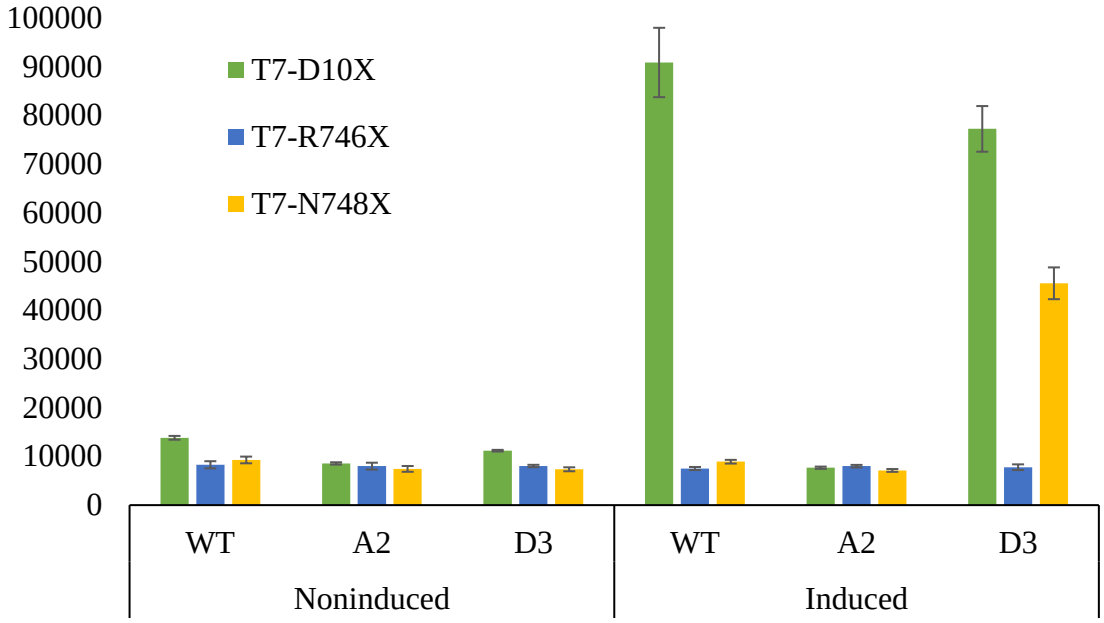

Figure 4-8: NCAA recognized promoter characterization

GFP assays to determine the activity of promotors identified after two CPR rounds of positive selection and a PCFA negative selection against T7 RNAP variants. The promoter P_{D3} is recognized by the T7 RNAP variant T7-N748X with 3nY incorporation. P_{D3} is also recognized by wild-type T7 RNAP, with higher expression rates.

Chapter 5: Evolution with an Expanded Genetic Code

Evolution has been limited to the 22 proteinogenic amino acids of the natural genetic code. While several theories on the evolutionary origin of the code have been proposed, experimental exploration of these theories is largely restricted to computational modeling and phylogenetic evaluation. Advances in genetic code expansion, through the addition of noncanonical amino acids (NCAAs), provide tools to explore proposed mechanisms of genetic code evolution. Several recent reports demonstrate that expanded genetic codes can provide fitness advantages through the unique physiochemical properties of NCAAs or by creating new mutational routes for evolution. Here we report the first long-term bacterial evolution of a genetically unaltered wild-type bacteria with an ambiguously translated amber (UAG) stop codon. Our results reveal several adaptive mutations in the addiction cassette and across evolved genomes. Evolved cells also decreased amber suppression efficiency, and increased fitness as measured by growth rates.

5.1 Introduction

Terrestrial evolution has been confined to the natural genetic code comprising the 20 canonical amino acids as well as selenocysteine and pyrrolysine, two rare, naturally occurring amino acids. These amino acids are encoded by nucleotide codons, which are decoded in the ribosome using transfer RNAs (tRNAs). In total, 64 codons are available, of which 61 signal for codon-specific amino acids and three signal translational termination. The genetic code was established early in evolutionary history, and is largely shared among all life. While several theories exist for the evolution of the genetic code, experimental evaluation of proposed theories has largely been limited to computational modeling and genomic analysis. Proposed processes in the evolution of codon

arrangement and assignment include codon capture (Osawa and Jukes; Osawa et al., 1992), ambiguous intermediate codons (Schultz and Yarus, 1994), and genome streamlining (Andersson and Kurland, 1998). None of these evolutionary processes are yet observed in controlled laboratory conditions. Despite the lack of direct observation, the existence of alternative codes (e.g. mitochondrial genomes), which likely evolved from the standard codon table, provides evidence that the genetic code is not frozen in its current state (Knight et al., 2001).

Orthogonal translation systems (OTSs) comprising aminoacyl tRNA synthetase (aaRS)/tRNA pairs independent of the hosts translational machinery enable addition of new chemistries to the genetic code *in vivo*. These OTSs are typically engineered to incorporate noncanonical amino acids (NCAAs) at the amber (UAG) stop codon, thereby expanding the set of proteinogenic amino acids available to the host organism. The use of UAG as the NCAA codon results in an ambiguously read codon throughout the genome; during translation, amber codons can be read by release factor 1 and signal translational termination or be read as a cognate codon by the orthogonal tRNA and incorporate an NCAA into the translating protein. In bacteria, UAG codon ambiguity has significant fitness effects, and without a functional dependence on NCAA incorporation *E. coli* populations rapidly deactivated the encoded OTS (Tack et al., 2016; Wang et al., 2014). These fitness effects are due in part to the elongation of cellular proteins which naturally terminate with the UAG codon (Mukai et al., 2011). When UAG is read as a cognate codon with near 100% efficiency, *E. coli* is severely compromised or nonviable (Johnson et al., 2012; Mukai et al., 2011). The fitness cost of OTSs can be mitigated by reassigning a number or all genomic TAG codons to alternative stop codons (Isaacs et al., 2011; Mukai et al., 2015).

Evolutionary studies have revealed a number of adaptive approaches bacteria take to increase fitness over short evolutionary time periods (Gayán et al., 2016; Horinouchi et al., 2015; Lenski et al., 1991). Engineered OTSs can incorporate chemistries not typically available with canonical amino acids including covalent crosslinking, metal chelation, and redox chemistry. These new chemistries can enable functions previously inaccessible through adaptation and evolution. Pyrrolysine and selenocysteine are chemically unique proteinogenic amino acids which are irreplaceable in their natural context, demonstrating that expanded codes have evolved and become essential in nature. It was recently shown that T7 bacteriophage with access to an expanded genetic code would evolve to use 3-iodotyrosine in its genome to reach new fitness peaks (Hammerling et al., 2014). More recently, it was demonstrated that NCAAs assigned to UAG codons can provide new evolutionary paths to rifampicin resistance (Hammerling et al., 2016). These studies used engineered *E. coli* lacking release factor 1, which eliminates competition for the UAG codon, allowing for nearly cognate UAG translation in the ribosome. Evolutionary exploration of ambiguous codons in wild-type *E. coli* has not been demonstrated, likely due to the inactivation of OTSs during short evolutionary time periods (Tack et al., 2016; Wang et al., 2014).

5.2 Approach

5.2.1 Experimental Design

We wished to examine the long-term adaptation and evolution of *E. coli* strains addicted to an unnatural amino acid, 3-nitrotyrosine (3nY). As a chassis for evolution, we chose to use *E. coli* strain MG1655 because it is well-characterized, with a sequenced and annotated genome (Blattner et al., 1997). MG1655 is autotrophic for all 20 canonical amino acids allowing for robust growth in amino acid knockout media.

MG1655 transformed with a plasmid encoding a non-reverting β-lactamase (TEM-1.B9) previously selected to be dependent on 3nY was used as a starting point for selection6. However, since TEM-1.B9 with 3nY already conferred resistance to high levels of ampicillin this would have prevented progressive selections for improved accommodation and incorporation. We therefore further engineered TEM-1.B9 to use ceftazidime (CAZ) as a substrate. We altered the Ω-loop of TEM-1.B9 to include mutations W165Y, E166Y, and P1657G (**Figure 5-1**), which are known to change TEM-1 from a penicillinase to a cephalosporinase (Palzkill et al., 1994; Stojanoski et al., 2015). This new β-lactamase, TEM-1.B9YYG-3nY, conferred moderate resistance to CAZ in a 3nY-dependent manner (Fig. 1) at much lower concentrations commonly used in bacterial cultures (3-10 μg ml^{-1}, **Figure 5-2**). In addition to the β-lactamase the plasmid contained an *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase (aaRS) previously engineered to be specific for 3-iodo-L-tyrosine (3iY) (Sakamoto et al., 2009), but which has also been demonstrated to be highly compatible with 3nY (Tack et al., 2016) (**Figure 3-5**), and the corresponding *M. jannaschii* tyrosyl-tRNA with the anticodon engineered to be complementary to the UAG amber stop codon.

Finally, we also created a control using TEM-1.B9YYG with phenylalanine (TTT) at position 162 (TEM-1.B9YYG-Phe). Phenylalanine is the only canonical amino acid that produces a functional TEM-1.B9 β-lactamase when replacing 3nY162. TEM-1.B9YYG-Phe conferred CAZ resistance in a 3nY-indepednent manner.

Strains were passaged in three different mixtures of amino acids in MOPS-EZ Rich Defined Medium (RDM). The first mixture contained all 20 standard amino acids (condition 1, RDM-20), the second mixture lacked tyrosine (condition 2, RDM-19), and the third mixture lacked serine, leucine, tryptophan, glutamine, tyrosine, lysine, and glutamate (condition 3, RDM-13) (**Figure 5-3**). These seven amino acids represent all

amino acids coded by codons accessible through single nucleotide mutations from the UAG stop codon, and thus represented the most stringent selection condition, where any single mutation would not necessarily acquire a ready source for suppression. Each media condition was supplemented with 10 mM 3nY, matching the concentration of L-serine, the most abundant amino acid in RDM.

5.3 Design and Results

5.3.1 Conditions

We selected three clones with 3nY suppression dependence (TEM-1.B9YYG-3nY), and three clones that did not require suppression for growth (TEM-1.B9YYG-Phe). These strains were denoted as 3nY-A, 3nY-B, and 3nY-C, and Phe-A, Phe-B, and Phe-C. Each clone was passaged in each of the three different amino acid environments described above (e.g. 3nY-A1, 3nY-A2, 3nY-A3).

The cultures were passaged daily by inoculating 5 ml of RDM with 1 µL of overnight growth. This resulted in approximately 12.5 generations per daily passage. During the course of passaging, CAZ concentration was increased at a rate of 1 µg ml^{-1} per 100 generations to a final concentration of 22 µg ml^{-1} to provide evolutionary pressure and ensure enforcement of 3nY dependence. Progenitor clones proved largely incapable of growth in RDM-13 when supplemented with 3nY, even in the absence of CAZ (**Table 5-1** and **Table 5-2**), so for the first 10 passages in RDM-13 the media was supplemented with 25% RDM-19. After these initial 10 passages, all lines were capable of growth in RDM-13.

At the conclusion of 80 passages, corresponding to 2000 generations, we selected a single clone from each evolved line and condition and sequenced the bacterial genomes of the single clone as well as then entire bacterial culture using HiSeq4000 Illumina

platform. Selected cells were, in general, genotypically representative of the average bacterial population. We further characterized phenotypic parameters of the selected evolved and progenitor cells.

5.3.2 Growth Rates

We measured the growth rates of progenitor strains in all three media conditions and the growth rates of the evolved strains in the media in which they were evolved. Growth rates were measured with and without 3nY, and with and without 2 or 22 μg ml^{-1} CAZ, doubling times were calculated based on growth rates in the absence of CAZ and with or without 3nY (**Table 5-2**).

The growth rate of the wild-type MG1655 was slowed by differing amino acid supplements, as well as by the addition of 3nY to the media (**Table 5-1**). Cells containing the 3nY OTS had similar growth rates to wild-type MG1655 in RDM-20, irrespective of whether 3nY was added to the media. However, in RDM-19 and RDM-13 the addition of the 3nY in the presence of its OTS further suppressed growth rates. It appears as though both amino acid limitation and the presence of the unnatural amino acid impact metabolism, and this impact is enhanced if the unnatural amino acid can be incorporated. In other words, there are multiple components to cellular fitness in the presence of the unnatural amino acid, some of which depend on translation, and some of which do not.

Over the course of evolution, the growth rates of all lines increased in all conditions, with the exception of Phe-B3, which showed no growth in media lacking 3nY (discussed in **5.3.5.5**). The relative fitness effects associated with media conditions remained consistent even after evolution; growth rates were highest in RDM-20 without NCAA, and greater metabolic stresses with decreasing numbers of amino acid supplements resulted in decreases in growth rates. Surprisingly, though, large growth

deficits that were seen with RDM-13 could be largely overcome during evolution. As was the case with the wild-type strains, it appears that some growth deficits were due to incorporation of 3nY, and some not (compare the initial growth deficits and evolutionary recoveries of strains with and without the OTS under RDM-13 growth conditions), but both types could be largely repaired over the course of evolution.

Since the OTS used is capable of incorporation the NCAA 3-iodo-tyrosine (3iY) as well as 3nY, we also tested growth rates of all lines with 3iY. Growth rates of progenitor cells were less affected by 10 mM of 3iY than by 3nY (**Table 5-1** and **Table 5-2**), but trends were similar. Interestingly, lines evolved to accommodate 3nY appeared to equally well accommodate 3iY, indicating that many of the evolutionary adaptations are not amino acid specific (**Table 5-2**).

5.3.3 Retention of the OTS

Given the experimental set-up, the primary question we were addressing was whether strains that were addicted to the orthogonal translation system would retain it and utilize it differently or better than strains that were not addicted to the OTS. To evaluate the functionality of the OTS we measured UAG suppression efficiency with a GFP reporter system that contained a tyrosine (TAT), amber (TAG), or ochre (TAA) at position Y39 of GFPmut2. We tested the GFP reporters in each selected evolved clone, as well as the progenitor clones. All of the addicted cell lines remained capable of efficient amber suppression (>25% 3nY incorporation at UAG codons) (**Figure 5-4**, top). In contrast, three of the nine non-addicted cell lines (Phe-A1, Phe-B2, and Phe-C2) had lost UAG suppression capability (<1% 3nY incorporation) (**Figure 5-4**, bottom). Surprisingly, UAG suppression efficiency dropped in both 3nY addicted and 3nY independent lines, relative to parental controls.

Of the 3nY addicted lines, only 3nY-A2 acquired any mutations in the OTS, acquiring a T->G mutation in the tRNA processing region at position T(-28). In contrast, in the nonaddicted lines OTS mutations occurred in the Phe-A1, A2, B2, and C2 lines. Line Phe-A1 and Phe-B2 have 3nY-aaRS mutations. In line Phe-A1, a single nucleotide deletion creates a frameshift in the 3nY-aaRS and resulted in a TGA ochre stop codon at position 123 in the 3nY-aaRS, resulting in complete loss of UAG suppression (Fig 3). Line Phe-B2 contains an IS-1 mediated insertion, a mechanism that has previously been observed to inactive OTS machinery (Wang et al., 2014). Lines Phe-A2 and Phe-C2 had mutations in the tRNA. The Phe-A2 plasmid contained a G26C substitution in the hinge between D-stem and C-stem (**Figure 5-5**, blue), while the Phe-C2 plasmid has a G42A substitution in the C-stem of the tRNA. The latter mutation is predicted to compromise tRNA structure, as determined using mfold (Zuker, 2003) and tRNA prediction software (Schattner et al., 2005) (**Fig 5-5**, red). This tRNA mutation potentially explains the complete loss of UAG suppression in the Phe-C2 line (Fig 3).

Overall, either the presence of a suppressor tRNA or more specifically the incorporation of 3nY puts cells at a fitness disadvantage, and in the absence of addiction to the unnatural amino acid this trait can be readily lost. Given that strains that incorporate 3iY grow as well as strains that incorporate 3nY, the fitness disadvantage can be attributed more to suppression itself rather than a particular NCAA.

5.3.4 Evolution of antibiotic resistance

During the course of passaging the cultures, CAZ concentrations were increased to levels beyond the MIC of the progenitor strains (the initial MIC of lines was approximately 2-6 μg ml^{-1}, while the final CAZ concentration challenge was 22 μg ml^{-1}). Increased bacterial survival at increasingly higher CAZ concentrations indicated that

CAZ resistance was evolving in the lines. We measured the MIC of each evolved clone with and without 3nY (**Figure 5-2**). The 3nY addicted lines remained dependent on 3nY for improved ceftazidime resistance, while the MICs of TEM-1.B9YYG-Phe lines increased in a 3nY-independent manner.

All nine of the addicted cell lines acquired at least a single mutation in TEM-1.B9YYG-3nY during evolution, (**Table 5-3**). Many of these mutations are known or expected to be stabilizing mutations of TEM-1 β-lactamase (Bershtein et al., 2008; Jacquier et al., 2013; Kather et al., 2008; Perilli et al., 2005; Salverda et al., 2010). Some mutations are specific to TEM-1.B9, which originally included a number of substitutions relative to the wild-type enzyme that addicted it to 3nY. The new substitution T139I is more similar to the original wild-type residue, leucine, but does not reduce the 3nY dependence of the enzyme. Of the non-addicted lines β-lactamase genes from Phe-A1, B1, C1, A2, and C2 acquired β-lactamase mutations, while β-lactamase genes from Phe-B2, A3, B3, and C3 were unchanged during evolution. The higher rate of mutation for TEM-1.B9YYG-3nY relative to the wild-type β-lactamase gene indicates that the enzyme dependent upon the unnatural amino acid was not initially as fit as its wild-type counterpart, especially in the presence of increasing antibiotic concentrations.

The phenotypic impact of the β-lactamase mutations was further examined by taking plasmids purified from the evolved lines and transforming them into native MG1655 (**Figure 5-6**). Notably, plasmids from 3nY-B1, A2, B2, C2, and A3 yielded increases in MIC greater than 2-fold the original MIC. These plasmids all contained the β-lactamase mutations known or suggested to have stabilizing effects (**Table-5-3**). However, β-lactamase mutations cannot fully explain MIC changes, as plasmids 3nY-C2 and 3nY-C3 have the identical mutation (T138K) yet confer different MICs when transformed into MG1655.

Beyond mutations in the β-lactamase gene itself, increases in MIC may have also been the result of other plasmid or genomic mutations that altered antibiotic resistance through other mechanisms. For example, plasmid 3nY-C3 has a mutation in *repC*, which can affect copy number (Haring et al., 1985). In addition, several genomic genes accumulated mutations that are known to have a role in antibiotic tolerance. Four lines had mutations in *envZ*, a histidine kinase that regulates *ompF* and *ompC* expression, which in turn alter membrane porosity and have been tied to β-lactam resistance (Jaffe et al., 1982). In the same vein, line 3nY-C3 has an ompF mutation. Mutations to *cyaA* or *crp*, two related proteins directly or indirectly involved in transcriptional regulation, occurred in eight evolved lines (3nY-B1, B2, C2, A3, C3 and Phe-A2, C2, C3); inactivation of these genes has been shown to produce resistance to β-lactams (Jaffé et al., 1983; Ruiz and Levy, 2011). The *opgG* gene was mutated in 3nY-C1,A2 and Phe-B1, and is involved in conferring resistance to antibiotics (Hanoulle et al., 2004). Finally, lipopolysaccharide (LPS) expression has been tied ceftazidime treatments, especially at or near MIC levels, and several lines had mutations in LPS biosynthetic genes (Leying et al., 1992; Pagani et al., 1990).

Beyond the stabilization of TEM-1.B9YYG-3nY there were few changes specific for lines that were addicted to 3nY relative to those that were not addicted to 3nY. None of the lines appear to use an expanded genetic code to increase fitness through ceftazidime resistance.

5.3.5 Genomic adaptations

As anticipated, the number of mutations acquired by a strain scaled with the stringency of selection. Full genome sequencing and analysis revealed that the lines addicted to 3nY acquired more mutations than those not addicted (there were a total of

forty-nine ORF-affecting mutations in the nine TEM-1.B9YYG-3nY strains, and a total of forty mutations in the nine TEM-1.B9YYG-Phe strains). Additionally, mutations increased in knockout media, with 3nY and Phe totaling twenty-two mutations in RDM-20, thirty-one mutations in RDM-19, and thirty-six mutations in RDM-13 (**Table 5-4**, and **Table 5-5**). Several trends were seen among evolution conditions and genes affected during evolution. There were seven mutations that appeared across multiple lines, summarized in (**Table 5-6**).

5.3.5.1 Tyrosine transporter tyrP

The *tyrP* gene, a tyrosine specific permease (Andrews et al., 1991), was deleted or otherwise inactivated in ten of the eighteen evolved lines, including all six lines (with and without the addicted β-lactamase) evolved in RDM-13 and all three TEM-1.B9YYG-Phe lines evolved in RDM-19, as well as 3nY-C2. Inactivation of the *tyrP* gene in lines 3nY-C2, A3, B3, C3, Phe-A2, B2, A3, B3, and C3 involved IS-mediated knockouts of varying lengths, while line Phe-C2 contained a SNP that generated a stop codon (W357tag). In contrast, the tyrP gene remained unaffected in all six lines evolved in media supplemented with tyrosine (RDM-20). Since 3nY is structurally similar to tyrosine, it is possible that 3nY was using tyrP for cell entry, and thus deactivating *tyrP* increased fitness in media lacking tyrosine (RDM-19 and RDM-13). To test this hypothesis, we used the Keio knockout collection (Baba et al., 2006) to examine the fitness effects of a *tyrP* deficiency in RDM with and without 3nY. The results (**Figure 5-7**) show *tyrP* knockout has no effect on growth compared to wild-type *E. coli* in RDM-20 and RDM-19, but significantly delays growth in RDM-13, both with and without 3nY. Thus, *tyrP* does not appear to mediate 3nY toxicity, at least as measured by growth conditions.

5.3.5.2 Methionine transporter mtr

Another transporter (*mtr*, specific for typtophan) (Brown, 1970) was the third most mutated gene amongst evolved lines. Mutations of *mtr* occurred in five of the six lines evolved in RDM-13, including all expect Phe-B3. Again, in contrast the mtr gene remained intact in all lines evolved in tryptophan rich media (RDM-20 and RDM-19). We used the *mtr* knockout from the Keio knockout collection (Baba et al., 2006) to explore the possibility that *mtr* involved in 3nY-induced fitness effects. Growth curves again show *mtr* knockouts did not increase cellular fitness in RDM-13 with 3nY (**Figure 5-7**).

5.3.5.3 cyaA/crp mutations

The second most common mutational target during evolution was adenylate cyclase *cyaA*, which activates the transcriptional regulator *crp*, which was also mutated in one evolved line. The *cyaA*/*crp* regulation system activates catabolic pathways for While in total eight lines had mutations to these genes associated with transcriptional regulation of pathways involved in lactose, galactose, and citrate catabolism. Seven of the eight mutations to these genes were single nucleotide substitutions resulting in amino acid substitutions, and one was a 78-nucleotide deletion at the c-terminus of the protein, involved in regulation. Using the Keio knockout Δ*cyaA* strain, we saw similar growth in all evolutionary conditions, with perhaps the exception of RDM-13, where the Δ*cyaA* strain grew slightly quicker than other knockout strains or the wild-type *E. coli* (**Figure 5-7**). This is likely not representative of the actual effects of the *cyaA* mutations seen during evolution, since single mutations often reduce enzyme kinetics, but the enzyme remains active. Additionally, *cyaA* mutations were not restricted to lines evolved in RDM-13, but seen in all growth conditions, and so is likely not the mechanism by which RDM-13 lines evolved to grow in RDM-13.

5.3.5.4 A large genomic fragment knockout

Besides the single gene mutations described above, there was a significant deletion from the genome of ten evolved lines, including all lines grown in RDM-13, and four lines evolved in RDM-19. This fragment is mediated by IS-1, the fragment varies in length, exact deletion sites, and the total number of genes either entirely or partially removed, though all 10 began at the *uspC* gene. Seven of the 10 genome excisions ended in the *tyrP* gene (3nY-A3, B3, and C3, and Phe-A2, B2, B3, and C3) and was responsible for many of the *tyrP* mutations as discussed above. One ends at *yecH*, leaving *tyrP* intact (3nY-B2), and two end in *yecA*, entirely removing *tyrP* from the genome (3nY-C2, and Phe-A3). This genome modification may be responsible for the growth rate improvements seen in RDM-13 evolved lines, as it is shared amongst all six of them. Efforts to remove this segment from wild-type MG1655 are underway, and growth curves in RDM-13 will be used to determine the possible phenotypic advantages this genome excision may confer. Here are a list of the genes removed by this excision, and their function (if known). In summary, there is no known function for these proteins that would clearly bestow a phenotypic advantage when removed from the genome of bacteria evolved in RDM-13.

Universal stress protein C (*uspC)* is involved in UV tolerance and the transition to stationary phase during growth in rich media (Gustavsson et al., 2002; Siegele, 2005). Trehalose-6-phosphate synthase (*otsA*) and trehalose-6-phosphate phosphatase (otsB) are involved osmorgulation, and contribute to antibiotic tolerance and oxidative stress prevention (Kuczyńska-Wiśnik et al., 2015). The arabinose transporter subunits *(araH*, *araG*, and *araF*) are responsible for the cross-membrane transport of the sugar L-arabinose (Kolodrubetz and Schleif, 1981). Ferritin iron-storage complex (*ftnB*) and related protein (*ftnA*) are involved, or predicted to be involved, in iron storage (Hudson et

al., 1993). The predicted proteins *yecJ, yecH* and *yecR* have unknown function, but may play a role in cell envelope or flagellar formation (Paradis-Bleau et al., 2014; Stafford et al., 2005). Additionally, *yecA* has predicted metal binding properties (Serres et al., 2001). The small membrane protein (*azuC*) has been identified as a stress response protein, and is regulated by *crp* (Fontaine et al., 2011; Hemm et al., 2010). The tyrosine transporter (*tyrP*) and its impact in growth effects was described above.

5.3.5.5 An in-frame TAG codon in an essential gene

A single clone, Phe-B3, evolved in RDM-13 and initially not dependent on 3nY for growth, acquired an in-frame genomic TAG codon in the gene *lptD*, which is involved in LPS biosynthesis. This was initially apparent when plate-base MIC assays of Phe-B3 showed no growth. Repeated attempts resulted in no growth as well, and no MIC is reported for Phe-B3 in the absence of 3nY (**Figure 5-2**). Growth curve analysis showed that Phe-B3 grew only in the presence of 3nY (**Figure 5-8**). Phe-B3 was not addicted to 3nY when the experiment began, and could have deactivated the OTS. It has acquired a mutation that, for the moment, requires UAG suppression with 3nY. Full genome sequencing of the mixed population of the PheB3 strain revealed that this mutation was not common in the evolving culture, and is thus likely a permissive, transient mutation in the population.

5.4 DISCUSSION

We present here the genomic and phenotypic analysis of the wild-type *E. coli* strain MG1655 which was evolved for 2000 generations with an expanded genetic code. Our results indicate that a functional addiction to the NCAA, as conferred by encoding with TEM-1.B9YYG-3nY, is required to ensure the OTS remains functional throughout evolution. Even after 2000 generations of evolution, that adoption of the NCAA

throughout the genome is limited. Despite a pressure to evolve with the NCAA, bacteria preferred to downregulate NCAA incorporation and UAG suppression, and use the natural genome to overcome evolutionary pressures.

5.5 METHODS AND MATERIALS

5.5.1 Evolutionary Set Up

5.5.1.1 Addiction Plasmid

The plasmid used for ceftazidime resistance was derived from pMMB67EH. The TEM-1 β-lactamase was replaced with the 3nY addiction operon as described previously (Tack et al., 2016). The substrate specificity was changed from penicillins to cephalosporins by replacing residues 165-167 with a YYG sequence, using primers B9YYGf and B9YYGr for 3nY cassette, and B9YYGFf and B9YYGFr (**Table 5-7**)for Phe control plasmid, and assembled using the Gibson protocol(Gibson, 2009). Plasmids were cloned into TOP10 *E. coli* and plating on LB-agar with 10 mM 3nY and 2 μg ml^{-1} CAZ. Samples were sequenced verified at University of Texas core facilities using Sanger sequencing.

5.5.1.2 Media

MOPS-EZ Rich Defined Media (RDM, purchased from TEKnova) with the full complement of amino acids, as well as the knockout medias, were prepared according the manufacturers specifications. Ultrapure water was autoclaved and then 3nY was added to a concentration of 17.24 mM while hot, to facilitate dissolving. Once cooled, the 3nY supplemented water was used to complete RDM, and the entire preparation was filter sterilized with Nalgene Rapid-Flow SFCA filtration units. Media was stored at 4°C before use, and moved to 20°C one day before use.

5.5.1.3 Passaging

Wild-type *E. coli* MG1655 was made electrocompotent with serial 10% glycerol washes, and electroporated with the addiction plasmid or control plasmid, and plated on LB-agar with 2 μg ml^{-1} CAZ. Three colonies from each transformation were picked and grown in RDM+3nY with 2 μg ml^{-1} CAZ, overnight. Each culture was then used to inoculate three subcultures: one in RDM containing the full complement of 20 amino acids, one in RDM lacking tyrosine, and one in RDM lacking tyrosine, glutamine, lysine, glutamate, tryptophan, serine, and leucine. Cultures lacking the seven amino acids were incapable of growth in 3nY initially, but were capable when 25% of media was replaced with RDM lacking tyrosine, this growth condition was used for the first 100 generations, and then cultures were capable of growth in experimental conditions lacking the seven amino acids. Cultures were passaged every 16-24 hours by transferring 1 μl into 5 ml of media, and grown shaking at 37°C.

5.5.2 Genome Sequencing and Assembly

After the 2000 generations, 1 μl from each line was streaked on MOPS-EZ rich defined media with agar and 10 mM 3nY. Single colonies were selected and grown in 3 ml MOPS-EZ RDM with 10 mM 3nY with 22 μg ml^{-1} CAZ. Simultaneously, samples from the progenitor strains were grown in similar conditions but with 2 μg ml^{-1} CAZ. After overnight growth, genomic DNA purified using bacteria genome miniprep kit (purchased from Sigma-Aldrich). Bacterial genomic DNA were prepped and sequenced with 150 bp paired ends on a HiSeq 4000, achieving greater than 100x coverage across all samples. Raw reads were processed through trimming and adapter removal using trimmomatic (v0.32) (Bolger et al., 2014). Alignment of sequencing reads and variant calling was performed through the breseq workflow (v0.27.2) (Deatherage and Barrick, 2014).

5.5.3 Growth Curves

Culture were inoculated from glycerol stocks and grown in respective appropriate media. After overnight growth to full OD, 1 µl from overnight cultures were used to inoculate 200 µL of media with appropriate concentrations of CAZ and 3nY. A Tecan Infinite M200 pro microplate reader set to 37°C monitored absorbance at 600 nm for approximately 24 hours. Data for each well were trimmed from .1 to .2 OD_{600} after which the rate of logarithmic change in OD was calculated for each line. Keio collection strains were BW25113 as the wild-type *E. coli*, JW3778 (Δ*cyaA*), JW1895 (Δ*tyrP*), and JW3130 (Δ*mtr*), and were a gracious gift from the Jeffrey Barrick Lab at the University of Texas at Austin.

5.5.4 MICs

The 6 progenitor bacteria and the 18 selected clones from the evolved lines were grown overnight in 3nY enriched MOPS-EZ growth media supplemented with 2 µg ml^{-1} or 22 µg ml^{-1} of CAZ respectively. Overnight cultures were used to inoculate fresh MOPS-EZ rich defined media lacking 3nY and without CAZ, and were grown for five hours at 37°C in shaking incubator. Aliquots of 25 uL from each line (corresponding to ~10^7 cfu) were plated on LB-agar, with or without 10 mM 3nY, in triplicate. Plates were allowed a period to dry for 5 minutes, and ceftazidime E-test MIC strips (Biomerieux) were added to each plate. Plates were incubated for 16 hours at 37°C. MIC values are reported as the lowest MIC concentration on E-test strip, in which no bacterial growth occurred.

5.5.5 GFP Assays

The codon for Tyr39 in GFPmut2, under control of the *tacI* promoter with a *lacO* operator, was mutated to amber codon (TAG), tyrosine codon (TAT), or ochre nonsense

codon (TAA) using primers TAGf, TATf, TAAf, along with N39r for PCR amplification. Amplicons were concentrated and desalted using QIAquick PCR purification kit, and resultant PCR product was circularized using the isothermal Gibson protocol, followed by the addition of DpnI to digest original PCR template. Gibson product was electroporated into TOP10 *E. coli*, and plated on LB-agar supplemented with kanamycin. Clones were grown, and plasmid DNA isolated using QIAprep Spin Miniprep Kit. Each sequence was confirmed by DNA Sanger sequencing at the University of Texas Core Facility.

Glycerol stocks of the eighteen evolved clones, as well as the six progenitor clones, were used to inoculate 2 ml of LB media, supplemented with 10 mM 3nY, as well as .02% glucose, in 2 μg ml^{-1} CAZ, and grown to saturation overnight. A 5 μl aliquot from each was used to inoculate 5 ml of LB, 10 mM 3nY, and 2 μg ml^{-1} CAZ, and grown at 37°C for 3 hours, and made electrocompetent with serial washes of 10% glycerol. Samples were transformed with GFP (N39X) variants, recovered with LB media, and plated on LB-agar with 10 mM 3nY, .02% glucose, 50 μg ml^{-1} KAN, and 2 μg ml^{-1} CAZ, and grown overnight at 37°C.

Four samples from each plate were picked and grown overnight at 37°C in 500 μl LB with 10 mM 3nY, .02% glucose, 50 μg ml^{-1} KAN, and 2 μg ml^{-1} CAZ. A 2 μl aliquot of each overnight culture was used to inoculate 1 ml LB media supplemented with 50 μg ml^{-1} KAN, and containing either 1 mM 3nY, 1 mM 3-iodotyrosine, or with no supplemented NCAA. Cultures were grown at 37°C for 3 hours, then induced with 1 mM IPTG, and grown for another 5 hours at 37°C. Final cultures were centrifuged and washed twice at 2000*g* with 4°C PBS (50 mM phosphate, 300 mM NaCl, pH 8), and a 150 μl aliquot of washed culture were used to measure absorbance at 600 nm wavelength,

was well as GFP fluorescence, with measured emission of 525 nm wavelength, upon excitation at 475 nm, using Tecan Infinite M200 pro microplate reader.

5.6 TABLES AND FIGURES

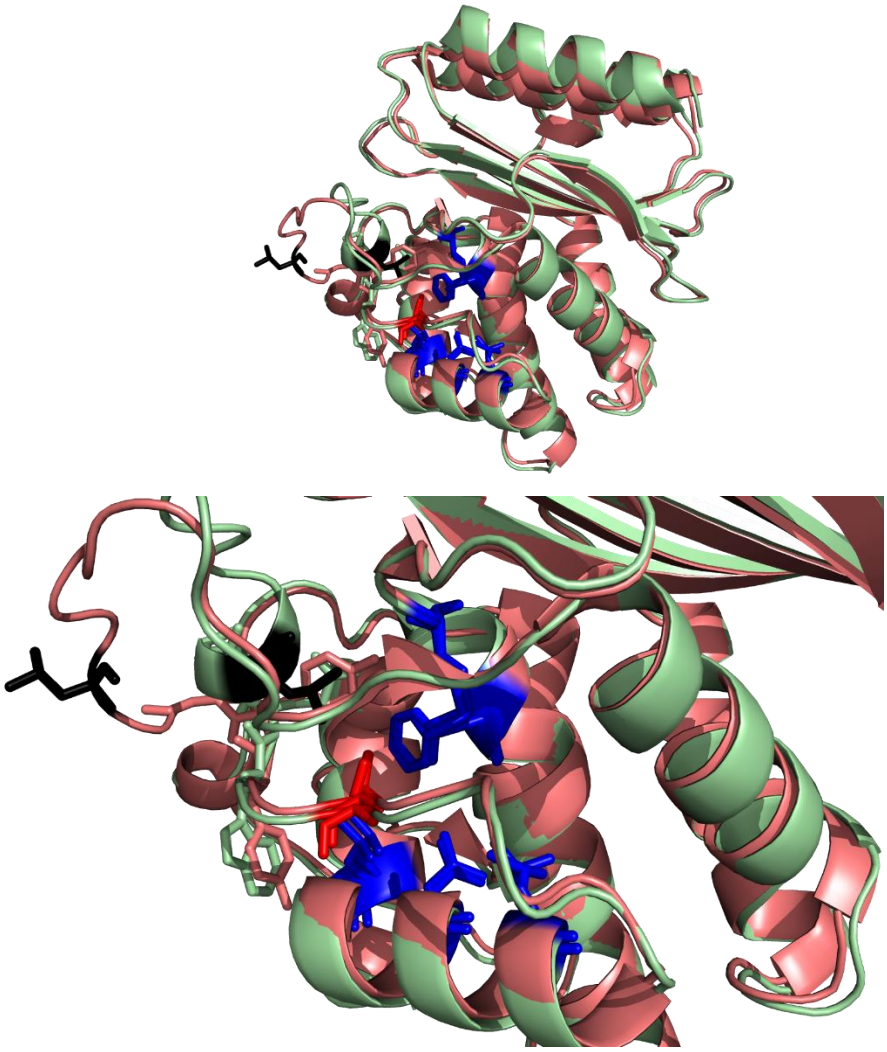

Figure 5-1: TEM-1YYG structure

Alignment of TEM-1 β-lactamase (green, PDB: 1XPB) that was used to engineered TEM-1.B9, and the TEM-1 Ω-loop mutant (salmon, PDB: 4rx3) that changes substrate specificity from lactams to cephalosporins. These mutations affect the loop structure and the pocket surrounding 3nY incorporation site L162 (red sticks).The engineered pocket (blue and black sticks) is mostly affected at L169 (black sticks). Despite this, the TEM-1.B9YYG-3nY enzyme was active, and 3nY dependent (**Figure 5-2**)

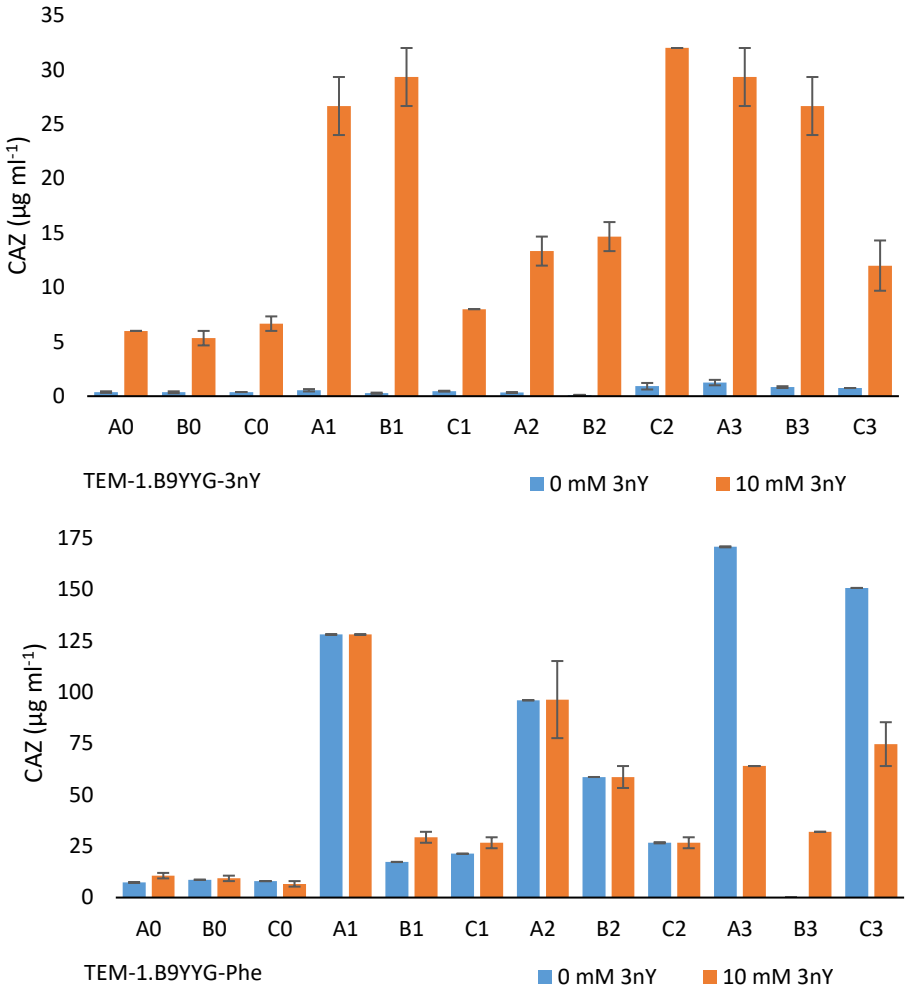

Figure 5-2: MICs of MG1655

Minimum inhibitory concentration (MICs) of ceftazidime determinations of all progenitor and evolved lines (A0, B0, and C0) with and without 3nY (orange and blue, respectively). Lines evolved with 3nY addiction (Top) increased MICs with 3nY while background in the absence of 3nY remained near zero. Nonaddicted lines saw higher increases in MIC over all.

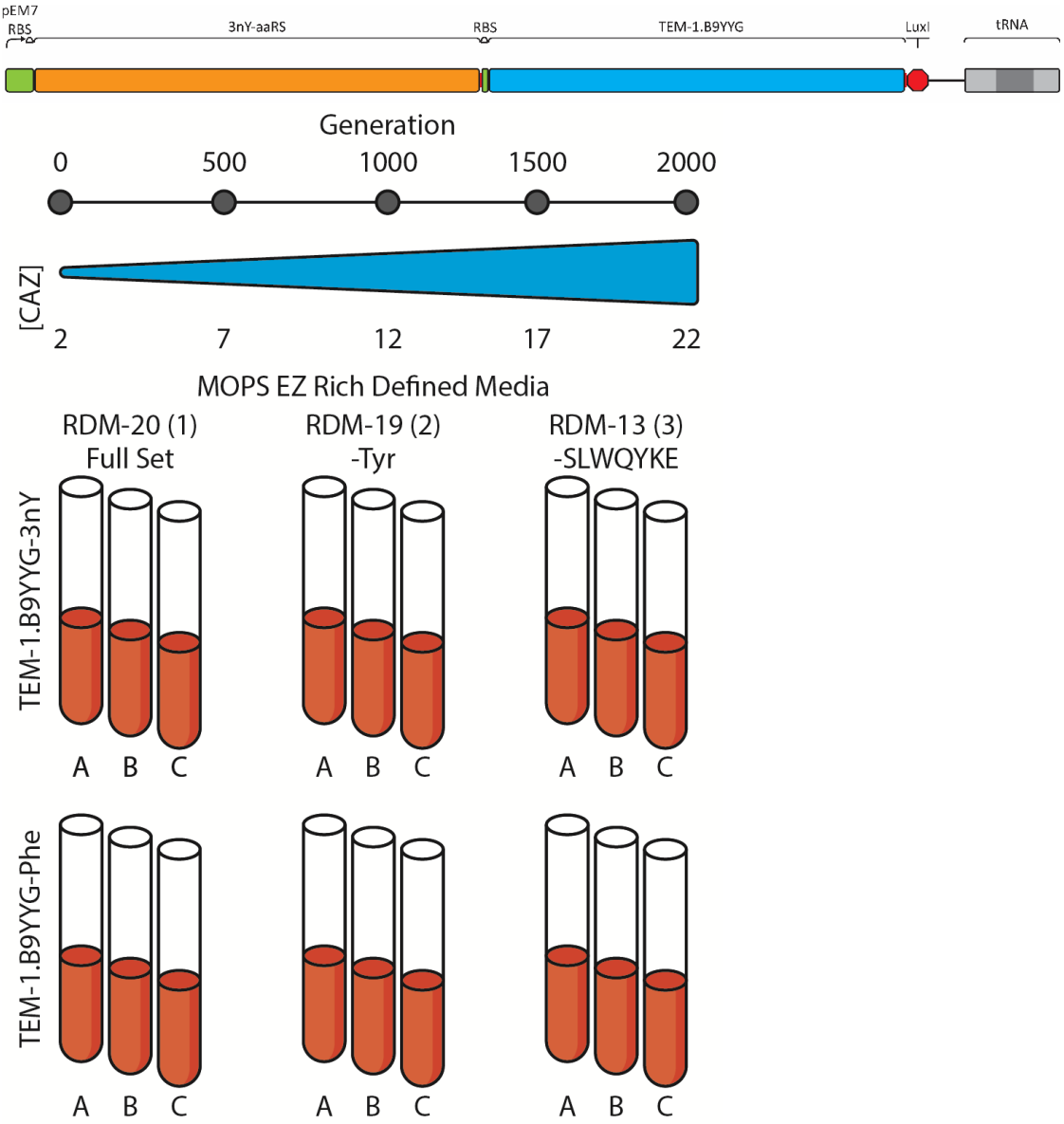

Figure 5-3: Evolutionary conditions

The addiction cassette included 3nY-aaRS, TEM-1.B9 variant, and tRNA. Three colonies containing TEM-1.B9YYG-3nY and three colonies containing TEM-1.B9YYG-Phe (A, B, and C for each) were selected, and each was grown in RDM-20, RDM-19, and RDM-13 with 10 mM 3nY and ceftazidime for 2000 generations. During the course of passaging, the ceftazidime concentration was gradually increased, beginning at 2 μg ml^{-1} and ending at 22 μg ml^{-1}.

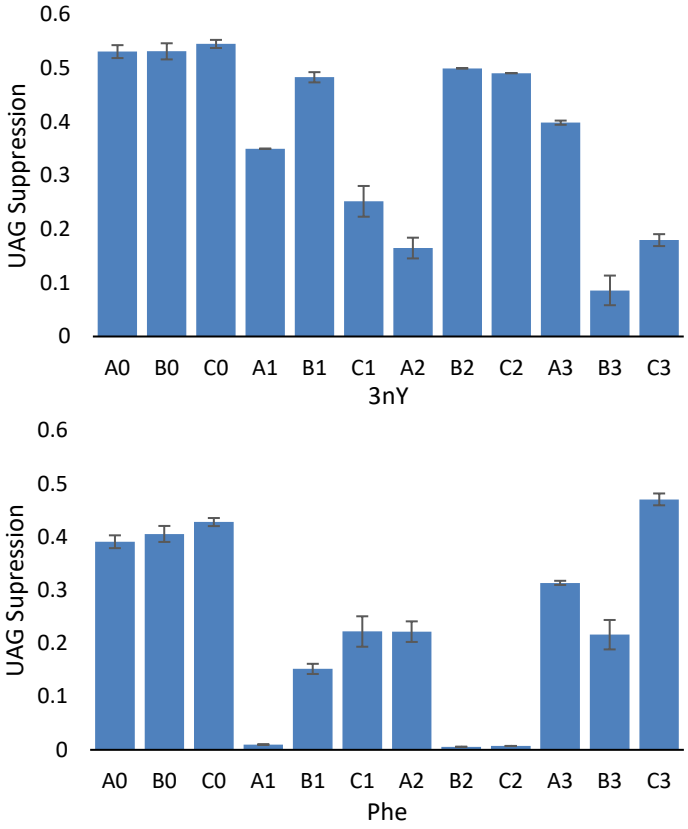

Figure 5-4: UAG suppression efficiency of evolved lines

GFPmut2 with a TAG codon replacing Tyr39 was expressed in all progenitor lines, as well all evolved lines to determine UAG suppression activity relative to a tyrosine (TAT) codon. All lines evolved with 3nY addiction maintained a functional level of UAG suppression (3nY-A1-C3). Three lines (Phe-A1, B2, and C2) from the nonaddicted evolved lines eliminated UAG suppression activity (<1% of TAT39 GFPmut2 expression). Error bars represent s.e.m.

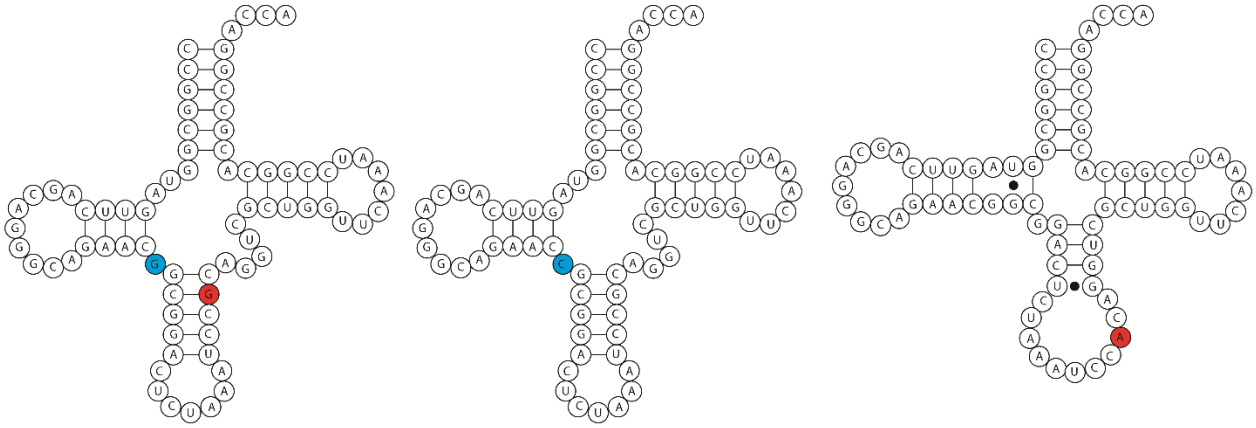

Figure 5-5: Structure of tRNAs

Computational predictions of the tRNAs used and found in this study. Left is the generation zero tRNA, center is the structure of the Phe-A2 tRNA substitution G27C (blue) which is predicted not to affect tRNA structure. Right is the predicted structural perturbation of the Phe-C2 mutation, a G43A substitution (red).

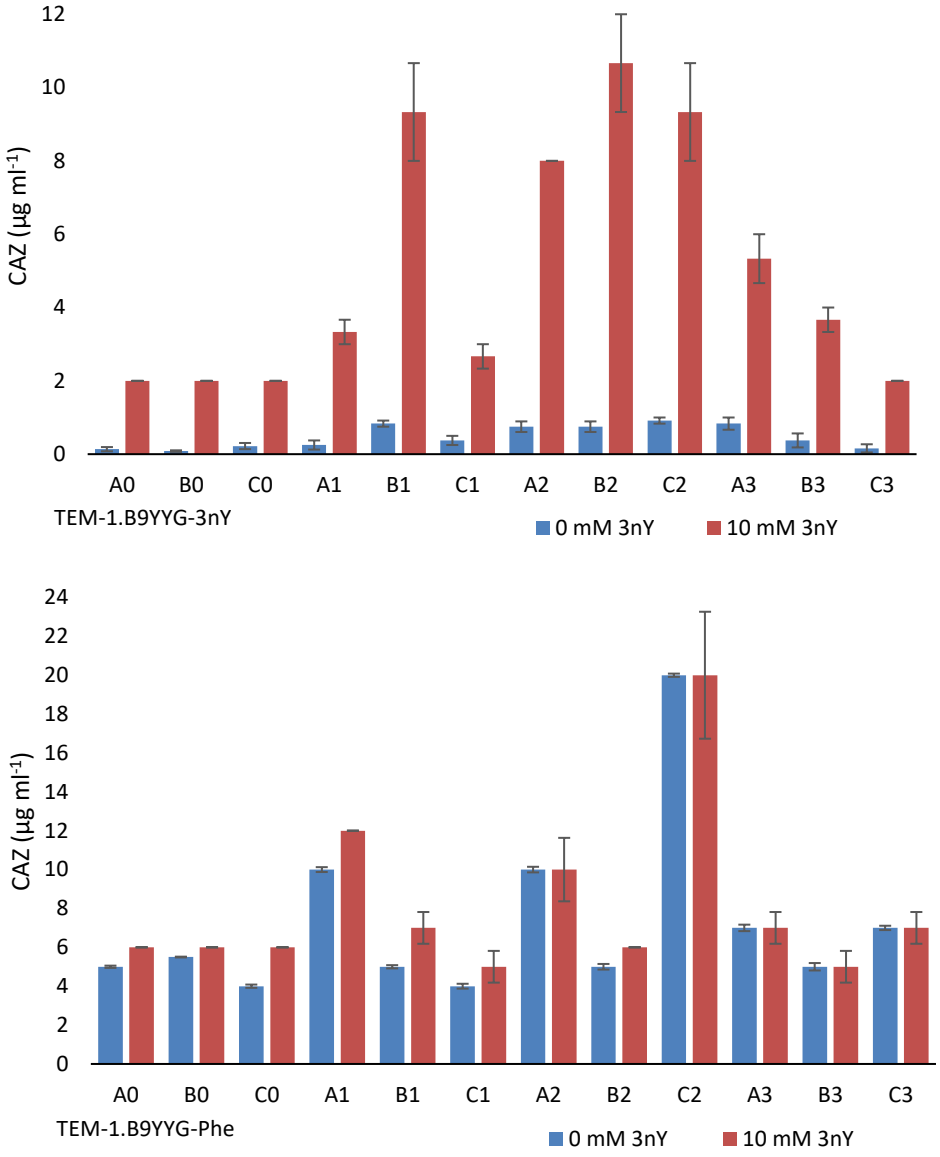

Figure 5-6: MICs of plasmid encoded resistance

MICs of MG1655 transformed with plasmid purified from each of the evolved line, as well as the progenitor cells. Reported values are average of biological triplicates, error bars represent s.d.

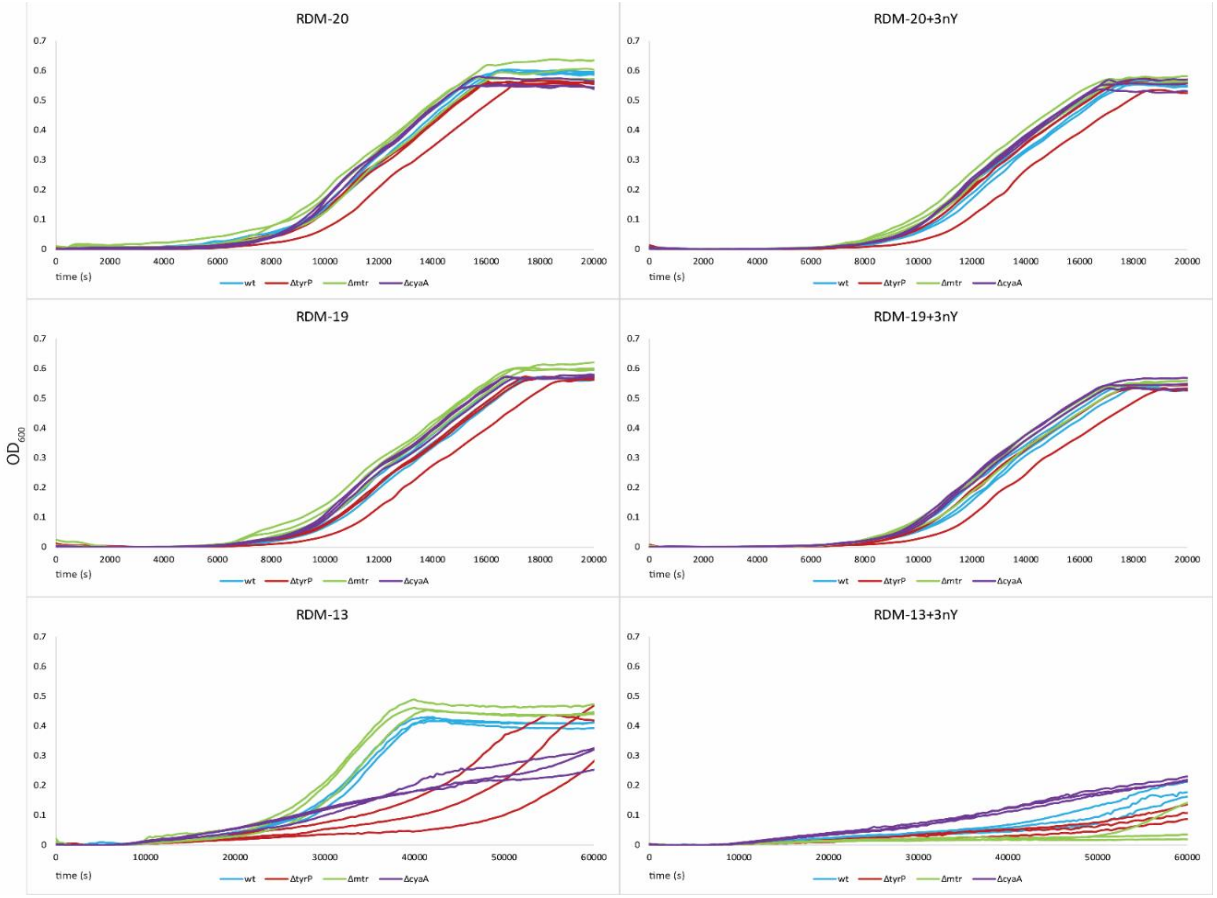

Figure 5-7: Keio knockout growth curves

Strains from the Keio knockout collection (Baba et al., 2006) were used to analyze the growth rates of lines with common knockouts in each media grown. Neither *cyaA* (purple), *mtr* (green), nor *tyrP (*red) knockouts significantly increased growth in RDM-13 when supplemented with 3nY.

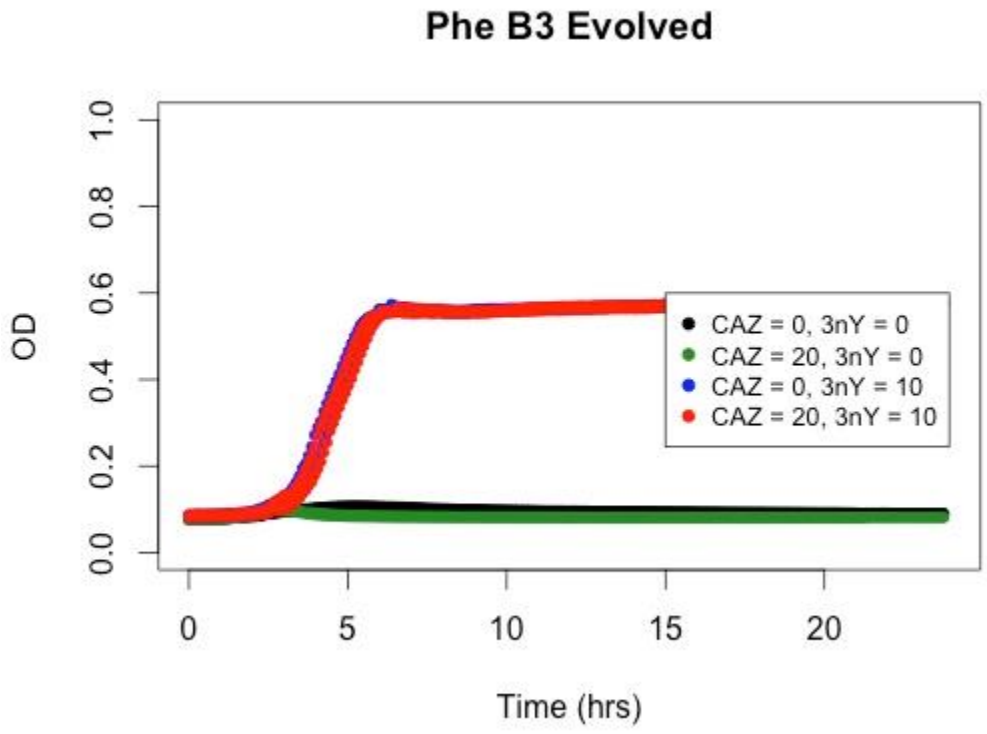

Figure 5-8: Phe-B3 growth rates

Evolved line Phe-B3 acquired a mutation resulting in an in-frame TAG codon in the essential gene *lptD*, which is involved in LPS biosynthesis. After 2000 generations, Phe-B3 became dependent on 3nY incorporation for growth, and had no growth in the absence of 3nY (green growth curves), but grew well in 3nY rich media (red and blue growth curves).

MG1655 Doubling Rate						
	Wild-Type			3nY-OTS		
	0 mM	10 mM 3nY	10 mM 3iY	0 mM	10 mM 3nY	10 mM 3iY
RDM-20	1.64	1.43	1.46	1.65	1.42	1.50
RDM-19	1.53	1.28	1.46	1.42	1.28	1.23
RDM-13	1.31	0.46	1.25	1.00	0.05	1.09

Table 5-1: Doubling rates of MG1655 wild-type

Calculated growth rates (hr^{-1}) of MG1655 in the amino acid environments used during evolution. MG1655 doubling rates decrease with the removal of amino acids and with the addition of 10 mM of 3nY, but are less affected by 10 mM 3iY. The impact on growth rates is increased in MG1655 encoding the 3nY addiction cassette used during evolution. Growth rates were calculated in the absence of antibiotics.

		Progenitor			Evolved		
		0 mM	10 mM 3nY	10 mM 3iY	0 mM	10 mM 3nY	10 mM 3iY
3nY	RDM-20	1.65	1.42	1.50	1.77	1.69	1.67
	RDM-19	1.42	1.28	1.23	1.69	1.53	1.48
	RDM-13	1.00	0.05	1.09	1.35	1.26	1.40
Phe	RDM-20	1.41	1.27	1.34	1.82	1.64	1.78
	RDM-19	1.30	1.19	1.17	1.71	1.60	1.66
	RDM-13	0.94	0.14	1.14	1.40	1.38	1.41

Table 5-2: Doubling rates of progenitor and evolved clones

Doubling rates (hr^{-1}) of MG1655 with addiction plasmid both before and after evolution. Over the course of evolution, cells largely recovered fitness in the evolutionary media.

Mutation	Lines	Effect
V33I	3nY-B3	
Q39K	Phe-B1	CAZ specific
G92D	3nY-A1,B1,B2	Stabilizing
T139I	3nY-C1 Phe-C2	TEM-1.B9 Specific
T140K	3nY-B2,C2, C3	
M152I	Phe-A1	TEM-1.B9 Specific
H153R	3nY-B1	Stabilizing
M155I	Phe-C1	
M182T	3nY-A2	Stabilizing
A184V	3nY-A3	
T200P	3nY-A1	
A224V	Phe-A2	Stabilizing

Table 5-3: Evolved TEM-1.B9YYG mutations

Listing of the TEM-1.B9YYG mutations and the lines they appeared in during evolution, as well as published effects of the mutation, if known. Every line of the 3nY addicted lines acquired at least one mutation in TEM-1.B9YYG-3nY.

		Tem-1.B9YYG-3nY		
		Gene	mutation	Location
RDM-20	3nY-A1	sdaC	9 nucleotide insert	925/1290
		yqiG	A>G	815/2439
		ebvZ	M58R	
	3nY-B1	ydeT	A>C	1108/1227
		waaO	+9 insert IS-1	531-539/1020
		mtdL	A324T	
		cyaA	-78	2420-2497/2547
	3nY-C1	opgG	Q314tag	
		clsA	+9 IS-1	1307-1315
		sdaC	A383V	
		yghG	G63V	
		envZ	F107C	
RDM-19	3nY-A2	opgG	+9 insert IS-1	385-393/1536
		ycjO	d87	250-336/882
		waaO	+4 insert IS-5	533-536/1020
		ycaA	d1	2033/2547
	3nY-B2	yahF	K208R	
		dosP	I282I	
		uspC-yecH	-10333 IS-1	
		lhgO	R346W	
		yqeK	L30(tga)	
		crp	F137L	
	3nY-C2	hofC	S27R	
		uspC-yecA	-12537 IS-1	
		ppx	-1	365/1542
		rrlD	C>T	2209/2904
		envZ	G85V	
		dnaA	-54	251-304/1404
		cyaA	R160L	
RDM-13	3nY-A3	flgJ	+9 insert IS-1	157-165/942
		uspC-tyrP	-10976	
		rpoD	P97L	
		mtr	-6	712-717/1245
		waaP	+4	413-416/798
		cyaA	D231V	
	3nY-B3	pykF	I264T	
		uspC-tyrP	-11298 IS-1	
		sanA	+9 IS-1	165-173/720
		mtr	-6	391-396/1245
		ompR	M57L	
		yibA	+9 IS-1	15-23/843
	3nY-C3	yaiT	A>C	682/1458
		ompF	+9 IS-1	323-331/1089
		cysB	-1	610/975
		uspC-tyrP	-11018 IS-1	
		mtr	-6	712/717/1245
		glpD	+9 IS-1	1109-1117/1506
		cyaA	P301L	
		yiiQ	K74taa	

Table 5-4: Genomic mutations in 3nY lines

List of the mutations found in each of the TEM-1.B9YYG-3nY evolved lines

		Tem-1.B9YYG-Phe		
		Gene	mutation	Location
RDM-20	Phe-A1	sdaC	+9 insert IS-1	1192-1200/1290
		waaQ	+9 insert IS-1	581-589/1035
		ybbP	A580E	
	Phe-B1	opgG	+9 IS-1	738-746/1656
		sdaC	W18tag	
		waaO	+4 IS-5	925-928/1020
	Phe-C1	sdaC	+1	798/1290
		envZ	S87Y	
		waaO	+5 IS-2	100-104/1020
		yidD	A14G	
RDM-19	Phe-A2	galU	H71L	
		uspC-tyrP	-11258 IS-1	
		garD	K343Q	
		tusD	G43G	
		cyaA	D231G	
	Phe-B2	opgH	H417D	
		uspC-tyrP	-11067	
		waaB	+4 IS-5	226-229/1080
	Phe-C2	caiC	I449S	
		quuD	+5 IS-2	233-237/384
		narK	S192S	
		tyrP	W357tag	
		cyaA	D184N	
		yihY	R208L	
RDM-13	Phe-A3	ftsW	L141I	
		galU	+12 IS-4	695-706/909
		uspC-yecA	-11714 IS-1	
		tyrA	M117I	
		kduI	M193I	
		mtr	-6	712-717/1245
	Phe-B3	lptD	Q557tag	
		uspC-tyrP	-11031 IS-1	
		yhcG	S136A	
		tnaA	+5 IS-2	303-304/1416
	Phe-C3	galU	+9 IS-1	81-89/909
		uspC-tyrP	-11280	
		mtr	L400R	
		mtr	-1	1011/1245
		cyaA	C120Y	
		rpsF	D13G	

Table 5-5: Genomic mutations in Phe lines

List of the mutations found in each of the TEM-1.B9YYG-Phe lines.

Gene	Lines	Discription	Identifier
tyrP	3nY-C2,A3,B3,C3 Phe-A2,B2,C2,A3,B3,C3	Tyrosine transporter	P0AAD4
uspC-yecH	3nY-B2,C2,A3,B3,C3 Phe-A2,B2,A3,B3,C3	See **Chapter 5.3.5.4**	
cyaA	3nY-B1,C2,A3,C3 Phe-A2,C2,C3	Adenylate cyclase	P00936
mtr	3nY-A3,B3,C3 Phe-A3,C3	Tryptophan transporter	P0AAD2
sdaC	3nY-A1,C1 Phe-A1,B1,C1	Serine transporter	P0AAD6
waaO	3nY-B1,A2 Ph3-B1,C1	(glucosyl)LPS α-1,3-glucosyltransferase	P27128
envZ	3nY-A1,C1,C2 Phe-C1	histidine kinase:osmotic regulation	P0AEJ4
galU	Phe-B2,A3,C3	UTP:glucose-1-phosphate uridylyltransferase	P0AEP3
opgG	3nY-C1,A2 Phe-B1	associated with synthesis of osmoregulated periplasmic glucans	P33136

Table 5-6: Genes with mutations in more than one evolved line

A list of the nine mutational hotspots during evolution, and the evolutionary lines in which mutations occurred in the listed gene.

Primer name	Sequence
B9YYGf	GATCGTTATTACGGCGAGTTGAATGAAGCCATACCAAACGACGAG
B9YYGr	ACTCGCCGTAATAACGATCCTAGCGAGTTACATGATCCCC
B9YYGFf	TCATGTAACTCGCTTTGATCGTTATTACGGCGAGTTGAATGAAGCCATACC
B9YYGFr	TAACGATCAAAGCGAGTTACATGATCCCCCATGTTGTGCATAAAAG
TATf	GAGGGTGAAGGTGATGCAACATATGGAAAACTTACCCTTAAATTTATTTGCACTACTGG
TAGf	GAGGGTGAAGGTGATGCAACATAGGGAAAACTTACCCTTAAATTTATTTGCACTACTGG
TAAf	GAGGGTGAAGGTGATGCAACATAAGGAAAACTTACCCTTAAATTTATTTGCACTACTGG
N39r	TGTTGCATCACCTTCACCCTCTC

Table 5-7: Primer sequences

The nucleotide sequences of primers used in this study

References

A. Caravella, J., Wang, D., M. Glaser, S., and Lugovskoy, A. (2010). Structure-Guided Design of Antibodies. Curr. Comput. - Aided Drug Des. *6*, 128–138.

Ai, H., Shen, W., Sagi, A., Chen, P.R., and Schultz, P.G. (2011). Probing Protein–Protein Interactions with a Genetically Encoded Photo-crosslinking Amino Acid. ChemBioChem *12*, 1854–1857.

Alfonta, L., Zhang, Z., Uryu, S., Loo, J.A., and Schultz, P.G. (2003). Site-Specific Incorporation of a Redox-Active Amino Acid into Proteins. J. Am. Chem. Soc. *125*, 14662–14663.

Allmang, C., and Krol, A. (2006). Selenoprotein synthesis: UGA does not end the story. Biochimie *88*, 1561–1571.

Anderson, J.C., and Schultz, P.G. (2003). Adaptation of an Orthogonal Archaeal Leucyl-tRNA and Synthetase Pair for Four-base, Amber, and Opal Suppression. Biochemistry (Mosc.) *42*, 9598–9608.

Anderson, J.C., Wu, N., Santoro, S.W., Lakshman, V., King, D.S., and Schultz, P.G. (2004). An expanded genetic code with a functional quadruplet codon. Proc. Natl. Acad. Sci. U. S. A. *101*, 7566–7571.

Andersson, S.G.E., and Kurland, C.G. (1998). Reductive evolution of resident genomes. Trends Microbiol. *6*, 263–268.

Andrews, A.E., Lawley, B., and Pittard, A.J. (1991). Mutational analysis of repression and activation of the tyrP gene in Escherichia coli. J. Bacteriol. *173*, 5068–5078.

Atkins, J.F., and Gesteland, R. (2002). The 22nd Amino Acid. Science *296*, 1409–1410.

Atkinson, G.C., Hauryliuk, V., and Tenson, T. (2011). An ancient family of SelB elongation factor-like proteins with a broad but disjunct distribution across archaea. BMC Evol. Biol. *11*, 22.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. J. Exp. Med. *79*, 137–158.

Ayyadurai, N., Prabhu, N.S., Deepankumar, K., Jang, Y.J., Chitrapriya, N., Song, E., Lee, N., Kim, S.K., Kim, B.-G., Soundrarajan, N., et al. (2011). Bioconjugation of l-3,4-Dihydroxyphenylalanine Containing Protein with a Polysaccharide. Bioconjug. Chem. *22*, 551–555.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Bacher, J.M., and Ellington, A.D. (2001). Selection and Characterization of Escherichia coli Variants Capable of Growth on an Otherwise Toxic Tryptophan Analogue. J. Bacteriol. *183*, 5414–5425.

Bacher, J.M., Bull, J.J., and Ellington, A.D. (2003a). Evolution of phage with chemically ambiguous proteomes. BMC Evol. Biol. *3*, 24.

Bacher, J.M., Bull, J.J., and Ellington, A.D. (2003b). BMC Evol Biol *3*, 24.

Bacher, J.M., Hughes, R.A., Wong, J.T.-F., and Ellington, A.D. (2004). Evolving new genetic codes. Trends Ecol. Evol. *19*, 69–75.

Bain, J.D., Diala, E.S., Glabe, C.G., Dix, T.A., and Chamberlin, A.R. (1989). Biosynthetic site-specific incorporation of a non-natural amino acid into a polypeptide. J. Am. Chem. Soc. *111*, 8013–8014.

Balaram, P. (1992). Non-standard amino acids in peptide design and protein engineering. Curr. Opin. Struct. Biol. *2*, 845–851.

Baldini, G., Martoglio, B., Schachenmann, A., Zugliani, C., and Brunner, J. (1988). Mischarging Escherichia coli tRNAPhe with L-4'-[3-(trifluoromethyl)-3H-diazirin-3-yl]phenylalanine, a photoactivatable analogue of phenylalanine. Biochemistry (Mosc.) *27*, 7951–7959.

Barderas, R., Desmet, J., Timmerman, P., Meloen, R., and Casal, J.I. (2008). Affinity maturation of antibodies assisted by in silico modeling. Proc. Natl. Acad. Sci. *105*, 9029–9034.

Barrell, B.G., Bankier, A.T., and Drouin, J. (1979). A different genetic code in human mitochondria. Nature *282*, 189–194.

Bentin, T., Hamzavi, R., Salomonsson, J., Roy, H., Ibba, M., and Nielsen, P.E. (2004). Photoreactive Bicyclic Amino Acids as Substrates for Mutant Escherichia coli Phenylalanyl-tRNA Synthetases. J. Biol. Chem. *279*, 19839–19845.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Berry, M.J., Banu, L., Chen, Y., Mandel, S.J., Kieffer, J.D., Harney, J.W., and Larsen, P.R. (1991). Recognition of UGA as a selenocysteine codon in Type I deiodinase requires sequences in the 3′ untranslated region. Nature *353*, 273–276.

Bershtein, S., Goldin, K., and Tawfik, D.S. (2008). Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. J. Mol. Biol. *379*, 1029–1044.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The Complete Genome Sequence of Escherichia coli K-12. Science *277*, 1453–1462.

Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., and Zinoni, F. (1991). Selenocysteine: the 21st amino acid. Mol. Microbiol. *5*, 515–520.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics btu170.

Bradford, P.A. (2001). Clin Microbiol Rev *14*, 933–951.

Brinkley, M. (1992). A brief survey of methods for preparing protein conjugates with dyes, haptens and crosslinking reagents. Bioconjug. Chem. *3*, 2–13.

Brown, K.D. (1970). Formation of Aromatic Amino Acid Pools in Escherichia coli K-12. J. Bacteriol. *104*, 177–188.

Bruce, A.G., Atkins, J.F., Wills, N., Uhlenbeck, O., and Gesteland, R.F. (1982). Replacement of anticodon loop nucleotides to produce functional tRNAs: amber suppressors derived from yeast tRNAPhe. Proc. Natl. Acad. Sci. U. S. A. *79*, 7127–7131.

Brustad, E., Bushey, M.L., Brock, A., Chittuluru, J., and Schultz, P.G. (2008). A promiscuous aminoacyl-tRNA synthetase that incorporates cysteine, methionine, and alanine homologs into proteins. Bioorg. Med. Chem. Lett. *18*, 6004–6006.

Burdine, L., Gillette, T.G., Lin, H.-J., and Kodadek, T. (2004). Periodate-Triggered Cross-Linking of DOPA-Containing Peptide−Protein Complexes. J. Am. Chem. Soc. *126*, 11442–11443.

Cabantous, S., Nguyen, H.B., Pedelacq, J.-D., Koraïchi, F., Chaudhary, A., Ganguly, K., Lockard, M.A., Favre, G., Terwilliger, T.C., and Waldo, G.S. (2013). A New Protein-Protein Interaction Sensor Based on Tripartite Split-GFP Association. Sci. Rep. *3*, 2854.

Cannon, M.B., and Remington, S.J. (2006). Re-engineering redox-sensitive green fluorescent protein for improved response rate. Protein Sci. Publ. Protein Soc. *15*, 45–57.

Charbon, G., Brustad, E., Scott, K.A., Wang, J., Løbner-Olesen, A., Schultz, P.G., Jacobs-Wagner, C., and Chapman, E. (2011). Subcellular Protein Localization by Using a Genetically Encoded Fluorescent Amino Acid. Chembiochem *12*, 1818–1821.

Chatterjee, A., Guo, J., Lee, H.S., and Schultz, P.G. (2013). A Genetically Encoded Fluorescent Probe in Mammalian Cells. J. Am. Chem. Soc. *135*, 12540–12543.

Chaudhury, S., and Gray, J.J. (2008). Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. J. Mol. Biol. *381*, 1068–1087.

Chaudhury, S., Berrondo, M., Weitzner, B.D., Muthu, P., Bergman, H., and Gray, J.J. (2011). Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. PLOS ONE *6*, e22477.

Cheetham, G.M.T., Jeruzalmi, D., and Steitz, T.A. (1999). Structural basis for initiation of transcription from an RNA polymerase–promoter complex. Nature *399*, 80–83.

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B.L. (2009). Design of therapeutic proteins with enhanced stability. Proc. Natl. Acad. Sci. *106*, 11937–11942.

Chin, J.W., Santoro, S.W., Martin, A.B., King, D.S., Wang, L., and Schultz, P.G. (2002a). Addition of p-azido-L-phenylalanine to the genetic code of Escherichia coli. J. Am. Chem. Soc. *124*, 9026–9027.

Chin, J.W., Martin, A.B., King, D.S., Wang, L., and Schultz, P.G. (2002b). Addition of a photocrosslinking amino acid to the genetic code of Escherichia coli. Proc. Natl. Acad. Sci. *99*, 11020–11024.

Cho, H., Daniel, T., Buechler, Y.J., Litzinger, D.C., Maio, Z., Putnam, A.-M.H., Kraynov, V.S., Sim, B.-C., Bussell, S., Javahishvili, T., et al. (2011). Optimized clinical performance of growth hormone with an expanded genetic code. Proc. Natl. Acad. Sci. *108*, 9060–9065.

Clark, L.A., Boriack-Sjodin, P.A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K.J.M., Jarpe, M., Liparoto, S.F., Li, Y., Lugovskoy, A., et al. (2006). Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. Protein Sci. *15*, 949–960.

Collins, D.W., and Jukes, T.H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. Genomics *20*, 386–396.

Conrad, B., Savchenko, R.S., Breves, R., and Hofemeister, J. A T7 promoter-specific, inducible protein expression system forBacillus subtilis. Mol. Gen. Genet. MGG *250*, 230–236.

Cooley, R.B., Feldman, J.L., Driggers, C.M., Bundy, T.A., Stokes, A.L., Karplus, P.A., and Mehl, R.A. (2014). Structural Basis of Improved Second-Generation 3-Nitro-tyrosine tRNA Synthetases. Biochemistry (Mosc.) *53*, 1916–1924.

Cowie, D.B., and Cohen, G.N. (1957). Biosynthesis by Escherichia coli of active altered proteins containing selenium instead of sulfur. Biochim. Biophys. Acta *26*, 252–261.

Cox, J.C., Lape, J., Sayed, M.A., and Hellinga, H.W. (2007). Protein fabrication automation. Protein Sci. *16*, 379–390.

Crick, F.H. (1966). Codon--anticodon pairing: the wobble hypothesis. J. Mol. Biol. *19*, 548–555.

Crick, F.H.C. (1958). On Protein Synthesis. Symp Soc Exp Biol *12*, 138–163.

Crick, F.H.C. (1968). The origin of the genetic code. J. Mol. Biol. *38*, 367–379.

D Mendel, V W Cornish, and Schultz, and P.G. (1995). Site-Directed Mutagenesis with an Expanded Genetic Code. Annu. Rev. Biophys. Biomol. Struct. *24*, 435–462.

Datta, D., Wang, P., Carrico, I.S., Mayo, S.L., and Tirrell, D.A. (2002). A Designed Phenylalanyl-tRNA Synthetase Variant Allows Efficient in Vivo Incorporation of Aryl Ketone Functionality into Proteins. J. Am. Chem. Soc. *124*, 5652–5653.

Deatherage, D., and Barrick, J. (2014). Identification of Mutations in Laboratory-Evolved Microbes from Next-Generation Sequencing Data Using breseq. In Engineering and Analyzing Multicellular Systems, L. Sun, and W. Shou, eds. (Springer New York), pp. 165–188.

Deiters, A., and Schultz, P.G. (2005). In vivo incorporation of an alkyne into proteins in Escherichia coli. Bioorg. Med. Chem. Lett. *15*, 1521–1524.

Dierks, T., Lecca, M.R., Schlotterhose, P., Schmidt, B., and von Figura, K. (1999). Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases. EMBO J. *18*, 2084–2091.

Dieterich, D.C., Lee, J.J., Link, A.J., Graumann, J., Tirrell, D.A., and Schuman, E.M. (2007). Labeling, detection and identification of newly synthesized proteomes with bioorthogonal non-canonical amino-acid tagging. Nat. Protoc. *2*, 532–540.

Dihazi, G.H., and Sinz, A. (2003). Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry. Rapid Commun. Mass Spectrom. *17*, 2005–2014.

Edwards, H., and Schimmel, P. (1990). A bacterial amber suppressor in Saccharomyces cerevisiae is selectively recognized by a bacterial aminoacyl-tRNA synthetase. Mol. Cell. Biol. *10*, 1633–1641.

Ellefson, J.W., Meyer, A.J., Hughes, R.A., Cannon, J.R., Brodbelt, J.S., and Ellington, A.D. (2014). Directed evolution of genetic parts and circuits by compartmentalized partnered replication. Nat. Biotechnol. *32*, 97–101.

Ellefson, J.W., Thyer, R., Wang, B., Gollihar, J., Forster, M.T., and Ellington, A.D. (2016). Addicting diverse bacteria to a noncanonical amino acid. Nat. Chem. Biol. *12*, 138–140.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

Ellington, A.D., and Szostak, J.W. (1992). Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. Nature *355*, 850–852.

Farady, C.J., Sellers, B.D., Jacobson, M.P., and Craik, C.S. (2009). Improving the species cross-reactivity of an antibody using computational design. Bioorg. Med. Chem. Lett. *19*, 3744–3747.

Feilmeier, B.J., Iseminger, G., Schroeder, D., Webber, H., and Phillips, G.J. (2000). Green Fluorescent Protein Functions as a Reporter for Protein Localization in Escherichia coli. J. Bacteriol. *182*, 4068–4076.

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., and Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. *6*, 99.

Fontaine, F., Fuchs, R.T., and Storz, G. (2011). Membrane Localization of Small Proteins in Escherichia coli. J. Biol. Chem. *286*, 32464–32474.

Fonze, E. (1995). Acta Crystallogr Biol Crystallogr *51*, 682–694.

Forné, I., Ludwigsen, J., Imhof, A., Becker, P.B., and Mueller-Planitz, F. (2012). Probing the conformation of the ISWI ATPase domain with genetically encoded photoreactive crosslinkers and mass spectrometry. Mol. Cell. Proteomics MCP *11*, M111.012088.

Forster, A.C., Tan, Z., Nalam, M.N.L., Lin, H., Qu, H., Cornish, V.W., and Blacklow, S.C. (2003). Programming peptidomimetic syntheses by translating genetic codes designed de novo. Proc. Natl. Acad. Sci. U. S. A. *100*, 6353–6357.

Francklyn, C. (2003). tRNA synthetase paralogs: Evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to de novo biosynthesis. Proc. Natl. Acad. Sci. *100*, 9650–9652.

Freeland, S.J., and Hurst, L.D. The Genetic Code Is One in a Million. J. Mol. Evol. *47*, 238–248.

Furter, R. (1998). Expansion of the genetic code: site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. Protein Sci. Publ. Protein Soc. *7*, 419–426.

Galardy, R.E., Craig, L.C., and Printz, M.P. (1973). Benzophenone triplet: a new photochemical probe of biological ligand-receptor interactions. Nature. New Biol. *242*, 127–128.

Gaston, M.A., Zhang, L., Green-Church, K.B., and Krzycki, J.A. (2011). The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. Nature *471*, 647–650.

Gautier, A., Juillerat, A., Heinis, C., Corrêa Jr., I.R., Kindermann, M., Beaufils, F., and Johnsson, K. (2008). An Engineered Protein Tag for Multiprotein Labeling in Living Cells. Chem. Biol. *15*, 128–136.

Gayán, E., Cambré, A., Michiels, C.W., and Aertsen, A. (2016). Stress-induced evolution of heat resistance and resuscitation speed in E. coli O157:H7 ATCC 43888. Appl. Environ. Microbiol. AEM.02027-16.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balázsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., et al. (2003). Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. J. Bacteriol. *185*, 5673–5684.

Gibson, D.G. (2009). Nat Methods *6*, 343–345.

Gieseg, S.P., Simpson, J.A., Charlton, T.S., Duncan, M.W., and Dean, R.T. (1993). Protein-bound 3,4-dihydroxyphenylalanine is a major reductant formed during hydroxyl radical damage to proteins. Biochemistry (Mosc.) *32*, 4780–4786.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science *309*, 1242–1245.

Goodarzi, H., Bennett, B.D., Amini, S., Reaves, M.L., Hottes, A.K., Rabinowitz, J.D., and Tavazoie, S. (2010). Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in E. coli. Mol. Syst. Biol. *6*, 378.

Goto, Y., Katoh, T., and Suga, H. (2011). Flexizymes for genetic code reprogramming. Nat. Protoc. *6*, 779–790.

Guo, J., Melançon, C.E., Lee, H.S., Groff, D., and Schultz, P.G. (2009). Evolution of Amber Suppressor tRNAs for Efficient Bacterial Production of Unnatural Amino Acid-Containing Proteins. Angew. Chem. Int. Ed Engl. *48*, 9148–9151.

Gustavsson, N., Diez, A., and Nyström, T. (2002). The universal stress protein paralogues of Escherichia coli are co-ordinately regulated and co-operate in the defence against DNA damage. Mol. Microbiol. *43*, 107–117.

Haig, D., and Hurst, L.D. (1991). A quantitative measure of error minimization in the genetic code. J. Mol. Evol. *33*, 412–417.

Hammerling, M.J., Ellefson, J.W., Boutz, D.R., Marcotte, E.M., Ellington, A.D., and Barrick, J.E. (2014). Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness. Nat. Chem. Biol. *10*, 178–180.

Hammerling, M.J., Gollihar, J., Mortensen, C., Alnahhas, R.N., Ellington, A.D., and Barrick, J.E. (2016). Expanded genetic codes create new mutational routes to rifampicin resistance in Escherichia coli. Mol. Biol. Evol. msw094.

Hanoulle, X., Rollet, E., Clantin, B., Landrieu, I., Ödberg-Ferragut, C., Lippens, G., Bohin, J.-P., and Villeret, V. (2004). Structural Analysis of Escherichia coli OpgG, a Protein Required for the Biosynthesis of Osmoregulated Periplasmic Glucans. J. Mol. Biol. *342*, 195–205.

Hanson, G.T., Aggeler, R., Oglesbee, D., Cannon, M., Capaldi, R.A., Tsien, R.Y., and Remington, S.J. (2004). Investigating Mitochondrial Redox Potential with Redox-sensitive Green Fluorescent Protein Indicators. J. Biol. Chem. *279*, 13044–13053.

Hao, B., Gong, W., Ferguson, T.K., James, C.M., Krzycki, J.A., and Chan, M.K. (2002). A New UAG-Encoded Residue in the Structure of a Methanogen Methyltransferase. Science *296*, 1462–1466.

Haring, V., Scholz, P., Scherzinger, E., Frey, J., Derbyshire, K., Hatfull, G., Willetts, N.S., and Bagdasarian, M. (1985). Protein RepC is involved in copy number control of the broad host range plasmid RSF1010. Proc. Natl. Acad. Sci. U. S. A. *82*, 6090–6094.

Harvey, B.R., Georgiou, G., Hayhurst, A., Jeong, K.J., Iverson, B.L., and Rogers, G.K. (2004). Anchored periplasmic expression, a versatile technology for the isolation of high-affinity antibodies from Escherichia coli-expressed libraries. Proc. Natl. Acad. Sci. U. S. A. *101*, 9193–9198.

Hashimoto-Gotoh, T. (2000). Gene *241*, 185–191.

Hemm, M.R., Paul, B.J., Miranda-Ríos, J., Zhang, A., Soltanzad, N., and Storz, G. (2010). Small Stress Response Proteins in Escherichia coli: Proteins Missed by Classical Proteomic Studies. J. Bacteriol. *192*, 46–58.

Herring, S., Ambrogelly, A., Polycarpo, C.R., and Söll, D. (2007). Recognition of pyrrolysine tRNA by the Desulfitobacterium hafniense pyrrolysyl-tRNA synthetase. Nucleic Acids Res. *35*, 1270–1278.

van Hest, J.C.M., Kiick, K.L., and Tirrell, D.A. (2000). Efficient Incorporation of Unsaturated Methionine Analogues into Proteins in Vivo. J. Am. Chem. Soc. *122*, 1282–1288.

Hinegardner, R.T., and Engelberg, J. (1963). RATIONALE FOR A UNIVERSAL GENETIC CODE. Science *142*, 1083–1085.

Hofmann, K., and Bohn, H. (1966). Studies on Polypeptides. XXXVI. The Effect of Pyrazole—Imidazole Replacements on the S-Protein Activating Potency of an S-Peptide Fragment1-3. J. Am. Chem. Soc. *88*, 5914–5919.

Horinouchi, T., Suzuki, S., Hirasawa, T., Ono, N., Yomo, T., Shimizu, H., and Furusawa, C. (2015). Phenotypic convergence in bacterial adaptive evolution to ethanol stress. BMC Evol. Biol. *15*.

Hortin, G., and Boime, I. (1983). [61] Applications of amino acid analogs for studying co- and posttranslational modifications of proteins. B.-M. in Enzymology, ed. (Academic Press), pp. 777–784.

Hudson, A.J., Andrews, S.C., Hawkins, C., Williams, J.M., Izuhara, M., Meldrum, F.C., Mann, S., Harrison, P.M., and Guest, J.R. (1993). Overproduction, purification and characterization of the Escherichia coli ferritin. Eur. J. Biochem. *218*, 985–995.

Hughes, R.A., and Ellington, A.D. (2010). Rational design of an orthogonal tryptophanyl nonsense suppressor tRNA. Nucleic Acids Res. *38*, 6813–6830.

Hwang, D.S., Yoo, H.J., Jun, J.H., Moon, W.K., and Cha, H.J. (2004). Expression of Functional Recombinant Mussel Adhesive Protein Mgfp-5 in Escherichia coli. Appl. Environ. Microbiol. *70*, 3352–3359.

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*, 272–275.

Ibba, M., Kast, P., and Hennecke, H. (1994). Substrate Specificity Is Determined by Amino Acid Binding Pocket Size in Escherichia coli Phenylalanyl-tRNA Synthetase. Biochemistry (Mosc.) *33*, 7107–7112.

Imburgio, D., Rong, M., Ma, K., and McAllister, W.T. (2000). Studies of Promoter Recognition and Start Site Selection by T7 RNA Polymerase Using a Comprehensive Collection of Promoter Variants. Biochemistry (Mosc.) *39*, 10419–10430.

Inagaki, Y., Ehara, M., Watanabe, K.I., Hayashi-Ishimaru, Y., and Ohama, T. (1998). Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. J. Mol. Evol. *47*, 378–384.

Isaacs, F.J., Carr, P.A., Wang, H.H., Lajoie, M.J., Sterling, B., Kraal, L., Tolonen, A.C., Gianoulis, T.A., Goodman, D.B., Reppas, N.B., et al. (2011). Precise Manipulation of Chromosomes in Vivo Enables Genome-Wide Codon Replacement. Science *333*, 348–353.

Jacquier, H., Birgy, A., Nagard, H.L., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., et al. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. Proc. Natl. Acad. Sci. *110*, 13067–13072.

Jaffe, A., Chabbert, Y.A., and Semonin, O. (1982). Role of porin proteins OmpF and OmpC in the permeation of beta-lactams. Antimicrob. Agents Chemother. *22*, 942–948.

Jaffé, A., Chabbert, Y.A., and Derlot, E. (1983). Selection and characterization of beta-lactam-resistant Escherichia coli K-12 mutants. Antimicrob. Agents Chemother. *23*, 622–625.

Johnson, D.B.F., Xu, J., Shen, Z., Takimoto, J.K., Schultz, M.D., Schmitz, R.J., Ecker, J.R., Briggs, S.P., and Wang, L. (2011). RF1 Knockout Allows Ribosomal Incorporation of Unnatural Amino Acids at Multiple Sites. Nat. Chem. Biol. *7*, 779–786.

Johnson, D.B.F., Wang, C., Xu, J., Schultz, M.D., Schmitz, R.J., Ecker, J.R., and Wang, L. (2012). Release Factor One Is Nonessential in Escherichia coli. ACS Chem. Biol. *7*, 1337–1344.

Jones, D.H., Cellitti, S.E., Hao, X., Zhang, Q., Jahnz, M., Summerer, D., Schultz, P.G., Uno, T., and Geierstanger, B.H. (2009). Site-specific labeling of proteins with NMR-active unnatural amino acids. J. Biomol. NMR *46*, 89.

Juillerat, A., Gronemeyer, T., Keppler, A., Gendreizig, S., Pick, H., Vogel, H., and Johnsson, K. (2003). Directed Evolution of O6-Alkylguanine-DNA Alkyltransferase for Efficient Labeling of Fusion Proteins with Small Molecules In Vivo. Chem. Biol. *10*, 313–317.

Kaiser, J.T., Gromadski, K., Rother, M., Engelhardt, H., Rodnina, M.V., and Wahl, M.C. (2005). Structural and functional investigation of a putative archaeal selenocysteine synthase. Biochemistry (Mosc.) *44*, 13315–13327.

Kamiyama, D., Sekine, S., Barsi-Rhyne, B., Hu, J., Chen, B., Gilbert, L.A., Ishikawa, H., Leonetti, M.D., Marshall, W.F., Weissman, J.S., et al. (2016). Versatile protein tagging in cells with split fluorescent protein. Nat. Commun. *7*, 11046.

van Kasteren, S.I., Kramer, H.B., Gamblin, D.P., and Davis, B.G. (2007). Site-selective glycosylation of proteins: creating synthetic glycoproteins. Nat. Protoc. *2*, 3185–3194.

Kather, I., Jakob, R.P., Dobbek, H., and Schmid, F.X. (2008). Increased Folding Stability of TEM-1 β-Lactamase by In Vitro Selection. J. Mol. Biol. *383*, 238–251.

Kato, Y. (2015). PeerJ *3*, e1247.

Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H., and Meiler, J. (2010). Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. Biochemistry (Mosc.) *49*, 2987–2998.

Keppler, A., Gendreizig, S., Gronemeyer, T., Pick, H., Vogel, H., and Johnsson, K. (2003). A general method for the covalent labeling of fusion proteins with small molecules in vivo. Nat. Biotechnol. *21*, 86–89.

Kim, J., and Yoon, M.-Y. (2010). Recent advances in rapid and ultrasensitive biosensors for infectious agents: lesson from Bacillus anthracis diagnostic sensors. Analyst *135*, 1182–1190.

Kim, C.H., Axup, J.Y., Dubrovska, A., Kazane, S.A., Hutchins, B.A., Wold, E.D., Smider, V.V., and Schultz, P.G. (2012). Synthesis of bispecific antibodies using genetically encoded unnatural amino acids. J. Am. Chem. Soc. *134*, 9918–9921.

Kirshenbaum, K., Carrico, I.S., and Tirrell, D.A. (2002). Biosynthesis of Proteins Incorporating a Versatile Set of Phenylalanine Analogues. ChemBioChem *3*, 235–237.

Knight, R.D., and Landweber, L.F. (1998). Rhyme or reason: RNA-arginine interactions and the genetic code. Chem. Biol. *5*, R215–R220.

Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. *2*, 49–58.

Kobayashi, K., Tsuchiya, M., Oshima, T., and Yanagawa, H. Abiotic synthesis of amino acids and imidazole by proton irradiation of simulated primitive earth atmospheres. Orig. Life Evol. Biosph. *20*, 99–109.

Kobayashi, T., Nureki, O., Ishitani, R., Yaremchuk, A., Tukalo, M., Cusack, S., Sakamoto, K., and Yokoyama, S. (2003). Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. Nat. Struct. Mol. Biol. *10*, 425–432.

Kolodrubetz, D., and Schleif, R. (1981). L-arabinose transport systems in Escherichia coli K-12. J. Bacteriol. *148*, 472–479.

Koonin, E.V., and Novozhilov, A.S. (2009). Origin and evolution of the genetic code: the universal enigma. IUBMB Life *61*, 99–111.

Kortemme, T., and Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. Proc. Natl. Acad. Sci. *99*, 14116–14121.

Kramer, E.B., and Farabaugh, P.J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA *13*, 87–96.

Kryukov, G.V., Kryukov, V.M., and Gladyshev, V.N. (1999). New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. J. Biol. Chem. *274*, 33888–33897.

Kuczyńska-Wiśnik, D., Stojowska, K., Matuszewska, E., Leszczyńska, D., Algara, M.M., Augustynowicz, M., and Laskowska, E. (2015). Lack of intracellular trehalose affects formation of Escherichia coli persister cells. Microbiol. Read. Engl. *161*, 786–796.

Kuroda, D., Shirai, H., Jacobson, M.P., and Nakamura, H. (2012). Computer-aided antibody design. Protein Eng. Des. Sel. *25*, 507–522.

Lajoie, M.J., Rovner, A.J., Goodman, D.B., Aerni, H.-R., Haimovich, A.D., Kuznetsov, G., Mercer, J.A., Wang, H.H., Carr, P.A., Mosberg, J.A., et al. (2013). Genomically Recoded Organisms Expand Biological Functions. Science *342*, 357–360.

Le Grand, S.M., and Merz, K.M. (1993). Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. J. Comput. Chem. *14*, 349–352.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K.W., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In Methods in Enzymology, M.L.J. and L. Brand, ed. (Academic Press), pp. 545–574.

Lee, H.S., Dimla, R.D., and Schultz, P.G. (2009a). Protein-DNA photo-crosslinking with a genetically encoded benzophenone-containing amino acid. Bioorg. Med. Chem. Lett. *19*, 5222–5224.

Lee, H.S., Spraggon, G., Schultz, P.G., and Wang, F. (2009b). Genetic Incorporation of a Metal-ion Chelating Amino Acid into Proteins as a Biophysical Probe. J. Am. Chem. Soc. *131*, 2481–2483.

Lee, S., Oh, S., Yang, A., Kim, J., Söll, D., Lee, D., and Park, H.-S. (2013). A Facile Strategy for Selective Incorporation of Phosphoserine into Histones. Angew. Chem. Int. Ed. *52*, 5771–5775.

Lemeignan, B., Sonigo, P., and Marlière, P. (1993). Phenotypic suppression by incorporation of an alien amino acid. J. Mol. Biol. *231*, 161–166.

Lenski, R.E., Rose, M.R., Simpson, S.C., and Tadler, S.C. (1991). Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. Am. Nat. *138*, 1315–1341.

Levine, M., and Tarver, H. (1951). Studies on Ethionine Iii. Incorporation of Ethionine into Rat Proteins. J. Biol. Chem. *192*, 835–850.

Leying, H.J., Büscher, K.H., Cullmann, W., and Then, R.L. (1992). Lipopolysaccharide alterations responsible for combined quinolone and beta-lactam resistance in Pseudomonas aeruginosa. Chemotherapy *38*, 82–91.

Leysath, C.E., Monzingo, A.F., Maynard, J.A., Barnett, J., Georgiou, G., Iverson, B.L., and Robertus, J.D. (2009). Crystal structure of the engineered neutralizing antibody M18 complexed to domain 4 of the anthrax protective antigen. J. Mol. Biol. *387*, 680–693.

Lieber, A., Kiessling, U., and Strauss, M. (1989). High level gene expression in mammalian cells by a nuclear T7-phase RNA polymerase. Nucleic Acids Res. *17*, 8485–8493.

Link, A.J., Mock, M.L., and Tirrell, D.A. (2003). Non-canonical amino acids in protein engineering. Curr. Opin. Biotechnol. *14*, 603–609.

Lippow, S.M., and Tidor, B. (2007). Progress in computational protein design. Curr. Opin. Biotechnol. *18*, 305–311.

Lippow, S.M., Wittrup, K.D., and Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. Nat. Biotechnol. *25*, 1171–1176.

Little, S.F., Leppla, S.H., and Cora, E. (1988). Production and characterization of monoclonal antibodies to the protective antigen component of Bacillus anthracis toxin. Infect. Immun. *56*, 1807–1813.

Liu, C.C., and Schultz, P.G. (2010). Adding New Chemistries to the Genetic Code. Annu. Rev. Biochem. *79*, 413–444.

Liu, B., Burdine, L., and Kodadek, T. (2006). Chemistry of periodate-mediated cross-linking of 3,4-dihydroxylphenylalanine-containing molecules to proteins. J. Am. Chem. Soc. *128*, 15228–15235.

Liu, C.C., Choe, H., Farzan, M., Smider, V.V., and Schultz, P.G. (2009). Mutagenesis and Evolution of Sulfated Antibodies Using an Expanded Genetic Code. Biochemistry (Mosc.) *48*, 8891–8898.

Liu, D.R., Magliery, T.J., Pastrnak, M., and Schultz, P.G. (1997). Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. Proc. Natl. Acad. Sci. U. S. A. *94*, 10092–10097.

Liu, J., Castañeda, C.A., Wilkins, B.J., Fushman, D., and Cropp, T.A. (2010). Condensed E. coli cultures for highly efficient production of proteins containing unnatural amino acids. Bioorg. Med. Chem. Lett. *20*, 5613–5616.

Luo, X., Wang, T.-S.A., Zhang, Y., Wang, F., and Schultz, P.G. (2016). Stabilizing Protein Motifs with a Genetically Encoded Metal-Ion Chelator. Cell Chem. Biol. *23*, 1098–1102.

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res. *25*, 1203–1210.

Lynch, M. (2006). Streamlining and Simplification of Microbial Genome Architecture. Annu. Rev. Microbiol. *60*, 327–349.

Macino, G., Coruzzi, G., Nobrega, F.G., Li, M., and Tzagoloff, A. (1979). Use of the UGA terminator as a tryptophan codon in yeast mitochondria. Proc. Natl. Acad. Sci. U. S. A. *76*, 3784–3785.

Mandell, D.J. (2015). Nature *518*, 55–60.

Mandell, D.J., and Kortemme, T. (2009). Computer-aided design of functional protein interactions. Nat. Chem. Biol. *5*, 797–807.

Maranhao, A.C., and Ellington, A.D. (2016). Evolving Orthogonal Suppressor tRNAs To Incorporate Modified Amino Acids. ACS Synth. Biol.

Marcatili, P., Rosi, A., and Tramontano, A. (2008). PIGS: automatic prediction of antibody structures. Bioinformatics *24*, 1953–1954.

Massey, S.E., and Garey, J.R. (2007). A Comparative Genomics Analysis of Codon Reassignments Reveals a Link with Mitochondrial Proteome Size and a Mechanism of Genetic Code Change Via Suppressor tRNAs. J. Mol. Evol. *64*, 399–410.

Matthaei, J.H., Jones, O.W., Martin, R.G., and Nirenberg, M.W. (1962). CHARACTERISTICS AND COMPOSITION OF RNA CODING UNITS*. Proc. Natl. Acad. Sci. U. S. A. *48*, 666–677.

Meyer, A.J., Ellefson, J.W., and Ellington, A.D. (2015). Directed Evolution of a Panel of Orthogonal T7 RNA Polymerase Variants for in Vivo or in Vitro Synthetic Circuitry. ACS Synth. Biol. *4*, 1070–1076.

Midelfort, K.S., Hernandez, H.H., Lippow, S.M., Tidor, B., Drennan, C.L., and Wittrup, K.D. (2004). Substantial Energetic Improvement with Minimal Structural Perturbation in a High Affinity Mutant Antibody. J. Mol. Biol. *343*, 685–701.

Miklos, A.E., Kluwe, C., Der, B.S., Pai, S., Sircar, A., Hughes, R.A., Berrondo, M., Xu, J., Codrea, V., Buckley, P.E., et al. (2012). Structure-Based Design of Supercharged, Highly Thermoresistant Antibodies. Chem. Biol. *19*, 449–455.

Misu, Y., Goshima, Y., and Miyamae, T. (2002). Is DOPA a neurotransmitter? Trends Pharmacol. Sci. *23*, 262–268.

Moayeri, M., and Leppla, S.H. (2004). The roles of anthrax toxin in pathogenesis. Curr. Opin. Microbiol. *7*, 19–24.

Mukai, T., Hayashi, A., Iraha, F., Sato, A., Ohtake, K., Yokoyama, S., and Sakamoto, K. (2010). Codon reassignment in the Escherichia coli genetic code. Nucleic Acids Res. *38*, 8188–8195.

Mukai, T., Yanagisawa, T., Ohtake, K., Wakamori, M., Adachi, J., Hino, N., Sato, A., Kobayashi, T., Hayashi, A., Shirouzu, M., et al. (2011). Genetic-code evolution for protein synthesis with non-natural amino acids. Biochem. Biophys. Res. Commun. *411*, 757–761.

Mukai, T., Hoshi, H., Ohtake, K., Takahashi, M., Yamaguchi, A., Hayashi, A., Yokoyama, S., and Sakamoto, K. (2015). Highly reproductive Escherichia coli cells with no specific assignment to the UAG codon. Sci. Rep. *5*, 9699.

Murgola, E.J. (1985). tRNA, Suppression, and the Code. Annu. Rev. Genet. *19*, 57–80.

Neumann, H., Peak-Chew, S.Y., and Chin, J.W. (2008a). Genetically encoding Nϵ-acetyllysine in recombinant proteins. Nat. Chem. Biol. *4*, 232–234.

Neumann, H., Hazen, J.L., Weinstein, J., Mehl, R.A., and Chin, J.W. (2008b). Genetically Encoding Protein Oxidative Damage. J. Am. Chem. Soc. *130*, 4028–4033.

Neumann, H., Wang, K., Davis, L., Garcia-Alai, M., and Chin, J.W. (2010). Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature *464*, 441–444.

Ng, J.Y., Boelen, L., and Wong, J.W.H. (2013). Bioinformatics analysis reveals biophysical and evolutionary insights into the 3-nitrotyrosine post-translational modification in the human proteome. Open Biol. *3*.

Ngo, J.T., and Tirrell, D.A. (2011). Noncanonical Amino Acids in the Interrogation of Cellular Protein Synthesis. Acc. Chem. Res. *44*, 677–685.

Noren, C.J., Anthony-Cahill, S.J., Griffith, M.C., and Schultz, P.G. (1989). A general method for site-specific incorporation of unnatural amino acids into proteins. Science *244*, 182–188.

Ohno, S., Yokogawa, T., Fujii, I., Asahara, H., Inokuchi, H., and Nishikawa, K. (1998). Co-Expression of Yeast Amber Suppressor tRNATyr and Tyrosyl-tRNA Synthetase in Escherichia coli: Possibility to Expand the Genetic Code. J. Biochem. (Tokyo) *124*, 1065–1068.

Ohtake, K. (2015). Sci Rep *5*, 9762.

Ohuchi, M., Murakami, H., and Suga, H. (2007). The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus. Curr. Opin. Chem. Biol. *11*, 537–542.

Osawa, S., and Jukes, T.H. Codon reassignment (codon capture) in evolution. J. Mol. Evol. *28*, 271–278.

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. Microbiol. Rev. *56*, 229–264.

Ostrov, N., Landon, M., Guell, M., Kuznetsov, G., Teramoto, J., Cervantes, N., Zhou, M., Singh, K., Napolitano, M.G., Moosburner, M., et al. (2016). Design, synthesis, and testing toward a 57-codon genome. Science *353*, 819–822.

Pagani, L., Landini, P., Luzzaro, F., Debiaggi, M., and Romero, E. (1990). Emergence of cross-resistance to imipenem and other beta-lactam antibiotics in Pseudomonas aeruginosa during therapy. Microbiologica *13*, 43–53.

Palzkill, T., Le, Q.-Q., Venkatachalam, K.V., LaRocco, M., and Ocera, H. (1994). Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of β-lactamase. Mol. Microbiol. *12*, 217–229.

Pantazes, R.J., and Maranas, C.D. (2010). OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. Protein Eng. Des. Sel. *23*, 849–858.

Pantoja, R., Rodriguez, E.A., Dibas, M.I., Dougherty, D.A., and Lester, H.A. (2009). Single-Molecule Imaging of a Fluorescent Unnatural Amino Acid Incorporated Into Nicotinic Receptors. Biophys. J. *96*, 226–237.

Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A., and Bernhardt, T.G. (2014). A genome-wide screen for bacterial envelope biogenesis mutants identifies a novel factor involved in cell wall precursor metabolism. PLoS Genet. *10*, e1004056.

Park, H.-S., Hohn, M.J., Umehara, T., Guo, L.-T., Osborne, E.M., Benner, J., Noren, C.J., Rinehart, J., and Söll, D. (2011). Expanding the Genetic Code of Escherichia coli with Phosphoserine. Science *333*, 1151–1154.

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

Pastrnak, M., Magliery, T.J., and Schultz, P.G. (2000). A New Orthogonal Suppressor tRNA/Aminoacyl-tRNA Synthetase Pair for Evolving an Organism with an Expanded Genetic Code. Helv. Chim. Acta *83*, 2277–2286.

Pelc, S.R. (1965). Correlation between coding-triplets and amino-acids. Nature *207*, 597–599.

Pelc, S.R., and Welton, M.G. (1966). Stereochemical relationship between coding triplets and amino-acids. Nature *209*, 868–870.

Peretz, Y., Levy, M., Avisar, E., Edelbaum, O., Rabinowitch, H., and Sela, I. (2007). A T7-driven silencing system in transgenic plants expressing T7 RNA polymerase is a nuclear process. Transgenic Res. *17*, 665–677.

Perilli, M., Mugnaioli, C., Luzzaro, F., Fiore, M., Stefani, S., Rossolini, G.M., and Amicosante, G. (2005). Novel TEM-type extended-spectrum beta-lactamase, TEM-134, in a Citrobacter koseri clinical isolate. Antimicrob. Agents Chemother. *49*, 1564–1566.

Petrosino, J., Cantu, C., and Palzkill, T. (1998). β-Lactamases: protein evolution in real time. Trends Microbiol. *6*, 323–327.

Pham, N.D., Parker, R.B., and Kohler, J.J. (2013). Photocrosslinking approaches to interactome mapping. Curr. Opin. Chem. Biol. *17*, 90–101.

Pimenova, T., Nazabal, A., Roschitzki, B., Seebacher, J., Rinner, O., and Zenobi, R. (2008). Epitope mapping on bovine prion protein using chemical cross-linking and mass spectrometry. J. Mass Spectrom. *43*, 185–195.

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G.M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S.S., Arruda, M. de, et al. (2006). Emergent Properties of Reduced-Genome Escherichia coli. Science *312*, 1044–1046.

Pratt, E.A., and Ho, C. (1975). Incorporation of fluorotryptophans into proteins of escherichia coli. Biochemistry (Mosc.) *14*, 3035–3040.

Raskin, C.A., Diaz, G.A., and McAllister, W.T. (1993). T7 RNA polymerase mutants with altered promoter specificities. Proc. Natl. Acad. Sci. U. S. A. *90*, 3147–3151.

Reese, M.G. (2001). Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput. Chem. *26*, 51–56.

Renfrew, P.D., Choi, E.J., Bonneau, R., and Kuhlman, B. (2012). Incorporation of Noncanonical Amino Acids into Rosetta and Use in Computational Protein-Peptide Interface Design. PLOS ONE *7*, e32637.

Richmond, M.H. (1962). The effect of amino acid analogues on growth and protein synthesis in microorganisms. Bacteriol. Rev. *26*, 398–420.

Roffler-Tarlov, S., Liu, J.H., Naumova, E.N., Bernal-Ayala, M.M., and Mason, C.A. (2013). L-Dopa and the Albino Riddle: Content of L-Dopa in the Developing Retina of Pigmented and Albino Mice. PLoS ONE *8*.

Rogerson, D.T., Sachdeva, A., Wang, K., Haq, T., Kazlauskaite, A., Hancock, S.M., Huguenin-Dezot, N., Muqit, M.M.K., Fry, A.M., Bayliss, R., et al. (2015). Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. Nat. Chem. Biol. *11*, 496–503.

Rong, M., He, B., McAllister, W.T., and Durbin, R.K. (1998). Promoter specificity determinants of T7 RNA polymerase. Proc. Natl. Acad. Sci. U. S. A. *95*, 515–519.

Rovner, A.J. (2015). Nature *518*, 89–93.

Roy, H., Becker, H.D., Reinbolt, J., and Kern, D. (2003). When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. Proc. Natl. Acad. Sci. *100*, 9837–9842.

Rudolph, B., Gebendorfer, K.M., Buchner, J., and Winter, J. (2010). Evolution of Escherichia coli for Growth at High Temperatures. J. Biol. Chem. *285*, 19029–19034.

Ruiz, C., and Levy, S.B. (2011). Use of functional interactions with MarA to discover chromosomal genes affecting antibiotic susceptibility in Escherichia coli. Int. J. Antimicrob. Agents *37*, 177–178.

Rush, J.S., and Bertozzi, C.R. (2008). New Aldehyde Tag Sequences Identified by Screening Formylglycine Generating Enzymes in Vitro and in Vivo. J. Am. Chem. Soc. *130*, 12240–12241.

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

Sakamoto, K., Murayama, K., Oki, K., Iraha, F., Kato-Murayama, M., Takahashi, M., Ohtake, K., Kobayashi, T., Kuramitsu, S., Shirouzu, M., et al. (2009). Genetic Encoding of 3-Iodo-l-Tyrosine in Escherichia coli for Single-Wavelength Anomalous Dispersion Phasing in Protein Crystallography. Structure *17*, 335–344.

Salverda, M.L.M., De Visser, J.A.G.M., and Barlow, M. (2010). Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. FEMS Microbiol. Rev. *34*, 1015–1036.

Santos, J., and Monteagudo, Á. (2011). Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. BMC Bioinformatics *12*, 56.

Santos, M.A., and Tuite, M.F. (1995). The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. Nucleic Acids Res. *23*, 1481–1486.

Sato, S., Mimasu, S., Sato, A., Hino, N., Sakamoto, K., Umehara, T., and Yokoyama, S. (2011). Crystallographic Study of a Site-Specifically Cross-Linked Protein Complex with a Genetically Incorporated Photoreactive Amino Acid,. Biochemistry (Mosc.) *50*, 250–257.

Saxinger, C., Ponnamperuma, C., and Woese, C. (1971). Evidence for the Interaction of Nucleotides with Immobilized Amino-acids and its Significance for the Origin of the Genetic Code. Nature *234*, 172–174.

Schattner, P., Brooks, A.N., and Lowe, T.M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. *33*, W686–W689.

Schenborn, E.T., and Mierendorf, R.C. (1985). A novel transcription property of SP6 and T7 RNA polymerases: dependence on template structure. Nucleic Acids Res. *13*, 6223–6236.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Schultz, D.W., and Yarus, M. (1994). Transfer RNA Mutation and the Malleability of the Genetic Code. J. Mol. Biol. *235*, 1377–1380.

Scolnick, E., Tompkins, R., Caskey, T., and Nirenberg, M. (1968). Release factors differing in specificity for terminator codons. Proc. Natl. Acad. Sci. U. S. A. *61*, 768–774.

Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T., and Riley, M. (2001). A functional update of the Escherichia coliK-12 genome. Genome Biol. *2*, research0035.

Sheppard, K., Yuan, J., Hohn, M.J., Jester, B., Devine, K.M., and Söll, D. (2008). From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Res. *36*, 1813–1825.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nat. Biotechnol. *19*, 751–755.

Shiota, T., Nishikawa, S., and Endo, T. (2013). Analyses of Protein–Protein Interactions by In Vivo Photocrosslinking in Budding Yeast. In Membrane Biogenesis, D. Rapaport, and J.M. Herrmann, eds. (Humana Press), pp. 207–217.

Siegele, D.A. (2005). Universal Stress Proteins in Escherichia coli. J. Bacteriol. *187*, 6253–6254.

Siles, E., Martinez-Lara, E., Núñez, M. i., Muñoz-Gámez, J. a., Martín-Oliva, D., Valenzuela, M. t., Peinado, M. a., Ruiz de Almodóvar, J. m., and Javier Oliver, F. (2005). PARP-1-dependent 3-nitrotyrosine protein modification after DNA damage. J. Cell. Biochem. *96*, 709–715.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins *34*, 82–95.

Sircar, A., and Gray, J.J. (2010). SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models. PLOS Comput Biol *6*, e1000644.

Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J.J. (2009). Toward high-resolution homology modeling of antibody Fv regions and application to antibody–antigen docking. Proteins Struct. Funct. Bioinforma. *74*, 497–514.

Smith, K.M., and Liao, J.C. (2011). An evolutionary strategy for isobutanol production strain development in Escherichia coli. Metab. Eng. *13*, 674–681.

Song, W., Wang, Y., Qu, J., Madden, M.M., and Lin, Q. (2008). A Photoinducible 1,3-Dipolar Cycloaddition Reaction for Rapid, Selective Modification of Tetrazole-Containing Proteins. Angew. Chem. Int. Ed. *47*, 2832–2835.

Stafford, G.P., Ogi, T., and Hughes, C. (2005). Binding and transcriptional activation of non-flagellar genes by the Escherichia coli flagellar master regulator FlhD2C2. Microbiol. Read. Engl. *151*, 1779–1788.

Staros, J.V., Wright, R.W., and Swingle, D.M. (1986). Enhancement by N-hydroxysulfosuccinimide of water-soluble carbodiimide-mediated coupling reactions. Anal. Biochem. *156*, 220–222.

Stec, B., Holtz, K.M., Wojciechowski, C.L., and Kantrowitz, E.R. (2005). Acta Crystallogr Biol Crystallogr *61*, 1072–1079.

Steer, B.A., and Schimmel, P. (1999). Major Anticodon-binding Region Missing from an Archaebacterial tRNA Synthetase. J. Biol. Chem. *274*, 35601–35606.

Stojanoski, V., Chow, D.-C., Hu, L., Sankaran, B., Gilbert, H.F., Prasad, B.V.V., and Palzkill, T. (2015). A Triple Mutant in the Ω-loop of TEM-1 β-Lactamase Changes the Substrate Profile via a Large Conformational Change and an Altered General Base for Catalysis. J. Biol. Chem. *290*, 10382–10394.

Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. Genes Dev. *12*, 599–606.

Studier, F.W., and Moffatt, B.A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol. *189*, 113–130.

Summerer, D., Chen, S., Wu, N., Deiters, A., Chin, J.W., and Schultz, P.G. (2006). A genetically encoded fluorescent amino acid. Proc. Natl. Acad. Sci. U. S. A. *103*, 9785–9789.

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*, 1122–1134.

Tack, D.S., Ellefson, J.W., Thyer, R., Wang, B., Gollihar, J., Forster, M.T., and Ellington, A.D. (2016). Addicting diverse bacteria to a noncanonical amino acid. Nat. Chem. Biol. *12*, 138–140.

Temiakov, D., Patlan, V., Anikin, M., McAllister, W.T., Yokoyama, S., and Vassylyev, D.G. (2004). Structural basis for substrate selection by t7 RNA polymerase. Cell *116*, 381–391.

Temme, K., Hill, R., Segall-Shapiro, T.H., Moser, F., and Voigt, C.A. (2012). Modular control of multiple pathways using engineered orthogonal T7 polymerases. Nucleic Acids Res. *40*, 8773–8781.

Théobald-Dietrich, A., Frugier, M., Giegé, R., and Rudinger-Thirion, J. (2004). Atypical archaeal tRNA pyrrolysine transcript behaves towards EF-Tu as a typical elongator tRNA. Nucleic Acids Res. *32*, 1091–1096.

Thyer, R., Robotham, S.A., Brodbelt, J.S., and Ellington, A.D. (2015). J Am Chem Soc *137*, 46–49.

Tsao, M.-L., Summerer, D., Ryu, Y., and Schultz, P.G. (2006). The Genetic Incorporation of a Distance Probe into Proteins in Escherichia coli. J. Am. Chem. Soc. *128*, 4572–4573.

Tuley, A., Wang, Y.-S., Fang, X., Kurra, Y., Rezenom, Y.H., and Liu, W.R. (2014). The genetic incorporation of thirteen novel non-canonical amino acids. Chem. Commun. *50*, 2673–2675.

Umeda, A., Thibodeaux, G.N., Zhu, J., Lee, Y., and Zhang, Z.J. (2009). Site-specific Protein Cross-Linking with Genetically Incorporated 3,4-Dihydroxy-L-Phenylalanine. ChemBioChem *10*, 1302–1304.

Umeda, A., Thibodeaux, G.N., Moncivais, K., Jiang, F., and Zhang, Z.J. (2010). A versatile approach to transform low-affinity peptides into protein probes with cotranslationally expressed chemical cross-linker. Anal. Biochem. *405*, 82–88.

Varani, G., and McClain, W.H. (2000). The G·U wobble base pair. EMBO Rep. *1*, 18–23.

Wang, L., and Schultz, P.G. (2001). A general approach for the generation of orthogonal tRNAs. Chem. Biol. *8*, 883–890.

Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*, 894–898.

Wang, J., Xie, J., and Schultz, P.G. (2006a). A Genetically Encoded Fluorescent Amino Acid. J. Am. Chem. Soc. *128*, 8738–8739.

Wang, L., Xie, J., and Schultz, P.G. (2006b). Expanding the genetic code. Annu. Rev. Biophys. Biomol. Struct. *35*, 225–249.

Wang, Q., Sun, T., Xu, J., Shen, Z., Briggs, S.P., Zhou, D., and Wang, L. (2014). Response and Adaptation of Escherichia coli to Suppression of the Amber Stop Codon. ChemBioChem *15*, 1744–1749.

Wang, Y.-S., Fang, X., Wallace, A.L., Wu, B., and Liu, W.R. (2012). A Rationally Designed Pyrrolysyl-tRNA Synthetase Mutant with a Broad Substrate Spectrum. J. Am. Chem. Soc. *134*, 2950–2953.

Wang, Y.-S., Fang, X., Chen, H.-Y., Wu, B., Wang, Z.U., Hilty, C., and Liu, W.R. (2013). Genetic Incorporation of Twelve meta-Substituted Phenylalanine Derivatives Using a Single Pyrrolysyl-tRNA Synthetase Mutant. ACS Chem. Biol. *8*, 405–415.

Wilson, C.G., Magliery, T.J., and Regan, L. (2004). Detecting protein-protein interactions with GFP-fragment reassembly. Nat. Methods *1*, 255–262.

Woese, C.R. (1964). UNIVERSALITY IN THE GENETIC CODE. Science *144*, 1030–1031.

Wold, F. (1981). In Vivo Chemical Modification of Proteins (Post-Translational Modification). Annu. Rev. Biochem. *50*, 783–814.

Wong, J.T. (1983). Membership mutation of the genetic code: loss of fitness by tryptophan. Proc. Natl. Acad. Sci. *80*, 6303–6306.

Xie, J., and Schultz, P.G. (2006). A chemical toolkit for proteins — an expanded genetic code. Nat. Rev. Mol. Cell Biol. *7*, 775–782.

Xie, J., Supekova, L., and Schultz, P.G. (2007). A Genetically Encoded Metabolically Stable Analogue of Phosphotyrosine in Escherichia coli. ACS Chem. Biol. *2*, 474–478.

Yamao, F., Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y., and Osawa, S. (1985). UGA is read as tryptophan in Mycoplasma capricolum. Proc. Natl. Acad. Sci. U. S. A. *82*, 2306–2309.

Yambe, H., Kitamura, S., Kamio, M., Yamada, M., Matsunaga, S., Fusetani, N., and Yamazaki, F. (2006). l-Kynurenine, an amino acid identified as a sex pheromone in the urine of ovulated female masu salmon. Proc. Natl. Acad. Sci. *103*, 15370–15374.

Yeung, P.K.K., Wong, F.T.W., and Wong, J.T.Y. (2002). Mimosine, the Allelochemical from the Leguminous Tree Leucaena leucocephala, Selectively Enhances Cell Proliferation in Dinoflagellates. Appl. Environ. Microbiol. *68*, 5160–5163.

Yin, Y.W., and Steitz, T.A. (2004). The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell *116*, 393–404.

Young, J.A.T., and Collier, R.J. (2007). Anthrax Toxin: Receptor Binding, Internalization, Pore Formation, and Translocation. Annu. Rev. Biochem. *76*, 243–265.

Yu, A.C.-S., Yim, A.K.-Y., Mat, W.-K., Tong, A.H.-Y., Lok, S., Xue, H., Tsui, S.K.-W., Wong, J.T.-F., and Chan, T.-F. (2014). Mutations Enabling Displacement of Tryptophan by 4-Fluorotryptophan as a Canonical Amino Acid of the Genetic Code. Genome Biol. Evol. *6*, 629–641.

Yuet, K.P., Doma, M.K., Ngo, J.T., Sweredoski, M.J., Graham, R.L.J., Moradian, A., Hess, S., Schuman, E.M., Sternberg, P.W., and Tirrell, D.A. (2015). Cell-specific proteomic analysis in Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A. *112*, 2705–2710.

Zhang, Y., Baranov, P.V., Atkins, J.F., and Gladyshev, V.N. (2005). Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies. J. Biol. Chem. *280*, 20740–20751.

Zhang, Z., Wang, L., Brock, A., and Schultz, P.G. (2002). The Selective Incorporation of Alkenes into Proteins in Escherichia coli. Angew. Chem. Int. Ed. *41*, 2840–2842.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.