ABSTRACT

Title of Document:	THE DEVELOPMENT AND APPLICATION OF TRANSPOSON MUTAGENESIS AND UNNATURAL AMINO ACID MUTAGENESIS TECHNOLOGIES
	Jia Liu, Doctor of Philosophy, 2012
Directed By:	Associate Professor Douglas Julin Department of Chemistry and Biochemistry University of Maryland, College Park

Recent three decades have seen a significant progress in the protein engineering field. Protein engineering has not only facilitated our understanding of the sequence-structure-function relationship of proteins, but also yielded numerous invaluable new proteins or protein variants for research, medical and industrial applications. The success of protein engineering largely relies on the ability to diversify protein sequence. Protein sequence diversification is usually achieved by directed or random mutagenesis technologies. This dissertation discusses two non-canonical protein sequence diversification techniques—transposon mutagenesis and unnatural amino acid (UAA) mutagenesis.

Transposon mutagenesis is based on an *in vitro* transposition reaction, where a transposon DNA randomly inserts into the gene encoding the protein of interest. Removal of the transposon sequence followed by subsequent manipulation generates desired

mutations. By engineering transposon DNA sequence and subsequent procedures, random "codon" deletions and substitutions can be introduced into random positions of a protein sequence. This dissertation describes how this technology can be applied to determine the minimal domain required for the fluorescence of green fluorescent protein (GFP) and improve the spectral properties of GFP.

In UAA mutagenesis, orthogonal aminoacyl tRNA and aminoacyl tRNA synthetase (aaRS) pairs are employed to reassign an amber stop codon (TAG) with an UAA. This dissertation describes how this technology can facilitate both enzymatic and nonenzymatic approaches to synthesize polyubiquitin chains.

Besides, two ongoing projects are introduced in the Appendices. The first project describes how UAA mutagenesis technology can be potentially used to genetically encode a thiotyrosine in a protein sequence, which can serve as a comparison probe. The second project intends to develop a rapid *in vitro* methodology to expand the functions of aaRS toward new UAAs.

THE DEVELOPMENT AND APPLICATION OF TRANSPOSON
MUTAGENESIS AND UNNATURAL AMINO ACID MUTAGENESIS
TECHNOLOGIES

By

Jia Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Douglas Julin   Associate Professor   Chair
T. Ashton Cropp   Associate Professor   Academic advisor
Dorothy Beckett   Professor
Nicole LaRonde-LeBlanc   Assistant Professor
Philip Bryan   Professor   Dean's representative

Dedication

To my mother Xianjie Sun and my father Xueping Liu,
thank you for your understanding and love.

Acknowledgments

First of all, I would like to thank my academic advisor Dr. Ashton Cropp. You have been not only an exceptional supervisor but also a great friend. You have been so patient to help me with all the problems during my experiments as well as many things beyond the research. It is during the graduate study in your lab that I have learned such a broad range of experimental skills. Now I do realize the importance of having diverse skills for the application of postdoctoral positions. I would always appreciate all the trainings I received in your laboratory. This experience will serve as a solid foundation for my future career. I am and would be always grateful to you for all your help and kindness. Although it was not easy to get through all these years, I had fun working with you, really. As I am moving to a new position, I hereby wish you the best for your family and research.

I am grateful to all my committee advisors. I really appreciate all the support and advice from you, as well as the knowledge I have learned from your classes. Thank you, Dr. Beckett, for helping me with my independent proposal. Thank you, Dr. Julin, for offering being my co-advisor. Thank you, Dr. LaRonde-LeBlanc, for all the help during your class, the rotation, and committee meetings. Thank you, Dr. Bryan, for offering being the Dean's Representative on my committee.

I also want to thank the entire Cropp lab, both past and present members. Special thanks to Kelly Daggett, Bryan Wilkins and Jianting Zheng, from whom I have learned so many experimental skills. Thanks to Xu Yang, Zijun Chen and Karina Herrara for making my time in the laboratory very enjoyable. Thank you, William Knight and Megan Reed for all the fun you have brought to me. I wish you all well in your endeavors.

Thanks to all my fellow graduate students. I would like to thank Katherine Connor, Urszula Nowicka, Poorni Adikaram and Joe Cannon for making my life in Maryland full of fun. I wish you all good luck for your future. I am also grateful to Sanghee Nah, Ying Lu and Yiling Luo, for all the good times with you. Special thanks to Yu Gu and Xiaochen Wang, with who I have been roommate for three and a half years. I have learned so much from you, received so much help from you, and more importantly had so much fun with you. Thank you, my dear friends. I am also grateful to all my Chinese friends, without who I could not have been able to get throught all these years easily. I wish you all the best for your following career and future lives.

Finally, I would like to thank my parents. Thank you, Mom and Dad, for all your support, understanding and love throughout these years.

Table of Contents

List of Tables

List of Figures

List of Schemes

Abbreviations

RNA	ribonucleic acid
tRNA	transfer RNA
mRNA	messenger RNA
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
RNase	ribonuclease
NTP	ribonucleotide triphosphate
dNTP	deoxyribonucleotide triphosphate
ATP	adenosine 5'-triphosphate
GTP	guanosine 5'-triphosphate
CTP	cytidine 5'-triphosphate
UTP	uridine 5'-triphosphate
cAMP	3'-5'-cyclic adenosine monophosphate
bp	base pairs
EDTA	ethylenediaminetetraacetic acid
PCR	polymerase chain reaction
kDa	kilodaltons
LB	Luria-Bertani broth
2×YT	2× Yeast extract and Trytone rich medium
SOC	super optimal catobilite-repression broth
SAM	*S*-adenosylmethionine
PMSF	phenylmethylsulfonyl fluoride

IPTG	isopropyl-β-D-1-galactopyronaside
OD	optical density
SDS	sodium dodecyl sufate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
UV	ultra violet
DTT	dithiothreitol
BSA	bovine serum albumin
ORF	open reading frame
MS	mass spectrometry
NMR	nuclear magnetic resonance
MW	molecular weight
UAAs	unnatural amino acids
aaRS	aminoacyl tRNA synthetase
FP	fluorescent protein
GFP	green fluorescent protein
YFP	yellow fluorescent protein
RFP	red fluorescent protein
EP-PCR	error-prone PCR
R-site	transposase recognition site
CDM	codon deletion mutagenesis
CSM	codon scanning mutagenesis
PTMs	post-translational modifications
SPPS	solid phase peptide synthesis

EPL	expressed protein ligation
Ub	ubiquitin
polyUb	polyubiquitin
PylT	pyrrolysyl tRNA
PylRS	pyrrolysyl tRNA synthetase
Boc	$^{\varepsilon}N$-*tert*-butyloxycarbonyl
EI-PCR	enzymatic inverse PCR
pTT	transposon-targeting plasmid
VMA	vacuolar membrane ATPase
VMAI-N	VMA intein N-terminal domain
VMAI-C	VMA intein C-terminal domain
BLA	β–lactamase
ORF	open reading frame
EGFP	enhanced GFP
GFP_{UV}	UV-optimized GFP
sgGFP	SuperGlo GFP
sfGFP	superfolder GFP
RID	random insertion and deletion
TriNEx	triplet nucleotide exchange
Ex	excitation
Em	emission
CDRs	complementarity determining regions
NCL	native chemical ligation

GOPAL	genetically encoded orthogonal protection and activated ligation
Cbz	carboxybenzyl
TFA	trifluoroacetic acid
Alloc	allyloxycarbonyl
MESNA	sodium 2-mercaptoethane sulfonate
Alloc-OSu	allyloxycarbonyl *N*-oxysuccinimide
DIEA	*N,N*-diisopropylethylamine
pBpa	*p*-benzoylphenylalanine
Onb	*o*-nitrobenzyl-oxycarbonyl
ESI-MS	electrospray ionization mass spectrometry
CSP	chemical shift perturbation
TyrT	tyrosyl tRNA
TyrRS	tyrosyl tRNA synthetase
CAT	chloramphenicol acetyl transferase
STA	streptavidin
IVC	*in vitro* compartmentalization
RBS	ribosome binding site
SerU	serine suppressor tRNA
PBS	phosphate buffer saline
TBS	Tris buffer saline

Chapter 1: Introduction

1.1 Directed Evolution: Evolving Biomolecules by Mimicking Natural Evolution

Staunch Darwinists attribute the complexity of living creatures to the natural algorithm of evolution—mutation and selection. Central to natural evolution is the concept of "fitness"—a species' ability to survive selection and reproduce itself in a given environment. Fitter species have larger chance to survive and propagate their "biological traits" to the whole population. Billions of years of natural evolution has eventually created a world of "the fitter ones" at all scale, from ecosystems down to single biomolecules. However, the journey of evolution has not yet ended. In a simple case of a protein molecule consisting of 100 amino acids, there are 20^{100} ($\approx 10^{130}$) possible variants—more than all the atoms in the universe. Even with billions of years' searching for the "fitter species", nature has explored only an infinitesimal fraction of all the possible solutions. There is no doubt that in this enormous unexplored space many biomolecules with dreaming functions are just waiting to be discovered.

Researchers are now applying the same concept—mutation and selection— to evolve biomolecules (RNA, DNA and proteins) under laboratory conditions. The artificial evolution of biomolecules is generally achieved by creating a library of variants and then selecting for improved variants. The variants with desired functional improvements can be subjected to next round of evolution. Therefore, the evolution pathway of the biomolecules is directed by artificially defined selection pressure. This experimental approach is given the name of "directed evolution".

Different from natural evolution, directed evolution experiments are no longer restricted to the improvement of biologically relevant functions. Instead, any function of

interest, as long as physically and chemically possible, can be pursued. Perhaps one of the best examples of artificially evolved biomolecules is green fluorescent protein (GFP). Roger Tsien won 2008 Nobel Prize in Chemistry for his work on engineering fluorescent proteins (FPs) for biological imaging (*1*). These FPs have been extensively engineered for their excitation/emission spectra, oligomerization state, solubility, pH stability, photostability and so on. These artificially evolved FPs may have little biological benefits in their native contexts but are serving as perfect biological imaging tools for research or medical purposes. Furthermore, laboratory evolution has dramatically shortened the time of molecular evolution from millions of years to decades or even years. Taking the FPs for example again, it took only two decades from its discovery to common laboratory use, and many of these FPs contain enhanced functions that nature failed to address during millions of years of evolution.
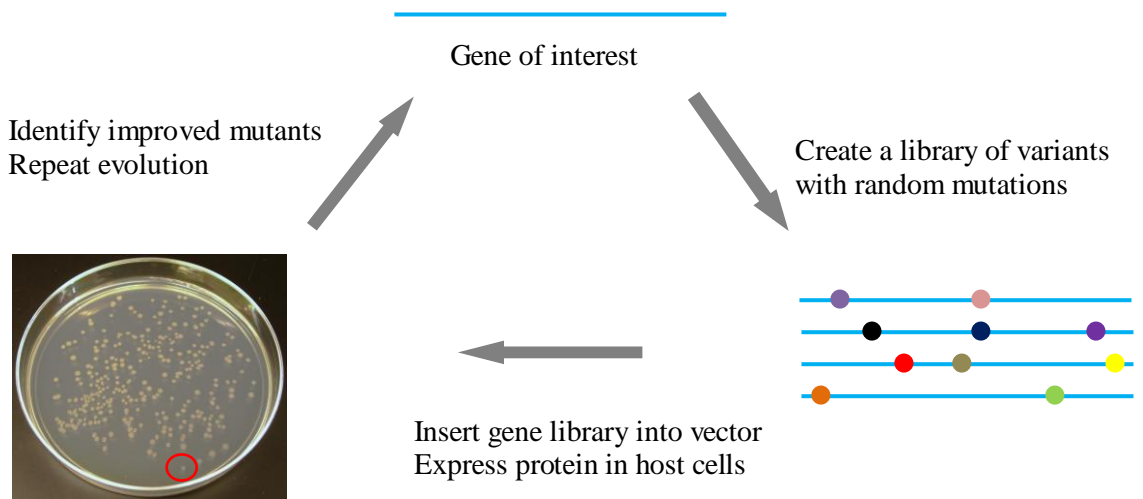
Figure 1.1. General procedure for laboratory evolution of biomolecules. The balls on the gene of interest represent introduced mutations

Laboratory evolution of biomolecules usually contains two major processes: i) creation of a mutant library ii) screening or selection of improved mutants (**Figure 1.1**). Numerous mutagenesis methods and screening/selection methods have been developed (*2, 3*). With an appropriate screening/selection method, improved mutants can be rapidly and accurately detected. However, it is the employed mutagenesis methods that decide if a beneficial mutation can even show up or not in the library of variants. The quality of a mutant library is critical to the success of an evolution experiment.

Unfortunately, there is no "single best" method to create mutant libraries for all purposes. Nearly all the mutagenesis methods have their own advantages and disadvantages. In fact, the advantage of a method under some circumstance may be a drawback when a different task is to be achieved. That is why a complicated evolution task is usually achieved by the combinatorial use of several different mutagenesis methods. Based on the information input, the methods of library creation can be divided into three types: i) rational design; ii) irrational design; iii) semi-rational design (**Figure 1.2**). The development of FP variants gives good examples how these three different approaches can be used to accomplish different goals. Below are introduced these three protein design methods following the stories of FPs.

Rational design requires considerable amount of information input (chemical, structural, and sometimes computational) as both the mutation positions and types need to be predicted. Once designed, these site-mutations can be introduced into proteins by site-directed mutagenesis (*4*). Although variations exist in different site-directed mutagenesis methods, most of them rely on the PCR amplification of the target gene using mutagenic primers carrying the desired nucleotides at the mutation sites. The

template DNAs (usually plasmids isolated from host cells) can be removed by digestion of endonuclease *Dpn* I, which cleaves only the plasmid DNAs with methylated adenosine but not newly amplified PCR product. For example, Tsien and co-workers solved the crystal structure of GFP in 1996, based on which they designed a T203Y mutation to tune the spectra of GFP (*5*). The introduced Y203 is in close proximity to the chromophore and stacks with its phenol ring, therefore stabilizing the π electons in the chromophore. The reduced energy gap between excited state and ground state of π electons result in the shift of emission to the red side. This T203Y variant is the first reported yellow FP (YFP).

Rational design

Site-directed mutagenesis

Irrational design

EP-PCR or DNA shuffling

Semi-rational design

Saturation mutagenesis

Figure 1.2. Rational, irrational and semi-rational approaches to create a library of variants. In rational design, both mutation positions and types are "fixed" or selected. Irrational design allows random types of mutations to occur at random positions. Semi-rational design generates random or selected mutations at fixed positions.

Irrational design requires least input of information and is particularly useful when the information (such as the structure of a protein of interest) is not easy to obtain, or the evolution task is too complicated to achieve using current knowledge. The two most widely used random mutagenesis methods are error-prone PCR (EP-PCR) (*6*) and DNA shuffling (*7*). The first technique, EP-PCR, is very similar to routine PCR process, but the conditions are modified such that DNA polymerases tend to introduce mis-matched base

pairs. Although the conditions of EP-PCR may have a lot of variations, such as the ratio of dNTPs (*8*), concentration of Mn^{2+} (*9*) or even different types of DNA polymerases (*10*), they all rely on the natural error rate of DNA polymerases. Therefore, the constructed libraries are inevitably marked by the mutation bias of the polymerases. Nevertheless, EP-PCR is perhaps the simplest random mutagenesis method one can imagine and is relatively time and cost-efficient as compared with other methods. One successful application of EP-PCR is the engineering of chromophore maturation of YFP (*11*). Soon after its discovery, YFP was found to mature very slowly *in vivo*, limiting its application as a biological probe. The chromophore maturation of FPs is a complicated process consisting of cyclization, dehydration and oxidation (*12*). Improving the maturation rate of FPs may not be easily achieved using rational protein design. Miyawaki and co-worker successfully enhanced the chromophore maturation rate of YFP by more than 10 fold using EP-PCR followed by fluorescence screening. The second technique, DNA shuffling, is a homologous recombination-based mutagenesis method. In a DNA shuffling experiment, a template DNA is first randomly fragmented by DNase I into 50~100 bp DNAs and these small fragments are then assembled in PCR based on their homologous region (*7*). The major advantage of DNA shuffling is that different members of proteins in the same family can be used as the templates together and homologous recombination during DNA shuffling will generate chimeric DNA sequences with unique properties. The widely used UV-enhanced GFP variant, GPF_{UV}, was the result of a successful DNA shuffling experiment (*13*). As this dissertation mainly focuses on mutagenesis methods generating "point-mutations", these recombination-based techniques will no longer be discussed below.

Semi-rational design can be seen as an "intermediate" between rational design and irrational design. Most semi-rational protein design experiments create mutant libraries using (multi-site) saturation mutagenesis (*14*), a derivative method of site-directed mutagenesis. In saturation mutagenesis, one or several mutation sites are chosen and defined or a mix of amino acid mutations are introduced into these positions, usually by PCR reactions with degenerate primers. Semi-rational design largely reduces the difficulty of mutation design as it no longer requires perfect prediction of mutation sites and types. Moreover, when these chosen mutation sites are structurally close (such as in the active site of an enzyme), the function of a protein can be dramatically improved. By using multi-site saturation mutagenesis in combination with EP-PCR, Tsien and co-worker have successfully developed monomeric red FPs (RFPs) with a wide range of spectra (*15*). Most of the currently used RFPs are directly from or derived from this work.

This dissertation will introduce two unique protein sequence diversification technologies that are very different from routine methods. The first one is transposon mutagenesis which relies on the random insertions of transposon DNA into a gene of interest in an *in vitr*o reaction. This method can be considered as an irrational approach to create mutant libraries. However, it does not rely on the natural error rate of DNA polymerase and thus, overcomes many limitations that are usually found in a PCR-based random mutagenesis method. The second technology in this disseration is unnatural amino acid (UAA) mutagenesis which can site-specifically incorporate a UAA into a protein sequence. This method can be considered as a rational approach to evolve protein functions. However, it is dramatically different from traditional rational design methods because it uses UAAs as new building blocks to evolve proteins to carry out chemistries

6

beyond those contained in the 20 natural amino acids. The basis of this method is the application of orthogonal aminoacyl tRNA/tRNA synthetase pairs that can assign unnatural amino acids into the amber stop codon (TAG) in a protein sequence.

1.2 Transposon Mutagenesis

A transposon is a DNA element that can be transported from one location in a genome to another. It is a ubiquitous genetic element that exists in bacteriophage, bacteria, plant, animals and human. Transposon has become an indispensable genetic tool in recent years. Classical *in vivo* transposition reaction has been applied as gene tagging and mapping techniques (*16, 17*) as well as DNA sequencing tool (*18*). As early as in 1980s, *in vitro* transposition reaction using Mu transposon DNA was developed (*19*). However, early *in vitro* reactions are as complex as the *in vivo* reactions as they require a number of accessory proteins and DNA cofactors, such as host-encoded DNA bending protein, phage-encoded cofactor protein that affects target specificity and transpositional enhancer DNA that is located in the phage genome (*20*).

With modified Mu transposon DNA and appropriate reaction conditions, a minimal reaction set-up requires only MuA transposase, donor DNA (transposon) and acceptor DNA (target plasmid) as the macromolecular components (*21-23*). In such mini-transposition reaction, MuA transposase first recognizes a specific sequence at each transposon end (transposase recognition site, R-site). This transposase-transposon complex, termed as transpososome, binds to the acceptor DNA (usually a plasmid), cleaves the double-stranded acceptor DNA and then catalyzes the joining reaction between donor DNA (transposon) and acceptor DNA (*21, 22*) (**Figure 1.3A**). Notably, the transposon DNA has been engineered such that it only needs one R-site at each end

7

A.

Mu transposon   MuA transposase   Transpososome

Target plasmid

Target plasmid with inserted transposon

B.

```
5′gatcTGAAG CGGCGC...GAAAA                   TTTTC...GCGCCG CTTCA     3′
3′     ACTTC GCCGCG...CTTTT  Selection marker AAAAG...CGCGGC GAAGTctag 5′
```

C.

```
...N₁N₂N₃N₄N₅...                    5′gatcTGA...TCAN₁N₂N₃N₄N₅...        ...N₁N₂N₃N₄N₅TGA...TCAN₁N₂N₃N₄N₅...
...N₁N₂N₃N₄N₅...      ...N₁N₂N₃N₄N₅ACT...AGTctag5′                       ...N₁N₂N₃N₄N₅ACT...AGTN₁N₂N₃N₄N₅...
```

Target DNA

Transpososome

Target DNA with transposon insertion

DNA repair in host cell generates 5-bp duplication

Figure 1.3. *In vitro* mini-transposition reaction. A) General procedure. B) Construct of Mu transposon. R-site sequences are shaded. C) Origin of the five-basepair duplication. Cyan color indicates the 5 bp sequence recognized by transposase.

for the reaction (**Figure 1.3B**). These two R-sites are inverted repeats or so called symmetric sequences. Flanking the R-sites are five nucleotides that can be modified to contain restriction sites without eliminating the ability of integration of Mu transposon (*24, 25*). These transposon DNAs are usually pre-cut by endonuclease *Bgl* II or *BamH* I to generate a 5' GATC overhang, which largely accelerates transposition (though not essential) by bypassing the otherwise necessary cleavage of transposon by transposase (*23*). The middle segment of the transposon is not involved in the transposition reaction and is usually a selection marker for selecting the target plasmids with transposon insertions (*23, 24*). Another feature of Mu transposition is that the acceptor DNA will have a five-basepair duplication along the transposon insertion site (*26, 27*). This is because Mu transpososome cleaves the acceptor DNA in a five-bp staggered manner and

8

then inserts the transposon sequence in between of the cleaved sequence (**Figure 1.3C**). After insertion, the transposon is flanked by a five nucleotide single-stranded sequence at each end. When transformed into *E. coli*, the single-stranded gaps will be repaired by *E. coli* DNA repair system, resulting in five nucleotide duplication. The transposition reaction will generate a library of plasmid each containing a random transposon insertion.

Chapter 2 describes three applications of the Mu transposition reactions. In the first experiment, a previously reported Mu transposon variant (MuDel transposon) (*25*) was applied to GFP to identify amino acid deletions that can retain fluorescence ("neutral" deletions). MuDel transposon-based deletion mutagenesis can generate triplet nucleotide deletion at a random position of a DNA sequence. This experiment aimed at generating a size-optimized GFP that is less likely to interfere with the native function of a fused protein. As the MuDel transposon does not have any control of reading frame, the random triplet nucleotide deletions sometimes affected two neighboring residues resulting in unexpected mutations in addition to the desired deletions. The second experiment aimed to solve this problem by engineering the Mu transposon to contain a DNA segment for reading frame selection. The outmost five nucleotides of this new transposon are asymmetric to facilitate downstream manipulation. Using this transposon, a new deletion mutagenesis method has been developed, named as "codon deletion mutagenesis" (CDM). This method allows for generation of up to five consecutive, in-frame codon deletions at a random position of a protein. CDM was performed on GFP and several **in-frame** "neutral" deletions were successfully identified. These two experiments together identified more than 40 residues (out of the total 238 amino acids) in GFP that are not essential for the fluorescence. In the third experiment, CDM approach

was adapted to develop a codon scanning mutagenesis (CSM) method that can generate up to three consecutive, in-frame amino acid substitutions at a random position of a protein. This asymmetric transposon-based CSM was performed on GFP. A series of spectra-shifted mutants were identified by screening just a small fraction of the mutant libraries.

1.3 Unnatural Amino Acid Mutagenesis

Proteins are the chief actors in living creatures, carrying out the majority of the biochemical reactions within cells. Many of the biochemical functions of proteins are achieved with only 20 amino acids, though in some rare cases noncanonical amino acids selenocysteine (*28*) or pyrrolysine (*29*) are also encoded for added functions of proteins. When it comes to chemistries beyond those contained in the 20 amino acids, proteins often require cofactors (such as metal ions, flavins and thiamine) or post-translational modifications (PTMs) (such as methylation, phosphorylation and glycosylation). These cofactor or PTMs confer on proteins new biophysical or biochemical properties to accomplish a broader range of biochemical reactions.

Nevertheless, none of the naturally occurring amino acids contain side chains that are highly reactive under mild or neutral conditions. For one reason, these amino acids are sufficient to support the lives of cells as proteins usually carry out their functions under controlled, specified micro-environment. Another reason is that highly reactive groups such as ketones, aldehydes and thioesters, if frequently present in the side chains of proteins, will possibly "over-heat" cellular machinery.

Although nature evolves proteins to achieve their functions with just 20 amino acids, researchers are not satisfied with the chemistries contained in this limited number of amino acids. Given the importance of proteins in chemistry, biology, medicine and pharmacy fields, it is desirable to be able to manipulate proteins with new building blocks for better understanding and application of proteins *in vivo* and *in vitro*. Unlike the first protein sequence diversification technology introduced above, the second technology, unnatural amino acid (UAA) mutagenesis diversifies protein sequence by adding new, artificial building blocks rather than randomizing protein sequence with the canonical 20 amino acids. In fact, a number of approaches are being pursued to incorporate unnatural amino acids (UAAs) into proteins. Below are introduced several different techniques for UAA incorporation.

One of the early methodologies is solid-phase peptide synthesis (SPPS) (*30*) which allows automated polypeptide synthesis by coupling the C-terminal carboxyl group of one amino acid with the N-terminal amine group of another. Using SPPS, UAAs can be incorporated into any position of a protein sequence just like natural amino acids. This method, however, is subject to size limitation. Peptides of 50~100 amino acids are difficult to synthesize in large quantities. This limitation can be partially alleviated by using native chemical ligation, where a small peptide containing C-terminal thioester can react with another peptide with N-terminal cysteine to assemble a larger peptide (*31*). Alternatively, the peptide containing C-terminal thioester can be produced by intein-mediated cleavage of recombinant proteins expressed from bacteria cells. The recombinant proteins containing C-terminal thioester can then be ligated with a synthetic peptide containing UAA at the desired position. This intein-based method, referred to as

expressed protein ligation (EPL) (*32*), largely facilitates the synthesis of long peptide chains using SPPS. Concerns about the incorporation of UAAs using SPPS and EPL includes the repeated protection and deprotection process and the non-native (*in vitro*) conditions of peptide synthesis, which may disrupt the native structures or functions of proteins (some proteins, for example, require chaperones to fold properly).

Another powerful yet technically challenging technique is based on tRNAs that are "pre-charged" with desired amino acids. In this method, truncated tRNAs are enzymatically ligated with chemically aminoacylated nucleotides (*33-35*). The "charged" tRNA is then used in cell-free translation system to incorporate UAAs into a specific position in response to a nonsense or frameshift codon. This method has also been adapted to inject or transfect the "charged" tRNA directly into living cells (*36, 37*). Although methods are potentially powerful, they are limited by the large-scale synthesis, instability and discontinuous supply of the aminoacylated tRNAs.

A third approach relies on natural or engineered aminoacyl tRNA synthetases to globally replace one amino acid in proteins with a close structural analog (*38-41*). In this method, a host strain auxotrophic for the amino acid to be replaced is first grown to mid-log phase as usual and then transferred to a minimal medium lacking the amino acid to be replaced but containing the analog amino acid of interest. Newly expressed proteins from the existing cells will have global incorporation of the desired amino acids. The drawbacks of this method are that the incorporated UAAs are limited to close structural analogs and that the UAA incorporation is not site-specific.

The UAA mutagenesis described in this dissertation is based on genetically encoded UAAs during translation process. In this method, an orthogonal tRNA and aminoacyl

tRNA synthetase (aaRS) pair is expressed from a plasmid in a host organism and then assign a UAA in response to a nonsense or frameshift codon in the gene of interest (**Figure 1.4**). Importantly, the orthogonal tRNA/aaRS pair can incorporate UAAs independent of endogenous tRNAs and aaRSs. That said, host tRNA is not aminoacylated with UAAs by orthogonal aaRS nor orthogonal tRNA is aminoacylated with any of the 20 canonical amino acids by endogenous aaRSs. The orthogonal tRNA/aaRS pairs are usually taken from another species, for example, *Escherichia coli* tyrosyl tRNA and aaRS pair is orthogonal in *Saccharomyces cerevisiae* (*42*) and *Methanococcus jannaschii* tyrosyl tRNA and aaRS pair is orthogonal to *E. coli (43)*. This strategy allows UAAs with various properties to be site-specifically incorporated into proteins in different organisms. So far, more than 70 UAAs have been successfully incorporated into *E. coli*, yeast and mammalian cells (*44*) and researchers have even gone to encode UAAs in living animals (*45*).
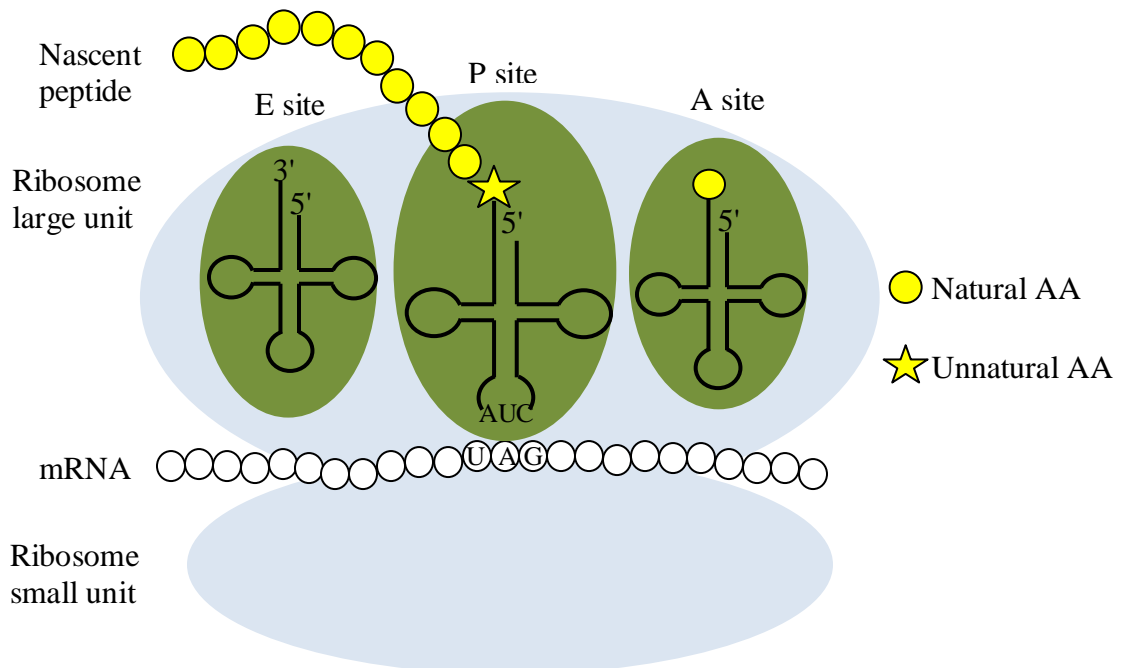
Figure 1.4. Site-specific incorporation of UAAs into proteins in response to amber stop codon (UAG).

Chapter 3 shows how UAA mutagenesis technologies can facilitate the enzymatic and nonenzymatic synthesis of polyubiquitin (polyUb) chains. In collaboration with Dr. Fushman's group, we have developed two approaches for synthesis of polyUb chains. Both of these two approaches rely on an orthogonal pyrrolysyl tRNA (PylT) and tRNA synthetase (PylRS) pair to incorporate $^{\varepsilon}N$-*tert*-butyloxycarbonyl-L-lysine (BocLys) into a targeted lysine position in Ub. The first method focuses on synthesis of homogeneous polyUb chains that require a linkage-specific Ub conjugating enzyme E2 (enzymatic approach). The second method does not involve E2 enzymes (non-enzymatic approach) and can generate polyUb chains of any linkage composition. Also included in this chapter is a simple yet highly efficient technique to improve protein production yield per unit UAAs used. This technique is particularly useful for large-scale production of Ub or other proteins containing UAAs.

Two ongoing projects are also included in this dissertation as appendices. Experiments in **Appendix I** aim to develop a general method for site-specific incorporation of thiotyrosine into proteins in *E. coli*. Tyrosine residues play important roles in the functions of many proteins. Besides serving as catalytic center, they can also act as regulatory signals. The ability to incorporate thiotyrosine will provide a convenient means to install a comparison probe at a desired tyrosine position in a protein.

Experiments in **Appendix II** aim at developing an *in vitro* system for the evolution of tRNA synthetases. *In vivo* evolution of tRNA synthetase is usually labor-intensive and requires considerable amount of UAAs. Moreover, the *in vivo* evolution does not allow the development of aaRSs for incorporating cytotoxic or impermeable UAAs whereas the *in vitro* approach is cell-free and thus, completely bypasses this limitation.

Chapter 2: Transposon Mutagenesis for Diversifying Protein Sequences

2.1 Introduction

This chapter shows how transposon mutagenesis technology can be used to diversify protein sequences and facilitate our understanding of proteins. The first two methods, MuDel transposon mutagenesis and codon deletion mutagenesis (CDM), share the same concept that removal of the random transposon insertions simultaneously deletes extra nucleotides from the target gene. The difference lies in that MuDel mutagenesis utilizes restriction digestion to remove the transposon (**Figure 2.1A**) whereas CDM relies on PCR process followed by digestion (**Figure 2.1B**). Another major difference is that CDM transposon (MuCDM) bears a reading-frame selection marker rather than simply a traditional selection marker as in the MuDel transposon. The third method, codon scanning mutagenesis (CSM), can be seen as advanced CDM where the process not only deletes codons from the target gene but also delivers new codons to fill in the codon "scar" (**Figure 2.1C**). Sections *2.1.1~2.1.3* will introduce the rationales of each mutagenesis method.

2.1.1 Rationale of MuDel deletion mutagenesis

MuDel deletion mutagenesis requires an engineered Mu transposon (MuDel) that contains an *Mly* I recognition site at each end. *Mly* I is a type IIS endonuclease and its cleavage site is four nucleotide away from its recognition site (**Figure 2.2A**). Other type IIS endonucleases used in the following experiments included *Bsg* I and *Bpm* I (**Figure 2.2A**). Type IIS restriction enzymes differ from the orthodox type II restriction enzymes, which are classified as type IIP enzymes (*46*), in that their recognition sites are
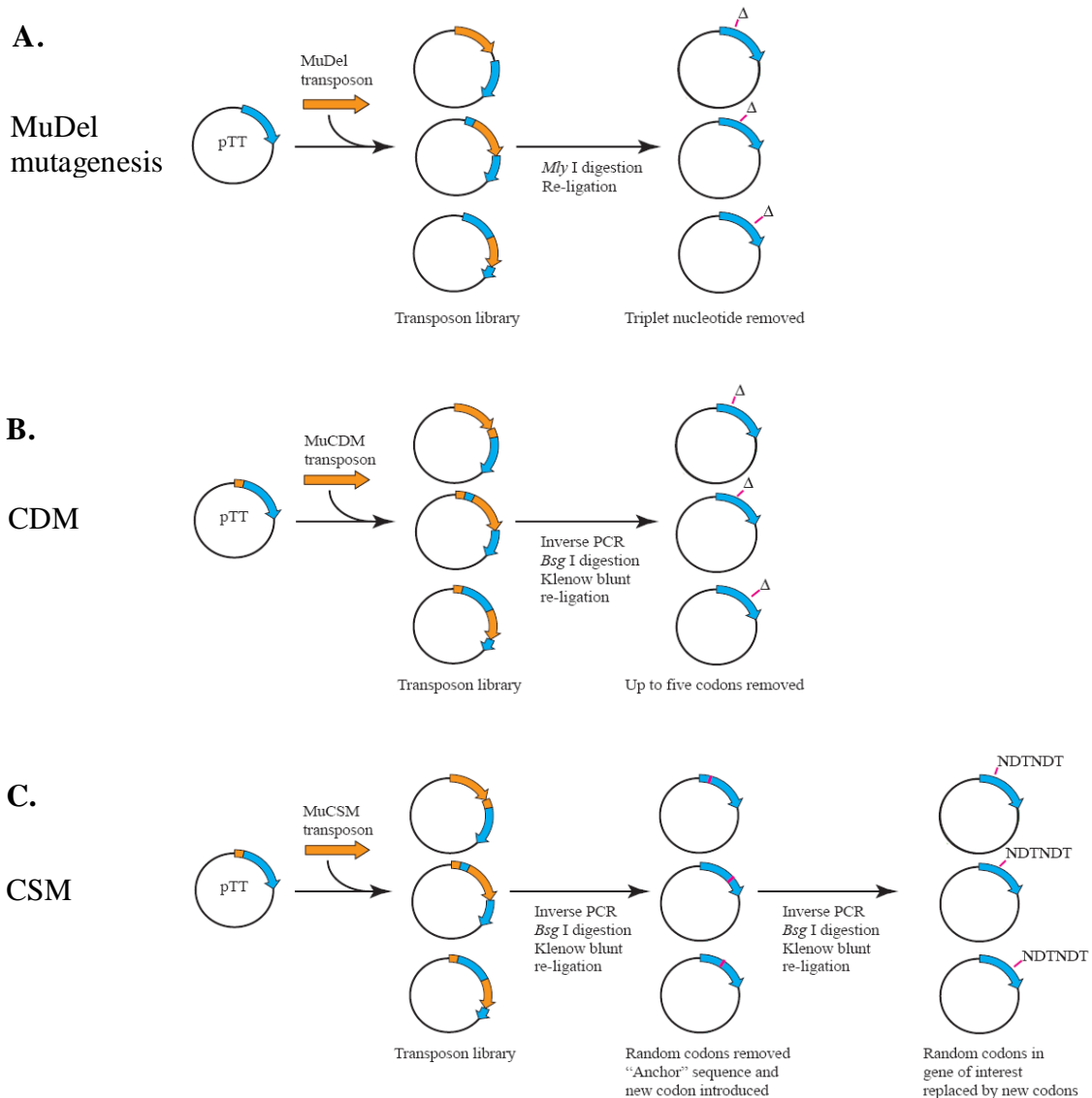
Figure 2.1. General procedures of transposon mutagenesis. **A**) MuDel transposon-based random deletion mutagenesis. **B**) CDM and **C**) CSM.

asymmetric rather than palindromic (see examples of type IIP enzymes in **Figure 2.2B**) and that their cleavage sites are located outside of their recognition sites. As mentioned in **Chapter I**, the five nucleotides flanking the R-sites of transposon can be modified to contain any restriction sites. Jones (*25*) has engineered these five nucleotides to contain an *Mly* I recognition site (**Figure 2.3A**). *Mly* I digestion removes the inserted transposons from the target gene, along with four additional nucleotides from each end. This results in

a net deletion of three nucleotides (5 bp duplication minus 4 bp deletion from each end).

The digestion product is a library of linearized plasmids each containing a random triplet

nucleotide "scar". Subsequent re-ligation gives back a library of circular plasmids with

random deletions. (**Figure 2.3B**).

A.

Mly I
```
5'...GAGTC(N)₅...3'
3'...CTCAG(N)₅...5'
```

Bsg I
```
5'...GTGCAG(N)₁₆...3'
3'...CACGTC(N)₁₄...5'
```

Bpm I
```
5'...CTGGAG(N)₁₆...3'
3'...GACCTC(N)₁₄...5'
```

B.

EcoR I
(5' overhang)
```
5'...GAATTC...3'
3'...CTTAAG...5'
```

EcoR V
(blunt end)
```
5'...GATATC...3'
3'...CTATAG...5'
```

Nsi I
(3' overhang)
```
5'...ATGCAT...3'
3'...TACGTA...5'
```

Figure 2.2. Recognition and cleavage patterns of restriction enzymes. **A)** type IIS. **B)** type IIP.

A.

B.
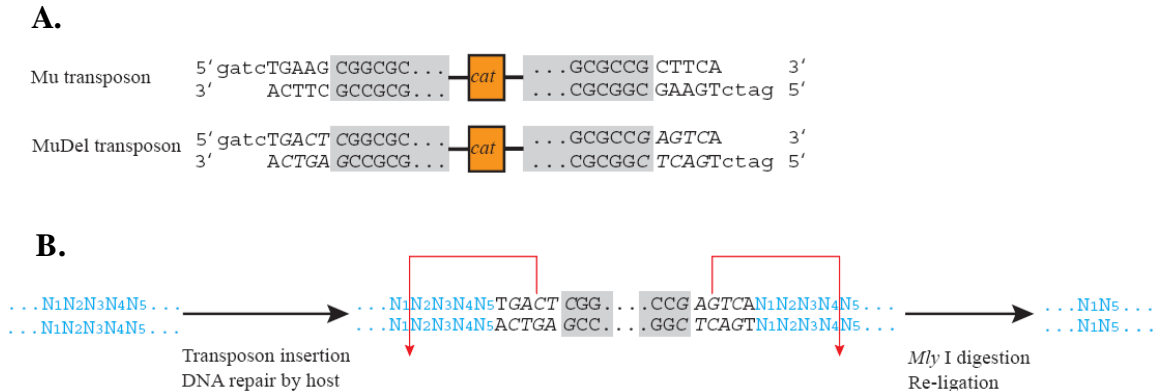
Figure 2.3. MuDel transposon-based deletion mutagenesis. **A)** Mu transposon and MuDel transposon with *Mly* I sites (italic letters) at the two ends. The selection markers for both transposons are chloramphenicol acetyl transferase (*cat*) gene. **B)** General procedure.

2.1.2 Rationale of codon deletion mutagenesis (CDM)

During the MuDel deletion mutagenesis, the importance of the control of reading frame was realized. To avoid unexpected mutations, we sought to develop a random mutagenesis method for generation of codon deletions. This method, named as codon deletion mutagenesis (CDM), allows for generation of up to five consecutive, in-frame codon deletions at a random position (**Figure 2.1B**). As compared with the MuDel mutagenesis, the major advantages of CDM are that 1) the deletions occur in frame and do not introduce unexpected mutations and that 2) up to five consecutive codons can be deleted from a random position. Key to this method is the development of an "all-in-one" Mu transposon (MuCDM) that contains asymmetric end sequences and a reading-frame selection element in the middle segment. The asymmetric end sequences serve as unique primer pairing sites for the downstream enzymatic inverse PCR (EI-PCR) (*47*) (see below for details). The reading-frame selection segment in transposon contains an intein-based dual reporter system (*48, 49*) which ensures that only in-frame transposon insertions can result in colonies on ampicillin plate (see below for details).

In CDM experiment, a gene of interest is first cloned into a transposon-targeting plasmid (pTT). A Tat signal peptide (which directs fused peptide into periplasm) and a *Saccharomyces cerevisiae* vacuolar membrane ATPase intein N-terminal domain (VMAI-N) are fused upstream to the gene of interest (**Figure 2.4A**). The C-terminal domain of VMA intein (VMAI-C) is carried by the MuCDM transposon, upstream fused to TEM-1 β–lactamase gene (*bla*, ampicillin resistance gene) (**Figure 2.4B**). VMAI-C and *bla* fusion genes do not bear a start codon such that their expression stringently depends on the insertion of transposon into an open reading frame (ORF). Only when the

18

MuCDM transposon is inserted in the same reading frame with the gene of interest, will VMAI-N and VMAI-C be present in the same reading frame and self-splice to assemble Tat signal peptide and TEM-1 β–lactamase, resulting in ampicillin resistance (**Figure 2.4C**). Therefore, the transformed cells containing in-frame transposon insertions can survive on plates supplemented with ampicillin. All the out-of-frame transposon insertions are purged from the library. The advantage of intein-based "head-and-tail" dual reporter system lies in that the function of β-lactamase will not be interrupted by N-terminal fusion peptide (*48*) (**Figure 2.4C**). Importantly, the ability to control the reading frame of transposon insertions guarantees that primers annealing to the transposon ends can install type IIS restriction sites in the correct reading frame for codon deletions.
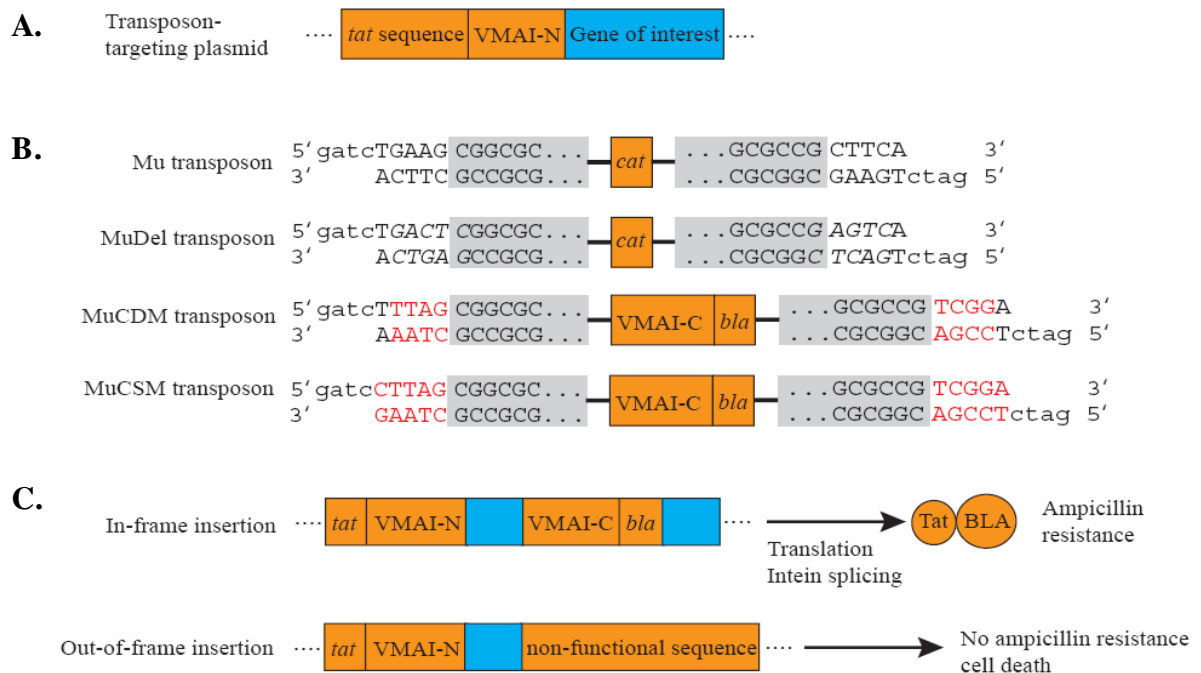
Figure 2.4. Construction of transposon-targeting plasmid (pTT) and asymmetric transposons MuCDM and MuCSM. **A**) pTT. **B**) Asymmetric transposons. **C**) Mechanism of intein-based "dual reporter" reading-frame selection system.

...$N_1N_2N_3N_4N_5$...
...$N_1N_2N_3N_4N_5$...

Transposon insertions
DNA repair by host cells

...$N_1N_2N_3N_4N_5$TTTAG CGGCGC...GCGCCG TCGGA$N_1N_2N_3N_4N_5$...
...$N_1N_2N_3N_4N_5$AAATC GCCGCG...CGCGGC AGCCT$N_1N_2N_3N_4N_5$...

End #1   End #2

EI-PCR:
double stranded DNA denatured
primers annealing to transposon ends

...$N_1N_2N_3N_4N_5$TTTAGCGGCGCACGAAAAACGCGAAAGC...
3'-AAATCGCCGCGTGCTT*GACGTG*CTTTCG-5'

End #1

5'-TTTCGCGTTTTT*GTGCAG*CCGTCGGA-3'
...AAAGCGCAAAAAGCACGCGGCAGCCT$N_1N_2N_3N_4N_5$...

End #2

PCR product: linear plasmid DNA

Bsg I

...$N_1N_2N_3N_4N_5$ TT TAGCGGCGCACGAA*CTGCAC*GAAAGC-3'
...$N_1N_2N_3N_4N_5$ AA ATCGCCGCGTGCTT*GACGTG*CTTTCG-5'

5'-TTTCGCGTTTTTC*GTGCAG*CCGTCGGA$N_1N_2N_3N_4N_5N_6$ N_7N_8 N_9...
3'-AAAGCGCAAAAAG*CACGTC*GGCAGCCT$N_1N_2N_3N_4N_5N_6$ N_7N_8 N_9...

Bsg I

Bsg I digestion
Klenow treatment to blunt 3' overhang
re-ligation to give back circular plasmid

...$N_1N_2N_3N_4N_5N_9N_{10}$...
...$N_1N_2N_3N_4N_5N_9N_{10}$...

Figure 2.5. Detailed procedure of CDM. The asymmetric end sequences are highlighted in red. *Bsg* I sites are in italic letters. Two codon deletion is shown as a demonstration.

The library of pTT plasmids with random, in-frame transposon insertions will serve as the template in the subsequent EI-PCR, where two primers anneal to the asymmetric ends of transposon in a back-to-back manner and amplify the whole plasmid except the transposon sequence (**Figure 2.5**). The PCR product is a linearized DNA where the majority of the transposon sequence no longer exists. The 5' sequence of each primer

20

introduces a unique *Bsg* I (type IIS endonuclease) recognition site. The cleavage site of *Bsg* I locates 16 nucleotides away from its recognition site. By placing the *Bsg* I site closer to or further away from the 3' of the primers, various number of nucleotides can be removed from the gene of interest. After Klenow treatment to blunt the 3' overhang and re-ligation, the plasmids become circular again and each will contain in-frame codon deletions in the gene of interest.

2.1.3 Rationale of codon scanning mutagenesis (CSM)

CSM adapts the approach described in CDM and aims to generate codon substitutions rather than deletions. CSM can replace up to three consecutive, in-frame codons at a random position of a protein with any desired codons. The Mu transposon for CSM (MuCSM) is nearly the same to MuCDM with only one nucleotide difference. MuCDM has a four nucleotide difference as the asymmetric ends whereas MuCSM contains a five nucleotide difference at the two ends (**Figure 2.4B**). This is designed so to reduce the non-specific binding of primers in the EI-PCR (see **Results and Discussion** section for details). Additionally, CSM has one more EI-PCR step than CDM. This is because new codons need to be delivered to the gene of interest. In the first EI-PCR, one primer introduces a *Bsg* I or *Bpm* I site that will remove in-frame codons from the gene of interest. The other primer delivers new codons to the gene of interest. Importantly, the sequence of this codon-carrying primer ("anchor" sequence) cannot be removed by *Bsg* I digestion and remains in the gene of interest. In the second EI-PCR, two primers anneal to the "anchor" sequence and each introduces a unique *Bsg* I site. Upon *Bsg* I digestion, all the foreign sequences will be removed, leaving only the new codons to fill in the codon "scar" created in the first EI-PCR (see **Figure 2.6** for detailed procedure).

```
...N₁N₂N₃N₄N₅CTTAGCGGCGCACGAAAAACGCGAAA...
        3'-GAATCGCCGCGTTTCAGCAACTAANHANHA-5'
              └─End #1─┘       └─────────────┘
                         "Anchor" and new codons
                         (do not pair with template)
```

Primer annealing in
first EI-PCR

```
                      Bsg I ┌─────────────────────────┐
              5'-TTTTTCGTGTGCAGTCGGA-3'                │
        ...GAAAGCGCAAAAAGCACGCGGCAGCCTN₁N₂N₃N₄N₅N₆N₇N₈N₉ N₁₀N₁₁ N₁₂N₁₃N₁₄...
                                    └─End #2─┘                      │
                                                                   ↓
```

Bsg I digestion
Klenow treatment
Re-ligation

↓

```
...N₁N₂N₃N₄N₅CTTAGCGGCGCAAAGTCGTTGATT NDTNDT N₁₂N₁₃N₁₄...
...N₁N₂N₃N₄N₅GAATCGCCGCGTTTCAGCAACTAA NHANHA N₁₂N₁₃N₁₄...
        └──────────────────────┘       └──────┘
           "Anchor" sequence          New codons
```

Second EI-PCR:
double-stranded DNA denatured
Primers annealing to "anchor"

↓

```
...N₁N₂N₃N₄N₅CTTAGCGGCGCAAAGTCGTTGATT NDTNDT N₁₂N₁₃N₁₄...
        3'-GAATCGCCGCGTAAATGACGTGCTTC-5'
                           └─Bsg I─┘
```

Primer annealing in
second EI-PCR

```
                Bsg I
        5'-TAACGTGCAGTTACAAGTCGTTGATT-3'
...N₁N₂N₃N₄N₅GAATCGCCGCGTTTCAGCAACTAA NHANHA N₁₂N₁₃N₁₄...
```

Bsg I digestion
Klenow treatment
Re-ligation

↓

```
...N₁N₂N₃N₄N₅NDTNDTN₁₂N₁₃N₁₄...
...N₁N₂N₃N₄N₅NHANHAN₁₂N₁₃N₁₄...
```

Figure 2.6. Detailed procedure of CSM. The asymmetric end sequences are highlighted in red. *Bsg* I sites are in italic letters. Green boxes indicate the new codons to be delivered. Di-NDT codon substitution is shown as a demonstration. In the first EI-PCR, an "anchor" sequence and the new codons are introduced by one primer and the other primer introduces a *Bsg* I site which removes the 5 bp duplication (from Mu transposon insertion) and additional 6 in-frame nucleotides. The second EI-PCR removes all the foreign sequences, leaving behind only the "NDTNDT" codons to fill in the 6 bp deletion gap. Note that the 2 bp 3' overhangs created by *Bsg* I is removed by Klenow treatment

22

2.1.4 Generation of size-minimized GFP using MuDel mutagenesis and CDM

This experiment aims to generate a size-optimized GFP that is less likely to interfere with the native function of a fused protein. *Aequorea victoria* GFP is generally considered as a size-minimized protein (*50-52*). Wild-type *A. victoria* GFP is a 27 kDa protein containing 238 amino acids. GFP and all its known variants adopt a β-can structure assembled by 11 antiparallel β-strands (*5, 53*). Most of the strands are connected by small loops consisting of one to four amino acids. Two larger loops appear at positions 129~143 and 189~197. The top and bottom "lids" composed mainly of residues 74-91 and 128-145, respectively (*5, 54*). The chromophore is located at the central α helix, surrounded by β-strands. The chromophore of wide-type GFP (wtGFP) is formed by residues Ser65, Tyr66 and Gly67, requiring no cofactors from host systems (*54, 55*).

Despite that GFP has become a commonly used laboratory tool, characterization of the functions of a target protein fused with GFP tag is not necessarily a simple task. GFP is usually considered as innocuous tag as it can fold independently and autocatalyze to form matured chromophore when fused to other proteins. However, GFP is, after all, a protein consisting of more than 200 amino acids. Fusion with GFP is still risky at disrupting the folding, native functions or intracellular locations of a target protein. Although rarely reported in literatures, some evidence exists that high level of GFP expression in certain locations is detrimental to cells (*56-58*) and that GFP fusion could change the intracellular location of a target protein (*59-61*). This problem could be alleviated by using a size-minimized GFP. Ideally, the flexible loops (which are very likely to interfere with other cellular proteins) can be removed or truncated without eliminating GFP fluorescence.

GFP is well tolerant to deletions at N or C terminus (*50, 51*). The minimal domain required for the fluorescence of wtGFP is 2~232 (*50*). An enhanced version of GFP (EGFP, mutation F64L/S65T) can still fluoresce with 5 amino acids deleted from N terminus or 10 amino acids deleted from C terminus (*51*). However, GFP is very sensitive to deletions at internal positions. Deletions of residues 132~139 or 191~195 in the two larger loops eliminated the fluorescence of EGFP (*51*). A targeted deletion analysis of the longest loop (129~143) in SuperGlo GFP (sgGFP, containing mutations F64L/S65C/I167T) showed that only two single deletions I128Δ and D129Δ could retain the fluorescence of GFP (*52*).

It is worth mentioning that previous deletion studies of GFP were all based on the whole-cell fluorescence. The underlying reason for the loss of fluorescence is still unclear. The intensity of whole-cell fluorescence of GFP reflects the overall contributions of expression level, folding robustness, efficiency of chromophore formation and intrinsic brightness (quantum yield and extinction coefficient) (*12*). One possibility is that those deletion mutants have lost considerable amount of conformational stability such that proteins cannot fold properly to catalyze the formation of chromophore. If this is the case, the lost fluorescence of the deletion mutants may be restored by folding-enhancing mutations (*62*). To understand if the deletions destroy the fluorescence of GFP by compromising its conformational stability, MuDel deletion mutagenesis (*25*) was performed on a UV-optimized GFP variant (GFP$_{UV}$) (*13*). The mutant library was screened for deletions that could retain GFP fluorescence ("neutral" deletions). Detailed characterization of these "neutral" deletions, indeed, suggested that the fluorescence loss was largely due to the compromised stability of proteins. Furthermore, incorporation of

folding-mutations largely recovered the fluorescence of these mutants and the fluorescence of the mutants with combined "neutral" deletions.

Encouraged by this experiment, we then set out to apply CDM to identify more "neutral" codon deletions. In total, more than 40 "neutral" deletions were identified using MuDel deletion mutagenesis and CDM. These results strongly indicated that GFP can be minimized, as opposed to the previous conclusion that GFP is a naturally size-optimized protein (*50-52*). Although CDM was developed in our experiment for minimizing GFP, it could have potential applications for engineering other proteins. Some protein functions can be improved by deleting certain regions or residues. In the case of streptavidin (159 amino acids), high affinity to biotin (K_d ~10^{-15} M) is only seen in the truncated or "core" streptavidin (residue 13~139) because the N- and C-terminal regions prevent its ligand, biotin, approaching the binding pocket (*63*). Another example is that site preference of MuA tranposase can be reduced by truncating its N-terminal sequence (*64*). Additionally, amino acid deletion is a general mechanism of sequence variation in immunoglobulin variable domain (*65*). Thus, CDM can be used as a general method for engineering protein functions.

2.1.5 Development of spectral variants of GFP using CSM

Traditional methods for diversifying protein sequences usually generate mutations at the nucleotide level. This largely compromises the diversity of a mutant library because of codon redundancy and the rarity of consecutive nucleotide mutations. A statistic analysis of the existing random mutagenesis methods have suggested that many of them are highly biased due to the limitation of generating mutations at the nucleotide level (*66*).

An ideal random mutagenesis method would allow one to substitute every amino acid of a protein sequence in a statistically equivalent manner.

Random insertion and deletion (RID) mutagenesis (*67*) is perhaps the first attempt to generate mutations at the amino acid level. RID allows random deletions of an arbitrary number of nucleotides from a gene of interest followed by insertions of any desired sequences. This method, however, requires multiple inter-molecular DNA ligations and can be difficult to employ in generating large libraries. Another important progress was made by Jones and co-workers, where they developed a MuDel transposon-based method to generate random triplet nucleotide substitutions (*68*) (referred to as TriNEx for triplet nucleotide exchange) or random domain insertions (*69*).

While both RID (*67*) and TriNEx can generate mutations in the unit of triplet nucleotides, our interest was to generate "codon mutations" in which the reading frame of the mutation is defined. In a previous study, we reported a process for replacing single, random in-frame codons with a new codon of choice (*49*). This method employs the same MuDel transposon described in the TriNEx method (*25, 68*) and integrates an intein-based "dual reporter" reading-frame selection system (*48*). This codon-based random mutagenesis method limits unwanted out-of-frame mutations resulting in a "pure" library and was perhaps the closest to achieve a non-redundant mutagenesis method (*68*). However, this method can generate only one mutation per round of mutagenesis and contains a technically challenging manipulation step (inter-molecular blunt-end ligation). Hence, construction of a large mutant library using this method may be difficult.

To overcome these problems, we developed an asymmetric transposon-based codon scanning mutagenesis (CSM) (*70*). Compared to the MuDel transposon-based single

codon mutagenesis method (*49*), the current method 1) can scan up to three consecutive codons rather than only one, 2) has user-friendly process of manipulation, and 3) has been largely optimized for high efficiency of mutagenesis. These features enable the construction of large-scale libraries. The major advantage of this method is that the mutant library is highly diverse so that beneficial mutants can be identified by screening just a small fraction of the mutants in the library. As a demonstration of this approach, we generated three libraries that scanned single, di- or tri-NDT degenerate codons (*71*), respectively, throughout the gene encoding superfolder GFP (sfGFP) (*72*). By screening only 500 clones from each library, we successfully identified several mutants that showed altered spectral properties.

2.2 Results and Discussion

2.2.1 Random deletion mutagenesis of GFP$_{UV}$ using MuDel transposon

Since early deletion analysis on GFP suggested that it is sensitive to amino acid deletions especially at internal positions (*50-52*), it is necessary to investigate the effect of amino acid deletions on GFP to answer the question if it is possible at all to generate a minimized GFP. A size-minimized GFP is most likely to be produced by incorporation of "neutral" deletions, amino acid deletions that can retain fluorescence. Therefore, MuDel transposon-based random deletion mutagenesis followed by fluorescence screening was performed on GFP$_{UV}$ to identify such "neutral" deletions first. These "neutral" deletions were then characterized to determine if the fluorescence loss was caused by compromised protein stability. This is particularly important because if this is the case, impaired GFP fluorescence may be restored by folding-enhancing mutations.

2.2.1.1 Construction of deletion library and identification of "neutral" deletions

The transposon reaction resulted in 38,000 individual colonies, sufficient for a full coverage of all the possible insertion sites. The collected pGFP$_{UV}$-MuDel transposon library was digested by *Kpn* I/*EcoR* I to isolate those transposon insertions inside the GFP$_{UV}$ gene (**Figure 2.7**). The diversity of library was maintained throughout the whole process.
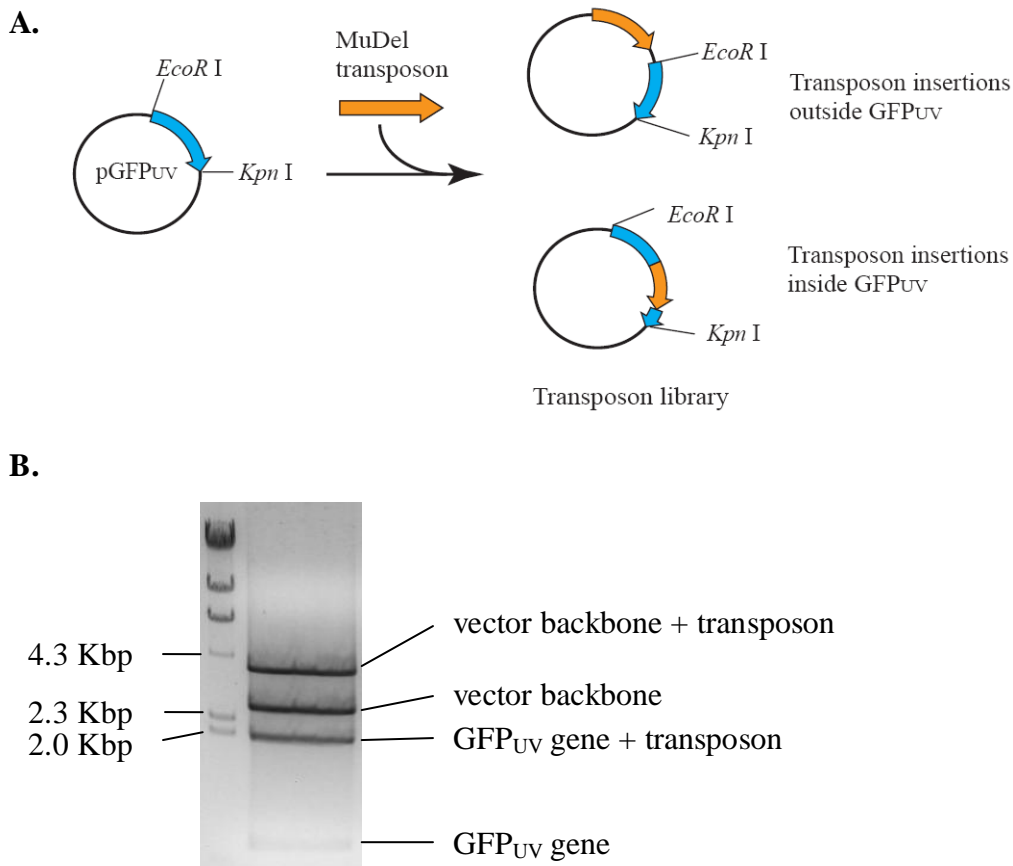
Figure 2.7. Isolation of transposon insertions inside GFP$_{UV}$ gene. **A)** *EcoR* I/*Kpn* I double digestion to isolate transposon insertions inside GFPUV. **B)** 1% agarose gel showing the four DNA fragments resulting from *EcoR* I/*Kpn* I digestion. A purified library can be obtained by gel purification and subsequent re-ligation of the DNA bands "vector backbone" and "GFP$_{UV}$ gene + transposon"

28

To screen the fluorescence of these mutants, cells were grown on plates for 48 h at 23 ℃. The in-plate fluorescence was then monitored under UV lamp by visual inspection. In total, 40 fluorescent mutants and 24 non-fluorescent mutants were picked from the plates and sequenced. Ninteen unique mutants were obtained from the 64 sequenced clones as some mutants carried identical deletions. A detailed analysis of the sequences of the mutants is shown in **Table S-1 (Appendix V)**.
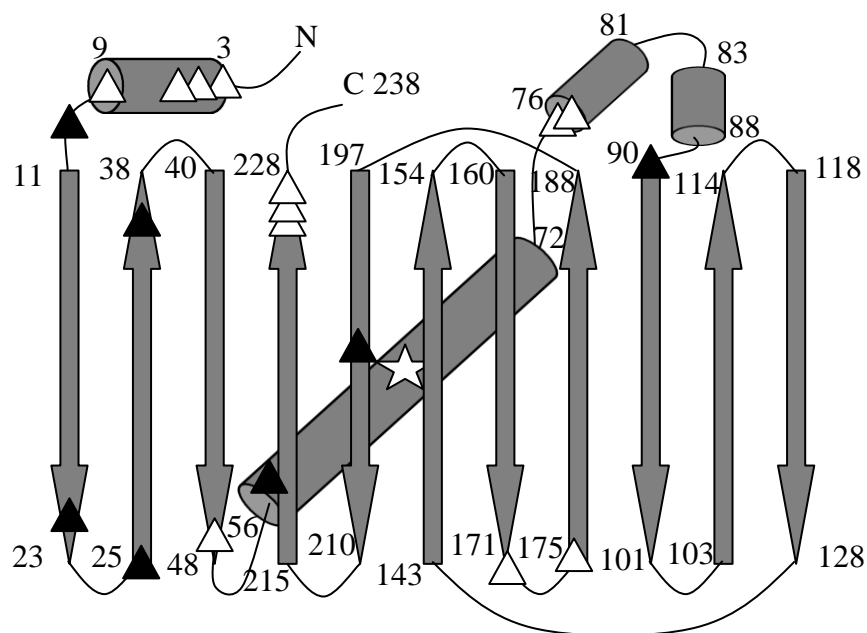
Figure 2.7. Deletion maps of GFP$_{UV}$. The displays are arrows for strands, cylinders for helices and star for chromophore. The numbers indicate the starting and ending positions of secondary structures. Deletions resulting in fluorescent and non-fluorescent phenotypes are indicated by white and black triangles, respectively. The fluorescent mutants include S2/K3→R, G4Δ, E5/D6→E, F8/T9→S, C48Δ, P75/D76→H, P75Δ, E172Δ, S175/V176→F, A225/A226→A, G228Δ and G228/I229→V. The non-fluorescent mutants include G10Δ, D21Δ, G24Δ, E34Δ, P58Δ, P89+E90→Q and T203Δ.

To verify the in-plate fluorescence, the identified 19 mutants were grown at 23 ℃ for 24 h in liquid culture. The results revealed that 12 of the deletion mutants were fluorescent and the other 7 were not (**Figure 2.8**). Of the 12 fluorescent mutants, 7 contained deletions at N- or C-terminus (S2/K3→R, G4Δ, E5/D6→E, F8/T9→S,

A225/A226→A, G228Δ and G228/I229→V) and 5 had deletions at internal positions (C48Δ, P75/D76→H, P75Δ, E172Δ, S175/V176→F). The internal deletions were all at the border of helices and strands. Of the 7 deletions that eliminated GFP fluorescence, one was found in the N-terminal region (G10Δ), one in the middle of a β-strand (T203Δ) and the other five mutants at the border of helices and strands (D21Δ, G24Δ, E34Δ, P58Δ, P89/E90→Q). Interestingly, mutant G10Δ could restore trace fluorescence after prolonged incubation (>48 hours) at 23 ℃. These results were consistent with previous analysis (*73*) and the general expectation that deletions within the secondary structure were unfavorable for retaining protein functions.

2.2.1.2 Characterization of GFP$_{UV}$ mutants with "neutral" deletions
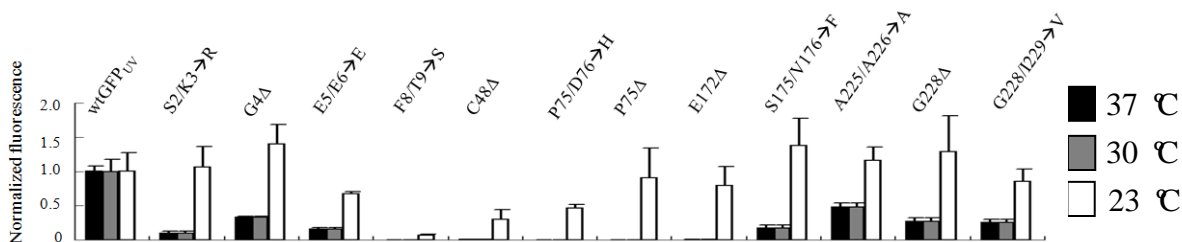
Figure 2.8. Whole-cell fluorescence of 12 mutants with "neutral" deletions under three different temperatures.

A size-minimized GFP could be most easily generated by sequential incorporation of these "neutral" deletions into wtGFP. However, characterization of these deletion mutants showed that their *in vivo* whole-cell fluorescence was all decreased when expressed at 37 ℃ (**Figure 2.8**). Interestingly, the reduced whole-cell fluorescence can be largely recovered at reduced temperatures (30 ℃ and 23 ℃). Some of these mutants were only fluorescent when expressed at 23 ℃. Importantly, the GFP fluorescence under different temperatures was not compared directly. Instead, the fluorescence of each

mutant was compared to the fluorescence of wtGFP$_{UV}$ under the same temperature. This "normalized" fluorescence was used to dissect the effect of these deletions on GFP fluorescence. The dramatic recovery of the fluorescence at low temperatures indicated that these mutants might encounter folding problems at high temperature.

A.

B.

C.
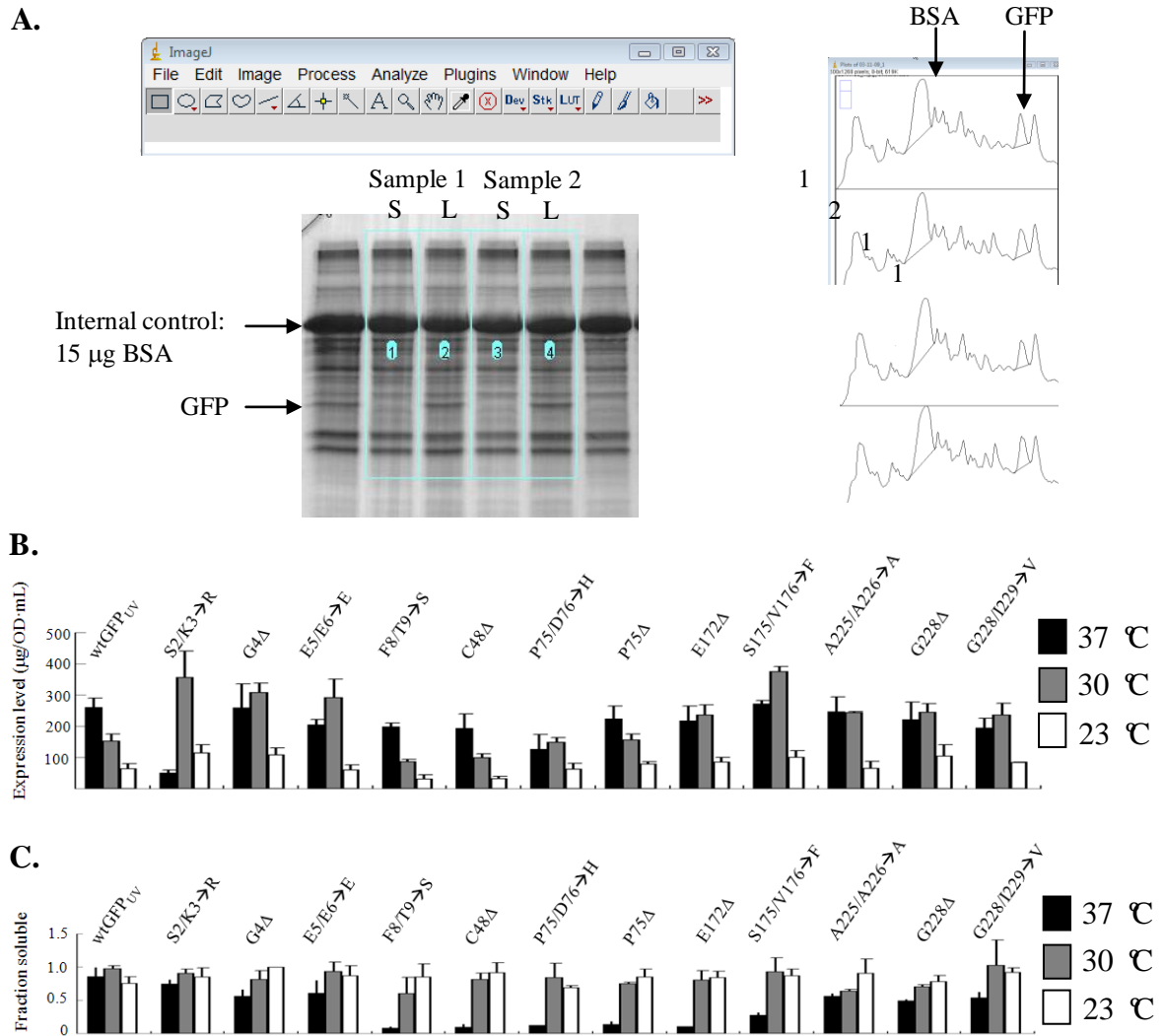
Figure 2.9. Characterization of *in vivo* expression and fraction soluble of deletion mutants. **A)** Quantification method using Image J. S: supernant; L: crude cell lysis. **B)** Overall GFP expression. **C)** Fraction soluble.

To further understand the cause of fluorescence loss, we characterized the *in vivo* expression and solubility of the 12 fluorescent mutants. For each mutant, the GFP overall expression (in crude cell lysis) and soluble portion (in supernatant of cell lysis) were analyzed by SDS-PAGE. The amount of protein present in each sample was estimated by comparing with an internal standard BSA of known concentration (**Figure 2.9A**). The quantification was carried out using Image J (http://rsbweb.nih.gov/ij/). The protein expression was presented in the unit of µg protein produced per OD_{600} per mL culture. Overall expressions of wtGFP$_{UV}$ and mutants were generally within 50% variation under the same temperatures (**Figure 2.9B**). The fraction soluble was the ratio of supernatant GFP and overall protein in whole-cell lysis. Remarkably, the soluble fraction of each mutant underwent a dramatic increase when the temperature was reduced, especially for those carrying internal deletions (**Figure 2.9C**). These results further indicated that these "neutral" deletions may disrupt protein folding at high temperatures.

To better understand the effect of these deletions on GFP, the mutants carrying internal deletions were purified and further characterized (**Figure 2.10**). These mutants were of particular interest, because only two internal deletions (I128Δ and D129Δ) were known to retain GFP fluorescence from previous deletion analysis (*52*).
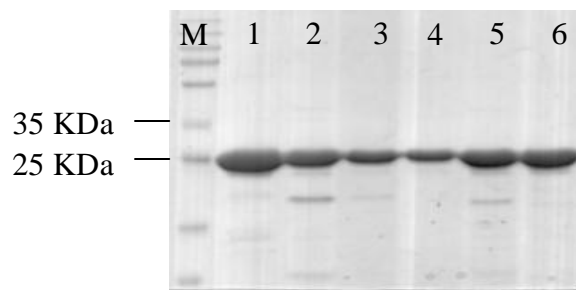
Figure 2.10. Purified mutants with internal deletions. Lane 1~6: mutants C48Δ, P75+D76→H, P75Δ, E172Δ, S175+V176→F, respectively, expressed at 23 ℃

GPF$_{UV}$ has a major excitation (Ex) peak at 397 nm and a minor Ex peak at 475 nm. When excited at 397 nm or 475 nm, the emission spectra of wtGFP$_{UV}$ and mutants were superimposable (**Figure 2.11A**). The excitation spectra of these mutants were very similar with the only difference being the ratio of the major and minor peaks (**Figure 2.11B**), which represents the relative abundance of the two ionization state of the chromophore. The neutral chromophore is responsible for the 397 nm peak whereas anionic chromophore gives 475 nm peak (*12*). The mutations that can alter the ratio of the two chromophore species have been observed before (*74*). For example, mutation S65T in EGFP (F64L/S65T) completes the ionization of the chromphore in the ground state and results in a single Ex peak at 475 nm.
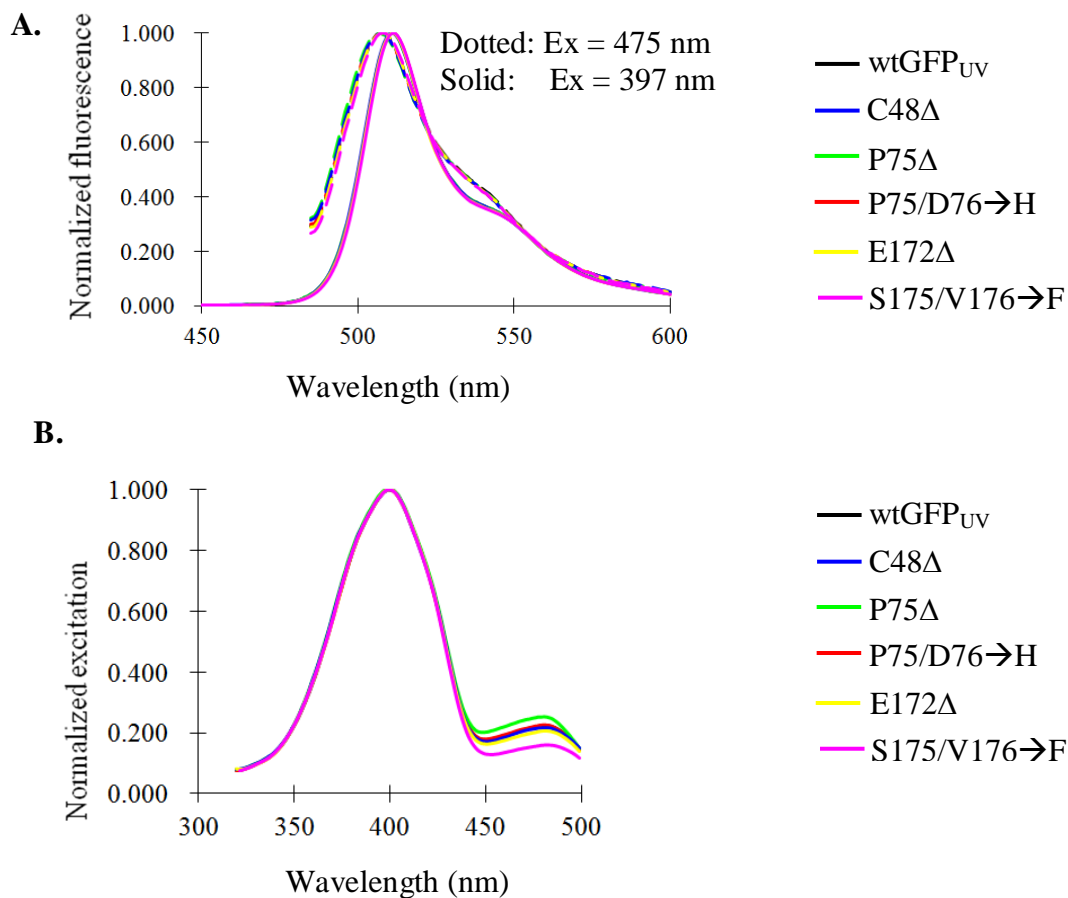
Figure 2.11. Spectra of mutants with internal deletions. **A**) Emission. **B**) Excitation.

The above results indicated that the chromophores of the five deletion mutants could be chemically similar to that of wtGFP$_{UV}$. To further characterize the properties of the chromophore, we determined the extinction coefficient at 397 nm (ε_{397}) and 475 nm (ε_{475}) and the quantum yield of each mutant (**Table 2.1**). While some deletions, such as S175/V176→F, had obviously altered ε_{475}, most deletions did not result in significant change in ε_{397}. Furthermore, the quantum yield of these mutants was nearly the same as that of wtGFP$_{UV}$. These results strongly supported the idea that these deletions did not disrupt the chemical structure of the chromophore. Therefore, decreased fluorescence loss observed in this study, presumably as well as those in previous studies (*50-52*), is very likely to arise from the compromised protein stability rather than damaged structures of chromophore.

Table 2.1. The effect of internal deletions on extinction coefficient and quantum yield

wtGFP$_{UV}$ and mutants	ε_{397} (M^{-1}cm^{-1})[a]	ε_{475} (M^{-1}cm^{-1})[a]	Quantum yield[a] (λ_{ex}=397 nm, %)
GFP$_{UV}$	29,600	7,500	0.79[b]
C48Δ	31,600	8,500	0.80
P75+D76→H	29,700	6,800	0.79
P75Δ	29,400	7,100	0.79
E172Δ	29,300	7,600	0.81
S175+V176→F	30,900	5,000	0.79

Note: [a] SD of each value is generally within estimated instrumental error ($\pm$2%)

[b] Patterson *et al*, 1997

In GFP, chromophore maturation and protein folding are coupled as the autocatalysis of the chromophore formation requires the residues involving in the chemistry to be in structural proximity (*12*). One result of the disrupted folding process is that the maturation of chromophore could be slowed down. In fact, the above results strongly supported this hypothesis because the five mutants had the same quantum yield values to wtGFP$_{UV}$ but decreased whole-cell fluorescence (**Figure 2.8**). These results suggested that the chromophore maturation of the five mutants may be interrupted.

A.   **B.**

wtGFP$_{UV}$
C48Δ
P75Δ
P75/D76→H
E172Δ
S175/V176→F
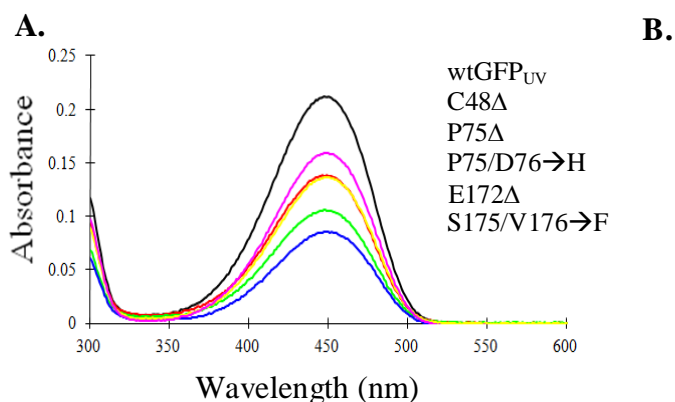
Figure 2.12. Efficiency of chromophore formation. **A**) Absorbance of 7 μM protein under "base denatured" condition. **B**) Calculated efficiency of chromophore formation. SDs are within instrumental error of ±2%.

To verify this conclusion, the efficiency of chromophore maturation of each sample was determined using the "based denatured" method (*75*). Under "based denatured" condition, all green-colored, matured chromophore has a standard ε_{447} of 44,000 $M^{-1}cm^{-1}$ (*75*). The concentration of matured chromophore can thus be obtained by Beer's Law using the experimentally determined absorbance (see **Materials and Methods** for details). Efficiency of chromophore maturation is equal to the ratio of matured protein and overall protein (see **Materials and Methods** for equations). In the current study, the efficiency of chromophore maturation for wtGFP$_{UV}$ was lower than the reported value of

1.0 (*76*). This was probably due to different expression conditions. Importantly, when compared with wtGFP$_{UV}$, all the five deletion mutants showed clearly reduced efficiency of chromophore maturation (**Figure 2.12**).

2.2.1.3 Restore fluorescence by incorporation of folding-enhancing mutations

As shown above, the "neutral" deletions reduced GFP fluorescence most likely by disrupting protein folding (as well as chromophore maturation). Next question would be if the decreased fluorescence can be recovered. Previous protein folding studies have shown that compromised protein stability can be rescued by incorporation of global folding-enhancing mutations (*62*). Being one of the most engineered proteins, GFP has many known folding mutations identified in previous study (*15, 72, 77*). Herein, we chose a central folding mutation F64L (*76*) and a distant folding mutation S30R (*72*) to rescue the fluorescence of the five mutants with internal deletions (**Figure 2.13**).
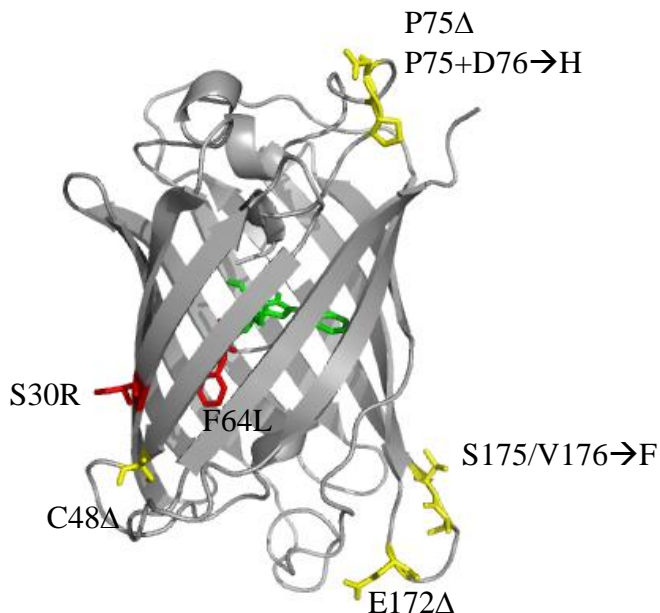
Figure 2.13. Folding-enhancing mutations F64L and S30R and the five internal deletions. The colors are yellow for deletions, red for folding- mutations and green for chromophore.

36
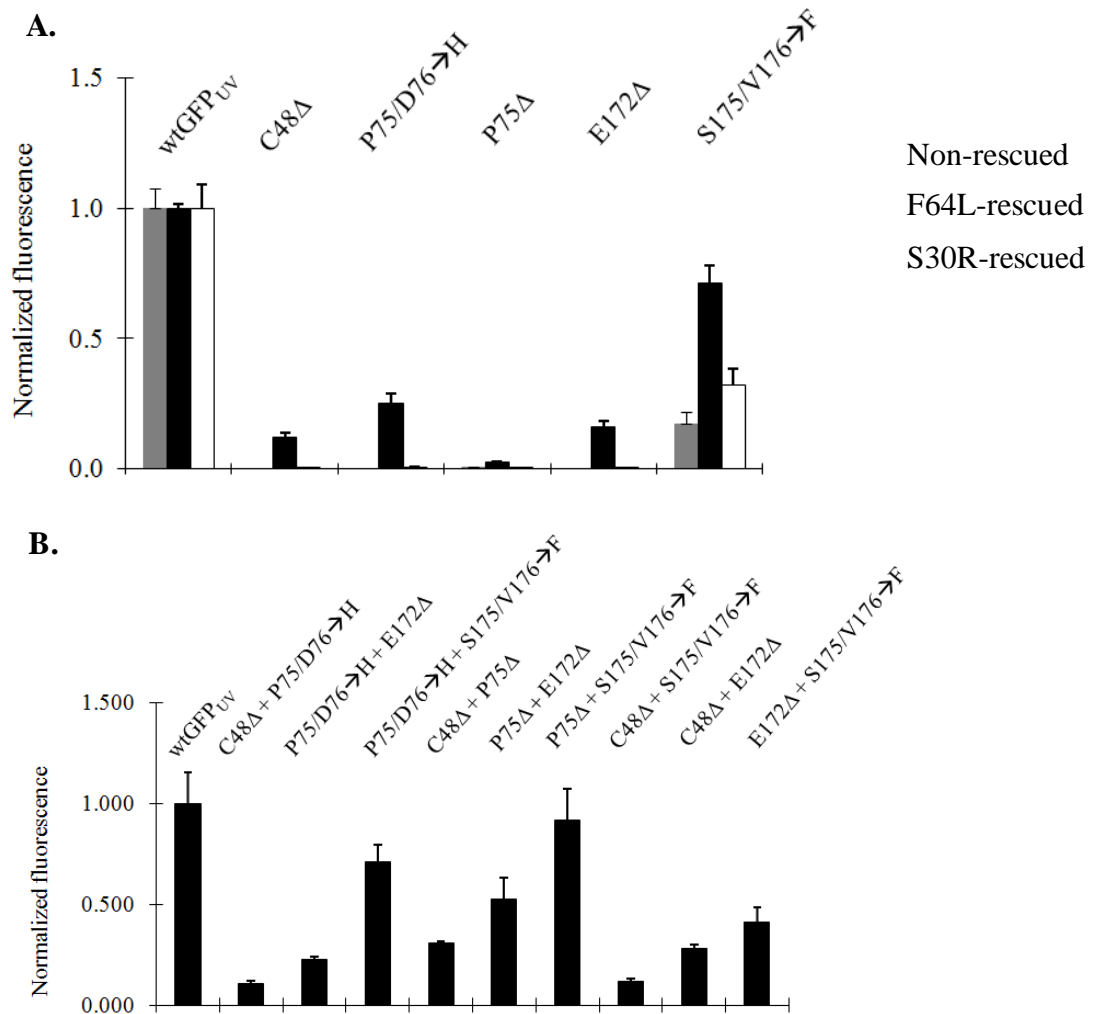
Figure 2.14. Folding-rescued whole-cell fluorescence. **A)** F64L and S30R rescued whole-cell fluorescence at 37 ℃. The rescued fluorescence of mutants is normalized to the fluorescence of wtGFP$_{UV}$ with corresponding folding mutations. **B)** F64L rescued the fluorescence of mutants with combined deletions at 23 ℃. The fluorescence of mutants is normalized to that of GFP$_{UV}$-F64L. Three individual replicates were performed for each sample.

F64L and S30R mutations were introduced into each of the five mutants individually. The whole-cell fluorescence at 37 ℃ was assayed for each new construct. Mutation S30R did not result in obvious fluorescence increase whereas F64L restored partial fluorescence of all the five mutants (**Figure 2.14A**). This is presumably because F64L is

a central folding mutation and may have larger impact on folding process than the distant folding-mutation S30R. Next was examined if the folding mutation F64L could rescue the fluorescence of mutants with combined deletions. This is particularly important because generation of a size-minimized GFP may require sequential incorporation of pre-determined "neutral" deletions into the same protein. All the possible mutants with double internal deletions were generated and tested for F64L folding-rescue. Remarkably, combination of any two of the five deletions eliminated GFP fluorescence at 23 ℃ and incorporation of F64L mutation successfully restored the fluorescence loss on all the ten double deletion mutants (**Figure 2.14B**). To the best of our knowledge, these mutants are the only known GFP variants that can retain fluorescence with multiple deletions incorporated at internal positions. However, mutation F64L was neither capable to rescue their fluorescence at 37 ℃ nor the fluorescence of any triple deletion mutants. This suggested that triple deletions may have overwhelmed the folding-rescue ability of F64L. Additional folding mutations may be necessary to recover the fluorescence of mutants with three or more deletions.

2.2.2 Codon deletion mutagenesis of GFP_{UV} using MuCDM transposon

MuDel deletion mutagenesis has two limitations that prevent rapid identification of "neutral" deletions. First, it lacks control of reading frame and thus some deletions were found with accompanying mutations. Second, each mutant in the library contains only one amino acid deletion. This motivated us to develop a novel random deletion mutagenesis that can generate multiple, random "codon" deletions, referred to as codon deletion mutagenesis (CDM) below. The Mu transposon used in this method is named MuCDM in the following section for convenience.

2.2.2.1 The development of an "all-in-one" asymmetric Mu transposon

Previously, transposons have been successfully employed to mutate sequence with foreign segments of DNA, therefore we hypothesized that they could also be used to generate new annealing sites for primers within a plasmid. The Mu transposon is known to accommodate nucleotide changes outside of the MuA transposase recognition sites (R-sites). These outside sequences are always symmetrical inverted repeats, however, and to the best of our knowledge there has been no attempt to generate asymmetric transposons. The forward and reverse ends of asymmetric transposon were designed to contain a four nucleotide difference with the outermost 3' nucleotide remaining symmetrical (**Figure 2.4B**). These asymmetric ends can force two different primers to be specifically used for whole-plasmid PCR amplification. As demonstrated in the following expeirments, a 4 bp difference in MuCDM transposon is sufficient for the specific annealing of carefully designed primers (see next section for details). The transposition efficiency of MuCDM was estimated to be ~1% by comparing a test plate supplemented with ampicillin and kanamycin to a control plate with only kanamycin. This efficiency is comparable to other Mu transposon variants, suggesting that the nucleotides flanking the transposase recognition sites need not be identical.

Another interesting observation was that MuCDM transposon could insert in the VMAI-N region on the pTT plasmid, as far as 60 residues upstream to GFP$_{UV}$. In these cases, the self-splicing of intein is disrupted by transposon insertions, but β-lactamase may still, though less likely, function with an N-terminal fusion peptide. Assuming that 60 residues of VMAI-N and 238 residues of GFP$_{UV}$ were the allowed positions for transposon insertions, 3,000 colonies were collected from the transposition reaction to

ensure a >95% chance of full coverage (*78*). In general, 1 μL of the transposition product could yield 1,500 colonies. Therefore, a single 20 μL transposition reaction should provide sufficient colonies to cover the full library of a very large protein.

The randomness of the transposon insertions was verified by *EcoR* I/ *Mly* I digestion of 20 randomly picked colonies (**Figure 2.15**). The transposon insertions were found to distribute throughout the GFP$_{UV}$ gene with colonies 13, 15, 17 and 19 containing N-terminal insertions and colonies 5, 10 and 11 bearing C-terminal insertions (the smaller band was difficult to visualize). Notably, transposon were inserted in VMAI-N in colonies 3, 8 and 18 (with a >780 bp smaller band). To investigate the distribution of transposon insertions, we did not purify the transposon library in this experiment. In any further applications, however, the library can be purified to isolate the transposon insertions in the target gene as described above (**Figure 2.7**).
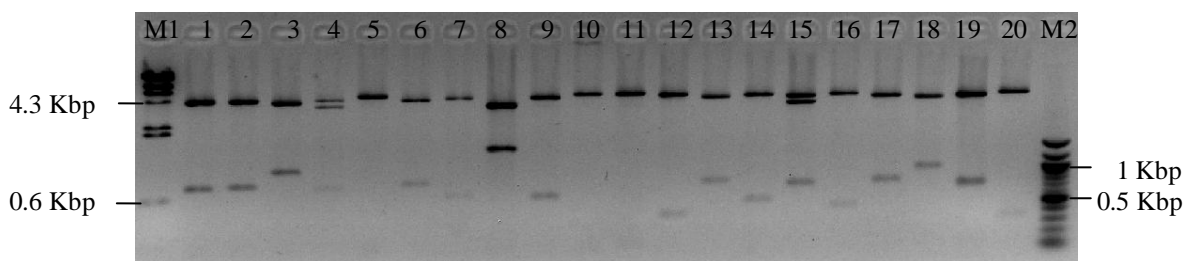
Figure 2.15. Verification of randomness of transposon insertions. Twenty random colonies from the transposition reaction were digested by *Mly* I/*EcoR* I. M1, λ *Hind* III DNA ladder; M2, NEB 100 bp DNA ladder. MuCDM transposon carries a Mly I site at the 3' end and pTT-GFP$_{UV}$ plasmid bears an *EcoR* I site at the C-terminus of GFP$_{UV}$. Upon *Mly* I/ *EcoR* I digestion, the N-terminal insertion of MuCDM transposon in GFP$_{UV}$ will result in 780/ 4,400 bp bands whereas the C-terminal insertion will result in 70/ 5,000 bp bands. If the transposon is inserted in the N-VMA sequence, >780 bp/ <4,400 bands are expected.

2.2.2.2 Optimization of conditions for enzymatic inverse PCR (EI-PCR)

In EI-PCR, two primers will specifically anneal to the transposon ends in a back-to-back manner and amplify the whole plasmid, except the transposon sequence, in an inverse direction (**Figure 2.5**). The starting and ending position of the PCR product were randomly distributed in the GFP_{UV} gene, depending on the position of transposon insertions. Therefore, the "head" of a PCR product could overlap with the "tail" of another product, resulting in "side product" with extended size (**Figure 2.16A**). This problem, however, was solved by simply reducing template concentration and cycling number (*79*). As a crude estimation, side product became predominant once the concentration of the desired PCR product exceeded 2 ng/μL (**Figure 2.16B**).
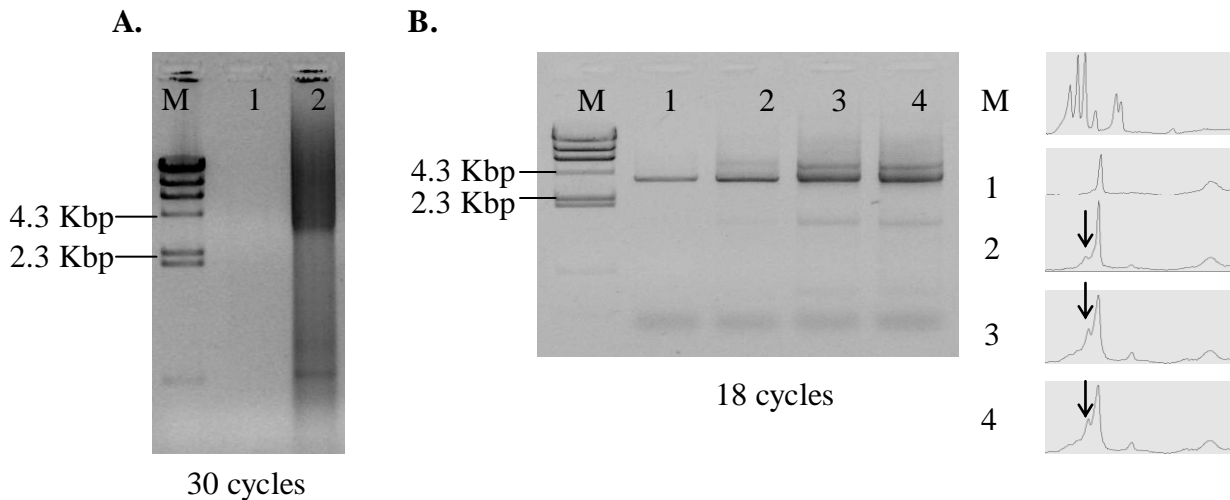
Figure 2.16. Testing template concentrations and cycling numbers of EI-PCR. **A**) Smeared PCR product occur when regular template concentration and cycling number are used. Lane 1, no template control; Lane 2, 0.1 ng/μL template. The cycling number is 30 for both. **B**) Optimization of template concentration. Lane 1~4: 5, 25, 125 and 250 pg/μL template, respectively. The cycling number is 18 for these four reactions. DNA gel is analyzed by Image J to determine the relationship between the concentration of desired product and the appearance of side product. Side product is indicated by arrow.

Another challenge of EI-PCR is that the primers need be carefully designed to prevent non-specific annealing (**Figure 2.17**). This is because MuCDM transposon has only four nucleotide asymmetric sequence at the two ends. The majority of the forward and reverse primer annealing sequence is symmetrical (**Figure 2.5**). That said, the four nucleotide difference provided the only basis for the primers to specifically anneal to one end but not the other. It was found that the length of these primers is very important. On one hand, the longer the primer is, the less weight this 4 bp difference will have in the whole primer. On the other hand, PCR process also requires primers to have sufficient nucleotides pairing with the template. Hence, the primer design must balance between efficient annealing and specific annealing. A variety of primers have been tested and the best ones (with efficient and specific annealing) identified are summarized in **Table 2.2**.
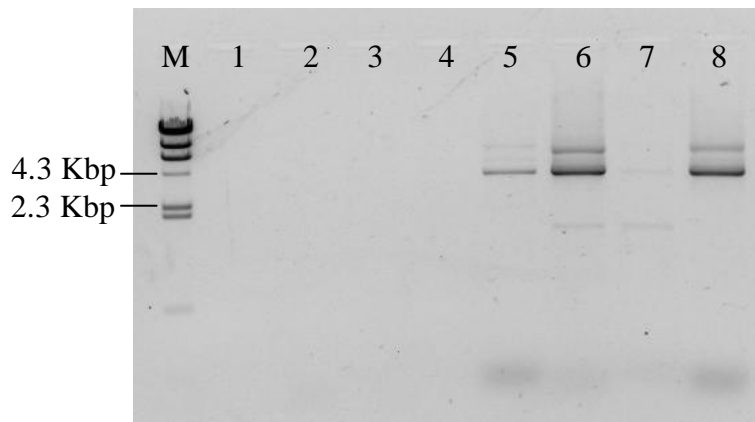
Figure 2.17. PCR reactions for testing primer design. Lane 1, no template control; lane 2, forward primer A only; lane 3, reverse primer B only; lane 4, reverse primer C only; lane 5, reverse primer D only; lane 6, A+B; lane 7, A+C; lane 8, A+D. Note that in lane 5, primer D non-specifically anneals to both ends of the transposon and results in a side product.

Table 2.2. List of primers for the EI-PCR of CDM

Primer names	Sequences
FWD8N	TTTCGCGTTTTT<u>GTGCAG</u>CCGTCGGA
FWD11N	TTTTTCGT<u>GTGCAG</u>TCGGA
REV0N	GCTTTC<u>GTGCAG</u>TTCGTGCGCCGCTAAA
REV3N	GCTTTCGCG<u>GTGCAG</u>GTGCGCCGCTAAA
REV6N	GCTTTCGCGTTT<u>GTGCAG</u>CGCCGCTAAA
REV9N	CGTTTTTC<u>GTGCAG</u>CGCTAAA

Note: 1. Asymmetric sequences highlighted in red.

2. *Bsg* I sites are underlined.

3. "FWD" and "REV" indicate forward and reverse primers, respectively.

4. The number in the primer names indicate the number of nucleotides that can be removed from target gene by *Bsg* I digestion. For examples, FWD8N means that 8 bp will be removed from the forward end of PCR product upon *Bsg* I digestion, including the 5 bp duplicate.

5. To delete a desired number of codons, more than one combination of primer pairs may exist. For example, a two codon deletion library can be generated using either the primer pair FWD8N/REV3N or FWD11N/REV0N.

2.2.2.3 Optimization of conditions for restriction digestion and Klenow treatment

Bsg I cleavage was >95% efficient under optimized conditions, with several final mutants containing uncut *Bsg* I sites. Fresh SAM and incubation of reaction in a PCR thermocylcer were critical for the complete cleavage of *Bsg* I. Klenow digestion was also >95% efficient following NEB's instructions. Short incubation time (~15 min) and quenching with EDTA before heat-inactivation were important to avoid overdigestion. Four nucleotide deletion and two nucleotide deletion were also observed in the sequenced mutants, which were presumably caused by overdigestion and incomplete digestion of Klenow, respectively.

2.2.2.4 Overall efficiency of CDM

Five deletion libraries of GFP$_{UV}$ were constructed by CDM, containing one to five consecutive codon deletions, respectively. To characterize the library quality, ~20 colonies were randomly picked from each library and sequenced (**Appendix V**). More than 90% of mutants in the library contained seamless, in-frame deletions, including those deletions occurring in the VMAI-N region (**Table 2.3**). The 8% undesired mutants were the results of incorrect *Bsg* I/ Klenow digestion as stated above. This mutagenesis efficiency is advantageous over many other approaches and is sufficient for any further applications. Importantly, the reproducibility of CDM was illustrated because the five deletion libraries were constructed following essentially the same procedure, with the only difference being the primer pairs used in EI-PCR.

Table 2.3. Quality of the five deletion libraries

Deletion Libraries	One	Two	Three	Four	Five	Overall
Total sequenced	20	21	21	20	22	104
Incorrect	2	1	2	1	2	8 (7.7%)
Correct in VMAI-N[a]	4	1	0	1	5	11 (10.6%)
Correct in GFP$_{UV}$	14	19	19	18	15	85 (81.7%)

Note: [a] As the pTTGFP$_{UV}$-MuCDM library was not purified to isolate transposon insertions inside GFP$_{UV}$ gene, the final deletions may occur in the upstream leading peptide.

Next the distribution of the 96 correct deletions was examined. This information is very important as large-scale deletion mutagenesis expeiments demand the coverage of the entire protein sequence. Strong site preference, if any, will be undesirable for the application of CDM. In **Figure 2.18**, the cumulative fraction of deletions is plotted with the deletion positions. In such a plot, uniform distribution should present a linear

relationship, because if deletions are evenly distributed throughout the entire protein sequence, the total number of deletions (cumulative deletions) should increase proportionally with the amino acid positions. If these deletions do not follow uniform distribution, experimental curves will deviate from the straight line. From **Figure 2.18**, the distribution of deletions from the five libraries was generally consistent with the uniform distribution with minor bias observed. This site preference of deletions may result from the intrinsic bias of MuA transposase as previously observed (*24*). Fortunately, recent studies have suggested that the site preference of MuA transposase can be reduced by truncating its N-terminal sequence or changing its working buffer (*64*). Another possible reason is the "dual reporter" reading-frame selection marker. Although the intein-splicing mechanism guarantees an identical "Tat+BLA" reporter regardless of the position of transposon insertions, the folding of VMAI-N and VMAI-C may be problematic depending on the peptide sequence between these two domains. Nevertheless, previous studies of this intein-mediated reading-frame selection system revealed a non-biased pattern of mutations.
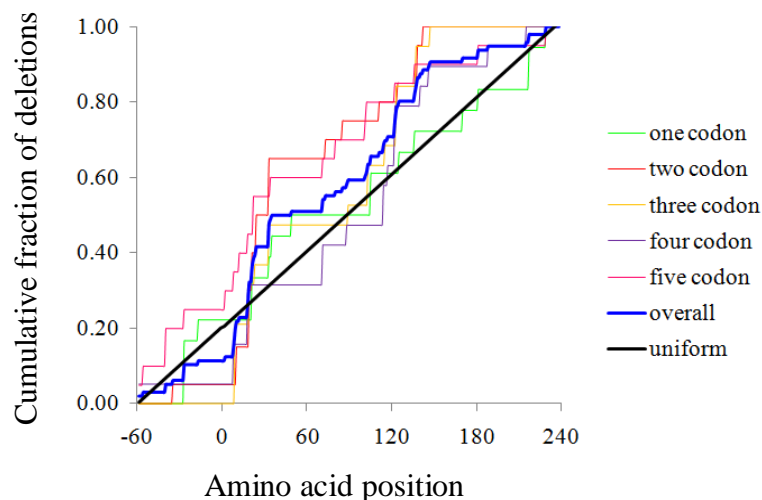
Figure 2.18. Distribution of deletions in GFP$_{UV}$ and VMAI-N, based on the 96 correct deletions. The first residue of GFP$_{UV}$ is numbered as 1. The straight and curved lines represent uniform distribution and experimental results, respectively.

45

2.2.2.5 Identified "neutral" deletions

The five deletion libraries were cloned into pET expression vector and screened for GFP fluorescence. In total, 24 non-fluorescent mutants and 12 fluorescent mutants were identified (**Figure 2.19**). The 12 fluorescent deletion mutants included 20 "neutral" deletions at internal positions (**Figure 2.20**). To the best of our knowledge, none of these residues have been reported before, nor can they be easily identified using traditional methods. Consistent with the results from MuDel deletion mutagenesis, these "neutral" deletions were located either in the loop structure or at the border region of helices/strands. These "neutral" deletions are usually expected to occur in the longest loop (*51, 52*), but our results suggested that small loops were well tolerant to deletions, as demonstrated in mutants D117Δ, G51Δ/K52Δ, E90Δ/G91Δ and G189Δ/D190Δ. Surprisingly, multiple deletions G51Δ/K52Δ, E90Δ/G91Δ, G189Δ/D190Δ, A154Δ/D155Δ/K156Δ, A226Δ/A227Δ/G228Δ/I229Δ and A226Δ/A227Δ/G228Δ/I229Δ/T230Δ could also retain GFP fluorescence. This result has largely expanded our knowledge of the tolerance of GFP to deletions as such consecutive amino acid deletions were usually thought to be very destabilizing to the compact β-can structure of GFP.
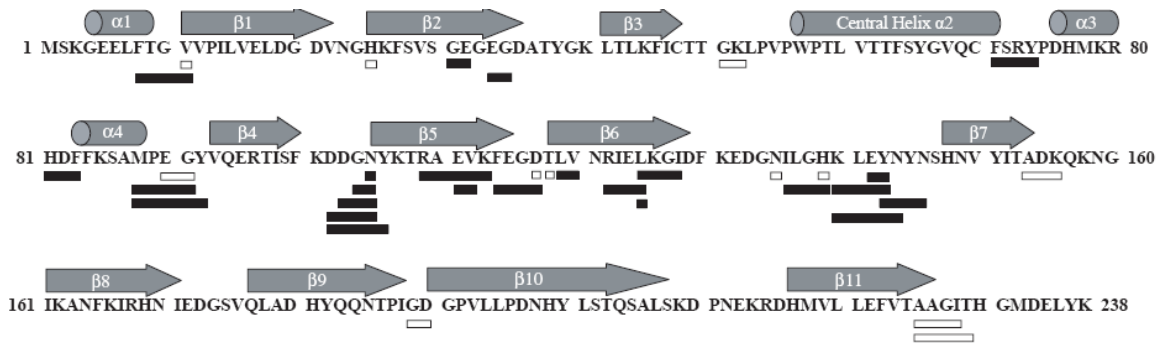
Figure 2.19. Identified deletions in GFP$_{UV}$. β strands and α helices are displayed as arrows and cylinders, respectively. The deletions that can retain fluorescence are shown in open boxes and those resulting in non-fluorescent mutants are filled boxes.
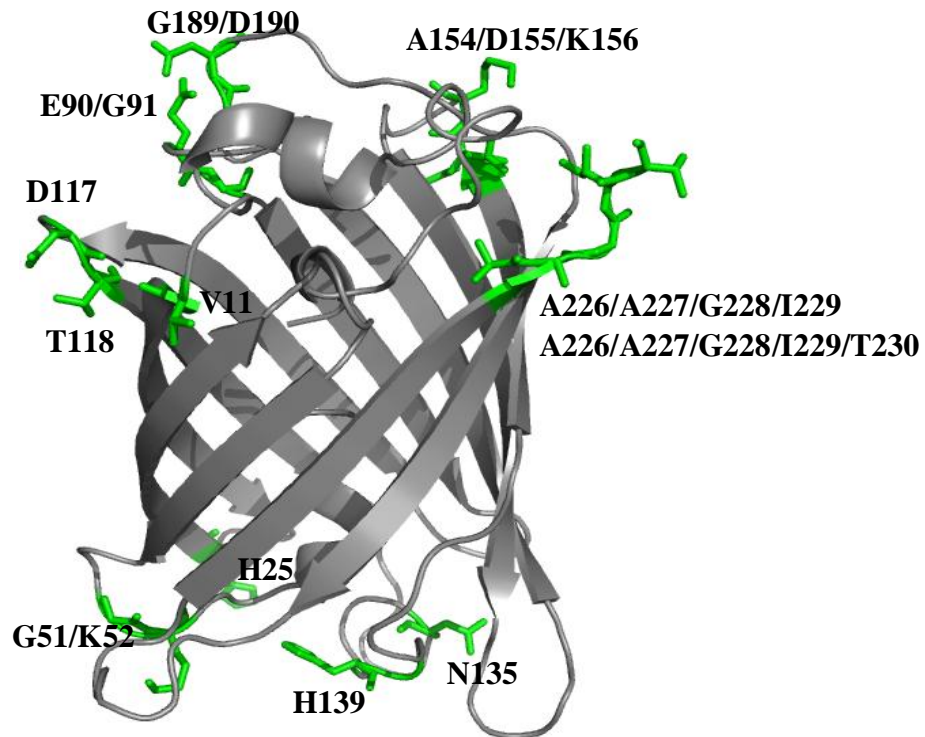
Figure 2.20. Structure of GFP$_{UV}$, showing the "neutral" deletions identified by CDM.

The liquid culture whole-cell fluorescence of the "neutral" mutants was also determined. These mutants showed little fluorescence when expressed at 37 ℃ but their fluorescence was significantly restored when expressed at 23 ℃ (**Table 2.4**). This is consistent with the conclusion from MuDel deletion experiment that fluorescence loss at higher temperature was probably caused by the disrupted folding process and/or chromophore maturation of GFP. Remarkably, by using MuDel deletion mutagenesis and CDM, more than 40 "neutral" amino acid deletions (from separate mutants) have been identified. These results strongly suggest that GFP can be minimized, as opposed to the previous conclusion that GFP is a naturally "size-optimized" protein (*50-52*).

Table 2.4. Liquid culture whole-cell fluorescence of the "neutral" deletion mutants

Deletion mutants	Arbitrary fluorescence
wtGFP$_{UV}$	1.00 ±0.13
V11Δ	0.36 ±0.07
H25Δ	0.09 ±0.00
D117Δ	0.17 ±0.05
T118Δ	0.05 ±0.00
N135Δ	0.10 ±0.01
H139Δ	0.14 ±0.00
G51Δ/K52Δ	0.05 ±0.00
E90Δ/G91Δ	0.07 ±0.00
G189Δ/D190Δ	0.22 ±0.06
A154Δ/D155Δ/K156Δ	0.08 ±0.00
A226Δ/A227Δ/G228Δ/I229Δ	0.17 ±0.01
A226Δ/A227Δ/G228Δ/I229Δ/T230Δ	0.26 ±0.01

Note: Fluorescence was determined in the context of pET28b(+) vector. The fluorescence assay was performed after induction at 23 °C for 18 hours. The fluorescence of mutants has been normalized to that of wtGFP$_{UV}$.

2.2.3 Codon scanning mutagenesis of sfGFP using MuCSM transposon

Codon scanning mutagenesis (CSM) is a non-redundant random mutagenesis method that allows generation of up to three consecutive, random codon substitutions within a protein sequence. This method advances the strategy used in CDM and, instead of generating in-frame deletions, it generates in-frame mutations. As generation of substitutions demands delivery of new codons into the gene of interest, one additional EI-PCR/digestion step is necessary. We further engineered asymmetric transposon, optimized the conditions of EI-PCR and restriction digestion to accommodate the changes in CSM. As a proof of principle, CSM was performed on superfolder GFP (sfGFP) (*72*), a highly stabilized GFP

variant (containing the following mutations as compared to wild-type jellyfish GFP: S30R/Y39N/S65T/F64L/F99S/N105T/Y145F/M153T/V163A/I171V/A206V). We created three mutant libraries, containing one, two or three consecutive, random codon mutations, respectively.

2.2.3.1 Optimization of asymmetric transposons, EI-PCR and digestion

In CDM, primers in EI-PCR may sometimes anneal non-specifically to both ends of transposon due to lack of asymmetric sequences (**Figure 2.17**). Hence, in CSM, transposon end sequence is further engineered to contain five asymmetric nucleotides (**Figure 2.4B**). This new transposon DNA, named as MuCSM, indeed alleviated the non-specific annealing problem as we found later in the EI-PCR. This MuCSM transposon has an insertion efficiency (~1%) similar to that of MuCDM transpson. This MuCSM tranposon again, demonstrates that the sequences flanking the R-sites need not be symmetrical. To verify the randomness of transposon insertions, 23 colonies were randomly chosen from the pTTsfGFP-transposon library and digested with *EcoR* I/*Mly* I (**Figure 2.21**). It is clear from the DNA gel that transposon insertions occurred throughout the entire sfGFP gene, suggesting that the asymmetric terminal sequences do not alter the randomness of transposon insertions. After transposition reaction, the pTTsfGFP-MuCSM transposon library was digested with *Nhe* I and *EcoR* I (which flanks sfGFP gene) to shuffle the sfGFP gene with transposon insertions into an expression vector pTrcHisA. This shuffling process allows isolation of transposon insertions inside sfGFP gene and also sets up the final mutant libraries in the context of an expression vector for direct functional assay.

The cleaned transposon library was used as the template in EI-PCR. The detailed procedure of CSM is shown in **Figure 2.6**. CSM has one additional EI-PCR/digestion step than CDM, because new codons need be delivered into the gene of interest. In the first EI-PCR, the forward primer introduces a *Bsg* I or *Bpm* I restriction site (see **Figure 2.2** for detailed information) that cleaves one or two codons from sfGFP gene (*Bpm* I site is only used in the forward primer in the first EI-PCR of one codon mutation. For some unknown reason, replacing this *Bpm* I site with *Bsg* I site caused PCR failure). The reverse primer supplies up to three new codons and an "anchor" sequence that will serve as the primer binding region for the second EI-PCR. Importantly, this "anchor" sequence remains in the target gene and can not be removed by digestion. Hence, *Bsg* I or *Bpm* I digestion of this PCR product followed by re-ligation yields a library of plasmids each containing a random codon "scar" filled by new codons and an "anchor" sequence. To remove the "anchor" sequence, the library from the first EI-PCR is subjected to a second EI-PCR where two primers anneal to the "anchor" in a "back-to-back" manner and amplify the whole plasmid (**Figure 2.6**). Subsequent *Bsg* I digestion, Klenow treatment and re-ligation generates a mutant library where each clone has up to three consecutive, random codon mutations. The primers used in CSM are summarized in **Table 2.5**.

We chose to deliver NDT degenerate codons (N= adenine (A)/cytosine (C)/guanine (G)/thymine (T); D=A, G, T) to sfGFP. NDT codon includes 12 codons, encoding 12 unique amino acids (Phe, Leu, Ile, Val, Tyr, His, Asn, Asp, Cys, Arg, Ser, Gly). This combination represents a balanced mix of polar and nonpolar, aliphatic and aromatic, and positively charged and negatively charged amino acids, excluding most of the structurally

similar amino acids. Hence, NDT codon can increase the chance of positive hits in screening by reducing the number of variants with identical or similar mutations (*71*).

In comparison with the previous MuDel transposon-based single codon mutagenesis (*49*), one of the most important advances of the current CSM method is that the number of codons to be mutated can be easily adjusted by using different primer pairs. This is achieved by precise placement of the type IIS resctriction sites in primers such that a certain number of nucleotides can be removed from the target gene. Moreover, challenging DNA manipulation steps, such as inter-molecular blunt-end ligation, are excluded from the current method.

Table 2.5. List of primers in EI-PCR of CSM

Name[a]	Sequence[b]	Use
FF1	TACTTTTTCTGGAGCCGTCGGA	First PCR, Fwd remove one codon.
FF2	TTTTTCGTGTGCAGTCGGA	First PCR, Fwd remove two codons.
FR1	AHNAATCAACGACTTTGCGCCGCTAAG	First PCR, Rev introduce one codon.
FR2	AHNAHNAATCAACGACTTTGCGCCGCTAAG	First PCR, Rev introduce two codons.
FR3	AHNAHNAHNAATCAACGACTTTGCGCCGCTAAG	First PCR, Rev introduce three codons.
SF0	TAACGTGCAGTTACAAGTCGTTGATT	Second PCR, Fwd cleave to new codons.
SF1	CATCGTGCAGATGCGCCGCTAAG	Second PCR, Fwd remove one codon.
SR	CTTCGTGCAGTAAATGCGCCGCTAAG	Second PCR, Rev cleave to target gene.

Note: [a] More than one primer pairs may exist for generation of a certain mutation. For example, two codon mutation can be achieved using either FF2/FR2 in the first PCR and SF0/SR in the second PCR or FF1/FR2 in the first PCR and SF1/SR in the second PCR.

[b] *Bsg* I and *Bpm* I sites are underlined. Asymmetric sequences are highlighted in red

2.2.3.2 Overall efficiency of CSM

To assess the quality of the library, 50 colonies were randomly chosen from each library and sequenced (see **Appendix V** for sequencing results). The frequency of clones containing seamless, in-frame mutations in single, di- and tri-NDT library was 47.6%, 65.0% and 63.6 %, respectively (**Table 2.6**). This suggests that CSM has sufficient mutagenesis efficiency for constructing large-scale libraries. The incorrect mutations generally arose from incomplete *Bsg* I/*Bpm* I digestion (resulting in uncut primer sequence) or overdigestion of Klenow (resulting in one additional nucleotide deletion).

Table 2.6. Mutagenesis Efficiency of CSM

Library	One	Two	Three
Total sequenced	50	50	50
Incorrect	21	20	28
Correct	29	30	22
% Correct	58.0	60.0	44.0

To analyze if these mutations were distributed uniformly along the sfGFP, the cumulative fraction of correct mutations were plotted against amino acid positions of sfGFP (**Figure 2.21**). The distribution of mutations in these three libraries shows similar trends. While mutations occur in all regions of the sfGFP protein, there is a mild preference for the N-terminal mutations, likely the result of transposon target sequence bias. Next was examined the amino acid composition of the introduced codons. The degenerate sequence "NDT" (12 codons/12 amino acids) is a balanced mix of codons (*71*). Ideally, these 12 codons can be delivered into a protein sequence with equal probability. In this study, however, we observed a bias of the introduced codons (**Figure 2.22**). The 81 correct mutants introduce 155 new codons. Thus, the expected average

number of each of the 12 amino acids is 13. The mutants in the current study are rich in

Arg and His mutations and short of Ile and Phe mutations. This codon bias most likely

arises from the synthesis of primers using automated mixing for degenerate

oligonucleotides.
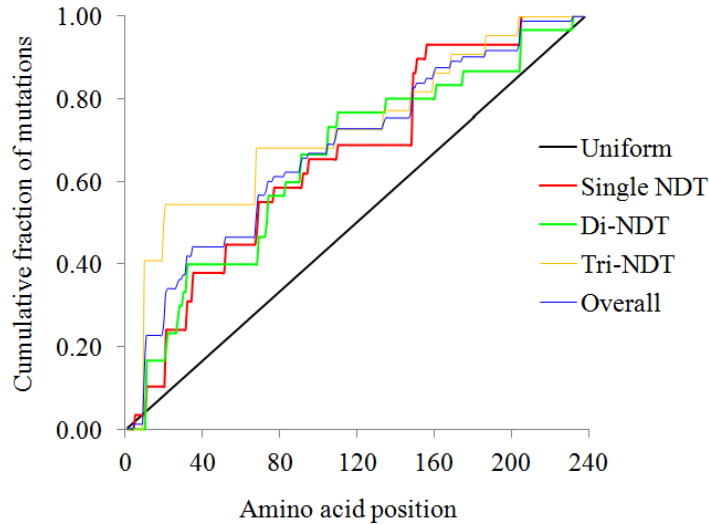
Figure 2.21. Distribution of mutations in sfGFP, based on 81 correct mutations. For consecutive mutations, positions of the first mutations are chosen for plotting the curve.
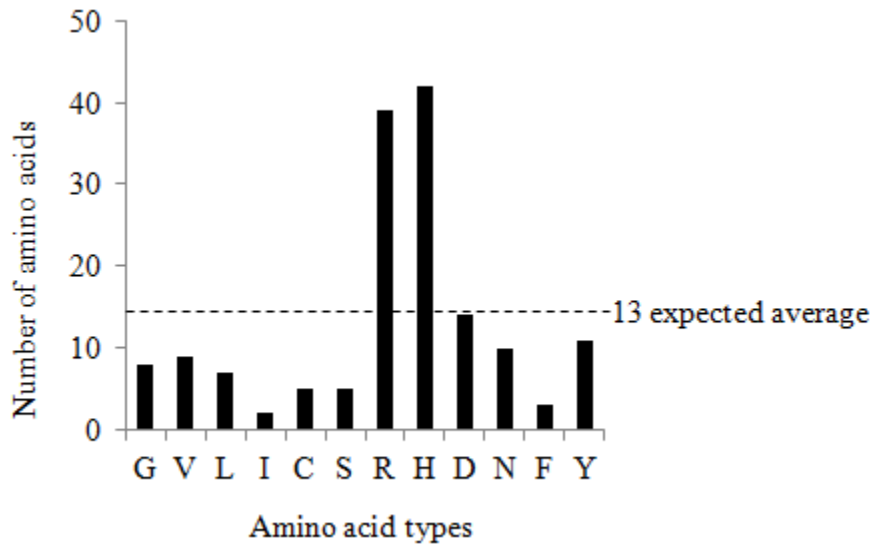
Figure 2.22. Introduced codon types in sfGFP, based on 81 correct mutants (introducing 155 new codons). The amino acids encoded by NDT codons are Gly (GGT), Val (GTT), Leu (CTT), Ile (ATT), Cys (TGT), Ser (AGT), Arg (CGT), His (CAT), Asp (GAT), Asn (AAT), Phe (TTT) and Tyr (TAT).

2.2.3.3 Identification of spectra-shifted mutants of sfGFP

As a proof of principle, CSM was performed on sfGFP to search for variants that exhibited altered spectral properties. Several spectra-shifted mutants were successfully identified by screening only 500 colonies from each library (**Figure 2.23**). The excitation peaks of these mutants shifted up to 15 nm to blue and up to 25 nm to red as compared to the 490 nm peak in original sfGFP (**Figure 2.23A**). For the emission peaks, up to 6 nm blue shift and up to 12 nm red shift were observed (**Figure 2.23B**). Interestingly, among those identified spectral mutations is a red-shift mutation S203Y, which has been reported in yellow fluorescent protein (YFP) (*5*). This tyrosine mutation at position 203 introduces an aromatic ring that overlays with the phenol ring in chromophore and thus stabilizes the excited π elections of chromophore.
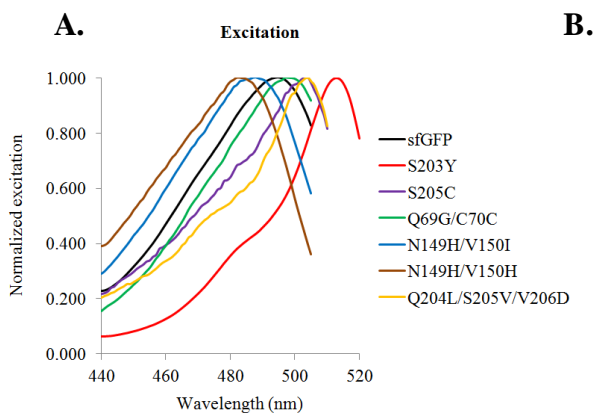
Figure 2.23. Normalized excitation and emission of spectra-shifted mutants. For excitation spectra (**A**), the emission wavelengths are 530 nm for mutant S203Y, 520 nm for mutants S205C and Q204L/S205V/V206D, and 515 nm for all other mutants. For emission spectra (**B**), the excitation wavelengths are 505 nm for S203Y and 490 nm for all other mutants. Note that all the mutations are referred to as in the context of sfGFP.

More interestingly, a S205D mutant was found to be 7-fold more fluorescent than the original sfGFP when excited under UV light but non-fluorescent when excited under visible light (**Figure 2.24**). Characterization of the excitation spectra showed that the S205D mutant bears a single peak at 395 nm as compared to the original sfGFP that has a minor peak at 395 nm and a major peak at 490 nm. The 395 nm and 490 nm excitation peaks are known to arise from the different ionization states of the chromophore (*80*). The neutral/protonated chromophore is responsible for the 395 nm peak. Upon light irradiation, the chromophore becomes deprotonated and can absorb light in the visible region (*80*). The enhanced fluorescence under UV light and the elimination of fluorescence under visible light in S205D mutant are likely due to the change of ionization states of the chromophore. The S205 residue in the sfGFP structure (PDB entry 2B3P) (*72*) projects toward the chromophore and the S205D mutation may alter the pK_a of this residue, locking it to an "all-protonated" state.
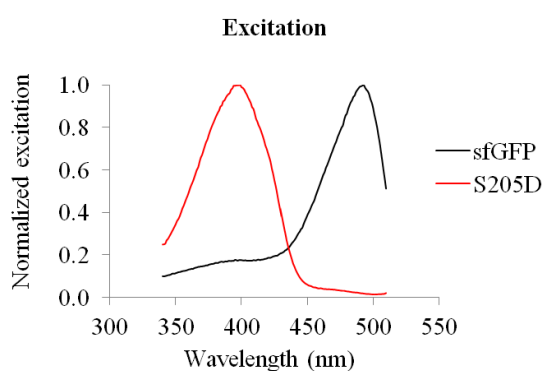
Figure 2.24. Normalized excitation and emission of S205D mutant. For excitation spectra (**A**), the emission wavelengths are 510 nm for both original sfGFP and S205 mutants. For emission spectra (**B**), excitation at either 395 nm or 480 nm was used.

In the current study, three factors play important roles in the successful identification of these spectra-shifted mutants. First, the highly stabilized template protein sfGFP (*72*) provides extra stability space for beneficial mutations to appear (*81*). Second, the NDT degenerate codon (12 codons/12 aa) increases the chance of a positive hit in screening by reducing codon redundancy (as compared to the conventional NNK codon, 32 codons/20 amino acids). Third and most importantly, our approach allows the direct construction of mutant libraries with high diversity on the amino acid level, bypassing the problems of codon degeneracy and rare chance of consecutive nucleotide mutations in traditional random mutagenesis methods that usually compromise the diversity of mutant libraries.

2.3 Conclusion and Future Applications

2.3.1 Generation of a size-minimized GFP

MuDel deletion mutagenesis and CDM experiments have determined more than 40 "neutral" deletions in GFP, suggesting that GFP is not a "naturally minimized" protein. Furthermore, the fluorescence loss caused by these "neutral" deletions can be restored by incorporation of folding-mutations. These results have pointed to the way to generate a size-minimized GFP: stepwise incorporation of "neutral" deletions followed by folding rescue. However, there are several general considerations (see **Figure 2.25** for proposed procedure). First, it is desirable to start with a stabilized template, such as sfGFP, which already has several folding mutations incorporated. Second, terminal deletions can be incorporated before internal deletions as GFP is known to tolerate terminal deletions (*51*). Third, folding rescue can be achieved by either adding a known folding mutation or globally optimizing the folding state of the truncated GFP by random mutagenesis. Last,

the final mini-GFP may require further optimization for oligomerization state and other properties before application.
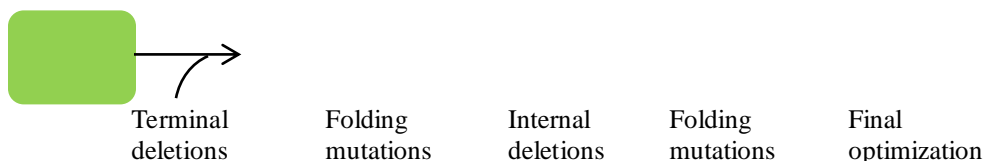
Figure 2.25. Proposed procedure for generation of a size-minimized GFP. The green and grey colors indicate fluorescent and non-fluorescent constructs, respectively.

2.3.2 Application of CDM

CDM is advantageous over many other random deletion mutagenesis methods because it i) generates random, in-frame codon deletions, ii) allows up to five consecutive codon deletions per round of mutagenesis and iii) is technically efficient (no challenging DNA manipulations). It can be used to engineer not only GFP but also many other proteins which have not been extensively explored for deletion engineering. Perhaps one of its most interesting applications is to study how length variations affect protein functions. Antibodies, for example, are known to have length variations in the complementarity determining regions (CDRs) and are thus good target for CDM. Another good reason of applying CDM on the CDRs of antibodies lies in that CDRs usually contain flexible loop structures which are responsible for antigen-binding specificity. As compared with β sheets, these loops may have less structural constraints and therefore, be more tolerant to random deletions. Importantly, different primer pairs can be combined in EI-PCR to

create a "mixed" library of mutants carrying one to five codon deletions. Outlined below

is a proposed experiment to engineer antibodies using CDM (**Figure 2.26**).
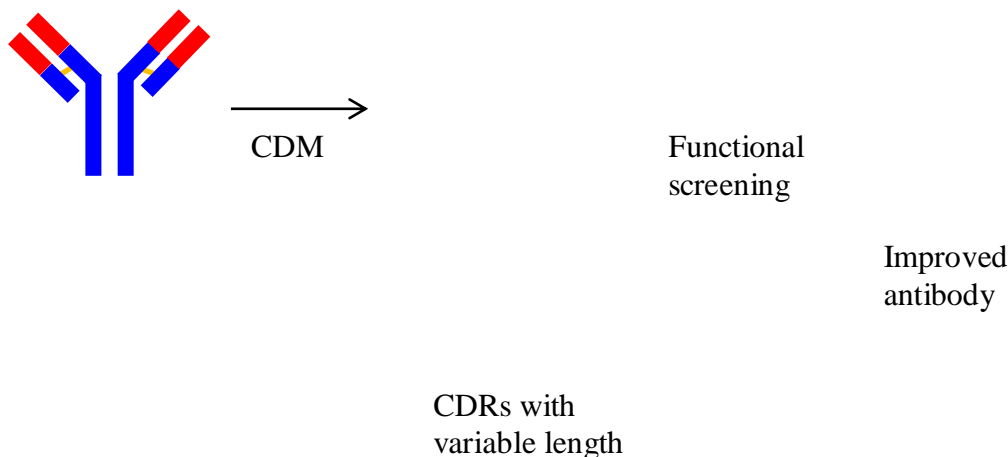
Figure 2.26. Proposed experiments of antibody engineering using CDM. The red color represents variable region while the blue color stands for constant region.

2.3.3 Application of asymmetric transposon-based CSM

CSM is a very powerful random mutagenesis method. It generates codon mutations and

the number of codons to be mutated can also be easily adjusted using different primer

pairs in the EI-PCR. These two features enable CSM to achieve complicated tasks.

Shown below are several potential applications of CSM.

2.3.3.1 Single, di- and tri- alanine scanning mutagenesis

Traditional alanine scanning mutagenesis (*82*) requires synthesis of different primer pairs

for alanine mutations at different positions. This method is basically to perform site-

directed mutagenesis on several sites of choice. Although this method is generally

successful, it requires predicting a specific region for scanning. The information beyond

the selected region is usually overlooked. Alanine scanning of an entire protein sequence

58

may be very difficult as it requires tremendous effort to synthesize the primers and generate all the mutants in individual reactions. CSM overcomes this problem as it allows full-length alanine scanning and can generate all mutants in a single reaction (**Figure 2.27**). With a proper screening/selection method, alanine mutants with interesting property changes can be identified rapidly. Furthermore, CSM can also generate two or three consecutive alanine mutations at once. This allows one to perform regional alanine scanning at random positions in a protein sequence.
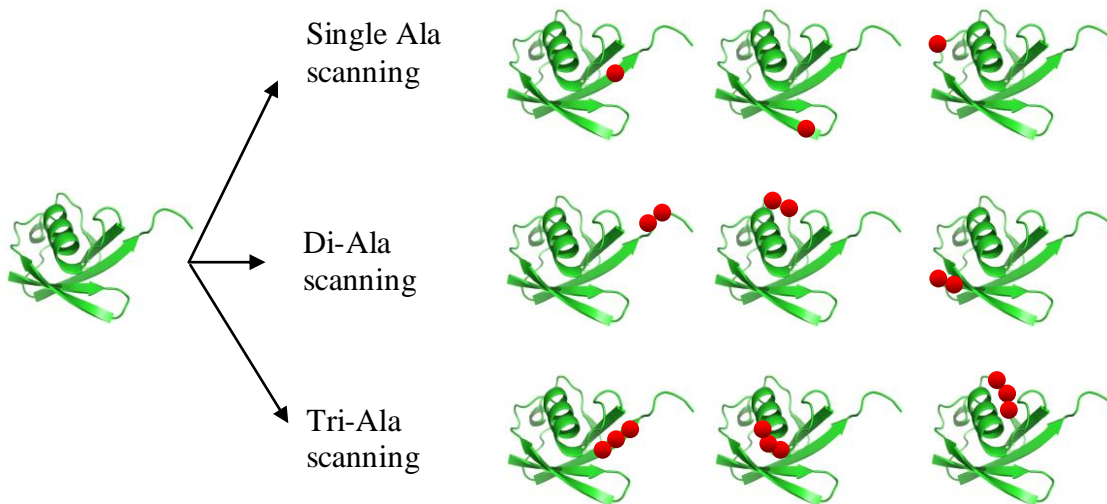
Figure 2.27. Alanine scanning mutagenesis using CSM. The red spheres represent alanine mutations.

2.3.3.2 Semi-rational twin-cysteine scanning mutagenesis

CSM can be also used to scan reactive amino acids throughout a protein sequence. One of such applications is twin-cysteine scanning mutagenesis to introduce disulfide bonds into proteins. Disulfide bonds play important roles in protein folding and stability, especially to those proteins secreted to the extracellular medium (*83*). Rational approach to stabilize protein with disulfide bonds may be very difficult as the two cysteine residues involving

in the disulfide bond need to be spatially close to each other with their side chains oriented properly. Furthermore, appropriate micro-environment (oxidizing conditions) is also an important consideration. CSM bypasses these limitations as it can randomly distribute cysteine mutations throughout the entire protein sequence.
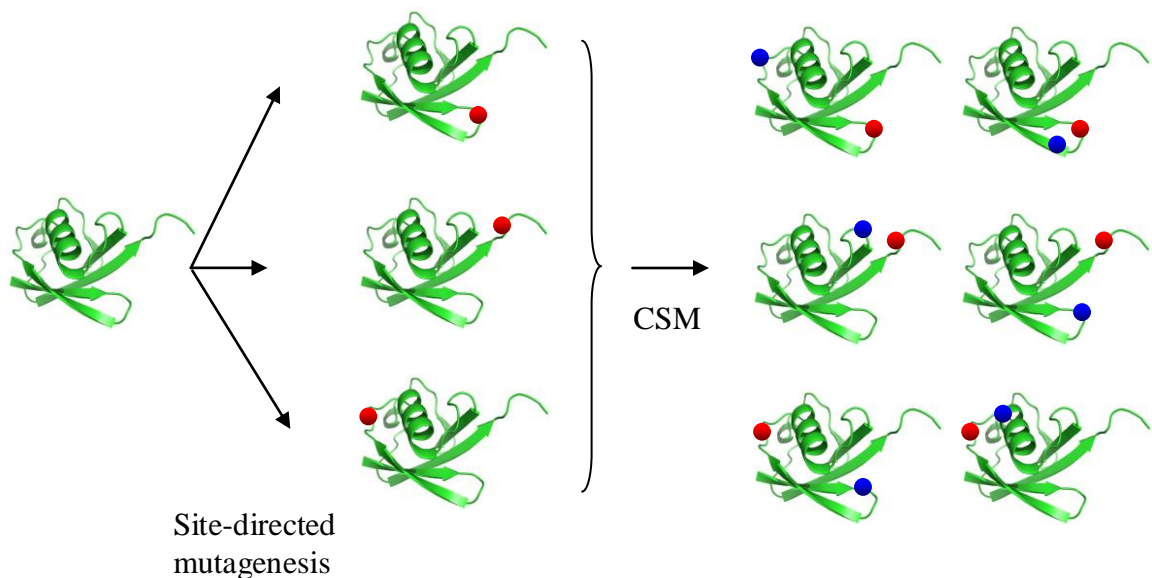
Figure 2.28. Semi-rational twin-cysteine scanning mutagenesis using CSM. The red spheres represent rationally introduced cysteine mutations by site-directed mutagenesis while the blue sphere represent randomly introduced cysteine mutations by CSM.

Perhaps the easiest and most efficient way to perform twin-cysteine scanning is a semi-rational approach with the combinatorial use of site-directed mutagenesis and CSM (**Figure 2.28**). First, several variants of a target protein can be generated using site-directed mutagenesis (rational design) to introduce the first cysteine at several preferred sites. This pool of variants can then be used as the starting template in CSM to introduce the second cysteine. This semi-rational approach has two advantages *versus* the irrational approach of merely performing two rounds of CSM. First, it is partially guided by our

knowledge and thus may increase the chance of disulfide bond formation and the chance of having a beneficial disulfide bond. Second, this semi-rational approach can create a library with more mutants containing desired mutations. CSM has a mutagenesis efficiency of ~50% (**Table 2.6**), meaning that two rounds of CSM will give a library where only ~25% of the mutants contain desired mutations. This is an important consideration especially when high-throughput screening methods are not available.

2.3.3.3 Generation of an "ultimate" mutant library

The most interesting feature of CSM is that the number of mutated codons can easily be adjusted by using different primer pairs. Although we reported the generation of up to three codon mutation, four or five codon mutations are also possible if proper primers can be designed. Significantly, these primers can be used separately as shown above or combined as a mix. For example, when primers FF1, FF2, FR1 and FR2 (**Table 2.5**) are used together in the first EI-PCR, four possible combinations can occur (**Figure 2.28A**). The combination of primer pairs allows generation of an ultimately diversified library that contains random codon deletions, insertions and substitutions (**Figure 2.28B**). To the best of our knowledge, such versatile libraries can not be easily constructed by any existing mutagenesis method. These libraries will largely expand our knowledge as it will provide the direct comparison of the effects of codon deletions, insertions and substitutions on protein functions.

A.

Choice of forward primer	————	Choice of reverse primer	Final library

One codon deletion
one codon substitution

FF1
(remove one codon)

FR1
(deliver one codon)

One codon substitution

FF2
(remove two codons)

FR2
(deliver two codons)

Two codon substitution

One codon substitution
one codon insertion

B.

substitution
deletion
insertion
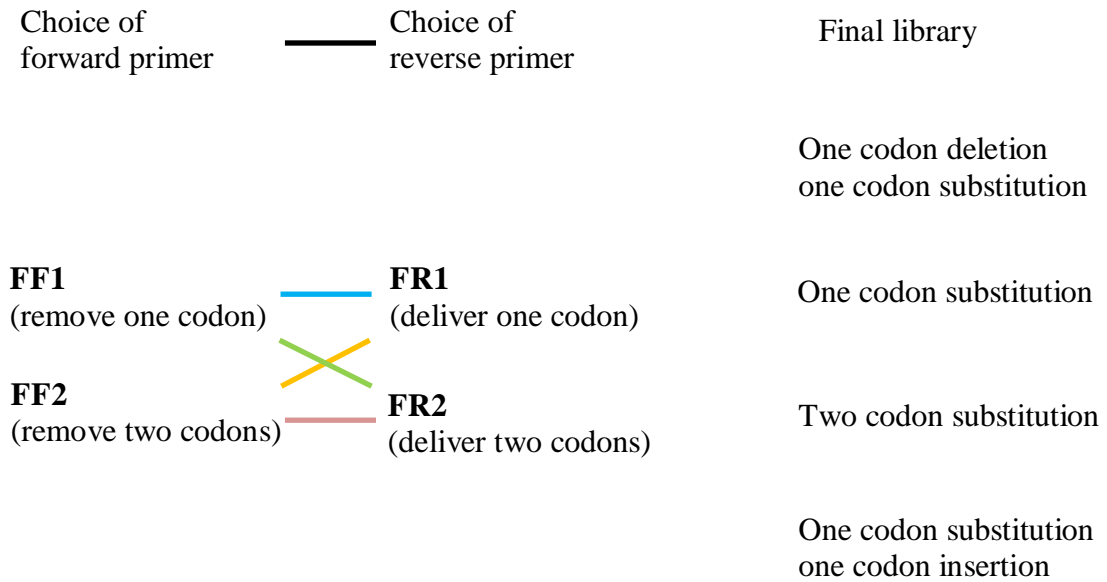
Figure 2.29. Generation of libraries containing random deletions, insertions and substitutions using CSM. **A**) Examples of libraries resulting from combined primer pairs. **B**) Generation of random deletions, insertions and substitutions in a single library.

2.4 Materials and Methods

2.4.1 Random deletion mutagenesis of GFP$_{UV}$ using MuDel transposon

2.4.1.1 Construction of deletion library of GFP$_{UV}$

pGFP$_{UV}$ vector (Clontech, Mountain View, CA) was used as the transposon-targeting plasmid in transposition reaction. All of the four *Mly* I sites (249, 2334, 2836 and 3322) were removed from pGFP$_{UV}$ by site-directed mutagenesis (see **Appendix III** for primer

sequences). The GFP$_{UV}$ protein expressed from this vector contains a 24 amino acid N-terminal fusion tag.

The MuDel transposon bearing chloramphenicol selection marker was PCR amplified as previously described (*25*). The PCR product was cloned into pCR2.1 TOPO vector (Invitrogen, Carlsbad, CA). Transposon DNA was released from this vector by *Bgl* II digestion and then gel-purified. The transposition reaction contained 570 ng of pGFPuv vector, 140 ng of transposon DNA, 1 unit of HyperMu MuA transposase (Epicentre Biotechnologies, Madison, WI), 50 mM Tris-acetate (pH 7.5), 150 mM potassium acetate, 10 mM magnesium acetate and 4 mM spermidine in a 20 μL solution. The reaction was kept at 30 ℃ for 4 h and then stopped by addition of SDS to 0.1% followed by heat-inactivation at 70 ℃ for 10 min. The reaction product (1 μL) was transformed into 50 μL of chemically competent GeneHogs *Escherichia coli* cells. The cells were recovered in SOC at 37 ℃ for 1 h and plated on LB agar containing 100 μg/mL ampicillin (Amp100) and 35 μg/mL chloramphenicol (Chl35) to select for the pGFP$_{UV}$ vector with transposon insertions. In total, 38,000 colonies were collected to build the pGFP$_{UV}$-MuDel library. These colonies were sufficient to ensure the full coverage of library diversity (*78*). The library DNA was extracted and digested by *EcoR* I/*Kpn* I (Fermentas, Glen Burnie, MD). As transposon could insert both in and outside of the GFP$_{UV}$ gene, *EcoR* I/*Kpn* I digestion generated four DNA fragments: GFP$_{UV}$, GFP$_{UV}$+transposon, vector backbone, vector backbone+transposon (**Figure 2.7B**). The 2.0 kb band (GFP$_{UV}$ + transposon) and and 2.6 kb band (vector backbone) were re-ligated to isolate those transposon insertions in the GFP$_{UV}$ gene. The ligation product was transformed into chemically competent GeneHogs *E.coli* cells. The plasmid DNA of the

cleaned transposon library was extracted and then digested by *Mly* I (Fermentas) as follows: 1.5 μg DNA, 10 unit of *Mly* I (New England Biolabs, Ipswich, MA), 100 μg/mL of BSA, 20 mM Tris-acetate (pH 7.9), 50 mM potassium acetate, 10 mM magnesium acetate and 1 mM DTT in a 100 μL reaction. The reaction was kept at 37 ℃ for 3 h and then gel-purified to isolate the linearized pGFP$_{UV}$ plasmid with random triplet nucleotide deletion. The gel-purified product was subjected to a blunt-end intramolecular ligation in a 20 μL reaction containing 300 ng DNA, 400 cohesive end units of T4 DNA ligase (NEB), 50 mM Tris-HCl (pH7.5), 10 mM MgCl$_2$, 10 mM DTT and 0.5 mM ATP. The ligation was performed at 16 ℃ for 16 h. This product was transformed into chemically competent GeneHogs *E. coli* cells. In total, 10,000 transformants were collected as the deletion library.

2.4.1.2 Screening of deletion mutants

The plasmid DNA of the deletion library was transformed into GeneHogs *E. coli* cells and then plated on LB agar (Amp100) at a density of ~500 colonies/plate. These plates were grown at either 37 ℃ for 14 h or 23 ℃ for 30 h. Screening of the fluorescent and non-fluorescent mutants was accomplished using a TC312E UV transilluminator (310 nm wavelength) (Spectronics, Westbury, New York) by visual inspection. In total, 40 fluorescent colonies were picked from the plates grown at either 37 ℃ or 23 ℃ and 24 non-fluorescent colonies were picked from the 23 ℃ plates. All the picked clones were verified by digestion and then sequenced at DNA Analysis Facility at Yale University (New Haven, CT).

2.4.1.3 Liquid culture whole-cell fluorescence

Plasmid DNAs of wtGFP$_{UV}$ and deletion mutants were transformed into GeneHogs *E. coli* cells and plated on LB agar (Amp100) at a density of ~200 colonies/plate. Single colonies with a diameter of ~0.5 mm were inoculated into 2 mL of liquid LB media (Amp100). Three replicates of each sample were grown at 37 °C, 30 °C and 23 °C for 14, 18 and 25 h, respectively. The *E. coli* cells of each sample were harvested by centrifugation, washed with 500 μL of TNG buffer (100 mM Tris-HCl, pH 7.5, 150 mM NaCl and 10% glycerol) (*72*) and then resuspended in 100 μL TNG buffer. A fraction of the resuspended cell solution was adjusted to an OD$_{600}$ of 0.150 $\pm$ 0.003 for fluorescence assay. The rest cell solution was stored at -20 °C for further experiments. Whole-cell fluorescence assay was performed using a model F4500 fluorescence spectrophotometer (Hitachi, Tokyo, Japan) with 397 nm for excitation wavelength and 509 nm for the emission wavelength. The background of empty GeneHogs cells was subtracted from each reading. The fluorescence of each deletion mutant was normalized to that of wtGFP$_{UV}$ grown under the same temperatures.

2.4.1.4 Expression level and fraction soluble analyzed by SDS-PAGE

The resuspended cells of wtGFP$_{UV}$ and deletion mutants were adjusted to an OD$_{600}$ of 0.100 with TNG buffer. For each sample, 300 μL of such cell solution was forced to two sequences of 10 pulse sonication with an interval of 3 min using a model 450 sonicator (Branson Ultrasonics, Danbury, CT) equipped with a 1/2 inch horn and a 1/8 inch tip with 50% power output and 50% duty time. The gel loading sample was prepared by mixing 15 μL sonicant (crude cell lysis) with 15 μg BSA protein (Sigma, St. Louis, MO) which served as an internal standard to quantify the GFP expression. The protein samples

were resolved in 12% acrylamide SDS-PAGE gels and analyzed by Image J (http://rsbweb.nih.gov/ij/). The overall GFP expression was quantified based on the density ratio of GFP band and BSA band on the protein gels as previously described (*72*) (**Figure 2.9A**). To calculate the soluble fraction of GFP, 15 μL sonicant was clarified by centrifugation and the supernatant was mixed with 15 μg BSA protein. The same analysis was performed as above to obtain the soluble GFP expression. The fraction soluble of each sample was the ratio of soluble GFP expression and overall GFP expression. Three individual replicates were performed for each protein sample.

2.4.1.5 Protein expression and purification

wtGFP$_{UV}$ and the five mutants with internal deletions (C48Δ, P75Δ, P75/D76→H, E172Δ, S175/V176→F) were cloned into pET28b(+) vector (EMD Chemicals, San Diego, CA) using primers GFP$_{UV}$_FWD and GFP$_{UV}$_REV (**Appendix III**). The recombinant DNAs were transformed into (DE3) *E. coli* cells (Stratagene, La Jolla, CA). A single colony was inoculated into 100 mL 2×YT media, induced with 1 mM of IPTG at OD$_{600}$ of 0.6 and then grown at 20 ℃ for 12 h. The cells were centrifuged, washed twice with binding buffer (100 mM HEPES, pH 7.5, 10 mM imidazole) and then resuspended in 30 mL binding buffer supplemented with 1 mM PMSF. The resuspended cells were sonicated and then centrifuged at 12,000 for 15 min to isolate the soluble fractions. The soluble proteins were run through Promega HisLink resins (Promega, Madison, WI), washed with wash buffer (100 mM HEPES, pH 7.5, 50 mM imidazole) and then eluted with elution buffer (100 mM HEPES, pH 7.5, 500 mM imidazole). The eluted proteins were dialyzed into two successive 200-fold TNG buffer. Each sample was purified to >90%

homogenity and the concentration was determined using Pierce BCA Protein Assay kit (Thermo Fisher Scientific, Rockford, IL).

2.4.1.6 Characterization of purified deletion mutants

The excitation and emission spectra were scanned using 6 μg/mL of purified proteins in TNG buffer. The emission wavelength for the excitation scan was 509 nm. The emission scan was recorded using excitation wavelengths of both 397 nm and 495 nm.

The concentration of proteins containing matured chromophore was determined using the "base-denatured" method (75). Briefly, 7 μM of purified proteins were denatured in 0.1 M NaOH (pH 13) for 5 min at 25 ℃. Absorption spectra from 300 nm to 600 nm were scanned using Cary 50 Bio UV-visible spectrophotometer (Varian, Palo Alto, CA). The concentrations of matured proteins were calculated using Beer-Lambert Law:

$$A = c \, l \, \varepsilon$$

where A is the experimentally determined absorbance at 447 nm, c is the concentration of matured protein that is to be obtained, l is the path length of the cuvette and ε is pre-determined extinction coefficient of matured GFP under base-denatured condition (44,000 $M^{-1}cm^{-1}$) (75). The efficiency of chromophore maturation ($E_{chromphore}$) or fraction of matured protein was calculated using the following equation:

$$E_{chromphore} = c_{matured} / c_{overall}$$

where $c_{matured}$ is the concentration of matured protein pre-determined using "base-denatured" method as above and $c_{overall}$ equals 7 μM in this experiment. Three measurements of the same protein preparation were performed with an estimated instrumental error of 2%.

The ε_{397} and ε_{495} of wtGFP$_{UV}$ and mutants were determined using Beer-Lambert Law:

$$A = c \, l \, \varepsilon$$

where A was the experimentally determined absorbance at corresponding wavelengths (397 nm or 495 nm), c is the concentration of matured GFP (as the efficiency of chromophore maturation is pre-determined, the concentration of matured protein can be calculated if the overall protein concentration is known) and l is the path length of the cuvette.

To determine the quantum yield of each mutant, protein samples were adjusted to an absorbance of 0.100 at 397 nm and then diluted 100 fold with water. The emission from 450 nm to 600 nm using an excitation wavelength of 397 nm were integrated and then compared to the integrated emission of wtGPF$_{UV}$, for which a quantum yield of 0.79 has been reported (*76*).

2.4.1.7 Fluorescence rescue by folding-enhancing mutations F64L and S30R

The central mutation F64L (*77*) and distant mutation S30R (*72*) were introduced into wtGFP$_{UV}$ and deletion mutants by site-directed mutagenesis (see **Appendix III** for the primer sequences). The whole-cell fluorescence of F64L- or S30R-rescued mutants at 37 ℃ was assayed as above. The fluorescence of deletion mutants with F64L or S30R mutations was normalized to that of GFP$_{UV}$-F64L or GFP$_{UV}$-S30R, respectively. Mutants with double internal deletions were constructed using site-directed mutagenesis (see **Appendix III** for the primer sequences). Mutations F64L and S30R were then introduced into the double deletion mutants. The rescued whole-cell fluorescence of double deletion mutants at 23 ℃ was assayed and normalized to the fluorescence of GFP$_{UV}$-F64L or GFP$_{UV}$-S30R, respectively.

2.4.2 Codon deletion mutagenesis of GFP_UV using MuCDM transposon

2.4.2.1 Construction of MuCDM transposon and transposon-targeting plasmid

The asymmetric Mu transposon for codon deletion mutagenesis (MuCDM) was constructed by amplifying the described *Mly* I TAG linker (*49*) with the primers 5'-AAA<u>AGATCT</u>TTAGCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGC GAAAACGGAAGGTATTCGCAATAATCTTAAT-3' and 5'-AAA<u>AGATCT</u>CCGACGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATG CGAAAATGAGTAAACGCGGTCTGACTC-3' (*Bgl* II site underlined, the four nucleotide difference highlighted in red and MuA transposase recognition sites in shadow). The PCR was performed using *Taq* DNA polymerase and the following PCR cycling conditions: 95 ℃ for 3 min, 11 cycles of 95 ℃ for 30 s, 66 ℃ (-1 ℃/cycle) for 30 s and 72 ℃ for 2 min, followed by 20 cycles of 95 ℃ for 30 s, 66 ℃ for 30 s and 72 ℃ for 2 min and final extension at 72 ℃ for 10 min. The PCR product was first cloned into pCR2.1 TOPO vector (Invitrogen) using TA-end cloning (easy cloning process). This PCR product was then transferred into the *Hind* III and *Xba* I site of pUC18 vector for long-term storage purpose. The MuCDM transposon was released from pUC18 vector by *Bgl* II digestion and then gel-purified using the QIAquick purification (Qiagen, Valencia, CA) prior to transposition reaction.

The plasmid pTTGFP_UV was constructed by cloning the GFP_UV gene (Clontech) into the previously described intein-targeting plasmid pIT (*49*). GFP_UV was fused downstream to the Tat signal peptide and vacuolar membrane ATPase (VMA) intein N-terminal domain (VMAI-N) (see **Appendix IV** for the full sequence). The GFP_UV insert sequence was confirmed by sequencing at Yale.

2.4.2.2 Transposition reaction

The transposition reaction was performed in a 10 μL reaction containing 420 ng of pTTGFP$_{UV}$, 125 ng of gel-purified MuCDM transposon, 1 unit of HyperMu MuA transposase (Epicentre), 50 mM Tris-acetate (pH 7.5), 150 mM potassium acetate, 10 mM magnesium acetate and 4 mM spermidine. The molar ratio of transposon DNA to target plasmid was approximately 1.5. The transposition reaction was incubated at 30 ℃ for 4 h and then stopped by 0.1% SDS followed by heat-inactivation at 70 ℃ for 10 min. The reaction was then cooled on ice and 2 μL of the product was transformed into 200 μlL electro-competent GeneHogs *E. coli* cells. The transformants were plated on LB agar supplemented with 50 μg/mL kanamycin (Kan50) and 40 μg/mL ampicillin (Amp40) and grown at 30 ℃ overnight. In total, 3,000 colonies were resulted and collected as the pTTGFP$_{UV}$-MuCDM library. Twenty random colonies were digested with *EcoR* I and FastDigest *Mly* I (Fermentas) to verify the randomness of transposon insertions.

2.4.2.3 Enzymatic inverse PCR (EI-PCR)

In this study, five deletion libraries were constructed, containing one to five codon deletions, respectively. The pTTGFP$_{UV}$-MuCDM library was used as the template in the EI-PCR. As the starting and ending position of the PCR product were randomly distributed in the GFP$_{UV}$ gene (depending on the position of transposon insertions), the "head" of a PCR product could overlap with the "tail" of another product, resulting in "side product" with extended size (see **Figure 2.16** for condition test of template concentrations). To avoid this type of homologous recombination of the PCR products, reduced template concentration and cycling numbers were used (*79, 84*). A variety of template concentrations and cycling numbers were tested and the optimum values were

determined. For each library, PCR was accomplished in four tubes of 100 μL reaction mixture containing 12.5 pg/μL template DNA, 0.5 μM of corresponding forward and reverse primers (see below), 0.2 mM dNTPs, 0.005 U/μL of Phusion DNA polymerase (NEB), 1 × Phusion HF buffer (NEB, providing 1.5 mM $MgCl_2$) and additional 2 mM $MgSO_4$. The primers for each PCR were carefully designed to avoid non-specific annealing. The optimized primer pairs were FWD8N/REV0N, FWD8N/REV3N, FWD8N/REV6N, FWD8N/REV9N and FWD11N/REV9N for one, two, three, four and five codon deletion libraries, respectively (**Table 2.2**). All five PCRs were performed in a standardized cycling condition: initial denaturation at 98 ℃ for 2 min, 18 cycles of 98 ℃ for 10 s, 59 ℃ for 30 s and 72 ℃ for 2.5 min and final extension at 72 ℃ for 10 min. The 400 μL PCR product of each library was pooled, purified using PCR purification kit (Qiagen) and eluted with 30 μL of 10 mM Tris-HCl buffer, pH 8.5.

2.4.2.4 *Bsg* I digestion and Klenow treatment

The purified PCR product was digested with *Bsg* I in a 50 μL reaction containing 250~400 ng DNA, 0.12 U/μL *Bsg* I (NEB), 80 μM SAM, 20 mM Tris-acetate, 50 mM potassium acetate, 10 mM magnesium acetate and 1 mM DTT. The reaction was incubated at 37 ℃ for 4 h in a PCR thermocycler. *Bsg* I was then heat-inactivated at 65 ℃ for 20 min. The reaction mixture was cooled to room temperature. dNTPs (Fermentas) were added into the reaction to a final concentration of 33 μM. Thereafter, 0.5 U Klenow (large fragment of DNA polymerase I) (NEB) was added to each 50 μL *Bsg* I digestion. The blunting reaction was incubated at 25 ℃ for exactly 15 min using a PCR thermocycler. The reaction was stopped instantly after incubation by addition of 1.1 μL of 0.5 M EDTA, pH 8.0, followed by heat-inactivation at 75 ℃ for 20 min. The 50 μL

reaction was then incubated with 4 U of *Dpn* I (NEB) at 37 ℃ for 80 min to destroy the template plasmid DNA (methylated). The digested DNA was gel-purified using QIAquick purification (Qiagen). This procedure typically resulted in 3~5 ng/μL gel-purified product.

2.4.2.5 Intra-molecular ligation and transformation

The blunt-end intra-molecular ligation was performed as follows: 2.5 ng/μL of gel-purified DNA, 0.1 Weiss unit/μL of T4 DNA ligase (Fermentas), 40 mM Tris-HCl (pH 7.8 at 25 ℃), 10 mM MgCl₂, 10 mM DTT and 0.5 mM ATP in a 10 μL reaction. The ligation was incubated at 16 ℃ for 8 h and then heat-inactivated at 70 ℃ for 6 min. The ligation product (1 μL) was transformed into 50 μL of electro-competent GeneHogs cells and plated on LB agar (Kan50) at 37 ℃ overnight. This yielded approximately 1,000 colonies/ μL ligation product. More than 5,000 colonies were collected for each library in this study. In total, 24 random colonies were picked from from each library and sequenced for assessing the library quality.

2.4.2.6 Screening of deletion mutants with various degree of fluorescence

The five deletion libraries were cloned into the *Hind* III/ *Nhe* I site of pET28b(+) vector. The ligation products were initially transformed into GeneHogs cells and more than 2,000 colonies were collected from each library. The plasmid DNA of each library was extracted from GeneHogs cells and then transformed into BL21 (DE3) cells (Stratagene Inc., La Jolla, CA) for analysis of fluorescence. The transformants were plated on LB agar (Kan50) and 1 mM IPTG and grown at 19 ℃ for 30 h. The screening was accomplished using a TC312E UV transilluminator (310 nm light) (Spectronics) by

visual inspection. From each library, 8 fluorescent and 8 non-fluorescent colonies were randomly picked for sequencing analysis.

2.4.2.7 Determination of liquid culture whole-cell fluorescence

Freshly transformed colonies of wtGFP$_{UV}$ and deletion mutants in the context of pET28b(+) vector were inoculated into 1 mL of liquid LB (Kan50) and grown overnight at 37 °C. The overnight culture (50 μL) was inoculated into 1 mL LB (Kan50) and grown at 37 °C for 4 hours. The culture was then induced with 1 mM IPTG and expressed at 20 °C for 18 h. The cells were harvested by centrifugation, washed twice with 500 μL of TNG buffer (*72*) and then resuspended in 50 μL TNG buffer. Whole-cell fluorescence assay was performed under an OD$_{600}$ of 0.150 ± 0.003 using a model F4500 fluorescence spectrophotometer (Hitachi). The excitation wavelength was fixed at 395 nm and the fluorescence intensity at 509 nm was recorded for each sample. The BL21 (DE3) cell background was subtracted from each reading. The fluorescence of each mutant was normalized to the fluorescence of wtGFP$_{UV}$. Four replicates of expression were performed for each sample.

2.4.3 Codon scanning mutagenesis of sfGFP using MuCSM transposon

2.4.3.1 Construction of MuCSM transposon and transposon-targeting plasmid

Asymmetric transposon MuCSM were constructed by PCR amplifying the described *Mly* I TAG linker (*49*) using primers 5'-AAAGGATCCTTAGCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGCGAAAACGGAAGGTATTCGCAATAATCTTAAT-3' AND 5'-AAAAGATCTCCGACGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGCGAAAATGAGTAAACGCGGTCTGACTC-3' (*Bgl* II and *BamH* I sites underlined, the

five nucleotide difference highlighted in red and MuA transposase recognition sites in shadow). This PCR was performed using *Taq* DNA polymerase with the following cycle conditions: 95 ℃ for 3 min, 11 cycles of 95 ℃ for 30 s, 66 ℃ (-1 ℃/cycle) for 30 s and 72 ℃ for 2 min, followed by 20 cycles of 95 ℃ for 30 s, 66 ℃ for 30 s and 72 ℃ for 2 min and final extension at 72 ℃ for 10 min. The PCR product was blunted and cloned into the pUC18 vector. Prior to transposition reactions, the transposon DNA was released from the vector by *BamH* I/*Bgl* II digestion, followed by gel-purification using a QIAquick gel-extraction kit (Qiagen, Valencia, CA). The pTTsfGFP plasmid was generated by cloning the sfGFP gene into the previously described intein-targeting plasmid pIT (*49*). The sfGFP gene used in this study is a synthetic gene that has been optimized for expression in *E. coli* and is devoid of the restriction sites required for the mutagenesis process in this study (*85*). Sequence details of MuCSM transposon and pTTsfGFP plasmid can be found in **Appendix IV**.

2.4.3.2 Generation of a sfGFP library with random transposon insertions

Transposition reactions were performed in a 20 μL reaction containing 450 ng pTTsfGFP, 125 ng gel-purified transposon DNA, 1 unit of HyperMu MuA transposase (Epicentre), 50 mM Tris-acetate (pH 7.5), 150 mM potassium acetate, 10 mM magnesium acetate and 4 mM spermidine. The reaction was incubated at 30 ℃ for 4 h and halted by the addition of 2 μL of 0.1% SDS followed by heat-inactivation at 70 ℃ for 10 min. The reaction was cooled on ice and 1 μL of the solution was transformed into 50 μL electro-competent GeneHogs cells. The transformants were recovered at 37 ℃ for 1 h and plated on LB agar (Kan50/Amp40). The plates were incubated at room temperature for 48 h. More than 10,000 colonies were resulted from the 20 μL

transposition reaction and were collected to build the pTTsfGFP-MuCSM library. From this transposon library, 23 colonies were randomly picked and digested with FastDigest *Mly* I and *EcoR* I (Fermentas) to verify the randomness of transposon insertions. The library DNA was extracted and digested with *Nhe* I/ *EcoR* I (Fermentas). This yielded four DNA fragments (pTT backbone-transposon, pTT backbone, sfGFP-transposon and sfGFP). The sfGFP-transposon fragment was isolated and ligated into the *Nhe* I/*EcoR* I site of a modified pTrcHisA expression vector (Invitrogen) in which the *Bsg* I and *Bpm* I sites had been removed (see **Appendix III** for primer sequences). In total, 30,000 colonies were collected from this ligation, sufficient to ensure the full coverage of library diversity (*78*). The DNA of pTrcHisA-sfGFP-MuCSM library was extracted and used as the template in the following EI-PCR.

2.4.3.3 EI-PCR

Single, di- and tri-NDT libraries were constructed using the same pTrcHisA-sfGFP-MuCSM library DNA as the template but different primer pairs in the PCR amplification. The optimum template concentrations and annealing temperatures were individually determined for each reaction. The PCR was performed in four tubes of 100 μL solution each containing **X** pg/μL library DNA (see below), 0.5 μM each of corresponding forward and reverse primers, 0.2 mM dNTPs, 0.005 U/μL of Phusion DNA polymerase (NEB), 1×Phusion HF buffer (NEB, providing 1.5 mM $MgCl_2$) and additional 2 mM $MgSO_4$. **X** indicates the optimum template concentrations and was 100, 125 and 200 pg/μL for single, di- and tri-NDT library, respectively. The PCR cycle conditions were: initial denaturation at 98 ℃ for 2 min, 18 cycles of 98 ℃ for 10 s, **Y** ℃ for 30 s and 72 ℃ for 2 min followed by a final extension at 72 ℃ for 10 min. **Y** represents the optimum

annealing temperatures and was 57, 59 and 65 ℃ for single, di- and tri-NDT library, respectively. The primer pairs were FF1/ FR1, FF2/ FR2 and FF2/ FR3 for single codon, di-codon and tri-codon libraries, respectively (**Table 2.5**). The PCR product was purified using QIAquick PCR purification kit (Qiagen).

2.4.3.4 Bsg I or Bpm I digestion and Klenow treatment

The purified PCR product of single NDT library was digested with *Bpm* I in a 50 μL reaction containing 500 ng DNA, 1 U FastDigest BpmI (Fermentas) and 1×FastDigest buffer (Fermentas). The reaction was incubated at 30 ℃ for 4.5 h followed by heat-inactivation at 65 ℃ for 20 min. The purified PCR product of di-NDT or tri-NDT library was digested with *Bsg* I in a 50 μL reaction containing 500 ng DNA, 6 U *Bsg* I (NEB), 80 μM SAM and 1× buffer 4 (NEB). The reaction was incubated at 37 ℃ for 4.5 h followed by heat-inactivation at 65 ℃ for 20 min. The digestion product of each library was treated with Klenow fragment by adding 1.7 μL of 1 mM dNTPs (NEB) and 0.5 U of Klenow fragment (NEB) into 50 μL heat-inactivated reaction solution. The Klenow treatment was performed at 25 ℃ for 15 min and stopped by addition of 1.1 μL of 0.5 M EDTA, pH 8.0, followed by heat-inactivation at 75 ℃ for 20 min. The Klenow-treated digestion product was then gel-purified using QIAquick gel-extraction kit (Qiagen).

2.4.3.5 Intramolecular ligation and transformation

The gel-purified digestion product (Klenow-treated) was subjected to an intramolecular ligation that contained 2.5 ng/μL DNA, 0.5 Weiss unit/μL T4 DNA ligase (Fermentas), 40 mM Tris-HCl (pH 7.8 at 25 ℃), 10 mM MgCl$_2$, 10 mM DTT and 0.5 mM ATP in a 10 μL reaction. The ligation reaction was incubated at 16 ℃ overnight and then heat-inactivated at 70 ℃ for 7 min. Several 10 μL ligation reactions were performed for each

library, pooled and then purified by ethanol precipitation. After ethanol precipitation, the purified ligation product was resuspended in water and then transformed into electro-competent GeneHogs cells. The transformants were recovered at 37 ℃ for 1 h and then plated on LB agar (Amp100). In total, 2×10^5 colonies were collected for each library and the DNA was extracted to be used as the template in the second EI-PCR.

2.4.3.6 Second round of EI-PCR, digestion and intramolecular ligation

The conditions for the second EI-PCR were generally the same as for the first one with, again, individually determined optimum template concentrations and annealing temperatures. The template concentrations were 200, 125 and 200 pg/μL for single, di- and tri-NDT library, respectively. The annealing temperatures were 67 ℃, 63 ℃ and 66 ℃ for single, di- and tri-NDT library, respectively. The primer pairs for single, di- and tri-NDT libraries were SF0/ SR, SF0/ SR and SF1/ SR, respectively (**Table 2.5**). The PCR product was purified, digested with *Bsg* I, treated with Klenow fragment, gel-purified, re-ligated and transformed as described above. More than 2×10^5 colonies were collected from each library and the library DNA was extracted for further characterization.

2.4.3.7 Screening for spectra-shifted mutants

From each of single, di- and tri-NDT library, 500 colonies were randomly chosen and inoculated into 96-well plates. Each well contained 1 mL LB media (Amp100) and 1 mM IPTG. The plates were grown at 37 ℃ for 12 h. The expression culture (50 μL) from each well was diluted with 250 μL distilled water and then loaded into 96-well fluorescence plates. The OD_{600} for each sample is generally between 0.15 and 0.2 under these conditions. The plates were screened using a SpectraMax M5 microplate reader

(Molecular Devices, Sunnyvale, CA) with 2 nm bandpass. Screening was performed using both 400 nm excitation wavelength with an emission range from 430 to 600 nm and 460 nm excitation wavelength with an emission range from 490 to 540 nm. The background fluorescence of GeneHogs cells containing empty pTrcHisA vector was subtracted from each sample reading. The calibrated spectral data were then normalized and compared to the original sfGFP spectrum. Candidate spectra-shift mutants were confirmed by three replicates of the whole-cell fluorescence assay as described above. The spectra-shifted mutants were then diluted to an OD_{600} of 0.2000 ± 0.0050 and the whole-cell spectra were recorded using a model F4500 fluorescence spectrophotometer (Hitachi). The obtained excitation and emission spectra were calibrated, normalized and compared with the normalized spectrum of original sfGFP. These spectra-shifted mutants were sequenced by Yale DNA sequencing facility (New Haven, CT).

Chapter 3: Enzymatic and Nonenzymatic Assembly of PolyUb Chains

3.1 Introduction

This chapter describes how unnatural amino acid (UAA) mutagenesis technologies can facilitate the enzymatic and nonenzymatic synthesis of polyubiquitin (polyUb) chains. Two approaches have been developed for the synthesis of polyUb chains. Both of these two approaches relied on an orthogonal pyrrolysyl tRNA (PylT) and tRNA synthetase (PylRS) pair to incorporate $^\varepsilon N$-*tert*-butyloxycarbonyl-L-lysine (BocLys) into a targeted lysine position in Ub. The first method focuses on the synthesis of homogeneous polyUb chain which requires a linkage-specific Ub conjugating enzyme E2 (and thus termed enzymatic approach). The second method utilizes a silver-based protein condensation technique rather than E2 enzyme and can generate polyUb chains of any linkage composition (and thus termed non-enzymatic approach). Also introduced in this chapter is a simple yet highly efficient technique for large-scale production of proteins containing UAAs, referred to as "condensed culture" method.

3.1.1 Biological functions of ubiquitin and polyUb chains

Ubiqutin (Ub) protein gets the name from its ubiquitous expression in all eukaryotic organisms (*86*). It is a 8.5 KDa regulatory protein consisting of 76 amino acids. Ub is a highly conserved protein among a variety of organisms. For example, human Ub is only three amino acid different from yeast Ub. Ub was discovery in the early 1980s by Aaron Ciechanover, Avram Hershko and Irwin Rose. The Nobel Prize in Chemistry 2004 was awarded to these three scientists "for the discovery of ubiquitin mediated protein degradation". During the past three decades, collaborative efforts from researchers

around the world have largely expanded our understanding of Ub. Now we know that Ub plays important roles not only in protein degradation but also in many other cellular processes, including gene transcription, DNA repair, cell cycle regulation and viral infection (*86*).

Ub carries out its biological functions by conjugating to a target protein. This conjugation process, referred to as ubiquitinylation or ubiquination, occurs through isopeptide bond formation between the ε-amino group of a lysine residue in a target protein and the C-terminal carboxyl group of Ub. Target proteins can be ubiquitinylated by Ub monomers (mono- or multi-Ub monomers) or polyUbs (**Figure 3.1A**). In polyUbs, Ub monomers are usually connected through isopeptide bond between the C-terminal carboxyl group of one Ub molecule (distal monomer) and the ε-amino group of one of the seven lysine residues of another Ub (proximal monomer). These polyUbs can have either the same linkage (homogeneous chains) or mixed linkages (**Figure 3.2B**).

Ub contains seven lysines K6, K11, K27, K29, K33, K48 and K63 out of 76 residues. Much of our understand about polyUb system comes from the K48 and K63-linked polyUb chains, where K48 or K63 in the proximal Ub involves in the isopeptide bond formation (*87, 88*). In fact, all of the seven lysines as well as the N-terminal methionine can serve as the distal Ub acceptors in nature. Different polyUb chains are known to have distinct biological functions. For example, long (tetramer or above) K48-linked polyUb chains can ferry ubiquinylated proteins to the proteasome for degradation, while K63-linked polyUb chains are generally thought to act as regulatory rather than poteolytic signals (*89-91*). However, our understanding of polyUb chains is quite fragmentary and systematic analyses are yet to be performed.

A.

Mono ubiquitinylation   Multi-mono ubiquitinylation   Poly ubiquitinylation

B.

Ub dimer (Ub$_2$)   K48-linked homogeneous Ub trimer (Ub$_3$)   K11, K48-mixed linkage Ub trimer (Ub$_3$)
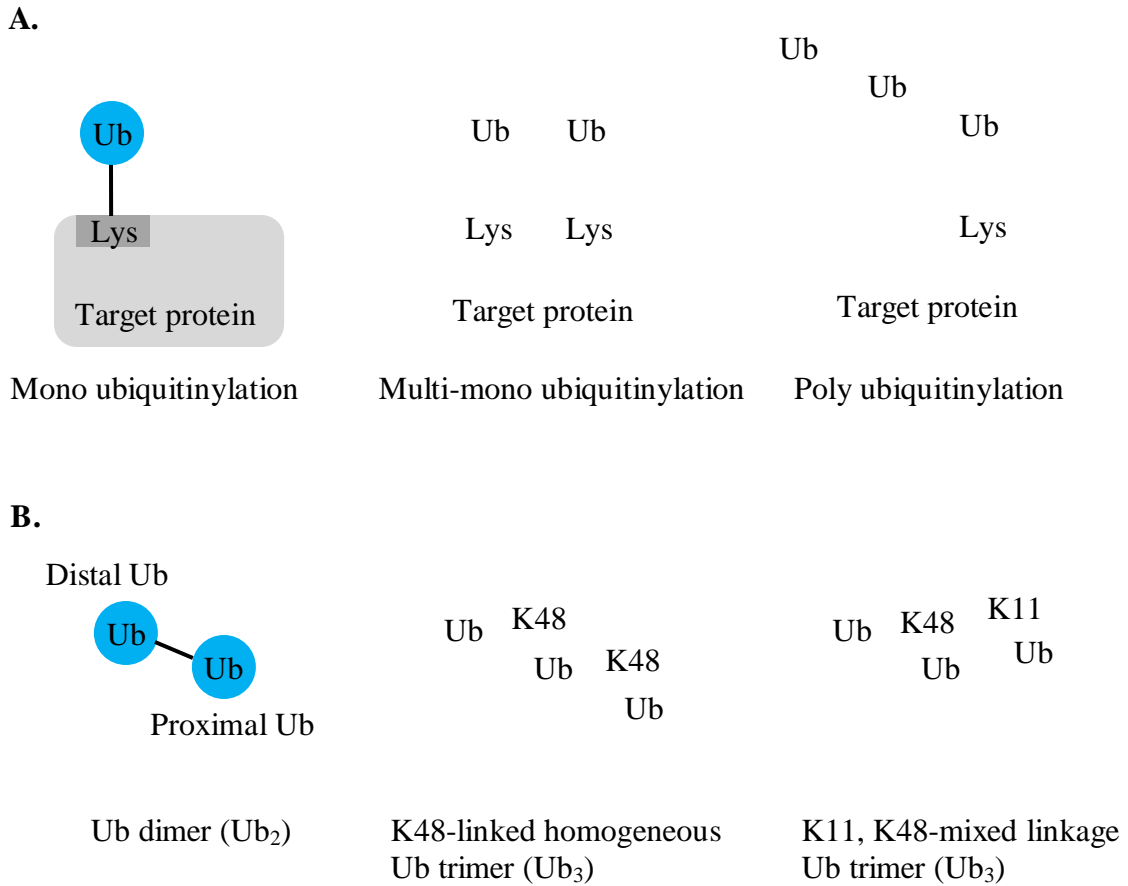
Figure 3.1. Ub conjugation complex. **A**) A target protein conjugated with mono-, multi-mono-, and poly-Ubs. **B**) Different polyUb chains.

In nature, polyUb chain synthesis requires a Ub activating enzyme E1, a linkage-specific enzyme E2 and a Ub ligase E3 (*87, 92*). Through an ATP-dependent mechanism, E1 activates the C-terminal G76 of Ub by coupling its carboxyl group to the thiol group of a cysteine residue of E1. This activation process yields an E1-Ub thioester intermediate, which will be then rendered to a conserved cysteine residue of a linkage-specific enzyme E2 by a transthiolation step. For E3 ligases, it is generally believed that there exist two major superfamilies, classified according to the mechanisms of thiotranslation. The first superfamily, often referred to as RING-domain E3s, serve as an

adaptor to position the E2-Ub thioester bond in close proximity to a substrate (monoUb, polyUb or other Ub-targeting proteins) for direct thioltranslation. The second class, termed HECT E3s, first accept the Ub molecule (from the E2-Ub conjugate) on to one of their conserved cysteine residues and then pass the Ub to a substrate. This generalization, however, is being challenged by the emerging experiments (*93*).

Studies of the relationship between the linkage, structure and function of the polyUb signals requires the ability to assemble these chains with natural connectivity, controlled length, defined linkage composition and in sufficient quantities (milligram scale). Until recently, this remained a challenging problem to all researchers in the Ub field. Traditional methods for the polyUb chain assembly can be generally classified into two categories: controlled enzymatic synthesis using chain-terminating mutations (see below for details) (*94, 95*) and total chemical synthesis followed by native chemical ligation (NCL) (see below) (*96, 97*), though several variations exist within each method. Recently, two UAA mutagenesis-based approaches were developed for synthesis of polyUb chains with natural connectivity and defined linkage composition (*98, 99*). These two methods have largely expanded polyUb synthesis methodologies, though they still have inherent limitations. The below sections will first introduce the principles of the above approaches and two novel methods (*100, 101*) developed by our group in collaboration with Dr. Fushman's group (at Department of Chemistry and Biochemistry, University of Maryland, College Park).

3.1.2 Controlled enzymatic assembly of polyUb chains

Controlled enzymatic assembly of polyUb chains relies on linkage-specific E2 conjugating enzymes. Homogeneous K11-, K48-, or K63-linked polyUbs have been

successfully assembled using identified E2 enzymes Ube2S, E2-25K, and Ubc13/Mms2, respectively (*94, 102, 103*). Like any enzymatic reaction, E2-catalyzed polyUb synthesis reactions are highly efficient, progressive and specific. However, the drawback is that these enzymatic processes do not have controls of the length of products, thus yielding mixed products with various lengths. A partial solution to this problem is to use a "chain terminator", a Ub monomer that halts further polymerization. Depending on the purpose of the experiments, "chain terminator" can be used for promixal, distal or both ends. Proximal "chain terminator" is usually a Ub mutant with two C-terminal glycine truncated (Ub_{1-74}) or with one additional aspartic acid residue (Ub_{D77}). Distal chain terminator is usually a Ub mutant with a arginine or cysteine mutation at the lysine position involving in the isopeptide bond formation. For example, to synthesize K48-linked Ub_2 (K48-Ub_2), Ub-K48C and Ub_{D77} can be used in a K48-specific E2 catalyzed reaction such that isopeptide bond only forms between Gly76 of Ub-K48C and Lys48 of Ub_{D77}. The blocked ends of polyUb can be chemically or enzymatically removed for further polymerization (**Figure 3.2**). For example, distal terminator Cys can be converted to a lysine analogue by ethyleneimine, resulting in a distal ε-amino group, while proximal terminator Asp77 can be removed by Ub C-terminal hydrolase to free the proximal Gly76 (**Figure 3.2**). This methodology, however, is limited by the non-native connectivity between Ub monomers and availability of linkage specific E2 enzymes. To solve these problems, we developed an enzymatic approach for synthesis of homogeneous polyUb chains and a non-enzymatic approach for synthesis of polyUb chains with any linkage composition (see below).

SH

Ub hydrolase
　　　　　　Ub

Ub

Proximally unblocked
for further extension

SH
SH

Ub
Ub

ATP, E1, E2 →
Ub

Ub
D77

Distal terminator
Ub-K48C

Doubly blocked Ub$_2$

+

H_3^+N

H_3^+N

Ub

S

D77
Ub

Proximal terminator
Ub$_{D77}$

Ethyleneimine
Ub

Ub

D77

Distally unblocked
for further extension
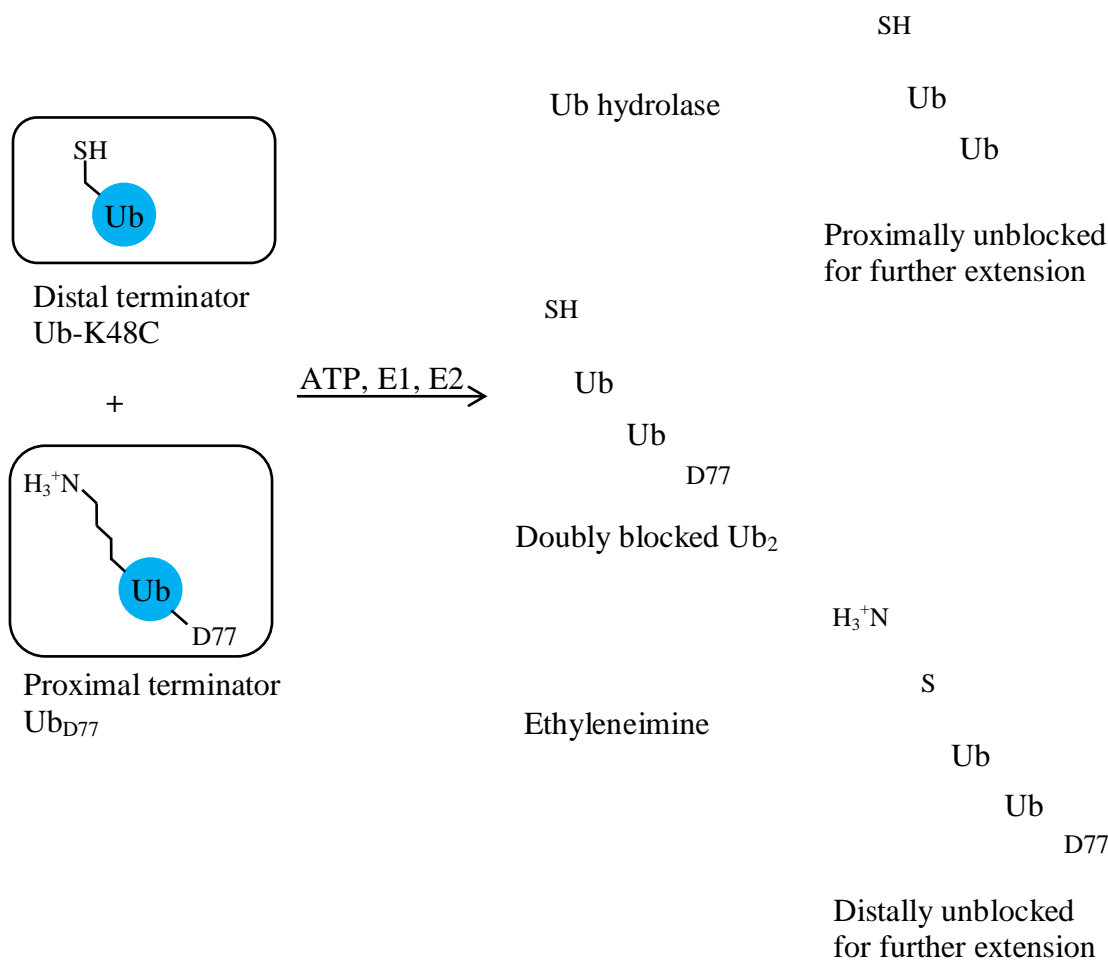
Figure 3.2. Controlled enzymatic synthesis of polyUb chains. Synthesis of K48-linked Ub$_2$ is shown as a demonstration. Proximally or distally unblocked Ub$_2$ can be further extended using the same approach.

3.1.3 Total chemical synthesis of polyUb chains

The limitations in controlled enzymatic synthesis have motivated several research groups to develop methods for total chemical synthesis of polyUb chains. The past decade has seen significant progress on this method (*96, 97, 104-107*). Ub$_2$ and even Ub tetramers (Ub$_4$) have been successfully made (*30*). Although variations exist within this method, they usually rely on solid phase peptide synthesis (SPPS) to synthesize a short peptide

(20~30 amino acid in length) followed by native chemical ligation (NCL) to assemble Ub monomer and oligomer. These methods have opened new ways to build polyUb chains and further expanded our understanding on how chemistries can be designed to accomplish enzymatically catalyzed reactions.

Figure 3.3 shows one approach to synthesize K48-Ub_2 using δ-mercaptolysine (δ-mercapto-Lys) as the "linker" between two Ub units (*106*). Two small peptides are first synthesized by SPSS for assembly of full-length Ub monomer. Proximal and distal Ub monomers share a common N-terminal peptide (Ub-N) that contains a thiolester at its C-terminus. In C-terminal peptide (Ub-C) of both proximal and distal Ubs, the first amino acid (position A46 in native Ub) is mutated to cysteine residue for NCL while proximal Ub-C additionally contains a protected δ-mercapto-Lys at K48 (**Figure 3.3A**). Full-length proximal and distal Ub monomers are assembled in NCL using the common Ub-N and their corresponding Ub-C peptides. After deprotection of δ-mercapto-Lys on the proximal Ub, the two Un monomers werere ligated by NCL again, resulting in a Ub_2 with native isopeptide bond (**Figure 3.3B**).

However, total chemical synthesis implementation is not readily available in every biochemical laboratory, thus limiting its broad applications. Moreover, total chemical synthesis methodologies require multiple protection/deprotection steps, and iterative NCL processes, thus yielding a small quantity of materials.

Due to the inherent limitations in the above two methods, UAA mutagenesis-based polyUb synthesis approaches have been developed (*99, 100, 108, 109*). Although there is no "one-fits-all" method, the UAA mutagenesis-based synthesis of polyUb chains can indeed overcome many limitations in the traditional methods.

A. Human Ub full sequence

```
1          10          20          30          40          50
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQL

51         60          70    76
EDGRTLSDYNIQKESTLHLVLRLRGG
```

Ub-N : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIF(thiolesterified C-terminus)

Ub-C (distal) : CGKQLEDGRTLSDYNIQKESTLHLVLRLRGG

Ub-C (proximal) : CG(protected δ-mercapto-Lys)QLEDGRTLSDYNIQKESTLHLVLRLRGG

B.
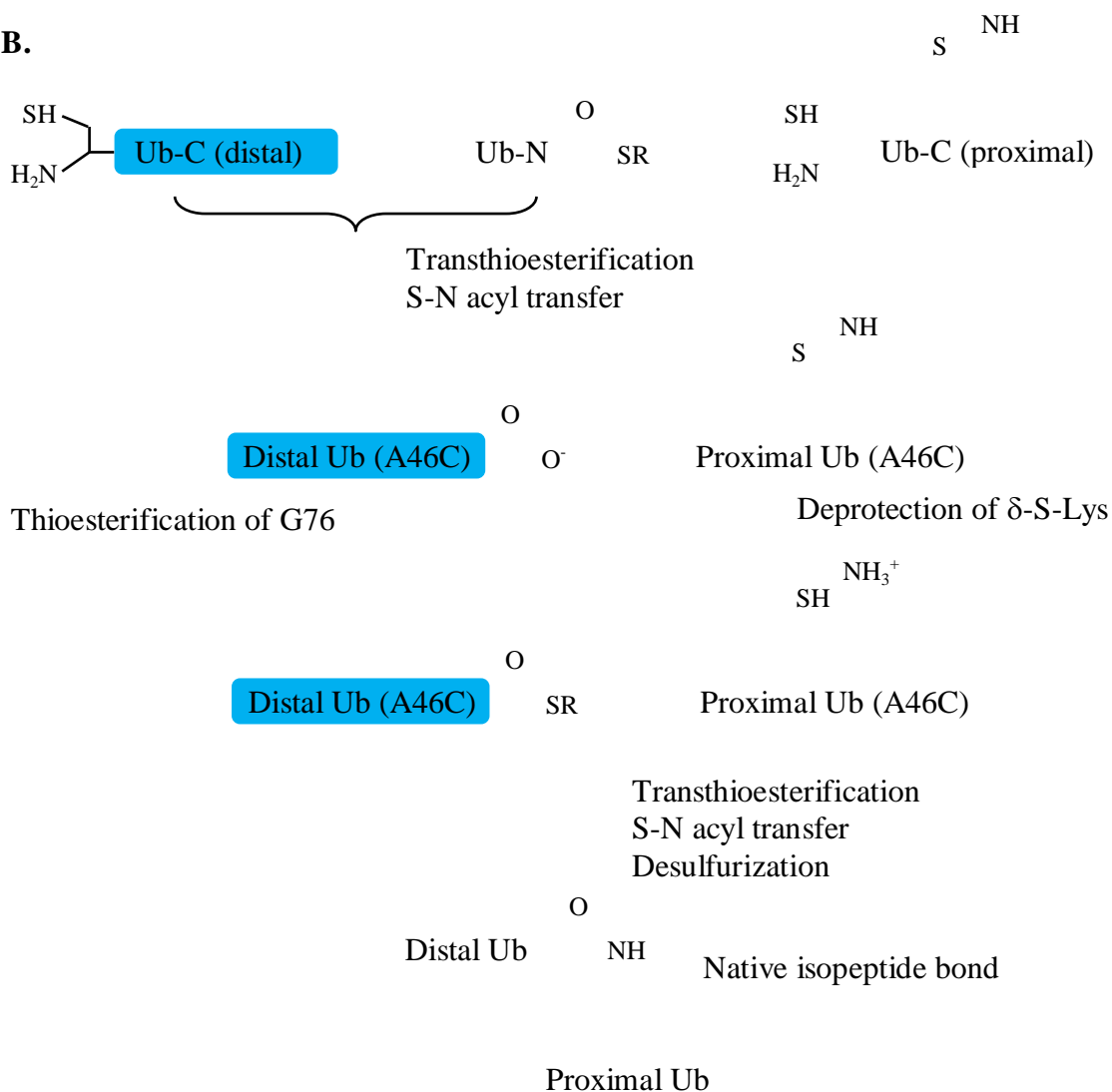
Figure 3.3. Total chemical synthesis of K48-linked Ub_2. **A**) N- and C-terminal peptides synthesized by SPSS for assembly of Ub monomer. Arrow indicates the joint position of the two peptides. **B**) Assembly of Ub_2 *via* a δ-mercapto-Lys linker

3.1.4 Synthesis of polyUbs using Ub units containing lysine analogues

Certain methanogenic archaeal or bacterial organisms, such as *Methanosarcina barkeri* and *Methanosarcina mazei*, incorporate pyrrolysine into some methyltransferases in response to a UAG codon carried in the genes of these enzymes (*29, 110, 111*). Soon after discovery, these pyrrolysyl tRNA (PylT)/aminoacyl tRNA synthetase (PylRS) pairs were adapted to express proteins in *E. coli*. These PylT/PylRS pairs proved to be able to orthogonally incorporate a pyrrolysine in response to a UAG codon (*112*). Later, PylRS was evolved to incorporate lysine derivatives into proteins in *E. coli* (*113*). Recent studies have successfully incorporated more than 10 UAAs into proteins using PylT/PylRS pairs (*114-116*) and emerging studies are rapidly expanding this UAA list (*117*).

Chin and co-workers have incorporated *tert*-butyloxycarbonyllysine (BocLys) into Ub and have developed a method to synthesize polyUb chains using Ub monomers containing BocLys (Ub-BocLys). This method, referred to as genetically encoded orthogonal protection and activated ligation (GOPAL), relies on a silver-mediated condensation (*118*) reaction to generate the isopeptide bond between two Ub monomers. The procedure of GOPAL is illustrated in **Figure 3.4**. The distal Ub is expressed from *E. coli* as a fusion protein with a C-terminal intein. Upon thiolysis of intein protein (*96*), a Ub-SR is directly obtained without further thioesterification step. The proximal Ub is expressed from *E. coli* using *M. barkeri* PylT/PylRS pair to incorporate BocLys at a target lysine position. Prior to condensation reaction, the distal monomer Ub-SR and proximal monomer Ub-BocLys need to be protected for free amine groups (N-terminal amines and side chain amines from the unprotected lysine residues). This protection is achieved using carboxybenzyl (Cbz) protecting group, yielding a distal Ub-SR with 8

Cbz groups (Ub-SR$_{Cbz8}$) and a proximal Ub-BocLys$_{Cbz7}$ (the target lysine site is protected by a Boc group already). The Boc group in Ub-BocLys$_{Cbz7}$ is then removed by trifluoroacetic acid (TFA) to generate a Ub with only the target lysine unprotected (Ub$_{Cbz7}$) for ligation. Remarkably, the Cbz groups in Ub$_{Cbz7}$ are left intact as the Boc deprotection conditions are orthogonal to that of Cbz groups. Ub-SR$_{Cbz8}$ and Ub$_{Cbz7}$ (with only the target lysine bearing a free ε-amino group) are subjected to a silver-mediated condensation reaction to generate an isopeptide bond between the C-terminal thioester of Ub-SR$_{Cbz8}$ and ε-amine of the target lysine on Ub$_{Cbz7}$. The Cbz protecting groups are then removed globally using an acid cocktail, yield a Ub dimer with a native isopeptide bond.
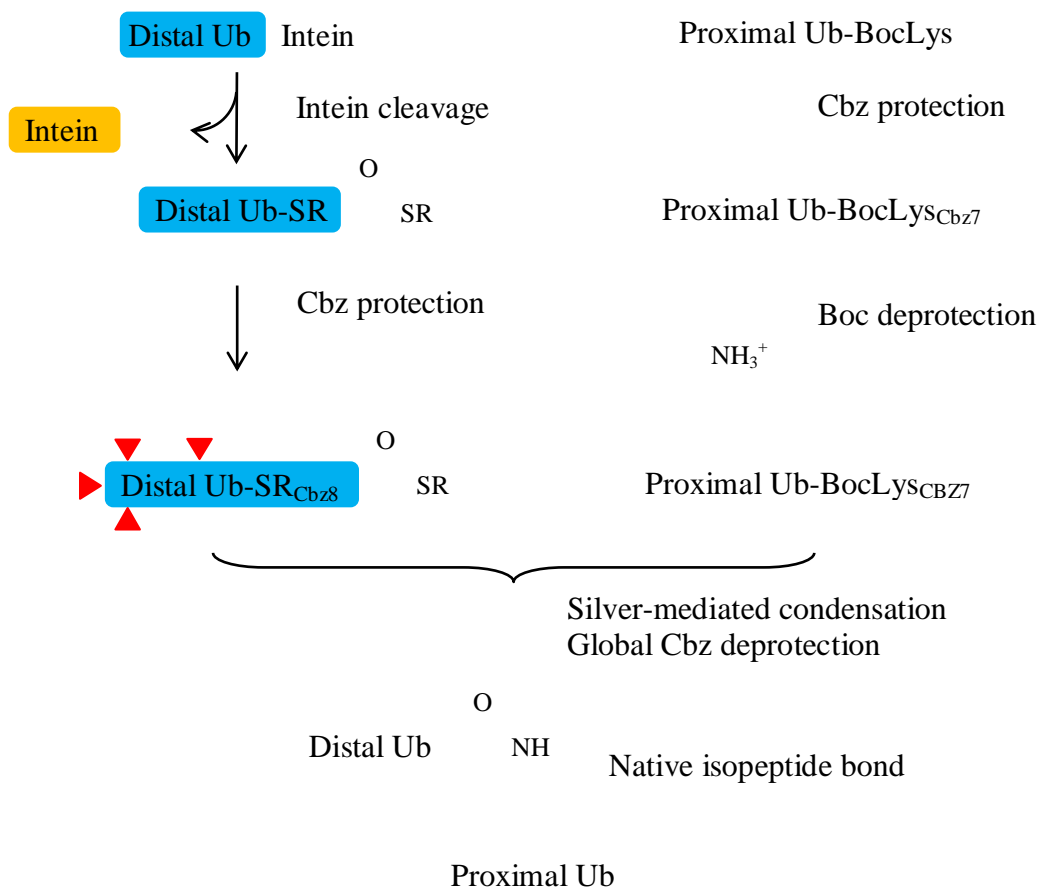
Figure 3.4. Procedure of genetically encoded orthogonal protection and activated ligation (GOPAL). The red sphere and triangle indicate Boc- and Cbz-protected amines, respectively

GOPAL allows generation of polyUb chains of any linkage composition without the need of linkage-specific enzyme E2. Unlike total chemical synthesis, Ub monomers in GOPAL are bacterially expressed proteins, bypassing the yield problem in SPPS and NCL. This method, however, requires extensive protection/deprotection of Cbz groups. The global deprotection of bulky Cbz group occurs at 15 different positions (in the case of a Ub$_2$) and does not go to completion easily in our hand even under denaturing conditions. This motivated us to develop a new method where the bulky Cbz group is substituted by a smaller allyloxycarbonyl (Alloc) group (see below).

In a more recent study, Chin and co-worker have evolved the *M. barkeri* tRNA/aaRS pair to directly incorporate δ-mercapto-Lys into a target lysine position in Ub. This Ub monomer can serve as the proximal unit to generate an isopeptide bond with a Ub-SR distal monomer. This method is advantageous over the total chemical synthesis method because the bacterially expressed Ub monomers can be readily produced in large quantity. This new method also bypasses the need of global protection/deprotection of Ubs using Cbz groups and thus more efficient. However, it does require a final desulfurization step to remove the thiol group which limits its application to ubiquitinylation of a substrate protein containing cysteine residues.

3.1.5 Traceless enzymatic synthesis of polyUbs using Ub-BocLys monomers

Traditional enzymatic synthesis of polyUb chains usually requires the conversion of a terminator cysteine to a lysine analogue for chain extension. This, as stated above, will generate a non-native isopeptide bond that may have distinct properties than the native isopeptide bond. To solve this problem, our group, in collaboration with Dr. Fushman's lab, sought to to establish a traceless enzymatic approach to synthesize polyUb chains.

This method is conceptually similar to traditional approaches, however, with the terminator cysteine or arginine residues replaced by a lysine analogue with removable protecting group. Despite that several PylRS variants have been evolved to incorporate a variety of lysine analogues into proteins, we chose to use the PylRS-Y384F mutant (*115*) to incorporate BocLys into Ub. PylRS-Y384F mutant can efficiently aminoacylate PylT with BocLys, which is a fundamental requirement for large-scale protein production. Moreover, Boc protecting group can be removed from a protein under mild conditions (2% TFA in water), leaving the Ub structure and function intact.

Figure 3.5 describes the rationale of traceless enzymatic synthesis of homogeneous polyUb chains. This method requires two building blocks: a proximal terminator and a Ub monomer containing BocLys (Ub-BocLys) at the desired position. The proximal terminator can be either Ub_{D77} or Ub_{1-74}. In the case of Ub_{D77}, native C-terminus (Gly76) can be generated by Ub hydrolase. Ub-BocLys can be prepared with high yield using the PylRS/PylT pair. By sequential addition of Ub-BocLys to the proximal terminator followed by TFA treatment, homogeneous polyUb chains can be synthesized with high efficiency.

Despite the many available methods, the traceless enzymatic synthesis of polyUb chains is nevertheless one of the most efficient approaches to generate homogeneous chains. The limitation of this method, however, is that linkage-specific conjugating enzymes E2 are available only for K11, K48 and K63 linkage. This motivated to develop an E2-independent (non-enzymatic) approach to synthesize polyUb chains with any linkage composition (see next section).
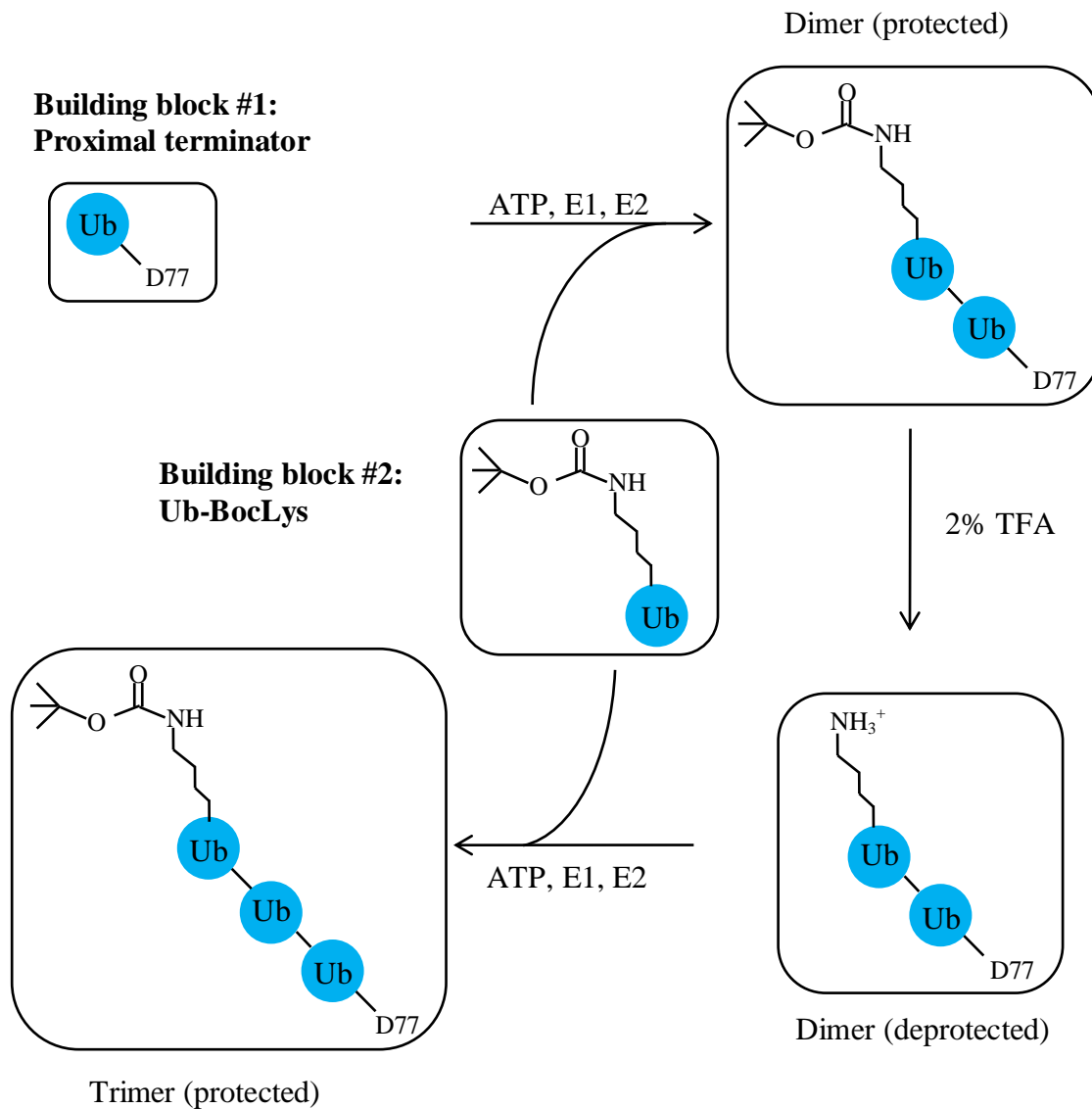
Figure 3.5. Procedure of traceless enzymatic synthesis of homogeneous polyUb chains. Protected trimer can be unblocked for further polymerization.

3.1.6 Nonenzymatic synthesis of polyUbs using Ub-BocLys monomers

This method was inspired by GOPAL where a silver-mediated condensation (*30*) is used to generate isopeptide bond between the distal and proximal Ub monomers. The procedure of Ub_2 synthesis using this nonenzymatic approach is shown in **Figure 3.6**.
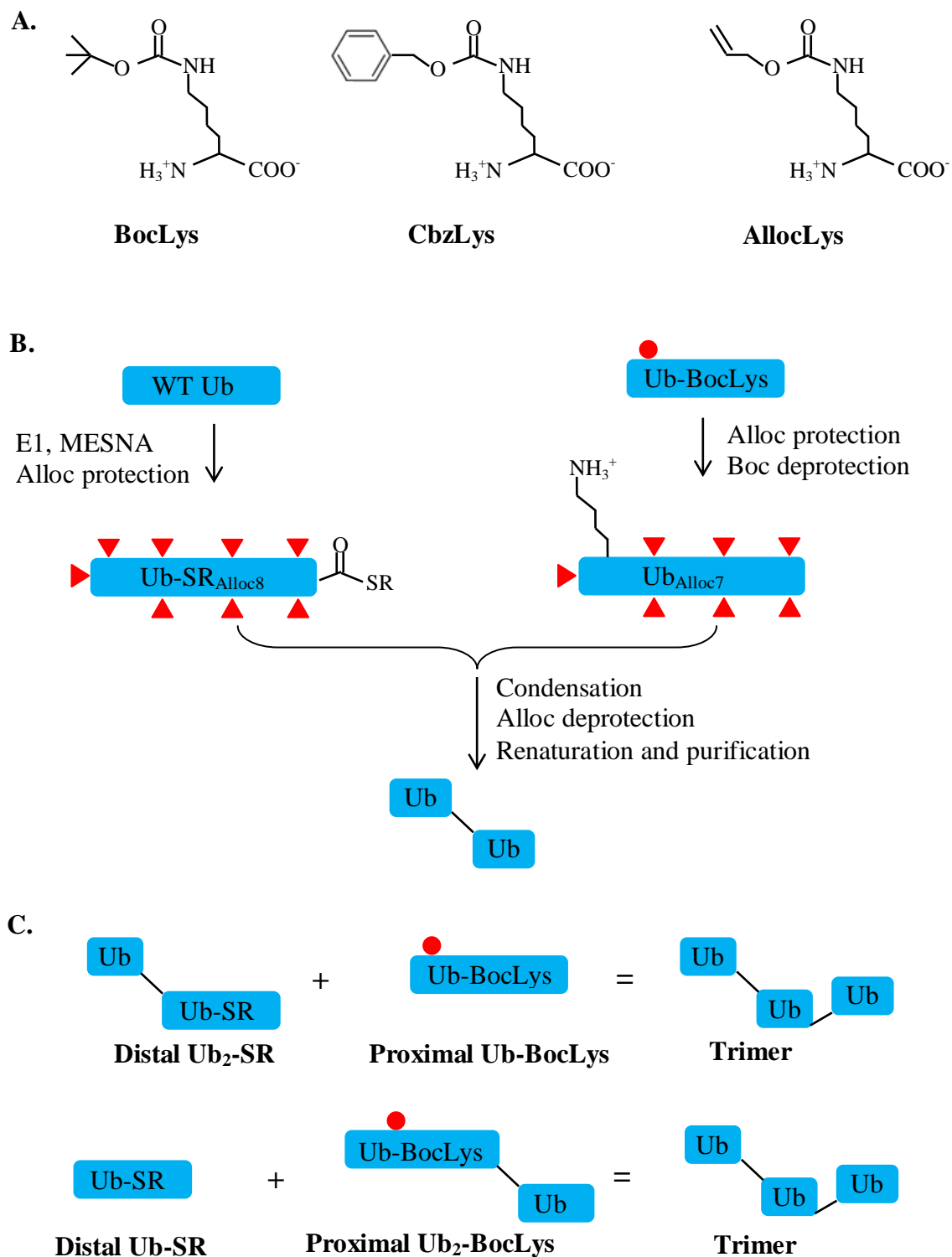
Figure 3.6. Generation of Ub$_2$ and Ub$_3$ using nonenzymatic synthesis approach. **A**) Structures of BocLys, CbzLys and AllocLys. **B**) Schematics of Ub$_2$ synthesis. **C**) Two approaches to synthesize Ub$_3$. The red spheres and triangles indicate Boc protected and Alloc protected amines, respectively.

To prepare the distal Ub monomer for condensation, WT Ub is first expressed from *E. coli* cells. The C-terminal carboxyl group of WT Ub is thioesterified by E1 Ub activating enzyme with sodium 2-mercaptoethane sulfonate (MESNA) supplemented as a thiol donor (*119*). All seven lysines as well as the N-terminal amine of this Ub-SR are then protected by allyloxycarbonyl (Alloc) protection group (**Figure 3.6A**) in the presence of allyloxycarbonyl oxysuccinimide (Alloc-OSu) as a Alloc group donor and *N,N*-diisopropylethylamine (DIEA) as a base. This results in Ub-SR$_{Alloc8}$ distal monomer for condensation reaction. The proximal Ub monomer for condensation reaction is prepared using Ub-BocLys as the template. Unprotected lysine residues as well as the N-terminal amine of this Ub-BocLys are protected with Alloc group as above, yielding Ub-BocLys$_{Alloc7}$. The Boc protection group of Ub-BocLys$_{Alloc7}$ is then removed by TFA treatment, exposing a single lysine residue accessible for ligation. Importantly, all the Alloc groups are kept intact as the deprotection condition of Alloc is orthoganol to that of Boc. This results in Ub$_{Alloc7}$ proximal monomer for condensation reaction. Ub-SR$_{Alloc8}$ distal monomer and Ub$_{Alloc7}$ proximal monomer are then ligated in a silver (AgNO$_3$)-meditated condensation reaction. The synthesized Ub dimer contains 15 Alloc protected lysines. Removal of the Alloc protecting groups using a ruthenium (II)-catalyzed reaction results in a native Ub$_2$.

To synthesize Ub$_3$ or higher polyUb chains, two strategies can be applied depending on the Ub$_2$ used in the condensation reaction. The first one is to use a proximal Ub-BocLys monomer and a distal Ub$_2$ that contains a thioesterified C-terminus on the proximal unit (**Figure 3.6C**). The thioesterified Ub$_2$ can be produced using the same protocol as for the Ub-SR. The second strategy is to use a distal Ub-SR and a proximal Ub$_2$ that contains a

BocLys at a desired lysine position on the distal unit (**Figure 3.6C**). This Ub_2-BocLys can be produced using the same protocol as for regular Ub_2 if Ub-BocLys, instead of WT Ub, is used as the starting distal monomer. Ub_4 or higher polyUb chains can be synthesized by either a convergent approach or a sequential approach. In the convergent approach, both distal and proximal unit are pre-assembled polyUb chains, for example, $Ub_2 + Ub_2 = Ub_4$. In the sequential approach, monomer Ub is added to a preassembled long polyUb chains, for example, $Ub_4 + Ub_1 = Ub_5$.

The major advantage of nonenzymatic method as compared with enzymatic method lies in the lack of requirement for linkage-specific conjugating enzyme E2 (nonenzymatic) and allows for the assembly of polyUb chains of any linkage composition. Our method has two major differences from GOPAL. First, the Ub-SR is generated by E1 activating enzyme, rather than by intein-mediated cleavage. In our hand, the E1 approach has higher efficiency than the intein approach. Furthermore, we use Boc/Alloc instead of Boc/Cbz as an orthogonal protecting group pair. This Alloc group is advantageous over Cbz because the deprotection of bulky Cbz groups in a well-folded protein may be more challenging than that of smaller Alloc groups.

3.1.7 Limitation of large-scale production of proteins containing UAAs

During the large-scale production of Ub-BocLys, it was realized that traditional methods for expressing proteins containing UAAs are not very efficient regarding to the usage of UAAs. The majority (>99%) of UAAs in the expression culture is unused. Instead, UAAs of high concentration (usually 2 mM) are supplemented into the culture just to ensure efficient cell uptake. To solve this problem, a simple yet highly efficient method was developed (**Figure 3.7**). In this method, *E. coli* cells are grown in regular rich media to

mid-log phase (OD_{600} of ~0.5). Prior to induction, these cells are centrifuged and resuspended in a smaller volume of fresh media containing inducing reagent and UAAs. For example, to achieve 20-fold condensation, cells grown in 1 L culture can be resuspended in 50 mL fresh media. This condensed culture approach was used to produce Ub-BocLys in large scale and turned out to be very efficient. To demonstrate the generality of this method, CbzLys and *p*-benzoylphenylalanine (pBpa) were also incorporated into Ub and sfGFP, respectively, using the same approach.

E.coli growth   Expression   High yields
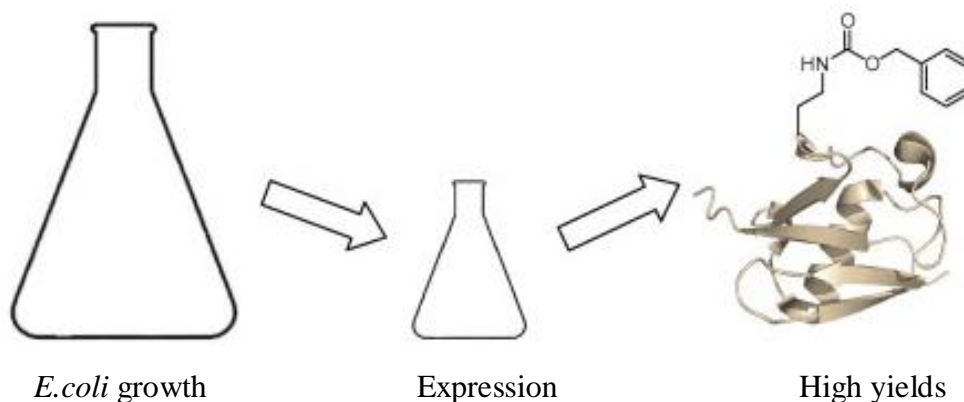
Figure 3.7. Expression of proteins containing UAAs using condensed culture.

3.2 Results and Disscussion

3.2.1 Traceless enzymatic synthesis of polyUb chains using Ub-BocLys

3.2.1.1 Construction of PylT and PylRS plasmids for BocLys incorporation

Two plasmid systems have been tested for incorporation of lysine analogues into Ub. The first one is a single plasmid system pSup plasmid (*120*) which carries both the tRNA and aaRS gene (**Figure 3.8A**). The second one is a double plasmid system (*43*) where the tRNA gene and the encoding gene for a target protein (myoglobin, for example) are carried on one plasmid pYC while the aaRS gene is encoded by another plasmid pBK

(**Figure 3.8B**). As both systems worked well (**Figure 3.8C**), we chose to use pSup for further studies because this single plasmid system is compatible with most *E. coli* expression vectors (*120*) and allows rapid expression test of a target protein in different vector context.
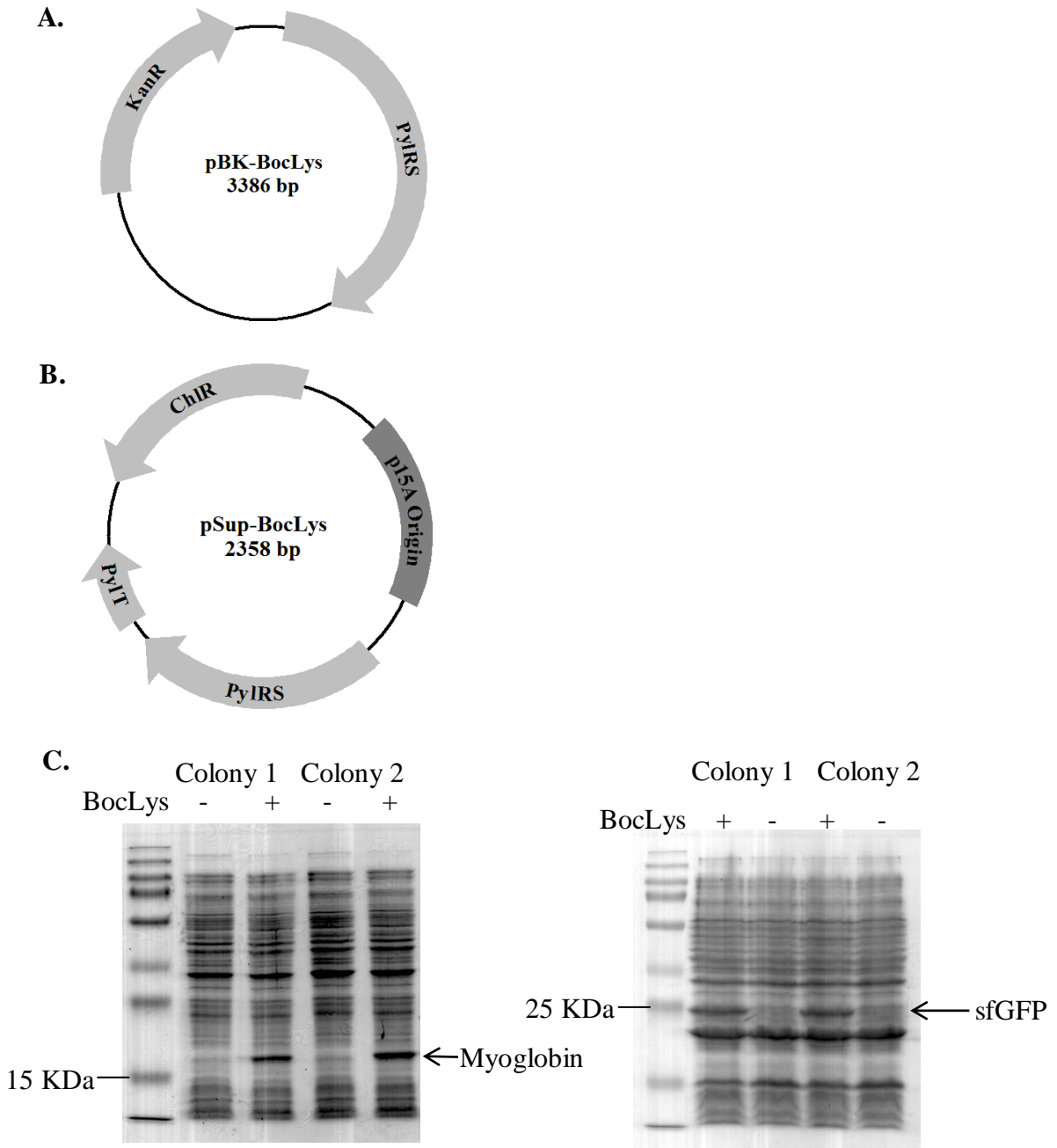
Figure 3.8. Plasmids for incorporation of BocLys into proteins. **A)** Double plasmid system. **B)** Single plasmid system. **C)** Expression tests on the two expression systems.

3.2.1.2 Construction of pSup plasmids for incorporation of other lysine derivatives

Several *Mm*PylRS variants have been developed to incorporate various lysine derivatives into proteins, including CbzLys (*116*) and photo-active amino acid *o*-nitrobenzyl-oxycarbonyl-L-lysine (OnbLys) (*121*). At the early stage of this project, all of these three amino acids were tested in Ub expression to gain the information about the efficiency of UAA incorporation. pET3a-Ub-K48tag (where Lys48 of Ub has been changed to a TAG stop codon) was used as the tester expression vector. The expression results suggested that all these three UAAs could be efficiently incorporated into Ub protein (**Figure 3.9**).
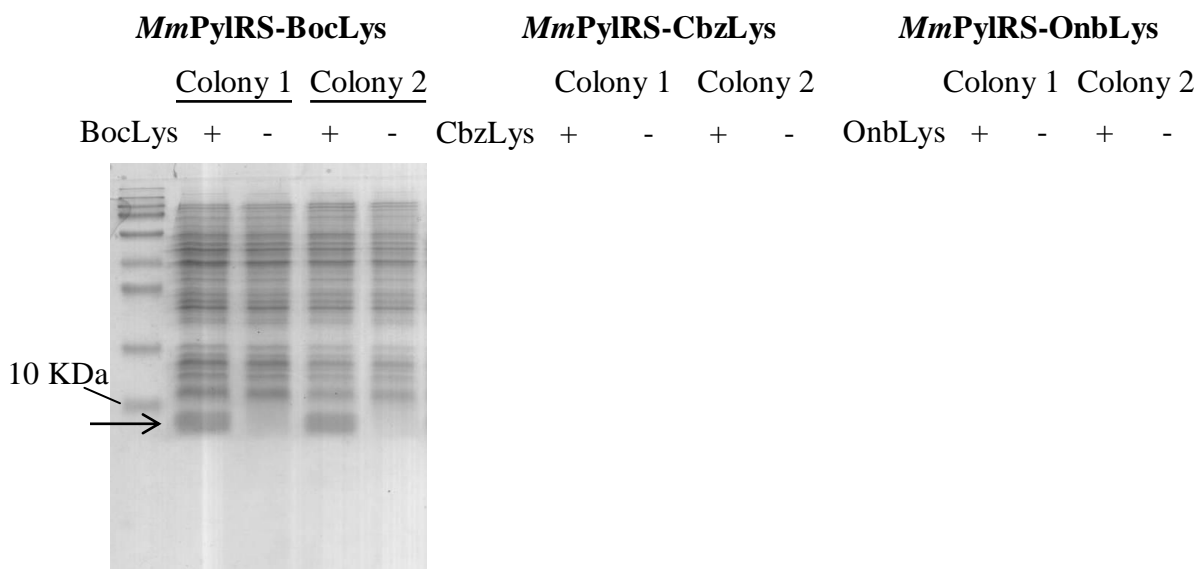
Figure 3.9. Expression tests for incorporation of BocLys, CbzLys and OnbLys into Ub-K48tag construct. Arrows indicate the UAA-dependent Ub expression.

Among these UAAs, BocLys was preferred over CbzLys as Boc deprotection condition is milder than that of Cbz. Moreover, BocLys had been successfully applied to synthesize polyUb chains (*99*). Ideally, if OnbLys can be incorporated into Ub with high fidelity, the synthesis of polyUb chains can be conducted by photochemical control. However, the OnbLys *Mm*PylRS variant was developed to incorporate OnbLys into

proteins in mammalian cells (*121*) and has not been demonstrated to work well for proteins in *E. coli* cells. The expression test (**Figure 3.9**) showed that there is little leaked expression (for example, aminoacylation of PylT with canonical lysine by this synthetase variant). However, mass spectrometry (MS) results (**Figure 3.10**) suggested that some of the Ub-OnbLys species might not have an intact caging (Onb) group. This was likely caused by post-translational modification, which was observed in UAA incorporation (*121*). Therefore, OnbLys was no longer considered for further method development.
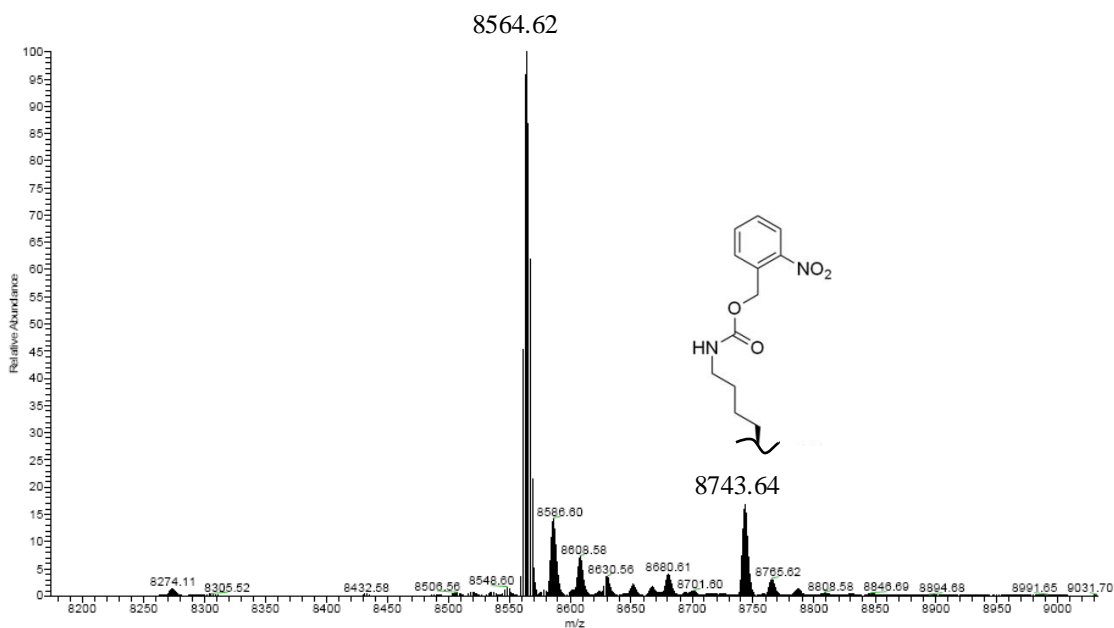
Figure 3.10. Deconvoluted ESI-MS result of Ub containing OnbLys. The expected mass for Ub-K48OnbLys and wild type Ub (K48) are 8744 and 8564, respectively.

3.2.1.3 Construction of Ub expression plasmid pTXB1-SynUb-K48tag

Expression of UAA-containing proteins usually has lower yield than expression of regular proteins. Individual factors in the expression system may need optimization to increase the yield of UAA-containing proteins. The previously used human Ub gene was first analyzed using Rare Codon Calculator (http://nihserver.mbi.ucla.edu/RACC). The

analysis result identified several rare codons for expression in *E. coli* cells (**Figure 3.10**). These rare codons might decrease protein yield or cause mis-incorporation of amino acids (as observed by NMR). A synthetic human Ub gene and homologous primers required to assemble this gene were designed using DNAWorks (www.helixweb.nih.gov) for optimum expression in *E. coli* cells. The DNA sequence of codon-optimized Ub is shown in **Figure 3.11**. This optimized Ub gene is referred to as SynUb (for synthetic Ub) in the following sections.

Red = rare Arg codons AGG, AGA, CGA
Green = rare Leu codon CTA
Blue = rare Ile codon ATA
Orange = rare Pro codon CCC

Human Ub (non-optimized)

```
atg cag atc ttc gtc aag acg tta acc ggt aaa acc ATA act CTA gaa gtt
gaa cca tcc gat acc atc gaa aac gtt aag gct aaa att caa gac aag gaa
ggc att cca cct gat caa caa AGA ttg atc ttt gcc ggt aaa cag ctc gag
gac ggt AGA acg ctg tct gat tac aac att cag aag gag tcg acc tta cat
ctt gtc tta AGA CTA AGA ggt ggt taa
```

Human Ub (codon-optimized)

```
atg cag ATT TTT GTG AAA ACC ctg acc GGC aaa acc ATT ACC CTG gaa GTG
gaa CCG AGC gat acc ATT gaa AAT GTG AAA GCG aaa att CAG GAT AAA gaa
ggc att CCG CCG gat CAG CAG CGT CTG ATT ttt GCG GGC AAA cag CTG GAA
GAT ggt CGT ACC ctg AGC gat tac AAT att cag AAA GAA AGC acc CTG cat
CTG GTG CTG CGT CTG CGT GGC GGC taa
```

Figure 3.11. Codon optimization of human Ub. The changed codons in SynUb gene are indicated by uppercased letters.

Next, expression tests were performed to determine if SynUb and PylRS variant Y384F could improve protein production. These expression tests were performed in the context of pET3a for convenient observation purpose (when expressed from pET3a, Ub can be well separated from background proteins when whole cell lysis is resolved on a protein gel). Expression tests were performed to understand 1) if the codon-optimized Ub gene can improve protein yield; 2) if Y384F mutation can improve the efficiency of

BocLys incorporation; 3) if BocLys incorporation depends on the position of "TAG" codon. The expression results (**Figure 3.12**) suggest that 1) codon-optimized Ub (SynUb) indeed improved expression efficiency (compare Lane 1 with Lane 9 or Lane 3 with Lane 11); 2) Y384F mutation enhanced BocLys incorporation (compare Lane 1 with Lane 3 or Lane 9 with Lane 11); 3) BocLys can be incorporated into K48 and K11 with equal or similar efficiency. It is worth mentioning that regular Ub (non-optimized) usually gives noisy NMR signal at rare arginine codon position due to mis-incorporation of amino acids while SynUb significantly alleviates this problem (confirmed by collaborator).
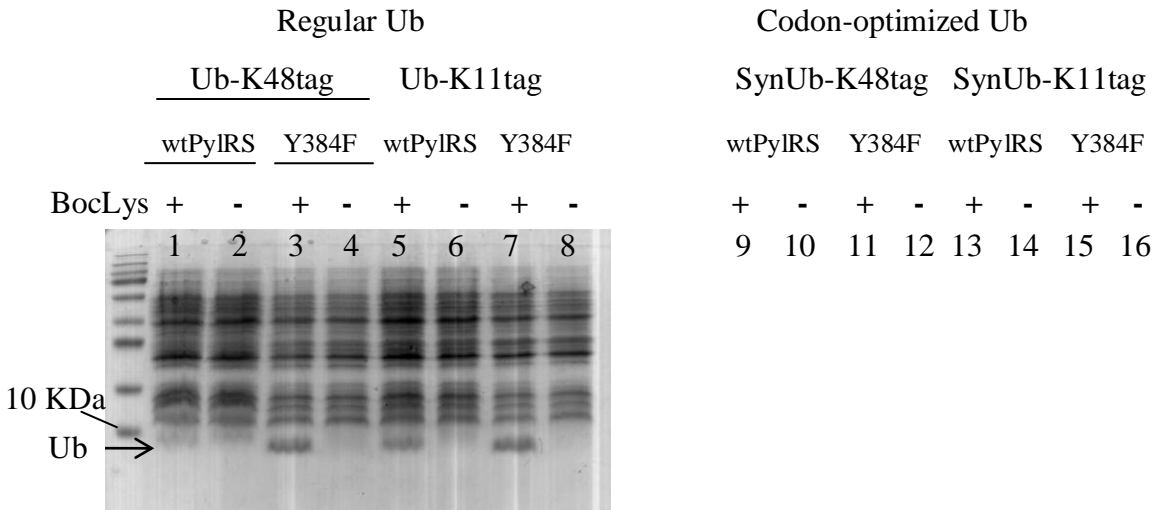
Figure 3.12. Expression tests for codon-optimized Ub (SynUb), PylRS variant Y384F and different "TAG" positions.

3.2.1.4 Expression and purification of Ub-K48BocLys, N^{15} labeled Ub_{1-74} and N^{15} labeled Ub_{1-74}-K48BocLys

Samples Ub-K48BocLys, N^{15}-labeled Ub_{1-74} and N^{15}-labeled Ub_{1-74}-K48BocLys were expressed and purified. The whole protein molecular weight (MW) of these samples was confirmed by electrospray ionization mass spectrometry (ESI-MS) (**Figure 3.13**). BocLys-containing Ub showed a minor peak with a mass corresponding to Boc-

deprotected Ub. This could be caused by sample handling during MS (samples were denatured/treated with acids prior to analysis). Leaked expression (natural lysine amino acid rather than BocLys had been incorporated) was unlikely the cause, because the expression results (**Figure 3.11**) suggested that the expression of Ub was strictly dependent on the supplement of BocLys.
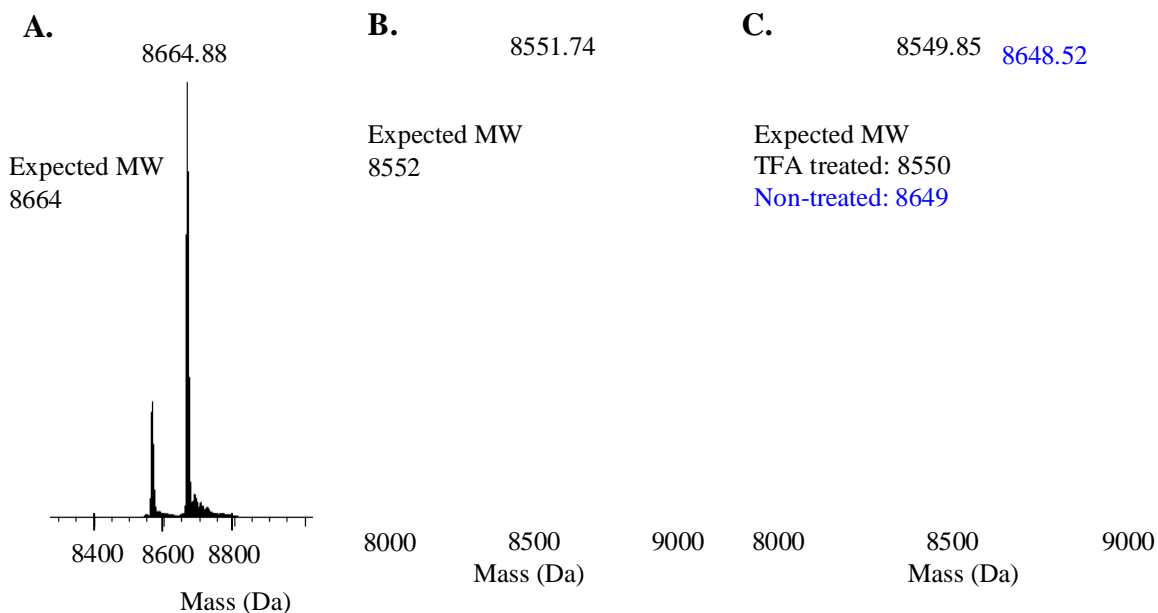
Figure 3.13. ESI-MS deconvolution of Ub monomer for whole-molecular weight determination. **A**) Ub-K48BocLys. **B**) ^{15}N Ub$_{1-74}$. **C**) ^{15}N Ub$_{1-74}$-K48BocLys. For ^{15}N labeled samples, 99%.incorporation of ^{15}N isotope is assumed to calculate the mass. Removal of the Boc group from Ub$_{1-74}$-K48BocLys results in 99 Da decrease in mass.

3.2.1.5 Deprotection of Boc protecting group

To confirm that Boc protecting group can be efficiently removed by TFA treatment, the ^{15}N Ub$_{1-74}$-K48BocLys sample was treated by 2% TFA, analyzed by ESI-MS and then compared with the non-treated sample (**Figure 3.13 C**). The MS result suggested that simple TFA treatment could completely remove the Boc protecting group. It is also worth mentioning that TFA-treated N^{15} Ub$_{1-74}$-K48BocLys is 2 Da less in mass than N^{15} Ub$_{1-74}$ because the BocLys amino acid used in this study is not ^{15}N labeled.

Another question is if the TFA treatment will disrupt the structure or function of Ub. To determine the effect of TFA treatment on the structure of Ub, TFA-treated N^{15} Ub$_{1-74}$-K48BocLys was analyzed by NMR. It is shown that the spectra of TFA-treated N^{15} Ub$_{1-74}$ K48BocLys and N^{15} Ub$_{1-74}$ can be overlayed (**Figure 3.14**). The only difference between these two spectra occurs at K48 position because BocLys (which became regular lysine after TFA treatment) is not ^{15}N labeled and thus does not give signals in NMR.
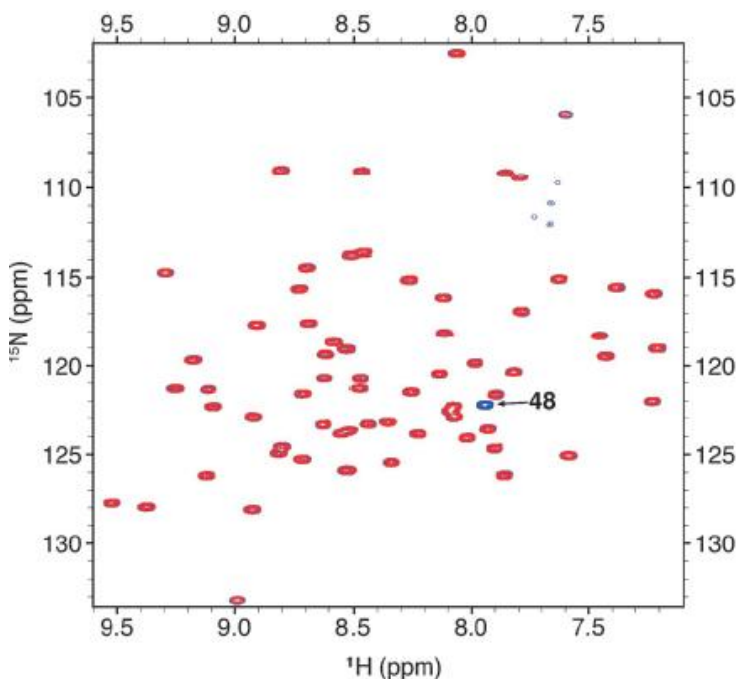
Figure 3.14. Overlay of ^{15}N–^{1}H TROSY-HSQC NMR spectra of TFA-treated Ub$_{1-74}$ K48BocLys monomer (red) and Ub$_{1-74}$ monomer (blue).

The effect of TFA treatment on Ub function was examined by subjecting the TFA-treated and non-treated Ub-K48BocLys sample to E1/E2-catalyzed polymerization reaction (**Figure 3.15**). K48BocLys mutant was first compared with the traditionally used distal chain terminator K48R variant. It is shown that K48BocLys can efficiently block chain polymerization from distal end. More importantly, TFA treated K48BocLys

generates a wild type-like monomer that allows chain polymerization from both distal and proximal ends. This suggests that TFA treatment does not impair the functionality of the deprotected lysine residue.
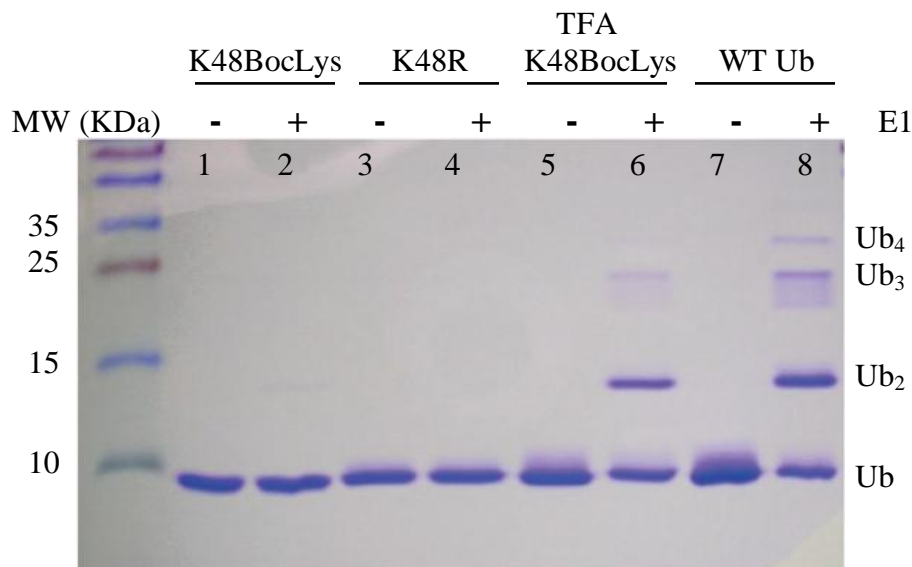
Figure 3.15. SDS-PAGE gels showing the corresponding polymerization reactions with and without E1 as controls.

3.2.1.6 Assembly of K48-linked Ub_2 and Ub_3 chains

Encouraged by the above experiments, Ub_2 was assembled using N^{15} Ub_{1-74} (proximal monomer) and Ub-K48BocLys (distal monomer) (**Figure 3.16**, Lane 1 and 2). Removal of the Boc protecting group from the synthesized Ub_2 allows further chain polymerization (**Figure 3.16**, Lane 4 and 5). These results indicated that Ub-K48BocLys could halt chain polymerization at Ub_2. This Ub_2 was purified and treated with TFA to remove the Boc protecting group. The MW of the Ub_2-K48BocLys sample before and after TFA treatment was confirmed by ESI-MS (**Figure 3.17**).
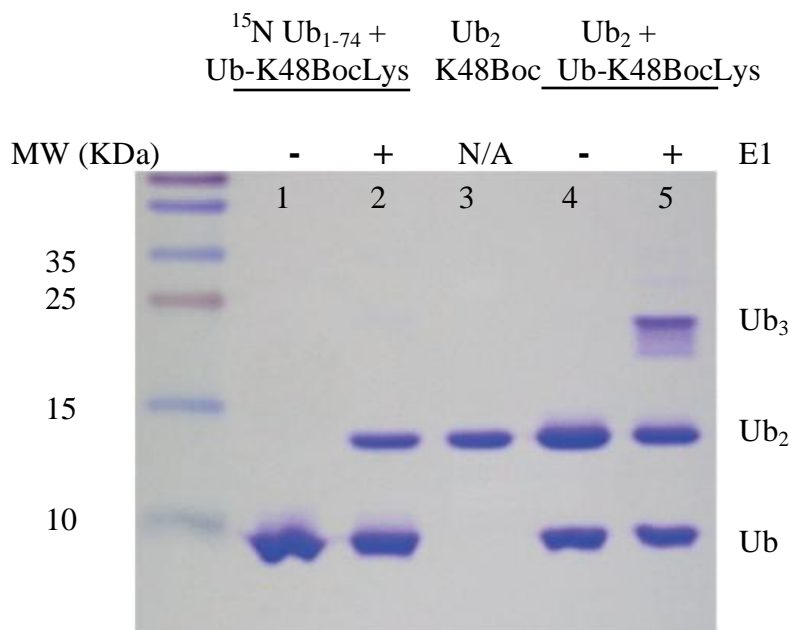
Figure 3.16. Ub_2 and Ub_3 synthesis. Combinations of ^{15}N Ub_{1-74} and ^{15}N Ub-K48BocLys halt polymerization at Ub_2 (Lane 1 and 2). Purified Ub_2-K48BocLys (Lane 3) can be iterated to produce Ub_3 (Lane 4 and 5).
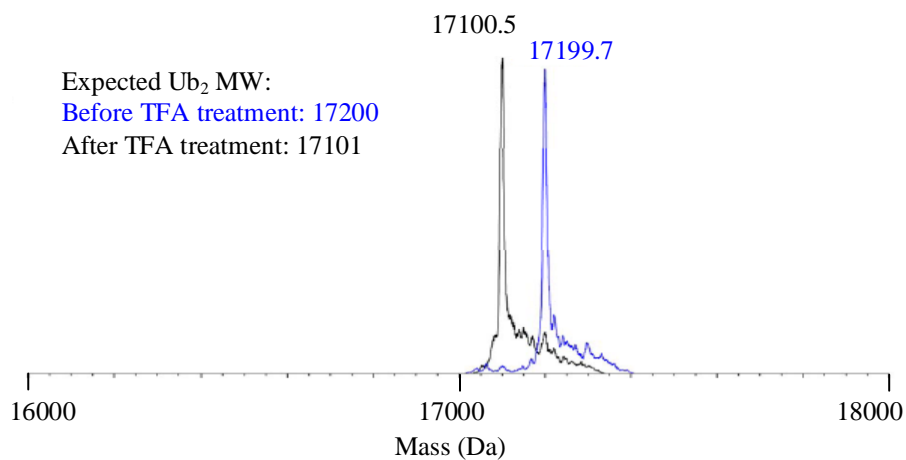
Figure 3.17. ESI-MS deconvolution of K48-linked di-Ub comprised of the Ub-K48BocLys monomer and ^{15}N-Ub_{1-74} monomer before (blue) and after TFA treatment (black). TFA treatment is expected to result in a mass loss of 99 Da due to the elimination of one Boc protecting group.

Although TFA treatment has been demonstrated to not disrupt the structure and function of Ub monomer, its influence on Ub_2 remains unclear as Ub_2 is a structurally and functionally distinct protein and might be fragile to TFA treatment. To test this, TFA-

treated ^{15}N-P Ub$_2$-K48BocLys (the proximal unit is ^{15}N labeled Ub$_{1-74}$) was purified, buffer-exchanged and then analyzed by NMR at pH 4.5 and pH 6.8. The obtained NMR spectra were overlayed with those of ^{15}N Ub$_{1-74}$ at pH 4.5 and pH 6.8, respectively.

The NMR spectra of Ub$_2$ show excellent spectral quality and chemical shift perturbation (CSP) (**Figure 3.18**), suggesting that TFA treatment did not irreversibly denature Ub$_2$. Furthermore, comparison of the spectra at pH 4.5 and pH 6.8 suggest that the conformational properties of Ub$_2$ are fully retained. K48-Ub$_2$ is known to exhibit two properties: i) a hydrophobic Ub/Ub interface centered at residues L8, I44 and V70 (*122, 123*); ii) a pH-controlled equilibrium between a closed conformation (predominant at pH 6.8) and open conformation(s) (predominant at pH 4.5) (*122, 124-126*). These properties can be readily determined using NMR spectroscopy (*122*). Our NMR data (**Figure 3.18**) demonstrate that the synthesized Ub$_2$ exhibits all of the above features. At pH 6.8, many residues in the proximal Ub showed significant spectral perturbations with respect to monomeric ^{15}N Ub$_{1-74}$ (**Figure 3.18D**). The perturbations center around the hydrophobic-interface residues L8, I44 and V70, in strong agreement with the closed conformation of K48-Ub$_2$ (*122*). At pH 4.5 (**Figure 3.18C**), significant chemical shift perturbations are only observed for residues near the site of the K48 isopeptide linkage. The lack of perturbations for residues centered around L8 and V70 indicates the absence of the hydrophobic interface at low pH, in full agreement with the previous observations for K48-linked Ub$_2$ (*122*). The CSPs of the proximal Ub unit in the synthesizd Ub$_2$ and the dynamic nature of Ub$_2$ under different pH are very similar to the characteristics of K48-Ub$_2$ preapred by other means (**Figure 3.18E and F**).
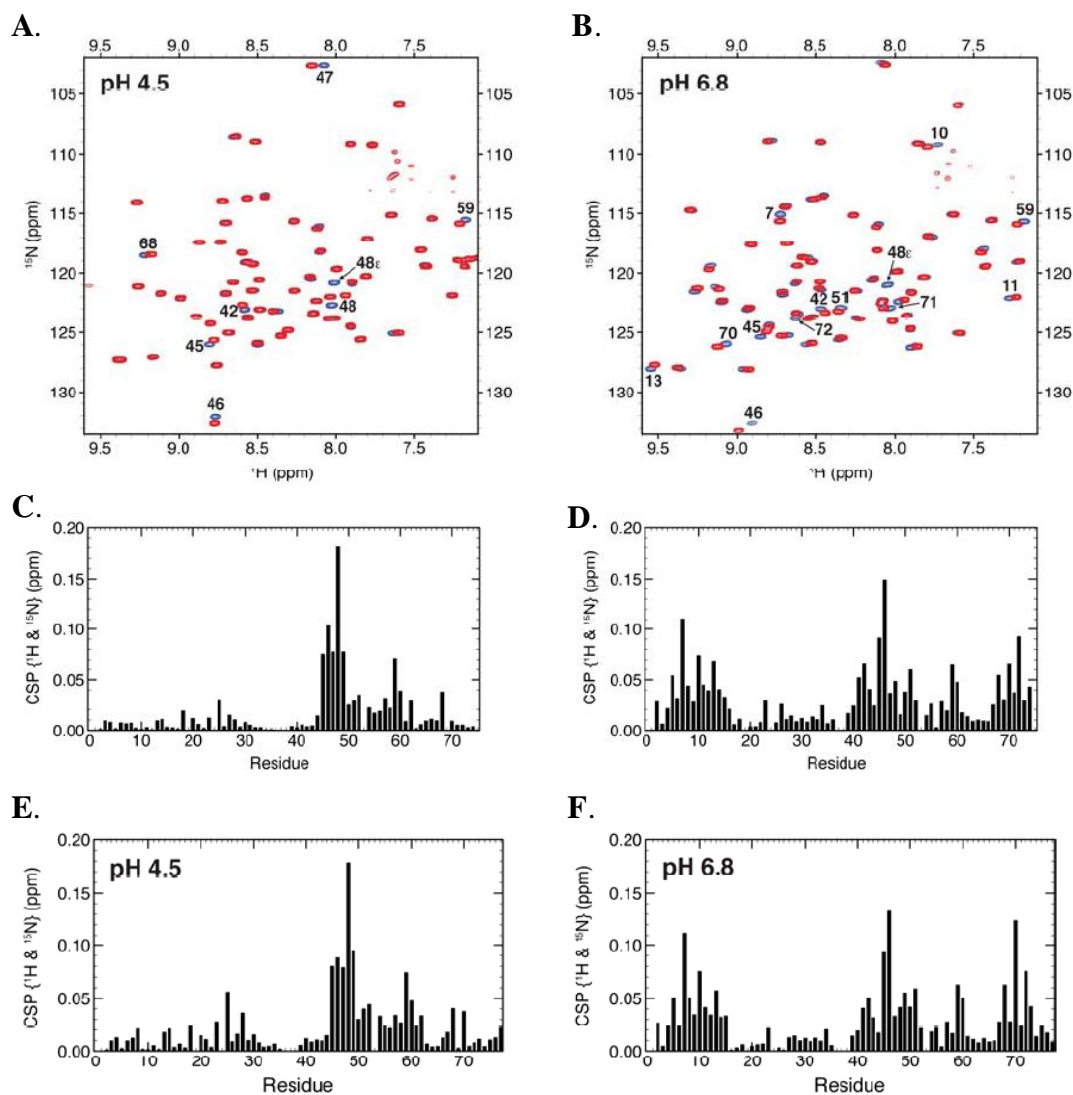
Figure 3.18. NMR characterization of K48-Ub$_2$. ^{15}N-^{1}H TROSY-HSQC spectra of the proximal ^{15}N Ub$_{1-74}$ (blue) in TFA-treated Ub$_2$-K48BocLys and monomeric ^{15}N Ub$_{1-74}$ (red), obtained at pH 4.5 (**A**) and pH 6.8 (**B**). CSPs of the proximal Ub$_{1–74}$ unit in Ub$_2$ *versus* Ub$_{1–74}$ monomer are determined at pH 4.5 (**C**) and pH 6.8 (**D**), respectively. Corresponding CSPs (**E**) and (**F**) of K48-Ub$_2$ constructed by proximal ^{15}N Ub$_{D77}$ and distal Ub-K48R using traditional methods are shown as comparison.

This new method is preferred over the exsiting enzymatic synthesis of polyUb as it generates "all native" polyUb chains. Since this methodology is limited by the available linkage-specific E2 enzymes, an E2 enzyme-free method was also developed to assemble polyUb chains of any linkage composition.

3.2.2 Nonenzymatic synthesis of polyUb chains using Ub-BocLys monomers

3.2.2.1 Preparation of ligation precursors Ub-K11BocLys, Ub-K33BocLys and Ub-SR

Although the efficiency of UAA incorporation into different "TAG" position may vary depending on the context of DNA sequence, our expression system (pSup-BocLys and pTXB1-SynUb) allows generation of Ub-BocLys with moderate yield (~10 mg/mL) for six of the seven lysine variants. The only low-yield (~3 mg/mL) construct is Ub-K63BocLys where a considerable amount truncated protein (Ub_{1-63}) is observed. The low efficiency of BocLys incorporation at position 63 is likely caused by the inefficient suppression of K63tag stop codon by PylT.

Here, we focus on K11- and K33-linked polyUb chains as a demonstration for the nonenzymatic synthesis of polyUb chains. Ub-K11BocLys and Ub-K33BocLys were produced as above and confirmed by ESI-MS. These BocLys-containing Ubs will serve as the proximal Ub monomer. The C-terminus of the distal monomers must be first activated with a thioester functional group for the ligation reaction (**Figure 3.6B**). Traditionally, Ub-SR is made *via* intein-mediated cleavage reaction (*99*). Unfortunately, this approach is slow (each cleavage reaction takes 24~40 h), necessitates HPLC purification (removal of an unwanted byproduct of non-cleaved Ub), and produces Ub-SR in low yield (only 10 mg of pure Ub-SR from expression in 2 L culture). An alternative approach is to react WT Ub with Ub-activating E1enzyme in the presence of MESNA (*119*). This approach turned out to be very efficient in our hand. Complete thioesterification of Ub C-terminus occurs after 6 h of treatment with the E1 enzyme and MESNA, as confirmed by ESI-MS (**Figure 3.19A**). Subsequent treatment with glacial acetic acid precipitates E1 and lowers the pH to maintain Ub-SR stability. Finally, buffer

exchange removes other reaction components to yield a pure Ub-SR. This method produces Ub-SR at high yield, because the starting protein, WT Ub, is very easy to express and purify in large amounts (50~100 mg from1 L culture). Importantly, the same procedure works nicely for activating Ub-BocLys as well as all preassembled polyUb chains (see below).
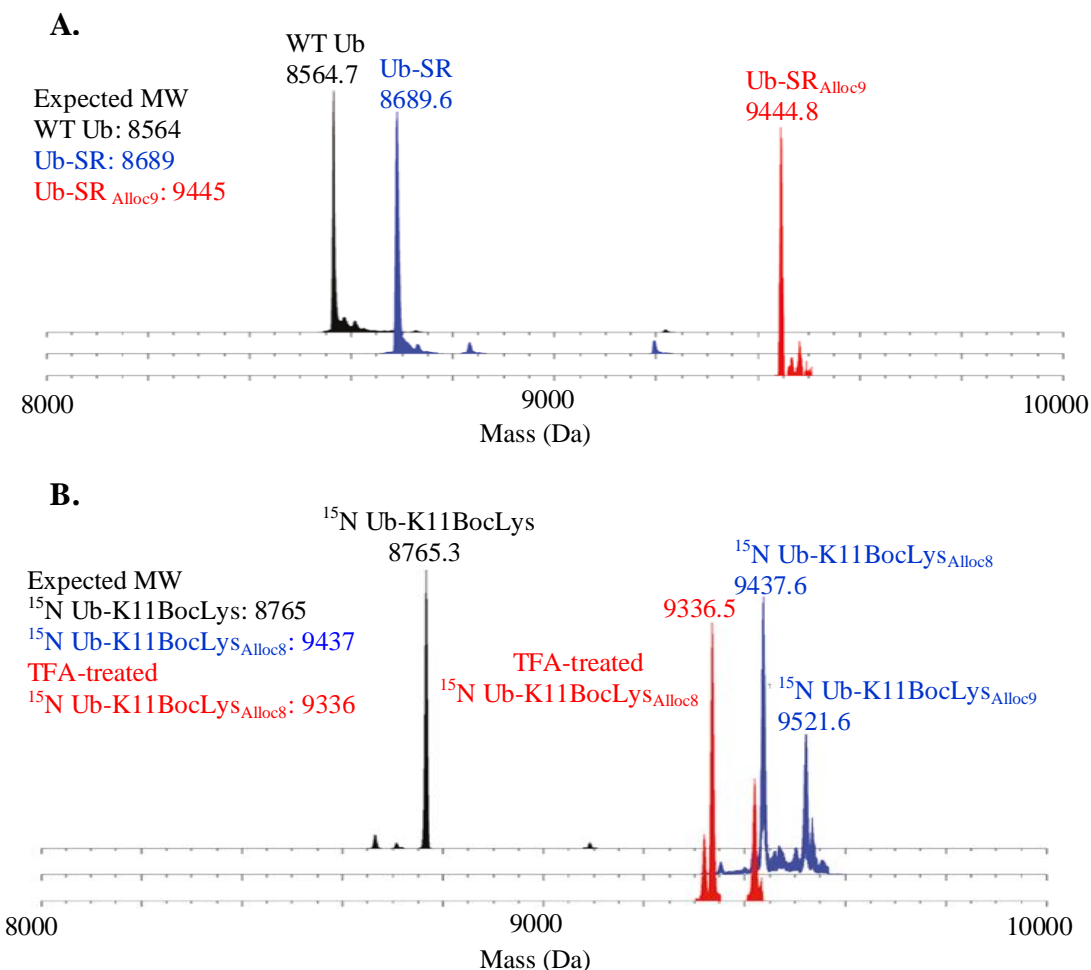
Figure 3.19. ESI-MS deconvolution to confirm thioesterification, Alloc protection and Boc deprotection. **A**) Preparation of distal monomer for condensation reaction. Thioesterfication of the C-termuns of WT Ub results in a 125 Da increase in mass for the addition of 2-mercaptoethane sulfonate (MES). Alloc protection adds 9 Alloc groups (84 Da each) and fully protects all free amines. **B**) Preparation of proximal monomer for condensation reaction. Alloc protection of [15]N Ub-K11BocLys adds 8 or 9 Alloc groups. Deprotection of this [15]N Ub-K11BocLys_Alloc8 removes the 100 Da Boc group while leaving all Alloc groups intact.

3.2.2.2 Alloc protection and Boc deprotection to prepare Ub monomers for

condensation reaction

To generate Ub_2, both of distal Ub-SR and proximal Ub-BocLys need be protected by

Alloc protecting group to produce Ub-SR$_{Alloc8}$ and Ub-BocLys$_{Alloc7}$. The Alloc protected

distal and proximal monomers were analyzed by ESI-MS (**Figure 3.19**) and the complete

protection of free amines in both monomers was confirmed. Interestly, one additional

Alloc protecting group was observed in both monomers. This was presumably caused by

the protection or partial protection of the single histidine (His68) in Ub. This

overprotection, however, will not cause any problem in the present method as the

deprotection of Alloc group can go to completion efficiently (see below). Ub-

BocLys$_{Alloc7}$ is then Boc-deprotected to produce Ub_{Alloc7} for condensation reaction.

3.2.2.3 Silver-mediated condensation reaction

Ub-SR$_{Alloc8}$ and Ub-BocLys$_{Alloc7}$ were subjected to a silver-mediated condensation

reaction (*118*). The ligation converted ~30% of the reacting Ub monomers into Ub_2

(**Figure 3.20**), consistent with the reported yield (*99*).
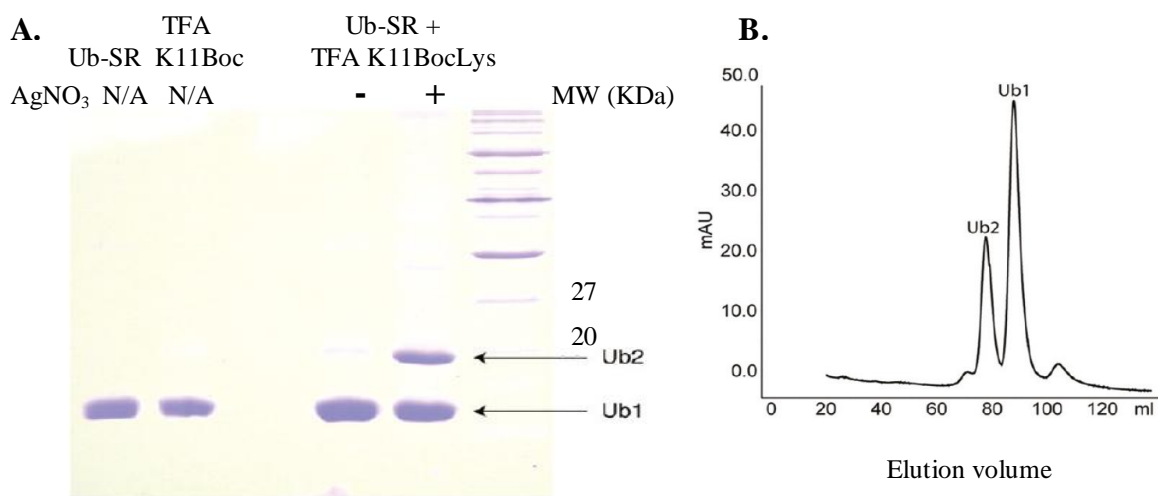
Figure 3.20. Silver-mediated condensation reaction. **A**) Formation of Ub_2. **B**)
Purification of the Ub_2 by size exclusion chromatography.

3.2.2.4 Global deprotection of Alloc groups and protein renaturation

Central to the success of this method is the complete Alloc deprotection. Recently, it has been shown that Alloc protecting groups can be efficiently removed using relatively mild conditions in neutral aqueous solution using thiophenol and a ruthenium (Ru) catalyst chloro-pentamethylcyclopentadienylcyclooctadiene-ruthenium(II) ([Cp*Ru(cod)Cl]) (*127*). However, application of these exact conditions to Ub protein, which can have up to nine Alloc groups attached to it simultaneously, failed to completely remove all Alloc groups. To determine the correct conditions for Alloc deprotection, multiple experimental conditions were tested on a ^{15}N-labeled sample of WT Ub (**Figure 3.21**). After many rounds of optimization, the correct amounts of Ru catalyst, thiophenol, and water to use in the Alloc deprotection reaction were determined.
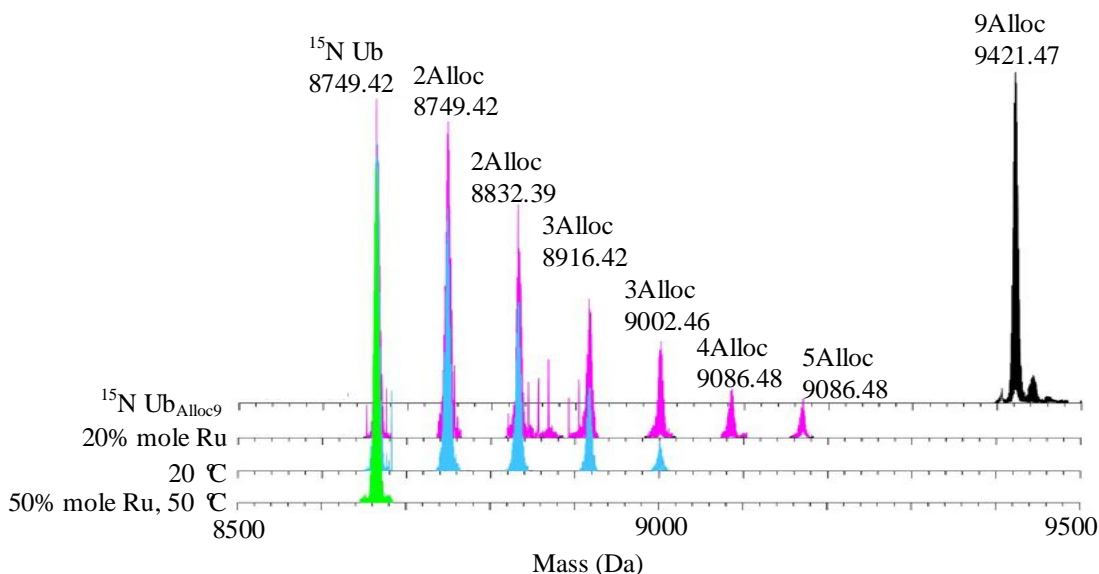
Figure 3.21. Deconvoluted ESI-MS for condition optimization of Alloc deprotection reactions. [15]N WT Ub is used as a test protein. Fully protected protein with nine Alloc groups was subjected to deprotection reactions with variable conditions. Optimal Alloc deprotection conditions require the use of 50% mol Ru catalyst, 50 equivalent of thiophenol, 30% H_2O and incubation at 50 °C for 2 h (green). If less catalyst (magenta) or lower temperature (sky blue) are used, the Alloc deprotection reaction does not go to completion, resulting in a distribution of multiple Ub species containing a variable number of Alloc groups.

The most critical features to the success of the reaction are (i) near-stoichiometric amounts of Ru catalyst (50% mol) relative to the moles of amines present, (ii) at least 30% H_2O (v/v), and (iii) temperature raised to at least 50 °C (denaturing condition). When any of these conditions was not met, Ub species typically still contained a variable number of Alloc protecting groups, as revealed by ESI-MS analysis (**Figure 3.21**). For example, lowering the amount of Ru metal in the reaction did not yield full deprotection. Under correct conditions, the deprotection reaction typically reached completion after a maximum of 2 h. Allowing reactions to proceed overnight did not change the ESI-MS results.

As the Alloc deprotection is performed under denaturing conditions, the final product needs be renatured for further characterization. In the present study, all the synthesized polyUb chains could be renatured efficiently. Renatured Ub_2 was characterized by ESI-MS to verify the complete deprotection of Alloc group (**Figure 3.22**). Importantly, Ub_2 can be either ^{15}N labeled at the distal unit (^{15}N-D Ub_2, **Figure 3.22A**) or at the proximal unit (^{15}N-P Ub_2, **Figure 3.22B**). The ability to label different Ub units in polyUb chains allows detailed NMR studies on the whole polyUb chains. Furthermore, as mentioned above, Ub_3 or higher polyUb chains can be polymerized from Ub_2 from either proximal end or distal end (**Figure 3.6C**). If Ub_3 is to be polymerized from the distal end of Ub_2, the synthesis of pre-assembled Ub_2 requires using a Ub-BocLys (rather than WT Ub) distal monomer for thioesterification reaction, such that the final Ub_2 will contain a BocLys at the distal Ub. It is very important to demonstrate that this BocLys is able to survive the Alloc protection/deprotection. ESI-MS results (**Figure 3.22B**) confirmed that this BocLys is indeed kept intact after Alloc protection/deprotection.

A.

K11-linked ^{15}N-D Ub$_2$

17210.9

Expected MW:
17210

17000   18000

Mass (Da)

B.

K11-linked ^{15}N-P Ub$_2$-K11BocLys

17312.2

Expected MW:
17310

17000   18000

Mass (Da)
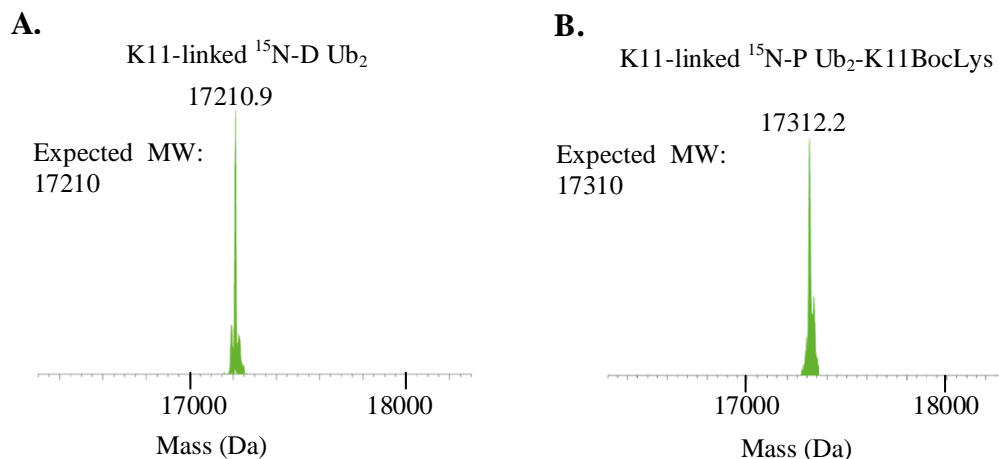
Figure 3.22. Deconvoluted ESI-MS spectra of Ub$_2$. **A**) K11-linked ^{15}N-P Ub$_2$. **B**) K11-linked ^{15}N-D Ub$_2$ containing a BocLys at K11 of the distal Ub.

3.2.2.5 Assembly and characterization of K11- and K33-linked Ub$_2$

With the ability to selectively label either the proximal or distal unit of Ub$_2$, both ^{15}N-P and ^{15}N-D versions of K11- and K33-linked Ub$_2$ were synthesized for detailed NMR studies. The NMR spectra of these polyUb chains revealed an excellent spectral dispersion of NMR resonances, indicative of a fully renatured Ub$_2$. No minor peaks were observed, indicating that the chemically ligated Ub$_2$ is free of contaminating species.

   K11-Ub$_2$ was first synthesized because its structural properties can be easily compared with the enzymatically synthesized K11- Ub$_2$ (by K11-specific E2 conjugating enzyme Ube2s). K11 linkages could be as abundant as K48 and K63 linkages (*128*) and K11-linked polyUb chains appear to act as both regulatory and proteolytic signals (*102, 129*). ^{1}H-^{15}N TROSY-HSQC (*130*) NMR experiments were performed on K11- Ub$_2$ (as well as other polyUb chains throughout this dissertation) to detect primarily the signals from the backbone ^{1}H-^{15}N amide bonds (**Figure 3.23**).
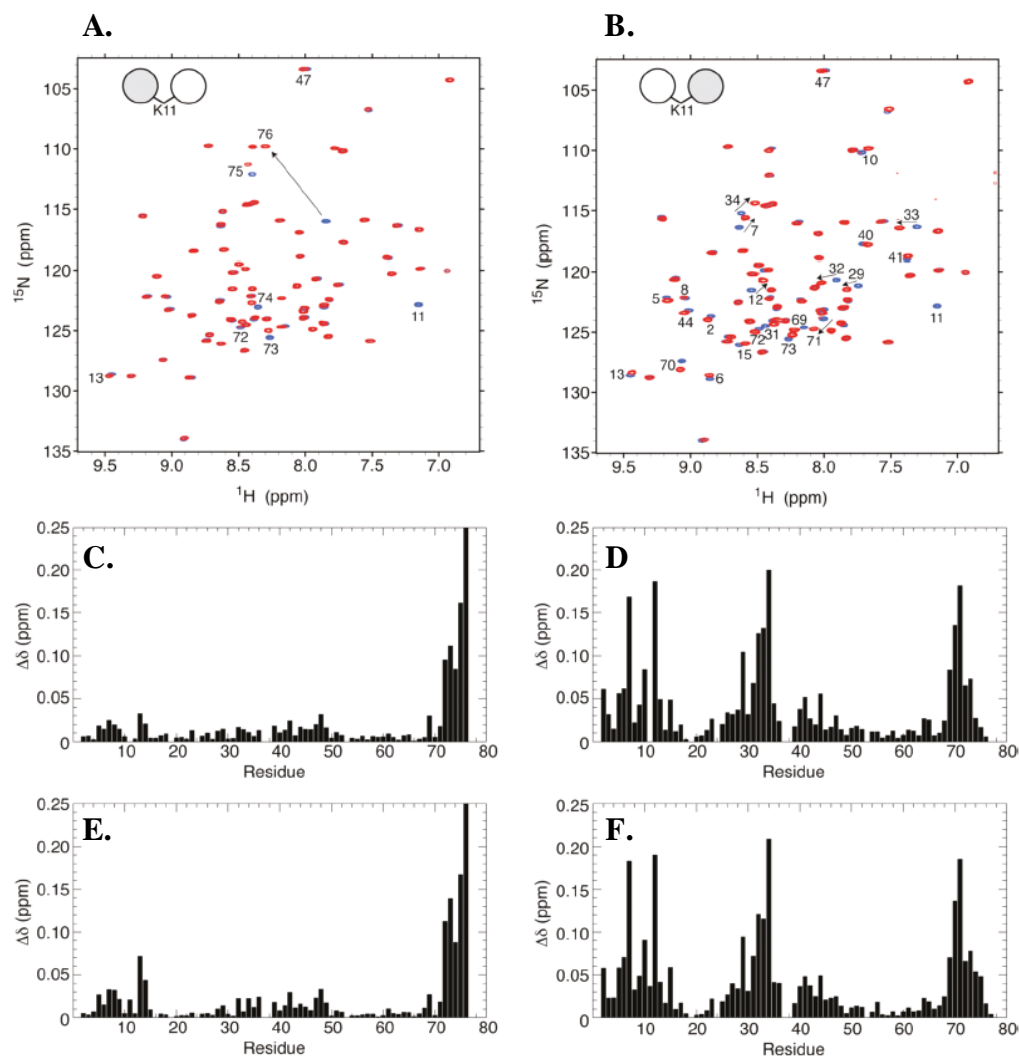
Figure 3.23. NMR studies of K11-Ub$_2$. **A)** and **B)** are ^{1}H-^{15}N TROSY-HSQC spectra of ^{15}N-D Ub$_2$ and ^{15}N-P Ub$_2$, respectively. The grey and white balls represent ^{15}N-labeled and non-labeled Ub units, respectively. **C)~F)** are the CSPs between each Ub$_2$ chains and corresponding unconjugated Ub monomer. For example, if proximal ^{15}N Ub$_{D77}$ is used to assemble Ub$_2$, ^{15}N-P Ub$_2$ will be compared with ^{15}N Ub$_{D77}$ rather than ^{15}N WT Ub. **C)** and **D)** are nonenzymatically synthesized ^{15}N-D Ub$_2$ and ^{15}N-P Ub$_2$, respectively, and **E)** and **F)** are enzymatically synthesized ^{15}N-D Ub$_2$ and ^{15}N-P Ub$_2$, respectively (from previous studies by Dr. Fushman's lab).

On the distal Ub, the largest signal shifts as compared with Ub monomer are observed at the C-terminal residues G75 and G76 which involve in the formation of isopeptide bond. The K11 backbone amide signal is not present in the ^{15}N-D Ub spectrum (**Figure 3.23A**) because Ub-K11BocLys was used as the distal precursor to assemble this Ub$_2$.and

BocLys incorporated at K11 is unlabeled and thus invisible in the ^{15}N NMR spectra. Likewise, the K11 backbone amide signal is absent also in the spectrum of ^{15}N-P Ub (**Figure 3.23B**).

To validate that the nonenzymatic approach itself will not change the properties of Ub$_2$, the spectra of these nonenzymatically synthesized K11-Ub$_2$ were compared with that of enzymatically synthesized K11-Ub$_2$. Unfortunately, the chain-terminating mutations carried by the enzymatically assembled K11-Ub$_2$ (distal Ub-K11R and proximal Ub$_{D77}$) complicate direct comparison of the spectra between the two Ub$_2$ chains. Therefore, we compared each Ub unit in these Ub$_2$ to its respective unconjugated Ub monomer. For example, if proximal Ub$_{D77}$ is used to assemble Ub$_2$, ^{15}N-P Ub$_2$ will be compared with Ub$_{D77}$ rather than WT Ub. The remarkable similarity between the CSPs of the nonenzymatically and enzymatically assembled K11-Ub$_2$ chains (compare **Figure 3.22 C** with **E**, and **D** with **F**) serves as a strong indicator that the two chains have identical structural and conformational properties.

Notably, the two Ub units in K11-linked Ub$_2$ show strikingly different CSP patterns (compare **Figure 3.23 C** with **D**). Significant CSPs are spread throughout the proximal Ub but almost absent in the distal Ub (except for the ligated C-terminus). This is in contrast with the CSP patterns observed in K48-linked (*122*) and K63-linked (*131*) Ub$_2$ chains. The CSP differences could be caused by 1) nonconvalent interdomain contacts, 2) chemical modification (isopeptide bond), or 3) a combination of both. The 3D structure of Ub suggests that the relatively big CSPs in the proximal Ub unit are very likely caused by the isopeptide formation. To test this hypothesis, we measured the CSPs between Ub-K11BocLys and WT Ub (**Figure 3.24B**), where the only difference between the two

proteins is whether a Boc protecting group is present on the side chain of K11. Indeed, the results show a CSP pattern remarkably similar to that in the proximal Ub of K11-Ub$_2$ (compare **Figure 3.23D** with **Figure 3.24B**). This suggests that the spectral perturbations observed in the proximal Ub of K11-Ub$_2$ are primarily caused by the chemical modification of the side chain, a result of the isopeptide formation.
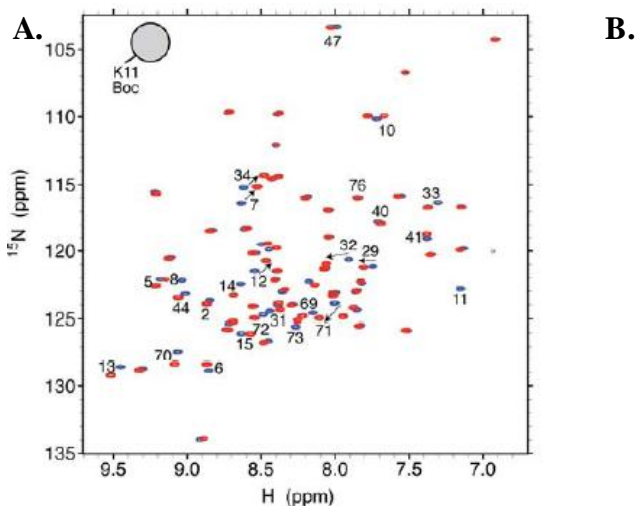
Figure 3.24. NMR studies of Ub-K11BocLys. **A)** ^{1}H-^{15}N TROSY-HSQC spectra of ^{15}N Ub-K11BocLys. **B)** CSP between Ub-K11BocLys and WT Ub monomer.

Encourage by the success of the synthesis of K11-linked Ub$_2$, we next assembled ^{15}N-P K33-Ub$_2$ to demonstrate the generality of our method. Again, the assembled Ub2 was confirmed by ESI-MS (**Figure 3.25**) and subjected to NMR analysis.
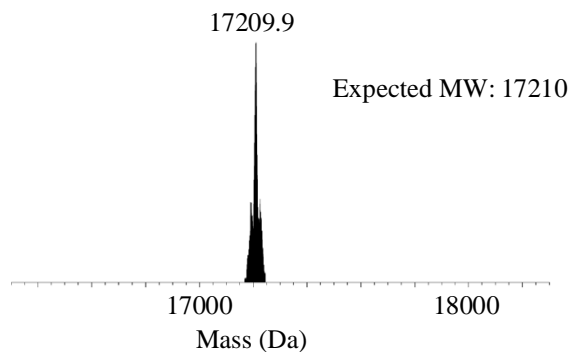
Figure 3.25. ESI-MS analysis of ^{15}N-P K33-linked Ub$_2$
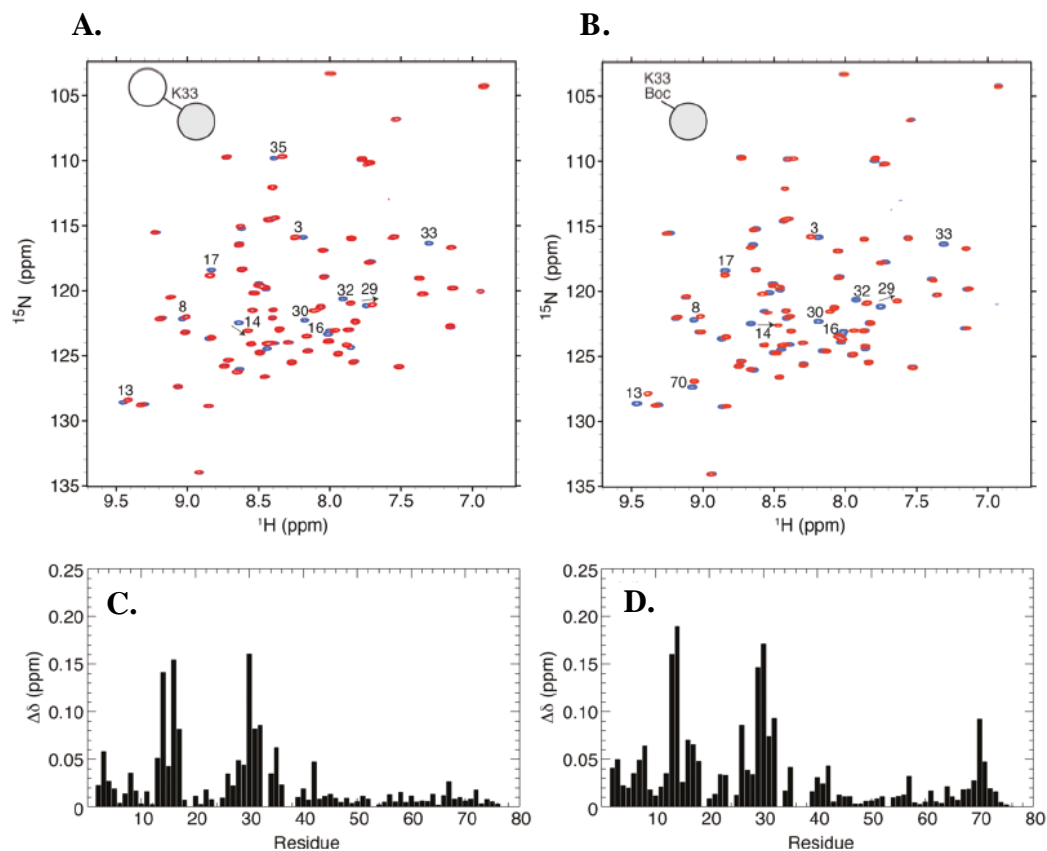
Figure 3.26. NMR studies of K33-Ub$_2$. **A)** and **B)** are ^{1}H-^{15}N TROSY-HSQC spectra of ^{15}N-P K33-Ub$_2$ and Ub-K33BocLys, respectively. **C)** and **D)** are the CSPs of the proximal Ub in K33-Ub$_2$ and Ub-K11BocLys (both *versus* WT Ub monomer), respectively.

 NMR characterization of the proximal Ub in K33-Ub$_2$ revealed an excellent spectral dispersion in both ^{15}N and ^{1}H dimensions (**Figure 3.26A**), strongly indicating that this Ub unit is well folded. The K33 signal in this Ub$_2$ chain is, again, absent due to the unlabeled BocLys. The CSP pattern of the proximal Ub in this K33-Ub$_2$ is distinct from that in the K11-Ub$_2$ (compare **Figure 3.26C** with **Figure 3.23D**). Interestingly, in K33-Ub$_2$, the CSPs are quite small for the hydrophobic patch residues L8, I44, and V70, and instead cluster around residues12~17 and 26~35 located close in space or adjacent to K33 (**Figure 3.26C**). This suggests that the majority of the spectral perturbations observed in

this Ub unit are a consequence of the chemical modification of K33 by the isopeptide linkage. To support this hypothesis, the spectra of Ub-K33BocLys and WT Ub were overlaid (**Figure 3.26B**) and the corresponding CSP is calculated (**Figure 3.26D**). Indeed, the CSP pattern of Ub-K33BocLys (*versus* WT Ub) is overall similar to that of the proximal unit in K33-Ub$_2$ (*versus* WT Ub). This suggests that the chemical modification is a major, if not all, cause of the CSP observed in the proximal unit in K33-Ub$_2$.

3.2.2.6 Assembly and characterization of K11-linked Ub$_3$ and K11/K33-linked Ub$_3$

The power of our approach lies in that pre-assembled Ub$_2$ can be easily extended to produce either homogeneous or mixed linkage Ub$_3$ or higher polyUb chains. With this method in hand, homogeneous K11-Ub$_3$ and mixed linkage K11/K33-Ub$_3$ chains were synthesized (**Figure 3.27**). In this dissertation, the nomenclature of mixed linkage polyUb chains follows the rule that the proximal linkages are placed in front of the distal linkages. Taking the K11/K33-Ub$_3$ for example, the middle Ub unit has its K33 linked to the C-terminus of the distal unit and its C-terminus linked to the K11 in the proximal unit. The yield of the condensation reaction for K11- Ub$_3$ and K11/K33- Ub$_3$ were 15% and 9%, respectively. The relatively low yields indicate that the conditions of these silver-based condensation reactions may require further optimization. The produced Ub$_3$ were purified by size-exclusion chromatography (**Figure 3.27B**) and then confirmed by ESI-MS as above. In both of these two Ub$_3$, the middle units were ^{15}N-labeled, enabling unit-specific NMR studies.

NMR spectra of K11-linked Ub$_3$ were first analyzed. The excellent spectral dispersion of both ^{1}H and ^{15}N resonances of the backbone amides (**Figure 3.28**) clearly indicates that the middle Ub is fully folded despite being tethered to two other Ub units in the

chain. The observed spectral differences between the middle Ub in K11-Ub$_3$ and Ub monomer are a composite of the perturbations observed above in the distal and proximal Ub units of K11-Ub$_2$ (**Figure 3.23**). This is not unexpected given that the middle Ub unit is both proximal and distal with regard to the two Ub units flanking it.
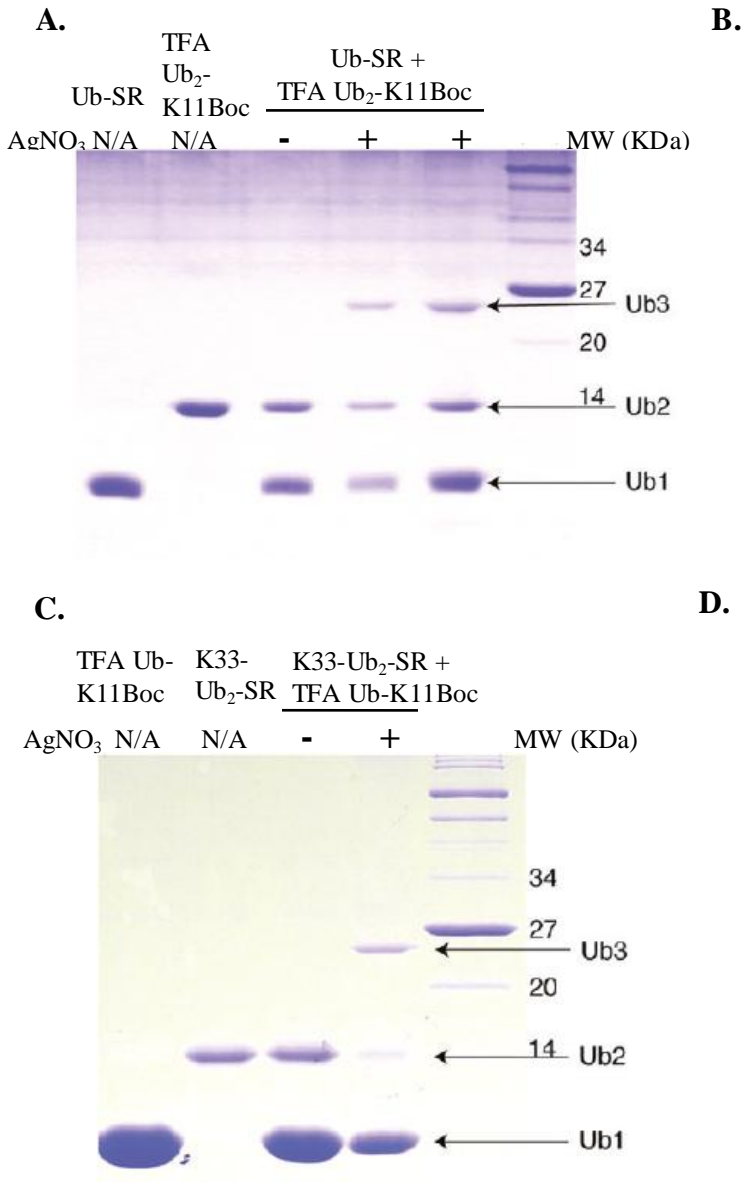
Figure 3.27. Assembly of K11-Ub$_3$ and K11/K33-Ub$_3$. **A)** and **C)** are coomassie-stained SDS-PAGE gel showing the formation of K11-Ub$_3$ and K11/K33-Ub$_3$, respectively. **B)** and **D)** are SEC purification of K11-Ub$_3$ and K11/K33-Ub$_3$, respectively.

The CSPs of the middle Ub in K11-Ub$_3$ *versus* WT Ub monomer (**Figure 3.28D**), distal Ub in K11-Ub$_2$ (**Figure 3.28E**), and proximal Ub in K11-Ub$_2$ (**Figure 3.28F**) were calculated. Remarkably, the CSP of this middle unit in Ub$_3$ *versus* the distal Ub in K11-Ub$_2$ (**Figure 3.28E**) is similar to the CSPs of the proximal Ub in K11-Ub$_2$ *versus* WT Ub monomer (**Figure 3.23D**). Likewise, the CSP of this middle unit in K11-Ub$_3$ *versus* the proximal Ub of K11-Ub$_2$ (**Figure 3.28F**) is similar to the CSPs of the proximal Ub in K11-Ub$_2$ *versus* WT Ub (**Figure 3.23C**). Again, this suggests that the oberserved CSPs of the middle Ub in K11-Ub$_3$ simply reflects the additive CSPs from the proximal and the distal units, meaning that no additional Ub/Ub interface forms when K11-Ub$_2$ is extended to K11-Ub$_3$. Structurally, this suggests that the proximal and distal Ub units in Ub$_3$ are distant from each other and no inter-domain "cross-talk" exists.
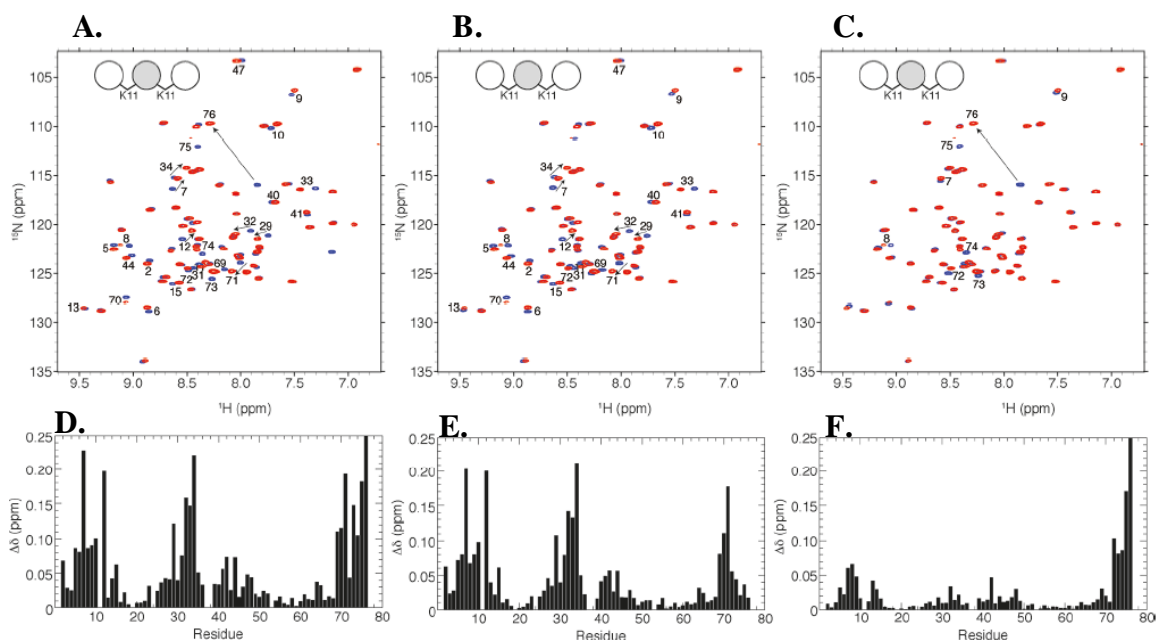
Figure 3.28. NMR studies of K11-linked Ub$_3$. **A)~C)** are overlaid ^{1}H-^{15}N TROSY-HSQC spectra of ^{15}N-M Ub$_3$ (red) with ^{15}N WT Ub (**A**), ^{15}N-D K11-Ub$_2$ (**B**), and ^{15}N-P K11-Ub$_2$ (**C**) in blue, respectively. The grey and white balls represent ^{15}N-labeled and non-labeled Ub units, respectively. **D)~F)** are the CSPs of the middle unit in K11-Ub$_3$ *versus* WT Ub (**D**), distal Ub in K11-Ub$_2$ (**E**) and proximal Ub in K11-Ub$_2$ (**F**).

Perhaps one of the most interesting features of the current method is that it allows generation of mixed linkage polyUb chains, which can be not easily obtained using traditional methods. We then set up to synthesize, for the first time, a mixed-linkage Ub_3 chain comprised of K11 and K33 linkages (**Figure 3.27C**). As stated above, this polyUb construct is named as K11/K33-Ub_3, designating that K11 is the first isopeptide bond counting from the proximal end. As above, the produced Ub_3 was SEC purified, confirmed by ESI-MS, and then analyzed by NMR (**Figure 3.29**).
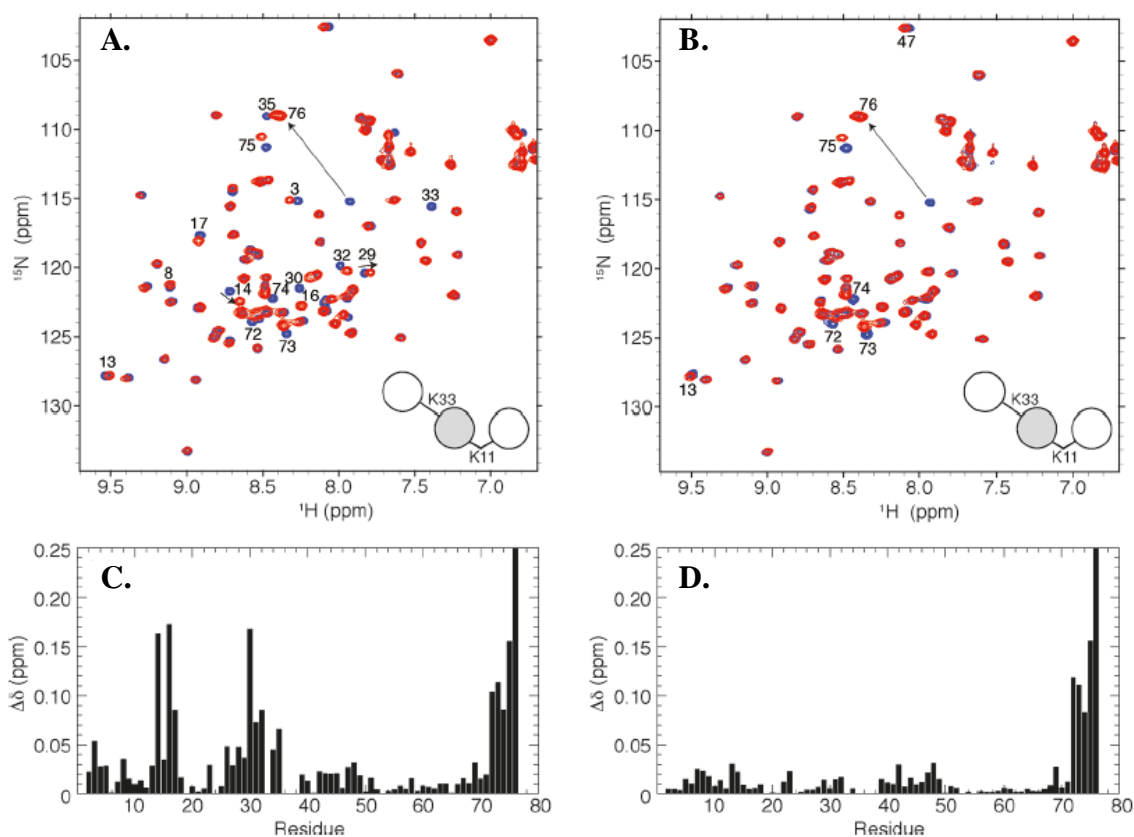
Figure 3.29. NMR studies of K11/K33-Ub_3. **A)** and **B)** are overlaid ^1H-^{15}N TROSY-HSQC spectra of ^{15}N-M Ub_3 (red) with ^{15}N WT Ub (**A**) and ^{15}N-P K33-Ub_2 (**B**). The grey and white balls represent ^{15}N-labeled and non-labeled Ub units, respectively. **C)** and **D)** are the CSPs of the middle unit in K11-Ub_3 *versus* WT Ub and proximal Ub in K33-Ub_2, respectively.

The nicely spread NMR signals of the middle Ub in K11/K3-Ub$_3$ chain strongly indicate that the structure of this Ub unit is intact. Similar to the middle Ub in K11-Ub$_3$, the spectral perturbations observed in the middle Ub of K11/K3 Ub$_3$ are roughly the additon of the perturbations in the respective Ub$_2$ constructs (proximal Ub in K33-Ub$_2$ in **Figure 3.26C**, and distal Ub in K11-Ub$_2$ in **Figure 3.23C**). This suggests that no additional inter-domain interaction exists in K11/K33-Ub$_3$.

As a proof of principle, we also assembled K11-Ub$_4$ using the pre-assembled Ub$_3$ as the distal unit precursor (to be thiolesterified) and a proximal Ub-K11BocLys. Shown in **Figure 3.30** is the product from condensation reaction, analyzed by coomassie-stained SDS-PAGE.
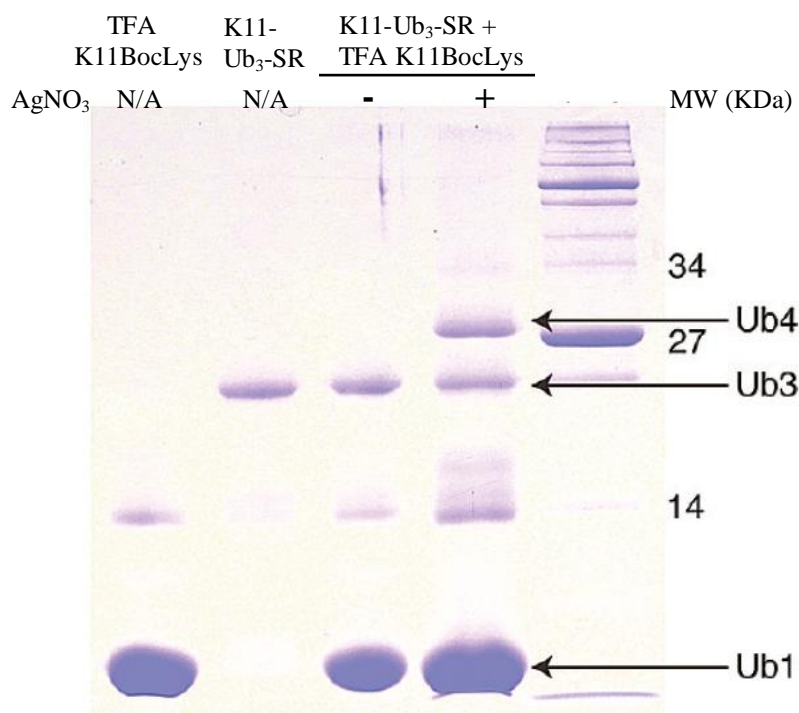
Figure 3.30. Coomassie-stained SDS-PAGE gel, showing the formation of K11-Ub$_4$.

Overall, the non-enzymatic approach described here is capable of assembling polyUb chains of any linkage composition and isotopic labeling scheme. polyUb chains up to trimer were successfully synthesized and characterized. With further optimized reaction conditions, tetramer or higher polyUb chain may be readily obtained. One limitation, however, is that large-scale production of polyUb chains requires a lot of Ub monomers containing BocLys incorporated at a desired lysine position. Traditional methods for expressing proteins containing UAAs require unnecessarily large amount of UAAs. This may prevent the large-scale production of Ub-BocLys in a cost-efficient way. Below is introduced a simple yet highly efficiently technique to solve this problem.

3.2.3 Condensed E. coli cultures for highly efficient production of proteins containing UAAs

3.2.3.1 Determination of optimum ratio for condensation

Condensed culture strategies have been used previously to minimize the cost associated with isotopic labeling of protein samples for NMR studies (*132, 133*). To test if the same strategy can be applied to express proteins containing UAAs, a 2 L *E. coli* culture was grown to express Ub-BocLys from pET3a vector. When the culture reached mid-log phase, the cells were harvested by low-speed centrifugation, resuspended in 80 mL (25-fold condensation) fresh media supplemented with corresponding antibiotics, inducing reagent and BocLys to induce the expression. The Ub-BocLys expression was analyzed on a gel. Robust expression was observed (**Figure 3.31**).

To demonstrate the generality of this methodology, two different UAA incorporation systems were tested for this condensed culture approach. In the first system, tester protein Ub was expressed from pET3a vector under the control of T7 promoter, and CbzLys was

site-specifically incorporated into K48 by *Mb*PylT/*Mm*PylRS variant (*115*). In the second system, sfGFP was expressed from pTrcHisA vector under the control of pTrc promoter, and *p*-benzoylphenylalanine (pBpa) was site-specifically incorporated into V150 by *Methanococcus jannaschii* tyrosyl tRNA/aaRS variant (*Mj*TyrT/*Mj*TyrRS) (*134*). Various condensation ratios ranging from $1\times$ (normal cell concentration) to $100\times$ have been surveyed for both systems (**Figure 3.32**).

BocLys   **+**   **-**

10 KDa
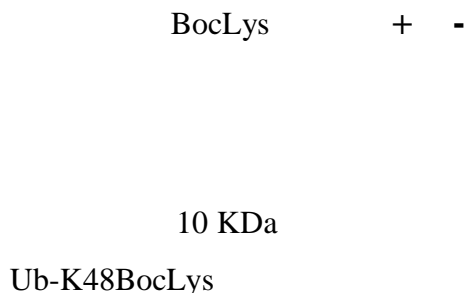
Ub-K48BocLys

Figure 3.31. Production of Ub-K48BocLys using condensed culture, with 2 mM BocLys.

In the case of Ub, clear protein expression can be seen in all samples except $100\times$ at which point the condensation experiment reaches a point of diminishing returns. This indicates that UAA can be added to cells at the time of induction and be utilized for protein expression at high cell densities. The lack of protein production in the absence of UAA indicates that the fidelity of amber codon suppression and incorporation is not affected by changes to culture density. We performed these analyses in triplicate and used the presence of a BSA internal standard (10 μg) to quantify protein yields and the relative efficiency of UAA usage under each scenario (**Table 3.1**). Using Ub-K48CbzLys as an example, we observed that the optimum protein production per unit amino acid occurs when cultures are condensed by a factor of 25 (**Table 3.1**). Under 25-fold condensation,

1.8 mg of Ub-K48CbzLys was produced from a 2 mL expression, representing a significant improvement in yield. Importantly, ~93% less UAA is needed. When expression cultures are not condensed ($1\times$), it is evident from the cell density and the observed protein quantities that there was some continued cell growth after IPTG induction but little overall increase in protein production.

A.

CbzLys   pBpa

B.

C.
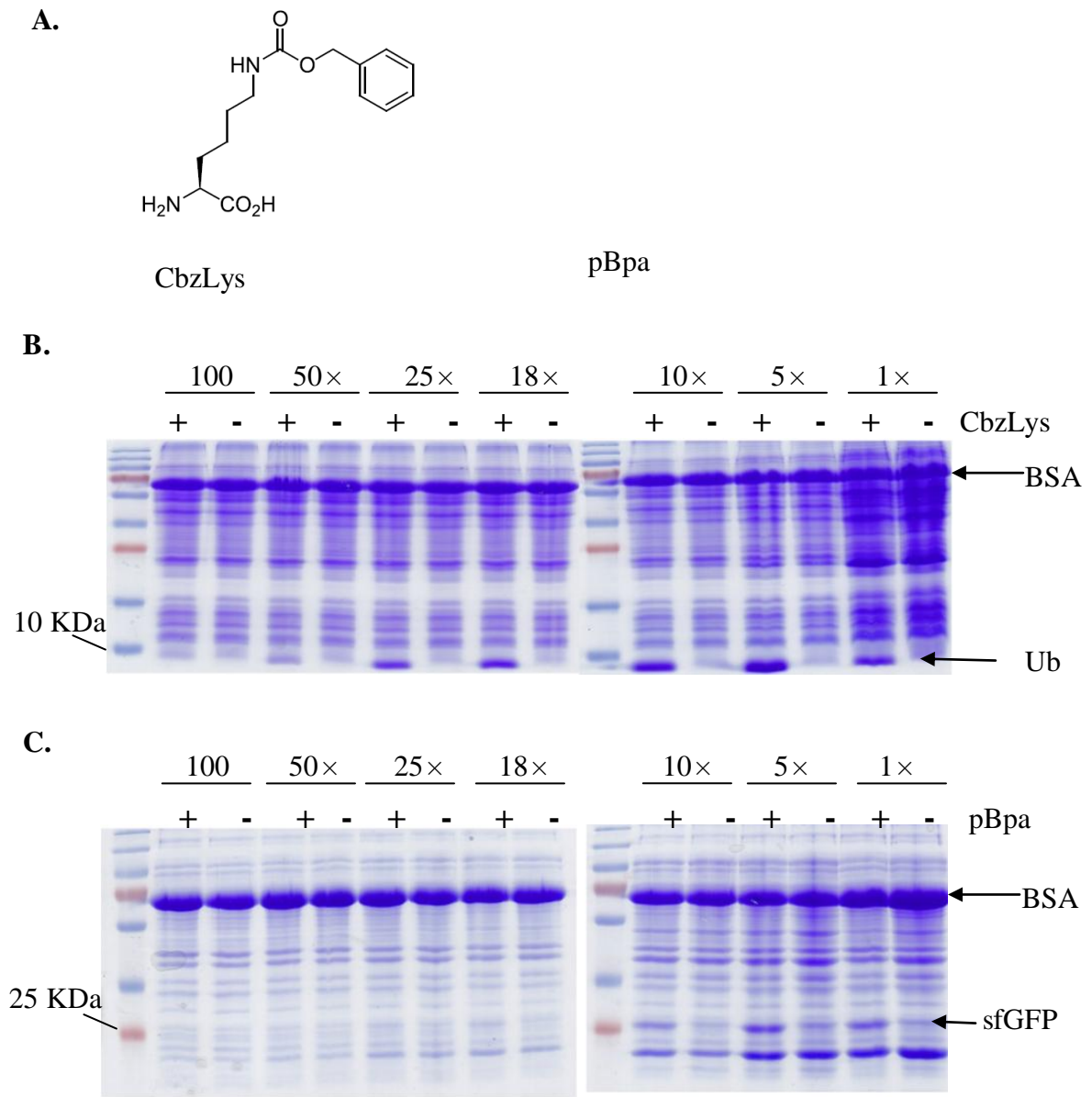
Figure 3.32. Determination of optimal condensation ratios. **A**) Structures of CbzLys and pBpa. **B**) Ub expression. **C**) sfGFP expression

In the case of sfGFP expression, there was very little protein observed by SDS–PAGE in condensation experiments above 10×. The reduced consumption of pBpa under 10× and 5×, however, is very evident from the calculation (**Table 3.1**). The lower levels of production observed can be ascribed to an overall lower productivity of this tRNA synthetase/tRNA pair, the relative lower expression of sfGFP in comparison to Ub, and/or less expression efficiency of pTrcHisA/ pTrc promoter as compared to pET/ T7 promoter. In addition, while we do not observe target protein production at the highest condensation ratios, it is possible that there is a low level of production, that is, not seen in a gel analysis but could be enriched by purification.

Table 3.1. Protein production per unit UAAs using condensed cultures

Condensation ratios	Proteins on gel (μg)	Total protein from 2 mL culture (μg)	Protein production per unit UAAs (nmol/μmol)[a]
Ub-K48CbzLys			
100×	ND	ND	ND
50×	1.3 ±0.3	1300	38
25×	3.6 ±0.5	1800	53
18×	4.0 ±0.7	1400	41
10×	6.3 ±1.7	1300	38
5×	7.4 ±1.8	740	22
1×	3.6 ±0.7	140	4.1
sfGFP-K150pBpa			
100×	ND	ND	ND
50×	ND	ND	ND
25×	ND	ND	ND
18×	0.7 ±0.1	250	2.3
10×	2.5 ±0.5	500	4.6
5×	3.8 ±0.5	380	3.5
1×	2.1 ±0.2	170	1.6

Note: [a] For 2 mL cultures, 4 μmol UAAs were used (final concentration 2 mM)

3.2.3.2 Characterization of purified proteins

Finally, "large scale" purifications of both Ub and sfGFP were conducted using the optimal condensed formats such that we could verify UAA incorporation by mass spectrometry. The condensed culture allowed us to induce 1 L of production cultures of Ub and sfGFP in 40 mL (25×) and 100 mL (10×) of media, respectively. These proteins were purified to >90% homogeneity resulting in large amounts of protein (>25 mg), a sample of which was subjected to in-gel tryptic digests and the resulting peptides analyzed by mass spectrometry. Upon examining the mass data we observed clear site-specific incorporation of CbzLys and pBpa into the correct tryptic fragments of Ub and sfGFP, respectively (**Figure 3.33**). Importantly we do not see incorporation of any endogenous amino acids at those positions, indicating again that suppression fidelity remains intact under very high cell density conditions.

In summary, the methodology presented here represents a simple yet effective way of maximizing the efficiency of UAA mutagenesis studies. This method requires no major changes to experimental conditions and is compatible with standing cell lines and plasmids. The observed improvement using this protocol varies with target protein, expression vector and UAA used. While it is sufficient to conserve UAA by culture condensation, it might be further enhanced when coupled with codon optimized genes. Ideally researchers will test a range of condensation conditions with a protein of interest as described here to optimize production.
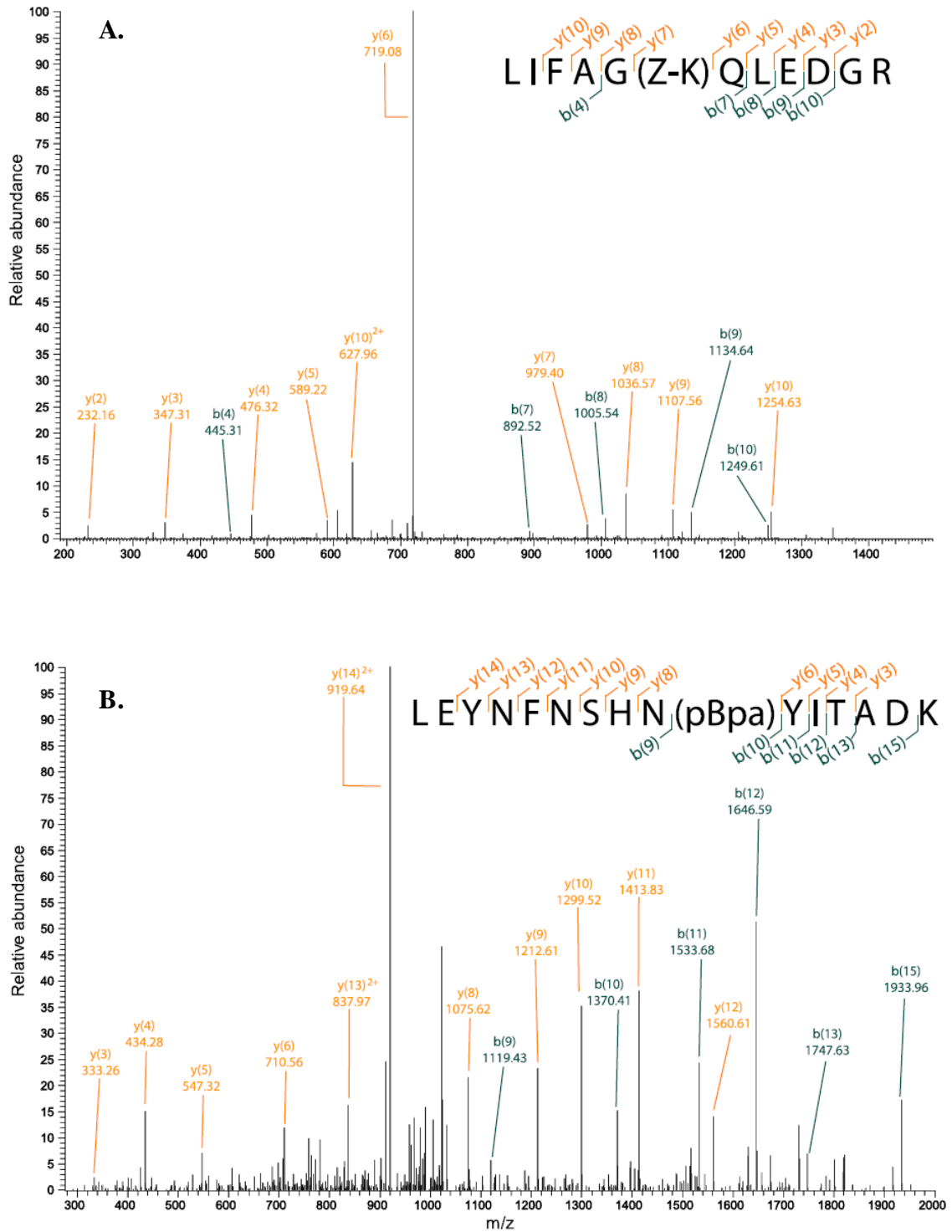
Figure 3.33. Mass spectral analyzes of purified proteins from condensed culture, showing the incorporation of UAAs. **A**) Ub-K48CbzLys. **B**) sfGFP-V150pBpa.

3.3 Conclusion and Future Applications

3.3.1 Traceless enzymatic synthesis of polyUb chains using Ub-BocLys

Enzymatic synthesis of polyUb chains using Ub-BocLys is advantageous over traditional methods as removal of Boc protecting group can result in a natural lysine residue which allows the subsequent formation of a natural isopeptide bond. Using this method, we successfully synthesized and characterized K48-linked Ub_2 and have demonstrated that synthesis of Ub_3 or higher polyUb chains is possible. Intriguingly, while many of the E2 enzymes remain uncharacterized, a growing number have been shown to be highly promiscuous and capable of targeting different lysine residues (*135*). Moreover it has been shown that the lysine-targeting preferences can be amplified by using blocked substrates, such as arginine mutants (*135*).

Using the strategy described here it might be possible to capitalize on promiscuous E2 enzymes and block multiple lysines simultaneously to direct the synthesis of atypical Ub linkages. For example, UBE2C has been determined to have conjugating activity toward both K11 and K48 linkages. Double blocking of K11 and K48 using Boc group and chain terminating mutations (Arg or Cys), followed by selective unblocking, could potentially direct UBE2C to produce polyUb chains of desired linkage composition (**Figure 3.34**). This strategy, however, may not result in polyUb chains with native residues because of the need of chain terminating mutations. The advantage of this "double blocking" strategy is that iterative protection/deprotection chemistries are not required and thus, a higher overall yield is expected.
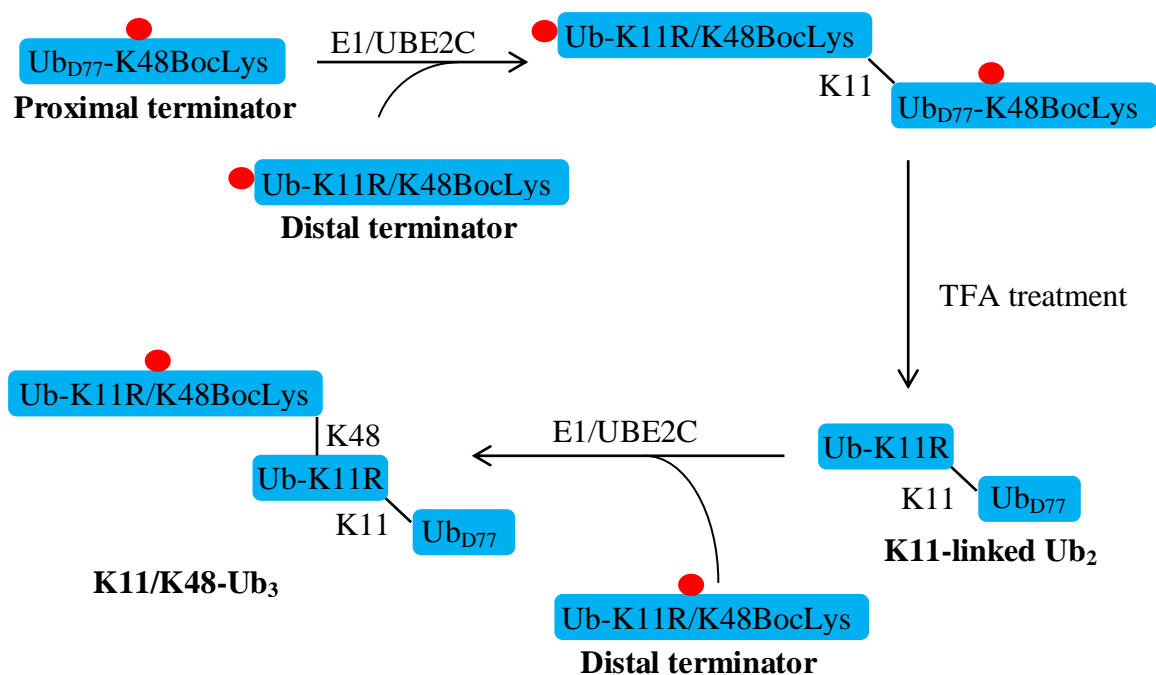
Figure 3.34. Proposed scheme of making K11-, K48-mixed linkage polyUb chains using a "double blocking" strategy by both chain terminating mutations and Boc group.

3.3.2 Nonenzymatic synthesis of polyUb chains using Ub-BocLys monomers

The dearth of native polyUb chains of any defined length, linkage composition, and isotopic labeling scheme has been a major hindrance to the ubiquitin field. To address this challenge, we have developed an affordable and widely accessible method for controlled, iterative nonenzymatic assembly of completely natural polyUb chains using recombinant monomers as the primary building blocks. Using this method, we synthesized homogeneously K11-Ub_2, Ub_3, and Ub_4, and K33-Ub_2, as well as a mixed linkage K11/K33-Ub_3. With this chain-assembly method in hand, it is now possible to generate and study essentially any polyUb chain, both homogeneously linked and with mixed linkages, to uncover the structure and receptor recognition of the polyUb signal and its processing by linkage-specific deubiquitinases. Importantly, our method can be

extended to (poly)ubiquitinate a target protein, with no need for specific E2 and E3 enzymes, and could also be used to form an isopeptide bond between Ub and virtually any two proteins.

Perhaps one of the most interesting experiments that can be accomplished with these atypical polyUb chains is to identify naturally occurring proteins that are involved in ubiquitination or deubiquitination processes. One possible approach is to capture potential target proteins by running phage displayed (*136*) human protein library through a desired polyUb chain that has been immobilized on a resin. Following MS proteomic analysis or DNA sequencing analysis the bound proteins or gene sequences will be revealed (**Figure 3.35**).
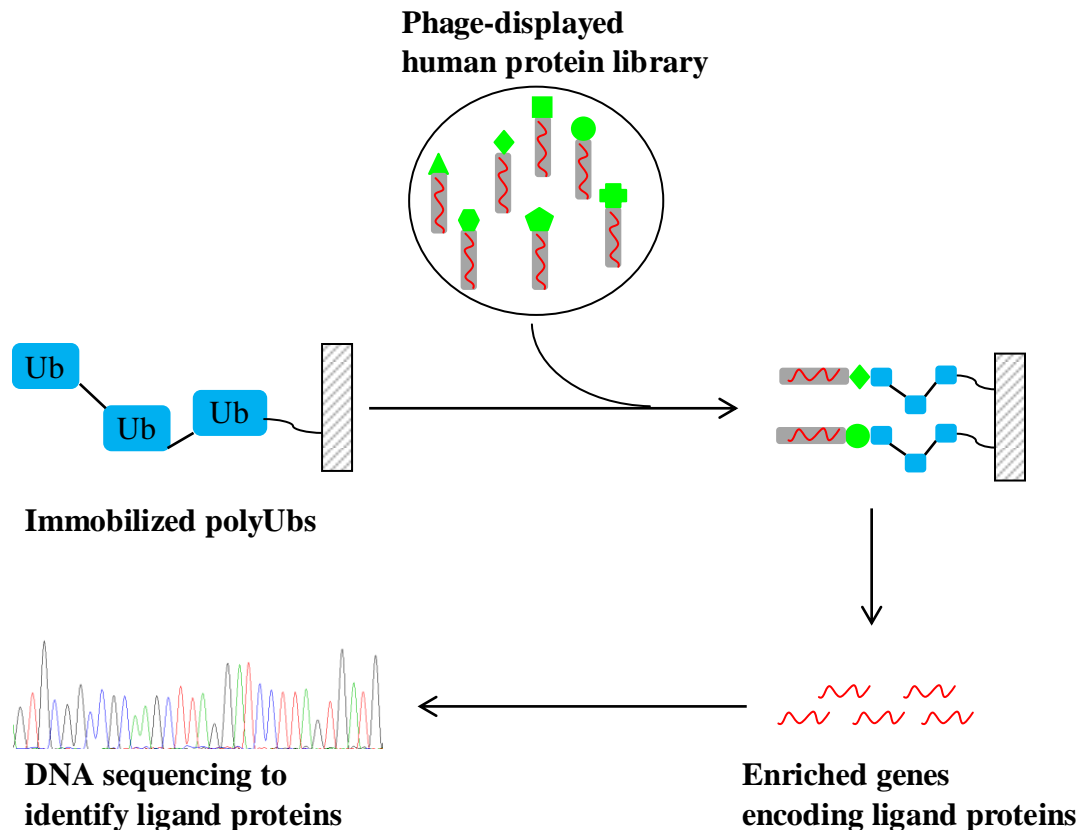
Figure 3.35. Identification of natural ligand proteins of polyUb chains from phage-displayed human protein library. Phage-displayed human proteins are shown in green.

3.3.3 Condensed E. coli cultures for highly efficient production of proteins containing

   UAAs

We have domenstrated that by using condensed culture method, protein production per unit UAAs can be increased by up to 10 fold. This will be particularly useful when the UAAs to be incorporated into proteins are not readily available. **Appendix I** includes one such example where condensed culture method was used to produce proteins containing *o*-nitrobenzyl thiol tyrosine (OnbSTyr). In addition to regular protein production, it is possible that condensed *E. coli* cultures could be used in genetic selections of novel aminoacyl-tRNA synthetase variants. Reporter proteins could be expressed in small volumes and then the cells diluted and selected. In the past such selections have been notorious for consuming large quantities of UAAs. Ultimately we believe that this approach will become standard practice in laboratories that are applying UAAs to solve biophysical problems, particularly in situations where the UAA in use must be chemically synthesized.

3.4 Materials and Methods

3.4.1 Traceless enzymatic synthesis of polyUb chains using Ub-BocLys

  3.4.1.1 Construction of PylT and PylRS plasmids for BocLys incorporation

The *Methanosarcina barkeri* pyrrolysyl tRNA (*Mb*PylT) encoding DNA (see **Apendix IV** for full sequence) was assembled using primers PylT#1~#6 (**Apendix III**). In this synthetic gene, PylT transcription is under the control of a upstream *proK* promoter (*137*), which has been reported to enhance the *in vivo* expression of tRNA (*120*). This *proK*-PylT fusion gene was PCR amplified using primers pSup_PylT_FWD and pSup_PylT_REV (**Appendix III**). The PCR product was digested with *ApaL* I/*Xho* I, gel-

purified, and ligated into pSup vector (*120*) cut with the same restriction enzymes to replace the original tRNA region. This new plasmid was named as pSup-PylT. The gene encoding *Methanosarcina mazei* pyrrolysyl tRNA synthetase (*Mm*PylRS) (see **Apendix IV** for full sequence) is PCR amplified from *M. mazei* genomic DNA (ATCC catalog number BAA-159D-5) using primers PylRS_FWD and PylRS_REV (**Appendix III**). The PCR product was digested with *Nde* I/*Nsi* I, gel-purified and ligated into pSup-PylT plasmid pre-treated with *Nde* I/*Pst* I (*Mm*PylRS encoding gene contains *Pst* I site. Therefore, *Nsi* I was used in primer to create a *Pst* I compatible end on the *Mm*PylRS PCR product). The recombinant plasmid was named as pSup-Pyl and was confirmed by DNA sequencing. A single mutation Y384F, which is known to enhance PylRS activity toward BocLys (*116*), was introduced into pSup-Pyl by site-directed mutagenesis using primers PylRS_Y384F_FWD and PylRS_Y384F_REV (**Appendix III**). The resultant plasmid was renamed as pSup-BocLys.

As a backup plan, the *Mb*PylT and *Mm*PylRS were also cloned into a double plasmid system. To clone the proK_PylT gene into pYC vector (*43*), pSup-BocLys was PCR amplified using primers pYC_PylT_FWD and pYC_PylT_REV (**Appendix III**). The PCR product was digested with *Pst* I/*Not* I, gel-purified and ligated into the *Pst* I/*Not* I site of pYC vector to replace the original tRNA. This new plasmid is named as pYC-PylT-Myo4tag, which also bears a myoglobin gene containing a "TAG" stop codon at amino acid position 4. To clone PylRS into pBK vector, pSup-BocLys vector was PCR amplified using primers PylRS_FWD and PylRS_REV (**Appendix III**). The PCR product was digested with *Nde* I/*Nsi* I, gel-purified and ligated into pBK vector pre-cut with *Nde* I/*Pst* I. The recombinant plasmid was named as pBK-BocLys

To confirm that the *Mb*PylT and *Mm*PylRS were functional in the new plasmids, expression tests were run to incorporate BocLys into proteins using the newly constructed plasmids. For the double plasmid system, the "tester" protein was the myoglobin carried by pYC-PylT-Myo4tag plasmid. For the single plasmid system, the "tester" protein was sfGFP from pTrcHisA-sfGFP150tag plasmid (*138*). This sfGFP gene was generated by overlapping PCR using synthetic primers (**Appendix III**). Val150 mutation was introduced by site-directed mutagenesis using primers sfGFP_V150tag_FWD and sfGFP_V150tag_FWD (**Appendix III**). pYC-PylT-Myo4tag/pBK-BocLys and pSup-BocLys/pTrcHisA-sfGFP150tag pairs were co-transformed into GeneHogs *E. coli* cells. pYC-PylT-Myo4tag/pBK-BocLys cells were plated on LB agar with 15 μg/mL tetracycline (Tet15) and Kan50. pSup-BocLys/pTrcHisA-sfGFP150tag cells were plated on LB agar (Chl35/Amp100). Two single colonies were picked from each plate, grown at 37 ℃ in 2×YT media with corresponding antibiotics to OD_{600} of ~0.5, and then induced with 1 mM IPTG and 2 mM $^\varepsilon N$-Boc-L-Lysine (Chem-Impex International, Wood Dale, IL) and grown overnight (~16 h). The whole-cell protein expression was analyzed by resolving whole cell lysate from 50 μL expression culture on 12.5% SDS-PAGE.

3.4.1.2 Construction of pSup plasmids for incorporation of other lysine derivatives

To generate *Mm*PylRS variant for CbzLys incorporation, one additional mutation Y306A (*116*) was introduced into the PylRS gene on pSup-BocLys plasmid by site-directed mutagenesis using primers PylRS_Y306A_FWD and PylRS_Y306A_REV (**Appendix III**). This new plasmid bearing the *Mm*PylRS for incorporation of CbzLys into proteins was named as pSup-CbzLys. To generate *Mm*PylRS variant for OnbLys incorporation, three additional mutations Y306M/L309A/C348A were introduced into the PylRS gene

on pSup-BocLys plasmid by sequential site-directed mutagenesis using primer pairs PylRS_Y306M_FWD/ PylRS_Y306A_REV (for both Y306M and L309A mutations) and PylRS_C348A_FWD/ PylRS_C348A_REV (**Appendix III**). This new plasmid was named as pSup-OnbLys.

To test if BocLys, CbzLys and OnbLys can be incorporated into Ub, pSup-BocLys, pSup-CbzLys and pSup-OnbLys were co-transformed with pET3a-Ub-K48tag plasmid into BL21 (DE3) *E. coli* cells, respectively. Two single colonies were picked from each plate, grown at 37 ℃ in 2×YT media (Amp100/Chl35) to OD_{600} of ~0.5 and then induced with 1 mM IPTG and 2 mM BocLys (Chem-Impex International), CbzLys (Acros Organics, Geel, Belgium) or OnbLys (see **Appendix VI** for synthesis details). After overnight expression, the cells from 100 μL culture were collected, resuspended in SDS loading dye, boiled at 95 ℃ for 10 min and then resolved on 12.5% acrylamide gel. Expression and purification of Ub-K48OnbLys were performed using IMPACT kit (NEB) as described in the section *3.4.1.4*.

3.4.1.3 Construction of Ub expression plasmid pTXB1-SynUb-K48tag

Codon-optimized human Ub gene was generated by overlap PCR using homologous primers SynUb #1~#6 (**Appendix III**). A two step PCR strategy was used, including an assembly PCR step and an amplification PCR step. Assembly PCR was performed in a 50 μL solution containing 0.2 mM dNTPs, 0.5 U of Phusion DNA polymerase (NEB), 1× Phusion HF reaction buffer (NEB) and 2 μM of each SynUb #2 through SynUb #9 primers. This mixture was thermal cycled as follows: initial denaturation at 98 ℃ for 30 s, 25 cycles of 98 ℃ for 10 s, 58 ℃ for 30 s and 72 ℃ for 20 s and final extension at 72 ℃ for 10 min. The synthetic Ub gene was then amplified in the following conditions: 0.2

mM dNTPs, 0.5 U of Phusion DNA polymerase (NEB), 1X Phusion HF reaction buffer (NEB), 1 μL of assembly PCR product and 2 μM of each SynUb #1 and SynUb #10 primers in a 50 μL solution, under the same cycling conditions. The PCR product was purified using QIAquick PCR purification kit (Qiagen) and digested with *Nde* I/*BamH* I. The digestion product was gel-purified and ligated into the *Nde* I/*BamH* I site of pET3a vector (EMD Chemicals Inc., San Diego, CA). The gene was confirmed by sequencing and then PCR amplified from pET3a vector using primers pTXB1_FWD and pTXB1_REV. The PCR product was digested with *Nde* I/*Sap* I, gel-purified and ligated into the *Nde*I/*Sap*I site of the pTXB1 vector (NEB) to generate pTXB1-SynUb plasmid. pTXB1-SynUb-K48tag plasmid was generated by site-directed mutagenesis using template plasmid pTXB1-SynUb and primers K48tag_FWD and K48tag_REV (**Appendix III**). Primers for introducing TAG codon into other lysine positions are also listed in **Appendix III**.

To test SynUb, Y384F mutation and different TAG positions, the following eight double transformations were performed using BL21 (DE3) chemically competent cells (1~4 for regular Ub and 5~8 for SynUb):

1) pSup-wtPylRS + pET3a-Ub-K48tag

2) pSup-BocLys (containing PylRS Y384F mutant) + pET3a-Ub-K48tag

3) pSup-wtPylRS + pET3a-Ub-K11tag

4) pSup-BocLys + pET3a-Ub-K11tag

5) pSup-wtPylRS + pET3a-SynUb-K48tag

6) pSup-BocLys + pET3a-SynUb-K48tag

7) pSup-wtPylRS + pET3a-SynUb-K11tag

8) pSup-BocLys + pET3a-SynUb-K11tag

A single colony was picked from each plate, grown at 37 °C in 2×YT media (Amp100/Chl35) to OD_{600} of ~0.5, induced with 1 mM IPTG and 2 mM BocLys (Chem-Impex International) and grown overnight (~16 h). After overnight expression, the cells from 100 μL culture were collected, resuspended in 10 μL 1×SDS loading dye, boiled at 95 °C for 10 min and then resolved on 12.5% acrylamide gel.

3.4.1.4 Expression and purificiation of Ub-K48BocLys, N^{15} labeled Ub_{1-74} and N^{15} labeled Ub_{1-74}-K48BocLys

Ub_{1-74} was expressed as previously described (*122*). To express Ub-K48BocLys, plasmids pSup-BocLys and pTXB1-SynUb-K48tag were double transformed into *E. coli* BL21(DE3) chemically competent cells. Expression was induced in regular or "condensed" culture (see section *3.4.3*). Briefly, a freshly transformed colony was grown in 2 L 2×YT media (Amp100/ Chl35) at 37 °C to OD_{600} of 0.5. The *E. coli* cells were pelleted, resuspended in 100 mL 2×YT media (Amp100/Chl35) containing 1 mM IPTG and 2 mM BocLys, and grown at 30 °C for 16 h. *E. coli* cells from the expression culture were harvested by centrifuge, washed three times with column buffer (20 mM HEPES, 500 mM NaCl, pH 8.5) and resuspended in 60 mL of lysis buffer (20 mM HEPES, 500 mM NaCl, 1 mM PMSF, pH 8.5). The resuspended cells were frozen and thawed three times, sonicated and then centrifuged to remove cell debris. Supernatant was applied to a column containing 35 mL of chitin-binding beads from IMPACT kit (NEB) which had been pre-equilibrated with 100 mL colume buffer. The beads with bound chitin-Ub fusion were washed with 400 mL column buffer. On-column cleavage was induced by flushing the column with 60 mL wash buffer containing 50 mM DTT. The column was

kept at 37 °C for 24 h. Cleaved Ub-K48BocLys was eluted from the column using 60 mL wash buffer. Proteins were concentrated using Amicon 3K NMWL centrifugal filter (EMD Millipore, Billerica, MA) and SEC purified. The overall yield of expression and purification was 10 mg protein per liter culture.

^{15}N-labeled Ub_{1-74} and Ub_{1-74}-K48BocLys protein was expressed using auto-inducing minimal media. pET3a-Ub_{1-74} was transformed into BL21(DE3) cells. pTXB1-SynUb$_{1-74}$-K48tag and pSup-BocLys were double transformed into BL21 (DE3) cells. A freshly transformed single colony from each plate was inoculated into 4 mL 2×YT media (Amp100/Chl35) to grow a starter culture. Next day, this starter culture was inoculated into a 2 L culture of auto-inducing minimal media containing 10 mM Na_2SO_4, 50 mM KH_2PO4, 50 mM Na_2HPO_4, 1 mM $MgSO_4$, 18 mM $^{15}NH_4Cl$ (Acros), 0.04% ferric citrate, 0.5% glycerol, 0.05% glucose, 0.2% lactose and 2 mM BocLys (Chem-Impex). This 2 L culture was grown at 30 °C for 24 h. Purification of ^{15}N Ub_{1-74} was performed as previously described (*122*) and purification of ^{15}N Ub_{1-74}-K48BocLys was performed using IMPACT kit as described above.

The MW of purified Ub-K48BocLys, ^{15}N Ub_{1-74} and ^{15}N Ub_{1-74}-K48BocLys samples were confirmed by ESI-MS at Maryland Proteomics Core Facility. High resolution MS of m/z 250-2500 were acquired with a JEOL AccuTOF-CS mass spectrometer in electrospray positive mode using flow injection. To determine the MW, spectra were deconvoluted using MagTran software (*139*) with the maximum charge set to 30. For dilute Ub_2 and Ub_3 samples, high-resolution mass spectra of m/z 400-2000 (or 600-4000 for protein with Alloc protection group) were acquired on a Thermo Scientific LTQ Orbitrap XL mass spectrometer using flow injection. The resolution was 60 000 at m/z

400. Deconvolution was carried out using either the Xtract program in Xcalibur software or MagTran with a maximum charge set to 30.

3.4.1.5 Deprotection of Boc protecting group

The Boc protecting group of Ub-K48BocLys was removed by adding trifluoroacetic acid (TFA) to a final concentration of 2%, followed by incubation at 37 °C for 4 h. The protein sample was then re-equilibrated to buffer containing 20 mM sodium phosphate, pH 6.8 using a 3K NMWL concentrator (Millipore). No appreciable amount of protein was lost from TFA treatment.

NMR samples of ^{15}N Ub$_{1-74}$ and TFA-treated ^{15}N Ub$_{1-74}$ K48BocLys (150 μM~ 1 mM) were prepared in the 20 mM sodium phosphate buffer, pH 6.8, containing 8% ^{2}H$_2$O and 0.02% (w/v) NaN$_3$. These and all the following NMR studies were performed on an Advance III 600 spectrometer equipped (Bruker Biospin, Billerica, MA) with a TXI cryoprobe using standard or in-house pluse sequences at 23 °C. ^{15}N-^{1}H HSQC and TROSY-HSQC spectra were acquired with spectral widths of 6010 Hz and 2100 Hz for the ^{1}H and ^{15}N dimensions, respectively. A total of 160 t_1 increments were collected with 2048 complex points in each. Assignments for Ub$_{1-74}$ and TFA-treated Ub$_{1-74}$ K48BocLys were determined by comparison with previously published spectra of human Ub (*122*). Spectra were processed using NMRPipe (*140*) and analyzed using Sparky (*141*).

Chain polymerization reactions on K48BocLys variant and TFA-treated K48BocLys for testing the functionality of deprotected lysine were performed as previously described (*122*) (see next section for reaction details).

3.4.1.6 Assembly of K48-linked Ub_2 and Ub_3 chains

Ub_2 assembly reactions were performed using 10 mg each of distal Ub-K48BocLys and proximal ^{15}N Ub_{1-74} in a reaction solution containing 50 mM Tris-HCl (50% base), 5 mM $MgCl_2$, 10 mM phosphocreatine (TCI), 4 mM ATP, 1 mM DTT, 0.6 unit/mL each of creatine phosphokinase and inorganic pyrophosphatase, 150 nM E1 enzyme and 10-20 μM E2-25K enzyme. The reactions were incubated overnight at 37 °C. The reactions usually went complete with little or no reactant left, as judged by a post-reaction SDS-PAGE gel. Product was purified using both cation and SEC to > 60% yield, which is comparable to the yield using a reported method (*95*). Assembly of Ub_3 followed the same procedure with the TFA-treated Ub_2-K48BocLys being used as the proximal monomer and Ub-K48BocLys as the distal monomer.

ESI-MS of TFA-treated and non-treated Ub_2 was performed as above. NMR analyses of ^{15}N Ub_{1-74} and Ub_2-K48BocLys (with proximal Ub_{1-74} ^{15}N labeled) were performed under pH 4.5 and pH 6.8, respectively, using the same NMR solutions as above with the only difference being that 50 mM ammonium acetate buffer was used for pH 4.5 solution. Chemical shift perturbations (CSPs) were determined using identical buffer and temperature conditions. CSPs were calculated as:

$$CSP = [(\Delta\delta H)^2 + (\Delta\delta N/5)^2]1/2,$$

where $\Delta\delta H$ and $\Delta\delta N$ are chemical shift differences for ^1H and ^{15}N, respectively. CSPs are plotted as a function of the residue number.

3.4.2 Nonenzymatic synthesis of polyUb chains using Ub-BocLys monomers

3.4.2.1 Preparation of ligation precursors Ub-K11BocLys, Ub-K33BocLys and Ub-SR

Plasmids pTXB1-SynUb-K11tag and pTXB1-SynUb-K33tag were co-transformed with pSup-BocLys into *E. coli* BL21 (DE3) chemically competent cells as detailed above. ^{15}N-labeled WT Ub and ^{15}N-labeled Ub-BocLys were expressed in *E. coli* using autoinducing minimal media with ^{15}NH$_4$Cl and were purified as described above.

Ub variants with the thioester group at the C-terminal G76 (Ub-SR) were generated from respective Ub monomers (unlabeled or ^{15}N labeled) using the following procedure. A 1 mL solution of 5 mg/mL Ub was prepared in 20 mM sodium phosphate buffer (pH 8.0) with 10 mM ATP, 10 mM MgCl$_2$, 100 mM sodium 2-mercaptoethane sulfonate (MESNA) and 250 nM of Ub-activating enzyme E1. The mixture was incubated at 37 ℃ for 6 h and then stored at 4 ℃. The E1 enzyme was precipitated by adding a drop of glacial acetic acid to the solution. The protein solution was buffer- exchanged into distilled H$_2$O containing 0.4% TFA and subsequently lyophilized. Ub-SR generation was confirmed by ESI-MS as above.

3.4.2.2 Alloc protection and Boc deprotection to prepare Ub monomers for
condensation reaction

The Alloc protection reaction took place by initially dissolving 5 mg of each Ub monomer in 450 μL of DMSO for 15~20 min on a platform rocker. To this solution were added 17 μL of diisopropylethylamine (DIEA) and 75 μL of freshly made 40 mg/mL Alloc-OSu (TCI America, Portland, OR) solution in DMSO. The reaction was proceeded at least 1 h at room temperature on the platform rocker. Complete Alloc protection was monitored by ESI-MS. For each 100 μL of protein solution, the protein was precipitated

into a small white pellet by mixing with ~2 mL of ice-cold ether, vortexing for 15 s, and centrifuging for 10 min at 4 ℃. The top organic layer was removed after centrifugation, and two more rounds of ether precipitation were conducted. The pellet was then air-dried for 15~20 min.

To remove the Boc protecting groups, 5 mg protein pellet from the previous step was dissolved in 500 μL of ice-cold 3:2 TFA/dH$_2$O solution and left to react for 2~3 h at 4 ℃. Complete dissolution occurred after 15 min and can be aided by gently pipetting the solution. Removal of Boc group (MW loss of 100 Da) was monitored by ESI-MS. For every 100 μL of solution, the protein was precipitated by three rounds of extraction with cold ether as described above. Typically, the protein precipitated instantly after initial ether addition, and the resulting white pellet was air-dried for 15~20 min.

3.4.2.3 Silver-mediated condensation reaction

Approximately 2 mg of each of the two protected Ub monomers were dissolved in DMSO and then added together to a total volume of 90 μL as originally described (*99*). DIEA (4 μL), hydroxysuccinimide (H-OSu) (1 μL of fresh 390 mg/mL solution dissolved in DMSO), and AgNO$_3$ (1 μL of fresh 57 mg/mL solution dissolved in DMSO) were added to the Ub monomers, mixed, and incubated at room temperature in the dark for at least 20~30 h. Formation of ligated Ub$_2$ species was confirmed by SDS-PAGE. The solution was subjected to three rounds of ether precipitation as described above and then air-dried. The pellet typically was slightly yellowish in color.

3.4.2.4 Global deprotection of Alloc groups and protein renaturation

Ub was deprotected using 50 mol % chloro-pentamethylcyclopentadienyl-cyclooctadiene-ruthenium (II) ([Cp*Ru(cod)Cl], from Sigma-Aldrich, St. Louis, MO)

and 50 equivent of thiophenol, relative to the number of moles of protected amines in the solution (counting all lysine, histidine residues and the N-terminal amine). For example, for 4 mg of total Ub monomer and assuring slight excess over nine amine protection sites on each Ub molecule, the reaction pellet was dissolved in 420 μL of DMSO to which were added 240 μL of H_2O, 116.2 μL of 20 mM [Cp*Ru(cod)Cl] (freshly dissolved in DMSO), and 23.8 μL of neat thiophenol. The resulting dark-brown/black solution was divided into 200 μL aliquots in PCR tubes. The reactions were incubated in a thermal cycler at 50 ℃ for 2 h. After this, the tubes were allowed to cool for 10~15 min; the resulting solution was typically a dark orange cloudy solution with black precipitate. Approximately 100 μL of each reaction tube were aliquoted into 2 mL eppendorf tubes for cold ether precipitation as described above. Typically, at least 10 rounds of ether precipitation were necessary to form a small dark-brown pellet in each of these eppendorf tubes. After each round of centrifugation, the top organic layer was very carefully removed, so as to not disturb the aqueous bottom layer. When cold ether was added, vigorous vortexing for 15 s was performed to ensure proper mixing of the solution. After pellet formation, the pellet was allowed to air-dry for 10~15 min.

The above protein solution from Alloc deprotection was initially dissolved in a total of 500 μL to 1 mL of filtered denaturaton solution containing 6 M GdnHCl and 20 mM sodium phosphate, pH 6.8, for 10~15 min. The solution was centrifuged for 5 min to remove the undissolved precipitate. Renaturation solution (20 mM sodium phosphate, pH 6.8, 130 mM NaCl) was added to the denatured protein samples stepwise in 200 μL aliquot with 10~15 min interval, until the solution became clear. The final volume of renaturation solution is 10~15 mL. The clear solution was then transferred to 3K MWCO

dialysis tubing and dialyzed overnight against 2 L of the renaturation buffer at 4 ℃. The dialyzed protein samples were concentrated to <1 mL using Amicon Ultra-15 3K MWCO filter (Millipore) and then SEC purified. The fractions pertaining to the desired Ub species were collected, concentrated, and exchanged into NMR buffer (20 mM sodium phosphate, pH 6.8). Purity was assessed to be >99% based on coomassie-stained SDS-PAGE gel analysis. Typically from an input of 5 mg of Ub monomers, 0.7~1 mg of Ub_2 can be obtained. The final Ub_2 was analyzed by ESI-MS to confirm the MW.

3.4.2.5 Assembly and characterization of K11- and K33-linked Ub_2

In the current study, the K11-Ub_2 was assembled using a Ub-K11BocLys distal precursor in the thioesterification reaction. The final Ub_2 chains after Alloc deprotection were subjected to TFA treatment to produce native Ub_2. Non-treated Ub_2 with a BocLys at K11 was used as the proximal unit in the synthesis of Ub_3 chains. The K33-Ub_2 was assembled using WT Ub distal precursor in the thioesterification reaction. The conditions for individual steps are described above.

3.4.2.6 Assembly and characterization of K11-Ub_3 and K11/K33-Ub_3

To assemble K11-linked Ub_3 with the middle Ub ^{15}N-labeled (^{15}N-M K11-linked Ub_3), approximately 0.75 mg of ^{15}N-D K11-Ub_2-K11Boc was reacted with 1.25mg of Ub-SR. Both proteins were protected with Alloc groups, and K11-Ub_2 was treated with TFA to expose the ε-amine of K11 prior to reacting with Ub-SR. Immediately prior to the reaction, each protein was dissolved in 45 μL of DMSO. After the proteins were mixed together, the remaining ingredients (DIEA, AgNO3, and H-OSu) were added. Alloc deprotection was performed assuming 3 mg of Ub monomer with 10 Alloc-protected amines per Ub. This overestimation of the amount of Ub ensured complete Alloc

deprotection. The reaction pellet was dissolved in 315 μL of DMSO to which were added 180 μLof H_2O, 87.2 μL of 20mM[Cp*Ru(cod)Cl] (freshly dissolved in DMSO), and 17.9 μL of neat thiophenol. Protein was precipitated with ether, renatured, and purified as described above. A small aliquot (5 μL of 40 μM protein) was analyzed by ESI-MS to confirm that all Alloc groups on K11-Ub_3 were removed. Total amount of purified Ub_3 was determined to be 0.3 mg (15% yield).

To assemble K11/K33-Ub_3 with the middle Ub ^{15}N-labeled, 0.75 mg of ^{15}N-P K33-linked Ub_2 was reacted with 3mg of Ub-K11Boc. The C-terminus of K33-Ub_2 was thioesterified using MESNA and E1 enzyme as described above. Complete thioesterification was verified by ESI-MS. The protein was protected with Alloc groups as described above, with 3-fold excess of Alloc groups, and analyzed by ESI-MS. Ub-K11Boc protected with Alloc groups was treated with TFA as described above. Prior to condensation reaction, each protein was dissolved in 45 μL of DMSO. DIEA, $AgNO_3$, and H-OSu were added into protein solution in DMSO in the same amounts as described above. Subsequent Alloc deprotection was performed assuming a total of 5 mg of Ub monomer with 10 Alloc-protected amines per Ub. The reaction pellet was dissolved in 525 μL of DMSO to which were added 300 μL of H_2O, 145.3 μL of 20 mM [Cp*Ru(cod)Cl] (freshly dissolved in DMSO), and 29.8 μL of neat thiophenol. The reaction was run as described above. Ub proteins were then renatured and purified. Total amount of K11/K33-linked Ub_3 was 0.35 mg (9% yield). The apparent lower yield reflects the excess amount of unreacted Ub-K11Boc in the reaction.

K11-Ub_4 was assembled using pre-assembled K11-Ub_3 as the distal unit and Ub-K11BocLys as the proximal unit. Conditions for individual steps were the same as above.

3.4.3 Condensed E. coli cultures for highly efficient production of proteins containing

    UAAs

  3.4.3.1 Determination of optimum ratio for condensation

For expression test of Ub-K48BocLys, pET3a-SynUb and pSup-BocLys were co-transformed into *E. coli* BL21 (DE3). A single colony was picked to grow 5 mL of overnight starter culture. Next day the starter culture was inoculated into a 2 L 2×YT (Amp100/Chl35). This 2 L culture was grown to an OD_{600} of 0.5 and cells were centrifuged at 4,000 rpm for 10 min and then resuspended in 80 mL (25-fold condensation) fresh 2×YT (Amp100/Chl35) containing 1 mM IPTG and 2 mM BocLys. Prior to addition of BocLys, 5 mL of the culture containing no BocLys was transferred into a new tube and grown as a control for the UAA-dependent Ub expression. These cultures were grown at 30 ℃ for ~16 h. For each of the samples with and without BocLys, the cells from 100 μL culture were centrifuged, lysed, and analyzed by a 16% SDS-PAGE gel.

  To determine the optimum ratio of condensation, pSup-pBpa/pTrcHisA-sfGFP-V150tag and pSup-CbzLys/pET3a-SynUb-K48tag were co-transformed into GeneHogs and BL21 (DE3) *E. coli* cells, respectively. Single colonies were picked to grow a 1 L 2×YT culture (Amp100/Chl35). When the OD_{600} of the cultures reached 0.5, cells were centrifuged at 4,000 rpm for 10 min and resuspended in fresh 2×YT to generate a series of 2 mL cultures with a condensation ratio ranging from 1× to 100× (see **Table 3.1**). Antibiotics Amp100/Chl35 and inducing reagent IPTG were individually added into all these cultures. The 2 mL cultures were then divide into two 1 mL cultures. For one culture was added BocLys or pBpa, accordingly (the UAA-free cultures served as a

145

negative control). These 1 mL cultures were grown at 37 ℃ overnight. Next day, the cells from these cultures were harvested. These cells were resuspended in 100 μL water. Except for the 1× sample, 10 μL of each resuspended cells was resolved on a 16% SDS-PAGE gel. The cells in the 1× sample had substantial growth and only 5 μL of the resuspended cells were run on the gel for best resolution, which was taken into consideration when calculation was performed on the amount of expressed proteins. A BSA internal standard was included in each loaded sample to quantify the expressed proteins. Protein quantification was performed using Image J as shown in **Figure 2.9**.

3.4.3.2 MS characterizaiton of purified Ub-K48CbzLys and sfGFP-V150pBpa

Ub-K48CbzLys was expressed as a fusion to chitin binding domain in 25× condensed culture (as described above) by resuspending cells from one L culture into 40 mL of induction media. The protein was purified according to the above purification protocol of Ub-BocLys construct. sfGFP-V150pBpa was expressed in a 10× condensed culture by resuspending cells from 1 L culture in 100 mL of induction media. sfGFP-V150pBpa was purified according to the protocol in **Chapter 2** for His tag affinity purification.

For purified sfGFP and Ub proteins, about 0.5 μg was digested using ~0.01 μg of trypsin Gold (Promega) in 50 mM NH_4HCO_3 buffer, pH 7.8. The digestion reaction was incubated at 37 ℃ for 4 h 30 min. MS analysis of the tryptic digested sample was performed at Maryland Proteomics Core Faciltiy.

146

Appendix I: Site-specific Incorporation of Photoactivable Tyrosine Analogues into Proteins in *E. coli*

I-1 Introduction

Tyrosine residues occur at the active sites of a wide range of proteins and are frequently involved in redox or radical reactions. Besides serving as catalytic centers, tyrosines can be also post-tranlationally modified by phosphrylation (*142*) or sulfation (*143*) to act as a regulatory signal. Understanding the functions of these tyrosine residues is not an easy task as the properties of the tyrosines, such as pK_a, could be altered by surrounding micro-environment.

The biological functions of specific tyrosine residues in a protein are sometimes investigated by mutagenesis analysis, such as mutation into an inert amino acid (alanine, for example) or an amino acid with distinct properties. These mutagenesis studies, however, are limited by the structural constrains of proteins when the newly introduced amino acids have significantly different structures.

An alternative approach is to replace the targeted tyrosines with tyrosine analogues that have similar structures yet distinct properties such as acidity or redox potential (*144*). Traditional methods for incorporation of tyrosine analogues include metabolic labeling or total chemical synthesis. In the former method, *E. coli* starved of tyrosine was grown in the presence of tyrosine analogues to force the global incorporation of tyrosine analogues in all *E. coli* proteins (*145*). The obvious concern of this method is the lack of the control of the incorporation sites. The latter method relies on the solid phase peptide synthesis (*30*) followed by expressed protein ligation (*146*). Although this method allows the site-specific incorporation of tyrosine analogues, it is limited to small or medium-sized proteins with robust renaturating ability.

A more recent approach allows unnatural amino acids (UAAs) to be genetically encoded in a protein sequence (*44*). This method relies on the assignment of an UAA to an amber stop codon (TAG) in the gene of interest using an orthogonal tRNA and aminoacyl tRNA synthetase (aaRS) pair. The challenge, however, is to selectively incorporate tyrosine analogues while maintaining little or no incorporation of natural tyrosine. One strategy to overcome this problem is to "disguise" the tyrosine analogues with cleavable protecting groups. After the orthogonal tRNA/aaRS incorporate these "disguised" amino acids into proteins, the protecting groups can be removed to expose the desired amino acids to act as comparison probes.

Among all the protecting groups, photoremovable groups or so-called caging groups represents an ideal choice. The most commonly used caging group is based on *ortho*-nitrobenzyl moieties (*147, 148*), while other groups with improved photochemical properties have been also developed (*149-151*). These photoremovable groups have many advantages over other protecting group. First, caging group enables traceless protection of amino acids where removal of the photoremovable group or "decaging" retains the intact structure and function of the desired amino acids. Secondly, caging groups can inhibit the function of a "caged" amino acid if attached to its functional moiety. Subsequent decaging process restores the full function of the desired amino acids. This allows the spatiotemporal control of the biological functions of the proteins carrying the "caged" amino acids. Thirdly, light irradiation is a non-invasive technique that results in minimal disrruption of protein structure or cellular processes. Studies of protein functions *in vivo* or *in vitro* can therefore be carried out under native conditions.

Genetically encoded photocaged tyrosine together with photochemical control has been employed in many biological systems to understand or spatiotemporally control the functions of proteins (*148, 152-159*). While the photochemical control of tyrosine is a truly powerful biological tool, the goal of the current study is to incorporate a tyrosine analogue with enhanced properties as a comparison probe. In collaboration with Dr. Deiters research group at North Carolina State University, we have successfully incorporated fluorinated tyrosine into proteins in *E. coli* using an evolved *Methanococcus jannaschii* tyrosyl tRNA synthetase (*138*). Here, we intended to develop a general method for the incorporation of thiotyrosine into proteins in *E. coli*. As a thiol group is in close structural proximity to a hydroxyl group, it is expected to alter the activity of tyrosine without disturbing the structure of a host protein.

To incorporate *o*-nitrobenzyl-thiotyrosine into proteins in *E. coli*, we first tested the evolved *Mj*YRS variant E10 that is capable to selectively introduce *o*-nitrobenzyl-2-fluorotyrosine (OnbFTyr) (*138*). This E10 synthetase was identified from a *Mj*YRS mutant library that contain random or selectively diversified mutations at seven active-site residues (Y32, L65, H70, F108, D158, I159, and Ll62) (*160*). This library was then applied in a double-sieve selection experiment, based on suppression of a stop codon in the gene encoding chloramphenicol acetyl transferase (CAT) (positive selection) and barnase (negative selection). This selection resulted in variant synthetase that can incorporate OnbFTyr but not natural tyrosine into proteins in *E. coli*. Subsequent decaging process removes the protection group and exposes the 2-fluoroTyr for biological studies.

Test of this E10 synthetase with OnbSTyr for the current study indeed resulted in a moderate OnbSTyr-dependent protein production. However, MS results suggested that the caging group in the produced proteins had been reduced (see next section). Although the detailed reason of this modification *in vivo* is unknown, we speculated that it might be caused by the low-fidelity incorporation of UAAs by *Mj*YRS. Recent studies have suggested that PylT/PylRS could selectively incorporate tyrosine deratives such as *o*-benzyltyrosine into proteins in *E. coli*. The possibility of incorporaton of OnbSTyr using PylT/PylRS (or variants) is currently under investigation.

I-2 Preliminary Results

I-2-1 Expression, purification and MS characterization of OnbSTyr-containing sfGFP

   and Ub produced by MjYRS *variant E10*

Ub-K48OnbSTyr was purified using IMPACT kit (**Figure I-1**) and then analyzed by ESI-MS for MW determination. The expected mass of Ub-K48OnbSTyr is 8750 Da. However, the MS results revealed a major peak at 8720, which had 30 Da less than the expected mass (**Figure I-2**). This 30 Da mass loss was very likely due to the reduction of the *o*-nitro group (NO_2) to amine group (NH_2) as the Ub protein had been incubated with high concentration of DTT (50 mM) to induce the intein cleavage during protein purification.

OnbSTyr   +   -

10 KDa

Figure I-1. Purificaiton of Ub-K48OnbSTyr.
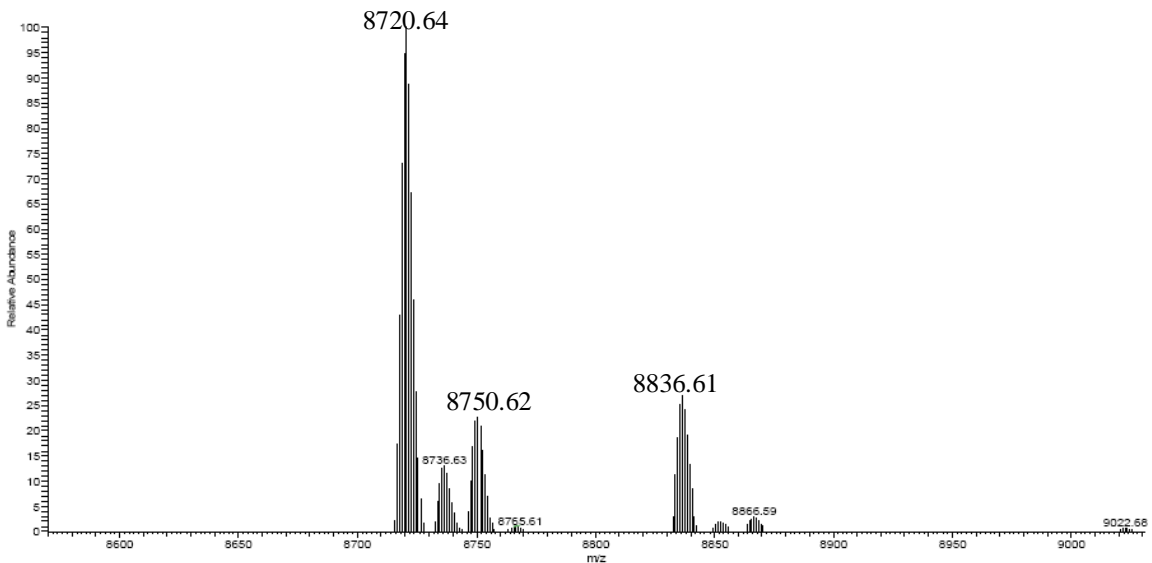
Figure I-2. Deconvoluted ESI-MS of Ub-K48OnbSTyr. Expected MW: 8750 Da.

151

To solve the potential nitroreduction problem, regular affinity tag expression and purification method were used where DTT is no longer a necessary component. We cloned SynUb-K48tag gene into pET11a vector to express Ub with a C-terminal His_6 tag. A pTrcHisA-sfGFP-Y66tag plasmid was also constructed. Ub-His-K48OnbSTyr and sfGFP-Y66OnbSTyr were purified using His tag affinity purification and then analyzed by SDS-PAGE gels (**Figure I-3**). Interestingly, a truncated $sfGFP_{1-66}$ was co-purified due to inefficient suppression of TAG codon and the N-terminal His tag. ESI-MS analysis of Ub-His-K48OnbSTyr, however, still displayed a major peak with 30 Da less mass (**Figure I-4**).

Figure I-3. Purificaiton of Ub-His-K48OnbSTyr and sfGFP-Y66 OnbSTyr.

Figure I-4. ESI-MS of Ub-His-K48OnbSTyr. Expected MW: 9573 Da.

I-2-2 Protein expression in the presence of nitroreductase inhibitor and incorporation of

o-aminobenzyl-thiotyrosine (OabSTyr) into Ub

The above experiment indicated that nitroreduction may not be caused by the DTT-mediated protein purification process. Instead, this problem may occur *in vivo*, likely caused by endogeneous nitroreductases. To solve this problem, we sought to test protein expression in the presence of a wide-range nitroreductase inhibitor dicoumarol (*161, 162*). Cell growth curve in the absence and presence of dicoumarol was first recorded to confirm that the supplement of dicoumarol would not affect cell growth (**Figure I-5**). Purification of Ub-His-K48OnbSTyr in the presence of dicoumarol, however, did not solve this problem, as analyzed by ESI-MS.
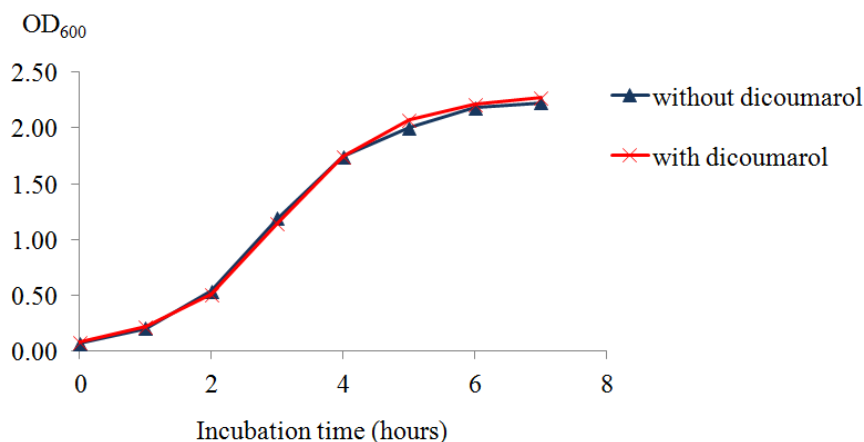
Figure I-5. Cell growth curve in the absence and presence of 500 μM dicoumarol.

The nitroreduction, if at all, could happen at either pre-translational or post-translational stage. In the former case, OnbSTyr gets reduced before being incorporated into proteins. That said, *o*-aminobenzyl-thiotyrosine (OabSTyr) needs to be a substrate of the E10 synthetase. If the reduction occurs at the post-translational level, OnbSTyr amino acid gets reduced after its incorporation into proteins. To test this hypothesis, OabSTyr was

153

synthesized and incorporated into Ub protein. The expression results suggested that OabSTyr was, indeed, a substrate of E10 synthetase (**Figure I-6**). Therefore, the observed nitroreduction problem could be explained as that OnbSTyr was first reduced to OabSTyr by *E. coli* cellular components and this reduce amino acid was then incorporated into proteins by E10 synthetase. This result also suggests that E10 synthetase has relatively low-fidelity toward OnbSTyr.
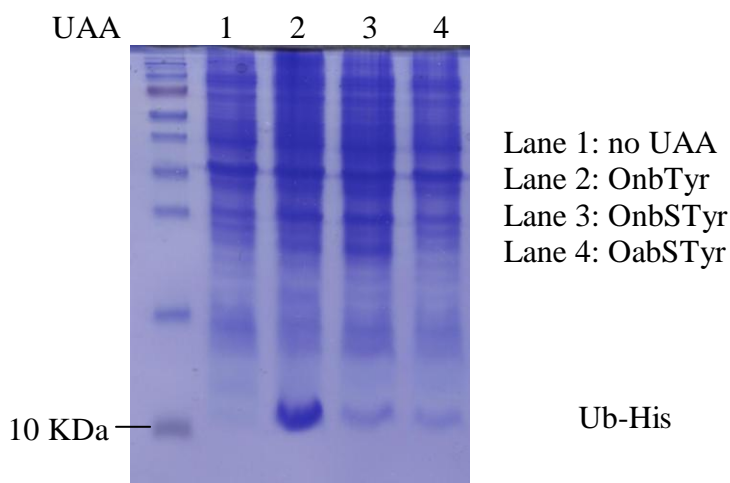
Figure I-6. Protein expression test for OabSTyr incorporation using E10 synthetase.

I-2-3 Incorporation of OnbSTyr into proteins using PylT/PylRS pair

As E10 synthetase (*Mj*YRS variant) seemed to have low-fidelity toward OnbSTyr, it may be advisable to test other tRNA/aaRS pairs. Recently, PylT/PylRS pair has been demonstrated to be able to incorporate a wide range of UAAs, including *o*-benzyl-tyrosine, into proteins in *E. coli* when PylRS contained two promiscuous mutations N346A/C348A (*117*). Several reported PylRS variants have been tested for their activity toward OnbSTyr, including WT PylRS, Y384F (BocLys) (*115*), Y306A/Y384F (CbzLys) (*115*), Y306M/L309A/C348A/Y384F (OnbLys) (*121*) and Y306M/L309A/Y384F

(referred to as Interm PylRS) which is an intermediate construct when OnbLys PylRS was generated from BocLys PylRS by introducing additional mutations. The two reported promiscuious mutations were also introduced into each of the above variants to test if they can enhance the activity of these synthetases. The PylRS variants with incorporated N346A/C348A mutations were named after the original variants with "+" suffixes (for example, BocLys+ stands for PylRS BocLys variant incorporated with the double mutations). The detailed mutations in each PylRS constructs are listed in **Table I-1**.

Table I-1. PylRS constructs tested for activities toward OnbSTyr

Numbers	PylRS synthetases[a]	Mutations				
		306	309	346	348	384
1	WT	Y	L	N	C	Y
2	BocLys	Y	L	N	C	F
3	CbzLys	A	L	N	C	F
4	OnbLys	M	A	N	A	F
5	Interm[b]	M	A	N	C	F
6	WT+	Y	L	A	A	Y
7	BocLys+	Y	L	A	A	F
8	CbzLys+	A	L	A	A	F
9	OnbLys+[c]	M	A	A	A	F
10	Interm+[c]	M	A	A	A	F

Note: [a] N346A/C348A mutations were incorporated in the five synthetases to create the "+" series of constructs.

[b] This was an intermediate construct when OnbLys synthetase was generated from BocLys synthetase.

[c] These two variants are containing the same mutations due to incorporation of N346A/C348A double mutations.

The in-plate fluorescence of these PylRS constructed was monitored under UV light to determine their activities toward OnbSTyr. Interestingly, synthetases OnbLys+ (#9) and Interm+ (#10) that carry the same mutations showed different fluorescence intensity (**Figure I-7**). This was likely due to some unexpected mutations in the synthetase gene, which can be verified easily by sequencing. Synthetases Interm (#5) and Interm+ (#10) are promising based on the plate fluorescence (**Figure I-7 B** and **D**).
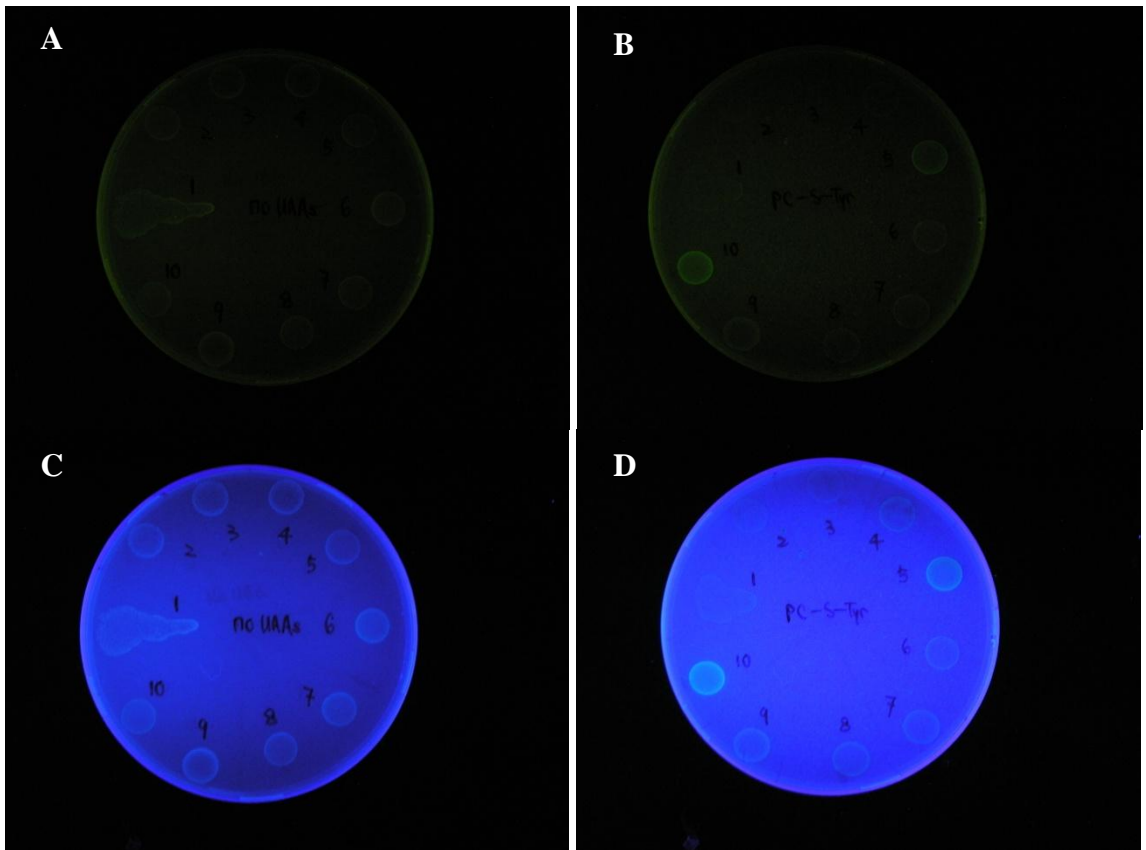
Figure I-7. Plate fluorescence of the constructed synthetases, showing their activities toward OnbSTyr. **A**) and **B**) were taken with light filter on camera and **C**) and **D**) were taken without filter. In **A**) and **C**), the LB agar plates contained 1 mM IPTG, but no UAA; in **B**) and **D**), the LB agar plates contained 1 mM OnbSTyr. The numbers indicate different synthetase constructs as shown in **Table I-1**.

I-3 Ongoing/Future Experiments

1. Purification of proteins containing OnbSTyr using synthetases #5 and #10

2. MS confirmation of the MW of the purified protein

3. Decaging experiments

I-4 Materials and Methods

I-4-1 Construction of plasmids for UAA incorporation

pTrcHisA-sfGFP-Y66tag was constructed by performing site-directed mutagenesis on previously reported pTrcHisA-sfGFP plasmid using primers sfGFP_Y66tag_FWD and sfGFP_Y66tag_REV (**Appendix III**). pTrcHisA-sfGFP-V151tag was constructed as above using primers sfGFP_V151tag_FWD and sfGFP_V151tag_REV (**Appendix III**). To clone SynUb gene into pET11a, SynUb gene was amplified by PCR using primers SynUb #1 and SynUb_pET11a_REV (**Appendix III**). The PCR product was digested with *Nde* I/*Nhe* I and then ligated into the pET11a vector treated with the same restriction enzymes. The Ub protein expressed from this pET vector contained a C-terminal His tag.

A series of pSup-PylRS constructs containing N346A/C348A double mutations (*117*) were generated by site-directed mutagenesis using primers PylRS_N346A/C348A _FWD and PylRS_ N346A/C348A _REV (**Appendix III**). These two mutations were incorporated into the following PylRS constructs: WT PylRS, Y384F (BocLys) (*115*), Y306A/Y384F (CbzLys) (*115*), Y306M/L309A/C348A/Y384F (OnbLys) (*121*) and Y306M/L309A/Y384F (Interm PylRS).

I-4-2 Expression and purification of Ub and sfGFP containing OnbSTyr

pSup-E10 (*138*) was co-transformed with pTXB1-SynUb-K48tag (*100*), pET11a-SynUb-His or pTrcHisA-sfGFP-Y66tag into BL21 (DE3) cells, respectively. A single colony

was picked from each plate to grow a 5 mL overnight starter culture. This starter culture was then inoculated into 100 mL 2×YT (Amp100/Chl35). The 50 mL culture was grown to an OD_{600} of 0.5, centrifuged, and resuspended in 10 mL fresh 2×YT media (10-fold condensation). The condensed culture was supplemented with Amp100/Chl35, 1 mM IPTG and corresponding UAAs to induce the expression. Ub-K48OnbSTyr was purified using IMPACT kit (NEB) as described (*100*). Ub-His-K48OnbSTyr and sfGFP-Y66OnbSTyr were purified using His tag affinity purification resins (Promega). The purified proteins were concentrated to 100 μL and then confirmed by SDS-PAGE gel. Ub-K48OnbSTyr expressed from pTXB1 vector was analyzed by ESI-MS at Maryland Proteomics Core Facility. Ub-His-K48OnbSTyr expressed from pET11a vector was analyzed by ESI-MS at Virginia Commonwealth University Proteomics Facility.

I-4-3 Scanning PylRS variants for activities toward OnbSTyr

To test the activities of the above PylRS variants toward OnbSTyr, pSup plasmids carrying these PylRS variants were co-transformed with pTrcHisA-sfGFP-V151tag into GeneHogs cells and plated on LB agar (Amp100/Chl35). A single colony was picked from each plate and grown to OD_{600} of 0.8. The culture (50 μL) of each variant was plated on two LB agar plates (Amp100/Chl35) containing 1 mM IPTG, no UAAs or OnbSTyr. The fluorescence of the cells on these plates was inspected under UV light.

Appendix II: Evolution of Aminoacyl tRNA Synthetases Using *In Vitro* Compartmentalization

II-1 Introduction

In vivo evolution of tRNA synthetase is usually laborious and notorious for consumption of a large amount of UAAs. Furthermore, tRNA synthetases can not be evolved *in vivo* to incorporate cytotoxic or cell-impermeable UAAs into proteins. To overcome these limitations, we intended to develop an *in vitro* system for evolving synthetase functions. This project is still being developed. Introduced below are the rationale of the proposed experiments and some preliminary results.

II-2 Rationale

This experiment relies on *in vitro* compartmentalization technique, or so-called "water-in-oil". Water-in-oil emulsions are created by mixing aqueous and oil phase solutions with the help of surfactant. The aqueous phase is cell-free transcription/translation system containing the DNA template. The ratio of genes to droplets is controlled such that most droplets contain no more than a single gene statistically.

To evolve the substrate specificity of tRNA synthetases, a DNA construct is designed to carry the gene library of a target synthetase and the ORF of streptavidin (STA) which bears an amber stop codon (TAG). During *in vitro* translation, **only the functional synthetase can suppress the stop codon in STA gene and result in a full-length STA**, which will bind to the biotinylated DNA template. This way, the functional synthetase gene is linked to STA protein that contains an affinity tag for purification thereafter. After several rounds of selection, the functional gene will be dramatically enriched in the whole population.
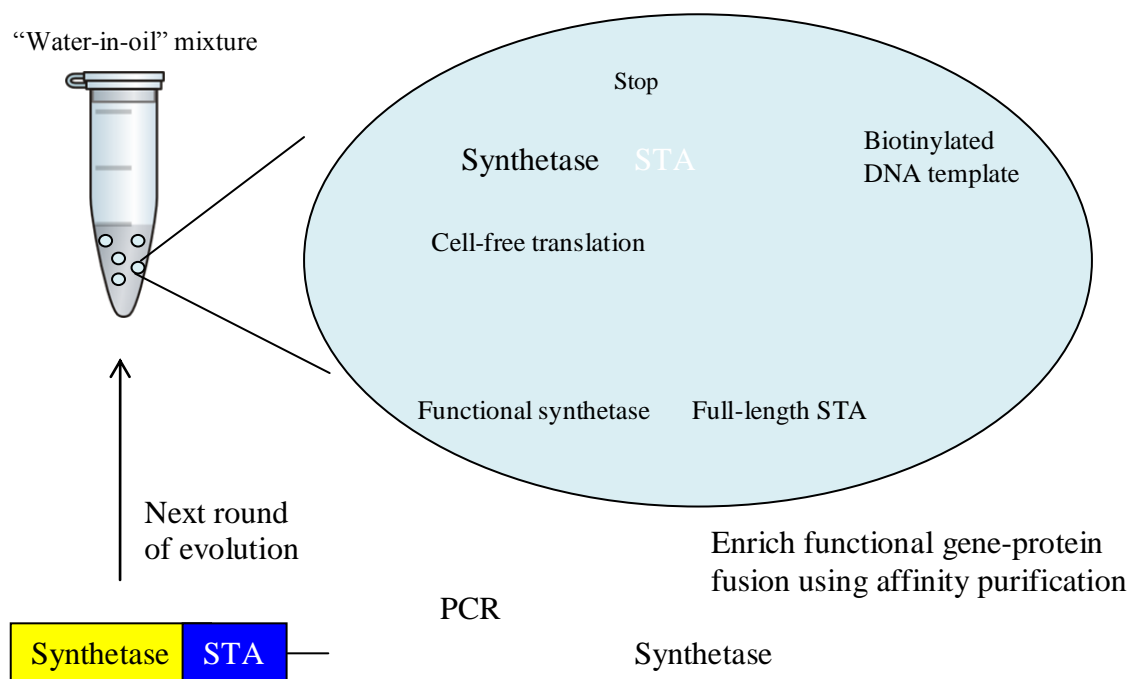
Figure II-1. Proposed experiment for *in vitro* evolution of synthetases.

II-3 Preliminary Results

II-3-1 IVC construct

The DNA template (IVC construct) used in the selection is a PCR product, amplified from a plasmid pIVC using biotinylated primers. The IVC construct contains the gene or gene library of a desired synthetase and STA gene that bears a stop codon. This linear DNA contains a single T7 promoter sequence but separate ribosome binding sites (RBS) such that a single messenger RNA (mRNA) will be used to translate two different proteins (see **Appendix IV** for DNA sequence of pIVC).

II-3-2 Cell-free translation using plasmid DNA

Three different cell-free translation systems have been tested: 1) NEB PURE kit which contained purified components required for protein translation; 2) Promega T7 High

Yield S30 Extract which is the *E. coli* cell extract optimized for expression of proteins under the control of T7 promoter 3) home-made S30 extract.

NEB PURE kit
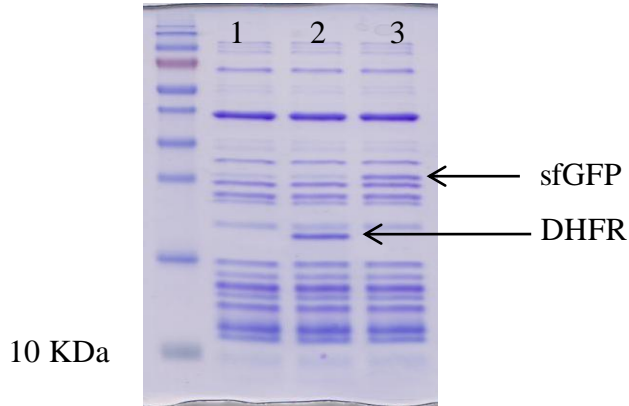
Figure II-2. *In vitro* sfGFP expression using NEB PURE kit. Lane 1: no DNA; Lane 2: 10 ng/μL dihydrofolate reductase (DHFR) containing plasmid, as a positive control; Lane 3: 10 ng/μL pRSET-sfGFP. The reactions were incubated at 37 ℃ for 2 h.

Figure II-3. *In vitro* expression of synthetases from pIVC constructs using NEB PURE kit, 5 μL reactions. Lane 1: no DNA; Lane 2~4: 10 ng/μL of pIVC-*Mj*YRS(WT)-STA, pIVC-*Mm*PylRS(BocLys)-STA(1TAG) and pIVC-*Mm*PylRS(BocLys)-STA(2TAG), respectively. STA expression was not observed due to the overlapping of protein bands.

Promega T7 High Yield S30 kit
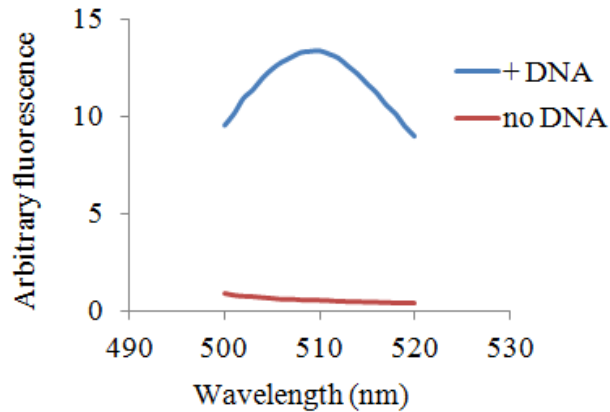
Figure II-4. *In vitro* expression of sfGFP, using Promega T7 High Yield S30 kit. Test, 20 µg/mL pRSET-sfGFP plasmid; negative control, no DNA. The reactions (10 µL) were incubated at 37 ℃ for 2 h. The GFP fluorescence was monitored on a fluorimeter with an Ex wavelength of 485 nm.

Interestingly, this kit did not work for linear DNA as shown by later experiments.

Home-made S30

The home-made S30 reactions were tested using both sfGFP reporter (**Figure II-5**) and luciferase reporter (**Figure II-6**). tRNA concentration was also optimized (**Figure II-7**).
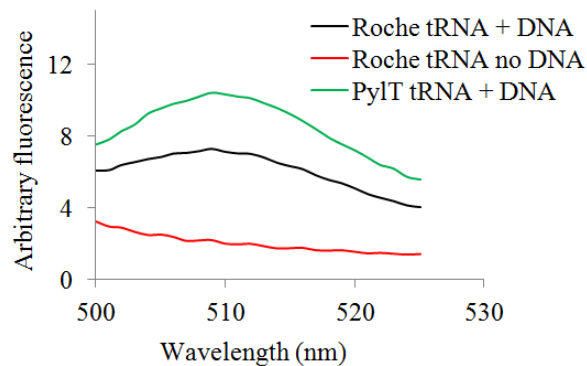
Figure II-5. *In vitro* expression of sfGFP using home-made S30, with 16 µg/mL pRSET-sfGFP plasmid and 0.17 mg/mL tRNA. The reactions (300 µL) were incubated at 37 ℃ for 2 h.

Figure II-6. *In vitro* expression of *Renilla* luciferase using home-made S30. Sample 1, no DNA; sample 2, 13 µg/mL *Renilla* luciferase control plasmid (Promega). The reactions (300 µL) were incubated at 37 °C for 2 h. The luminescence assay was monitored with a ~5 s delay and an integration time of 10 s.
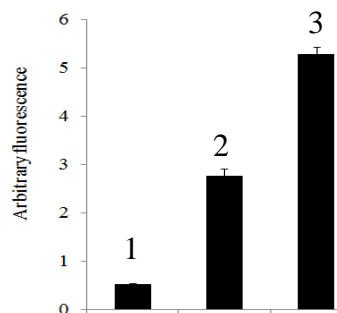
Figure II-7. Optimization of tRNA concentration in home-made S30 reaction, using 13 µg/mL *Renilla* luciferase control plasmid (Promega). Sample 1, no tRNA; Sample 2, 0.17 mg/mL total tRNA; Sample 3, 0.85 mg/mL total tRNA. Three replicates were performed for each sample.

II-3-3 Cell-free translation using linear DNA and emulsified reactions

As our experimental design requires the cell-free translation to be performed on biotinylated PCR product in emulsified condition, it is advisible to examine if linear DNA or "water-in-oil" has any effect on the translation reation. These experimental tests were performed using home-made S30 cell extract. It was evident that the home-made S30 reaction worked for both linear DNA template and emulsified reaction. However, both of these conditions decreased the translation efficiency by ~50%.

163
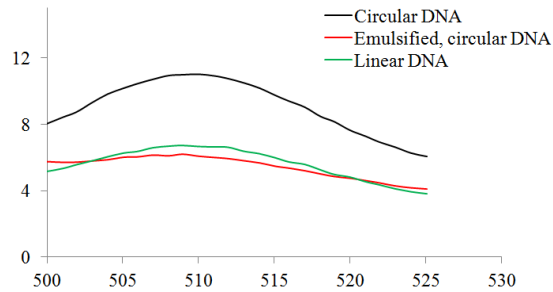
Figure II-8. Cell-free translation using linear DNA (6.7 μg/mL PCR product T7-sfGFP) and emulsified translation reactions (fluorescence assayed after ether extraction).

II-3-4 Isolation of E. coli total tRNA

We first focused on production of proteins containing UAAs using PylRS/PylT pairs. The PylT tRNA was prepared by two different methods: 1) isolation of total tRNA from an *E. coli* strain constitutively expressing PylT; 2) *in vitro* transcription (run-off) reaction.
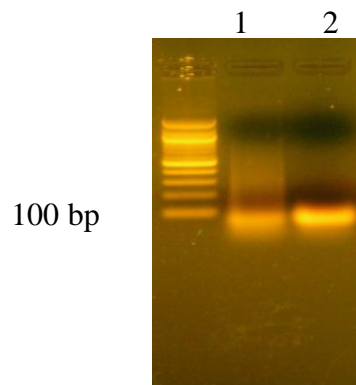
Figure II-9. Isolated total *E. coli* tRNA. Lane 1, 0.4 μg of commercial total *E. coli* tRNA (Sigma); Lane 2, 1.85 μg of home-made total *E. coli* tRNA.
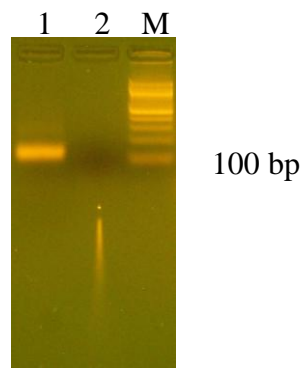
Figure II-10. Run-off transcription. Lane 1, with 5 mM NTPs; Lane 2, no NTPs.

164

II-3-5 Cell-free translation of proteins containing UAAs
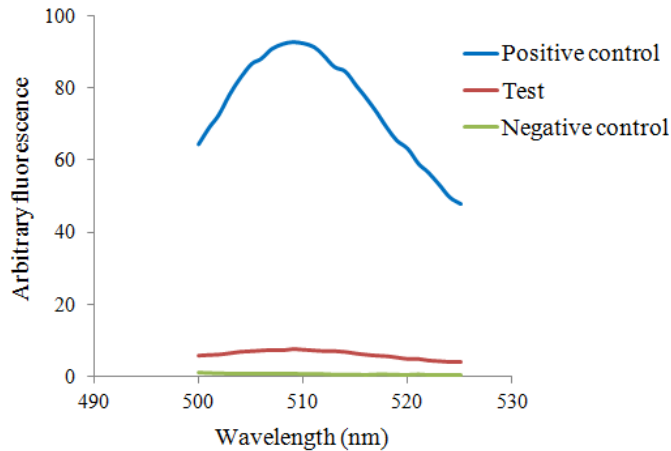
Figure II-11. *In vitro* expression of sfGFP containing BocLys at position 151, using NEB PURE reaction (10 μL). Positive control, 5 ng/μL pRSET-sfGFP (without stop codon); negative control, 5 ng/μL pRSET-sfGFP-V151tag, 73 ng/μL *Mb*PylT total tRNA, 5 ng/μL pIVC- *Mm*PylRS(BocLys)-STA(1TAG); test, the same as negative control, supplemented with 2 mM BocLys amino acid.

NEB PURE kit gave promising results. However, Promega S30 and home-made S30 did not show evident production of proteins containing UAAs.

II-3-6 Examination of binding specificities of nickel resins and antibodies

In the IVC experiment, the basis of the selection is that the functional (full-length) STA captures the biotinylated template DNA, and this DNA-protein fusion is enriched by His tag affinity purification after breaking emulsion (in the presence of excess free biotin to prevent the binding between excess functional STA and non-functional DNA). We were concerned that the functional DNA-protein fusion may be difficult to enrich by traditionally used nickel resin-based His tag purification as the negative charges on DNA may interefere with the positively charged resins.

To test this possibility, linear PCR product of IVC-*Mj*YRS(WT)-STA and IVC-*Mm*PylRS(BocLys)-STA(2TAG) (1:1 molar ratio) were incubated with nickel resins under varying NaCl concentrations. The resins were washed and then applied in PCR reactions to amplify any residing DNAs (**Figure II-12**). The results suggested that 1) high concentration of NaCl inhibits PCR reaction (from **Lane 3** and **4**) and that 2) nickel resins could indeed nonspecifically capture DNAs (from **Lane 9~12**).
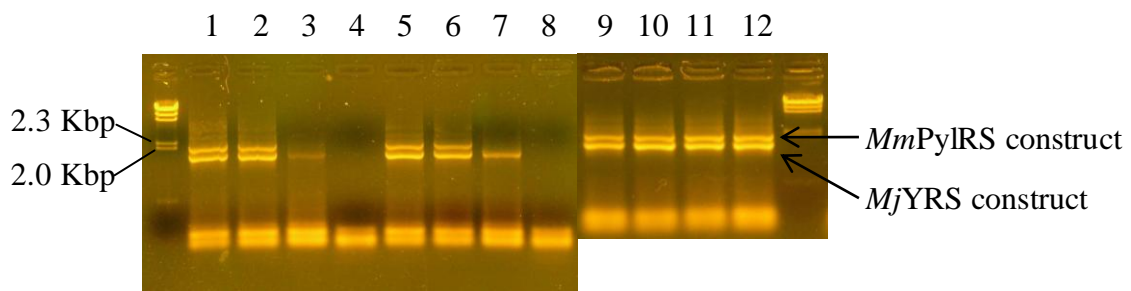
Figure II-12. PCR tests of non-specific binding between nickel resins and DNA templates, under varying NaCl concentrations. Lane 1~4: starting solution (prior to binding with resins); Lane 5~8: binding flow-through; Lane 9~12, eluted solution from resin. The NaCl concentrations are: Lane 1, 5 and 9, 50 mM; Lane 2, 6 and 10, 200 mM; Lane 3, 7 and11, 500 mM; Lane 4, 8, 12, 1 M.

We then intended to examine if non-specific binding also existed between anti-His antibodies and DNA templates. The binding between antibodies and His$_6$ tag is more specific as it is selective not only for charges but also the structure of the His$_6$ tag. We expected that antibody could solve the problem of non-specific binding. The antibody selection system was based on rabbit antibodies and protein A immobilized on magnetic beads that could bind IgG antibodies. Two different antibodies, anti-His tag and anti-FLAG tag, were examined. The results suggested that there were not many DNA residing on the Protein A magnetic beads. However, it was not clear if the antibody could bind to the protein A beads properly. Futher experiments are required.

1 2 3 4

2.3 Kbp
2.0 Kbp

*Mm*PylRS construct

*Mj*YRS construct

Figure II-13. PCR tests of non-specific binding between antibodies and DNA templates. Lane 1, wash solution, anti-His tag antibody; Lane 2, wash solution, anti-FLAG tag antibody; Lane 3, protein A beads, anti-His; Lane 4, protein Abeads, anti-FLAG.

II-3-7 IVC Model selection

Before the IVC is used for the evolution of synthetases, we set up control experiments to model the selection process. Functional gene (PCR product IVC-*Mj*YRS(WT)-STA) and non-functional gene (PCR product IVC-*Mm*PylRS(BocLys)-STA(2TAG)) were used in 1:1 ratio as the starting template in emulsified NEB PURE translation reactions. The functional gene has plain STA and thus can produce full-length STA which can capture the biotinylated PCR product. The non-functional DNA cannot produce full-length STA. Ideally, the functional gene-STA fusion should be readily enriched by His tag affinity purification. In our experiment, a 5-fold enrichment was indeed observed, as determined by Image J (**Figure II-14**). However, the conditions may need further optimizations because 1) there were non-specific PCR products and 2) fold enrichment was insufficient as compared with the 50-fold enrichment reported in previous study (*163*).
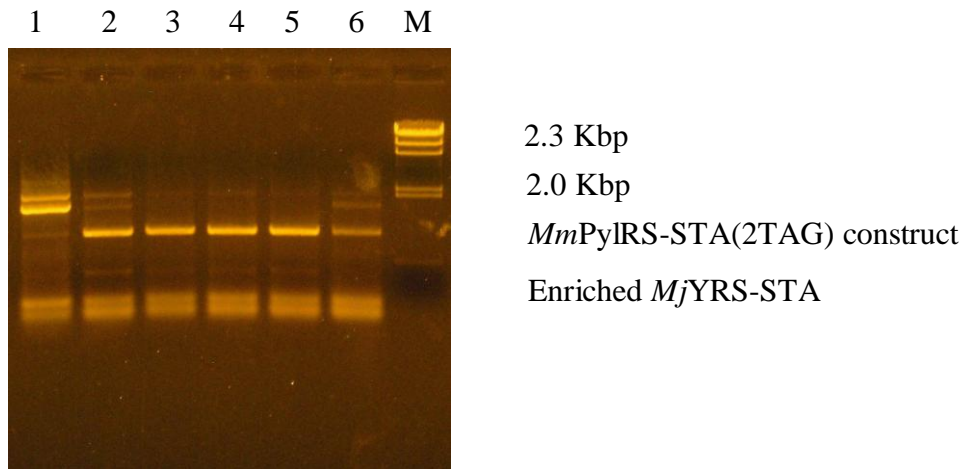
Figure II-14. IVC model selection. Lane 1, binding flow-through; Lane 2~5, wash 1~4, respectively; Lane 6, protein A beads bound with protein-DNA fusion.

II-4 Ongoing/Future Experiments

1. Repeat the *in vitro* expression of sfGFP-V151BocLys using NEB PURE kit

2. Optimization of IVC selection conditions (to achieve >10-fold enrichment)

3. Construct synthetase library using overlapping PCR

4. Selection of functional synthetase gene using IVC

II-5 Materials and Methods

II-5-1 Construction of pIVC plasmids containing synthetases and STA genes

WT *Mj*YRS and STA gene bearing no stop codon was amplified then cloned into pET24a vector. The WT *Mj*YRS gene in this plasmid has been replaced by *Mj*YRS variant for pBpa incorporation and PylRS Y384F variant for BocLys incorporation. To clone the *Mj*YRS-pBpa gene, PCR was performed on pSup-pBpa plasmid (*138*) using primers *Mj*YRS_pIVC_FWD and *Mj*YRS_pIVC_REV. The PCR product was digested with *Nde*

I/*Pst* I and ligated with the pIVC-*Mj*YRS(WT)-STA treated witht the same restriction enzymes. For PylRS(BocLys) synthetase, pSup-BocLys plasmid was PCR amplified using primers PylRS_FWD and PylRS_REV. The PCR product was digested by *Nde* I/*Nsi* I and then cloned into pIVC construct pre-cut with *Nde* I/*Pst* I. To introduce the TAG stop codons into STA gene, site-directed mutagenesis was performed on pIVC plasmids using primers STA_TAG_REV and STA_one TAG_FWD (for one stop codon) or STA_two TAG_FWD (**Appendix III**). In summary, the following plasmids were constructed: pIVC-*Mj*YRS(WT)-STA, pIVC-*Mj*YRS(pBpa)-STA, pIVC-*Mm*PylRS(BocLys)-STA, pIVC-*Mj*YRS(WT)-STA(1TAG), *Mj*YRS(pBpa)-STA(1TAG), pIVC-*Mm*PylRS(BocLys)-STA(1TAG), pIVC-*Mj*YRS(WT)-STA(2TAG), pIVC-*Mj*YRS(pBpa)-STA(2TAG) and pIVC-*Mm*PylRS(BocLys)-STA(2TAG).

II-5-2 Cell-free translation reactions using commercial kits

To test the efficiency of different translation systems, sfGFP reporter gene was cloned into pRSET vector to create pRSET-sfGFP tester plasmid.To create reporter gene for UAA incorporation, stop codon was introduced into V151 of sfGFP by PCR amplification of pRSET-sfGFP plasmid using primers sfGFP_V151tag_FWD and sfGFP_V151tag_REV.

For translation reaction using NEB PURE kit, 50 μg pRSET-sfGFP was added into 5 μL reaction solution containing 2 μL solution A and 1.5 μL solution B. The reaction was incubated at 37 ℃ on a PCR thermocycler for 2 h and the fluorescence was monitored by Cary Eclipse fluorescence spectrophotometer (Varian Inc., Palo Alto, CA). To test the expression of synthetases from pIVC constructs, 50 μg of pIVC-*Mj*YRS(WT)-STA, pIVC-*Mm*PylRS(BocLys)-STA(1TAG) and pIVC-*Mm*PylRS(BocLys)-STA(2TAG) were

added to 5μL of reaction solution, respectively. The production of synthetases was monitored by SDS-PAGE gel.

For translation reaction using Promega T7 S30 kit, 100 μg pRSET-sfGFP was added into 10 μL reaction solution containing 4 μL premix solution and 3.6 μL S30 extract. The reaction was incubated at 37 ℃ on a shaker for 2 h. The fluorescence was monitored.

II-5-3 Preparation of home-made E. coli *S30 cell extract*

The homemade *E. coli* S30 extract was produced using the reported protocols (*164, 165*). Freshly transformed BL21 (DE3) cells were grown in 2×YT media to an OD_{600} of 0.8. The cells were then harvested by centrifugation at 5,000 rpm for 10 min at 4 ℃. The cell pellet was washed twice with S30 buffer (10 mM Tris-acetate, pH 8.2, 60 mM potassium glutamate, 14 mM magnesium acetate, 1 mM DTT). The cell pellet was then weighed and stored at -80 ℃ for further application. To lyze the cells, 1 mL of S30 buffer was added per gram of cells. The cells were resuspended and then transferred to the 15 mL chamber of bead beater (Biospec Products, Bartlesville, OK) pre-filled to half way with cold 0.1 mm glass beads (Biospec). The outside chamber was filled with ice. The cells were lyzed by applying 20 pulses of 30 s vibration. The cell lysis were transferred into a centrifugation tube and then clarified by centrifugation at 30,000 g for 30 min twice. The supernatant was transferred into a new tube. To this supernatant was added 0.2 volume of pre-incubation solution containing 370 mM Tris-acetate, pH 8.2, 11.1 mM magnesium acetate, 5.5 mM DTT, 16.5 mM ATP (MP Biomedicals, Solon, OH), 50 μM each of the 20 natural amino acids (Acros), 105 mM phosphoenolpyruvate (Alfa Aesar, Ward Hill, MA) and 8.4 U/mL Calbiochem rabbit muscle pyruvate kinase (EMD Millipore, Rockland, MA). This solution was incubated at 37 ℃ for 80 min in dark to

consume all the endogenous mRNA. The solution became cloudy after incubation and was dialyzed using MWCO 6,000~8,000 dialysis tubing (Fisher Scientific, Pittsburgh, PA) against 100 volumes of S30 buffer for 40 min at 4 °C. The dialyzed solution was centrifuged at 3,000 g for 10 min at 4 °C. The supernatant was transferred into a new tube, aliquoted into 200 μL solution/tube and then flash-frozen. These cell extracts were kept at -80 °C.

II-5-4 Cell-free translation reactions using home-made S30 cell extract

The homemade S30 reaction contained 55 mM HEPES, pH 7.5, 1.7 mM DTT, 27.5 mM ammonium acetate, 208 mM potassium glutamate, 19.3 mM magnesium acetate, 3.5 mM each of the 20 natural amino acids, 68 μM folinic acid, 1.2 mM ATP, 0.85 mM each of GTP, CTP and UTP, 0.64 mM 3', 5'-cyclic adenosine monophosphate (cAMP), 0.25 mg/mL of creatine phosphokinase, 80 mM creatine phosphate, 0.17 mg/mL of *E. coli* total tRNA (Roche Applied Science, Indianapolis, IN), 0.1 mg/mL T7 RNA polymerase, 25% (v:v) of cell extract and 16 μg/mL DNAs. The reaction was incubated on a shaker at 37 °C for 4 h. For sfGFP reporter, the fluorescence was monitored directly after reaction. For luciferase reporter, 100 μL Renilla luciferase assay solution (99 μL assay buffer + 1 μL substrate) (Promega) was added to 50 mL S30 reaction and the luminescence was monitored on a luminometer.

To successfully perform this S30 reaction, the following stock solutions were prepared and added in the order as below: 1) nuclease-free water 2) 10× reaction buffer (550 mM HEPES, pH 7.5, 17 mM DTT, 275 mM ammonium acetate, 2.05 M potassium glutamate, 193 mM magnesium acetate); 3) 50 mM water soluble amino acid mix (Ala, Arg, Gly, His, Lys, Pro, Ser, Thr and Val), dissolved in water; 4) 50 mM of base soluble amino acid

171

mix (Ile and Phe), dissolved in 1 M NaOH; 5) 50 mM of acid soluble amino acid mix (Asp, Asn, Cys, Glu, Gln, Leu, Met, Trp and Tyr), dissolved in 1 M HCl; 6) 64 mM cAMP (100×), dissolved in water by dropwise addition of 1 M NaOH; 7) 6.8 mM folinic acid (100×), dissolved in water; 8) NTP mix containing 120 mM ATP and 85 mM each of GTP, CTP and UTP (100×), dissolved in water; 9) 1.6 M creatine phosphate (20×), dissolved in water; 10) 25 mg/mL creatine phosphate kinase (100×), dissolved in water; 11) 5 mg/mL home-purified T7 RNA polymerase (50×), in storage buffer (50 mM Tris-HCl, pH 7.9, 100 mM NaCl, 20 mM 2-mercaptoethanol, 1 mM EDTA, 50% Glycerol and 0.1% Triton X-100); 12) S30 cell extract, pre-thawed on ice (may take up to 1 h); 13) DNA template of high quality and cocentration, dissolved in nuclease-free water; 14) commercial or home-made *E. coli* total tRNA.

II-5-5 Cell-free translation using linear DNA and emulsified reactions

Home-made S30 was used for testing linear DNA template and emulsified reactions. For reactions with linear DNA, 6.7 mg/mL sfGFP PCR product (generated using primers IVC_FWD and IVC_REV, with T7 promoter and terminator) was used in translation reaction with other conditions the same as above. For emulsified S30 reaction, 16 μg/mL pRSET-sfGFP was used as template and the reaction was emulsified (see details below in Section *II-5-9*). The reaction was incubated as usual and the emulsion was removed by ether extraction (see Section *II-5-9*). The fluorescence of the ether-extracted solution was monitored.

II-5-6 Isolation of total tRNA

Total *E. coli* tRNA isolation was performed as reported (*166*). This protocol can be used for most *E. coli* strains including those containing suppressor tRNAs. A culture from the

desired strain was grown in 2×YT to an OD_{600} of 0.8 and then harvested by centrifugation. The cell paste was washed with distilled water for twice. To 10 g cell were added 17.5 mL 1% (w/v) NaCl and 23.75 mL water-saturated phenol (dissolve crystal phenol in distilled water; acidic pH is critical to keep DNA in the organic phase). This solution was incubated at room temperature with shaking for 30 min. To this cell lysis was added 2.375 mL chloroform. The mixture was incubated at room temperature with shaking for 15 min and cleared by centrifugation at 27,000 g for 30 min. The upper aqueous phase was transferred into new tubes (under acidic pH, DNA is protonated and more soluble in organic phase. RNAs which in general have lower pK_a than DNA due to the 2' hydroxyl group will be more soluble in aqueous phase). This aqueous solution was ethanol precipitated by adding 0.1 volume of 3 M sodium acetate, pH 5.2 and 2 volume of cold ethanol. Ethanol precipitation was performed at -20 °C for at least 2 h (overnight for best result). The precipitates were collected by centrifugation at 5,000 rpm for 20 min. The cell pellet was washed twice with 75% cold ethanol, air-dried until becoming transparent and then dissolved in 4 mL of 0.3 M sodium acetate, pH 7.0. To this dissolved pellet was added 6.5 mL isopropanol. This solution was kept on ice for 30 min and then centrifuged at 12,000 rpm for 10 min. The supernatant was transferred into a new tube (mRNA and rRNA are less soluble than tRNA. However, there might be low level 5 *S* rRNA contaminations). To this supernatant was added 5.56 mL isopropanol. The precipitates were collected by centrifugation at 12,000 rpm for 10 min, air-dried and then dissolved in 640 μL nuclease-free water (less volume can be used to increase the concentration of tRNA). The tRNA solution was aliquoted and stored at -20 °C. The quality of the prepared tRNA was assessed by running the samples on 2% agarose gels. OD_{260} and

OD$_{280}$ were determined to assess the purity and concentration (pure tRNA has an OD$_{260}$/OD$_{280}$ of ~2.0 and a concentration of 40 μg/mL for OD$_{260}$ of 1.0 if 1 cm path length is used for the measurement).

II-5-7 In vitro *transcription of tRNA (run-off reaction)*

Serine suppressor tRNA (SerU) was produced using run-off transcription. The DNA template was generated by overlapping PCR using primers SerU #1~3 (**Appendix III**). The PCR product was ethanol precipitated. The run-off reaction contained 40 mM TrisHCl, pH 7.5, 2.5 mM spermidine, 25 mM MgCl$_2$, 10 mM DTT, 5 mM each NTPs, 0.2 U/μL RNase inhibitor, 0.005 U/μL inorganic pyrophophatase, 50 μg/mL T7 RNA polymerase, 20 ng/μL SerU DNA template (~0.3 μM). The reaction was incubated at 37 ℃ overnight and the product was analyzed by 2% agarose gel.

II-5-8 In vitro *production of proteins containing UAAs*

In vitro translation of proteins containing UAAs has been tested using all the three above cell-free translation systems. For NEB PURE kit, 10 μL reactions were set up containing 4 μL solution A, 3 μL solution B, 73 ng/μL home-made total *E. coli* tRNA containing *Mb*PylT. The positive control contained 5 ng/μL pRSET-sfGFP DNA. Negative control contained 5 ng/μL pRSET-sfGFP-V151tag and 5 ng/μL pIVC-*Mm*PylRS(BocLys)-STA(1TAG). The test experiment contained the same components as in negative control, but is supplemented with 2 mM BocLys amino acid. Experiments on the other two translation systems followed the same procedure and DNA concentrations, with individual conditions as described above.

II-5-9 Examination of binding specificities of nickel resins and antibodies

pIVC-*Mj*YRS(WT)-STA and pIVC-*Mm*PylRS(BocLys)-STA(2TAG) were individually amplified in the PCR reactions containing $1 \times$ Phusion HF buffer (NEB), 0.2 mM dNTPs, 0.5 μM each of primers IVC_FWD and IVC_REV, and 0.01 U of Phusion DNA polymerase (NEB). The PCR products were biotinylated at both 5' and 3' ends. These PCR products were purified using QIAquick PCR purification kit (Qiagen). The concentrations were determined by running an agarose gel and comparing with a standard DNA with known concentration.

To test if there was non-specific binding between nickel resins and DNA, 15 ng each of IVC-*Mj*YRS(WT)-STA and IVC-*Mm*PylRS(BocLys)-STA(2TAG) linear PCR product (~1:1 molar ratio) were dissolved in 100 mM TrisHCl (pH 7.5) buffer containing 10 mM imidazole and varying concentrations of NaCl (50 mM, 200 mM, 500 mM and 1 M). These DNA solutions were incubated with 50 μL nickel resins (Promega) at room temperature for 30 min with shaking. The supernatant (binding flow-through) was then transferred to a new tube. The resins were washed twice with 100 mM TrisHCl buffer (pH 7.5) containing 50 mM imidazole. The washed resins were incubated with 100 mM Tris buffer (pH 7.5) containing 150 mM NaCl and 500 mM imidazole to elute all the residing DNAs from the resins. PCR reactions (20 μL) were performed on 1 μL each of the pre-binding solution, binding flow-through and eluted solution using primers IVC_FWD and IVC_REV. The PCR results were resolved on 1% agarose DNA gels.

To test if there was non-specific binding between anti-His antibody and DNA, 15 ng each of IVC-*Mj*YRS(WT)-STA and IVC-*Mm*PylRS(BocLys)-STA(2TAG) linear PCR product were mixed with 0.25 mg of rabbit anti-His or anti-FLAG antibody (Genscript,

Piscataway, NJ) and 5 μL settled Protein A magnetic beads (Genscript) in $1\times$ phosphate buffer saline (PBS). The mixture was incubated at 4 ℃ for 1 h with shaking. The supernatant (binding flow-through) was transferred into a new tube. The beads were washed with 500 μL PBS for three times. The first wash solution (1 μL) and 1 μL of the beads were used in PCR reactions as above to detect residing DNAs.

II-5-10 IVC Model selection

The emulsion oil contained 95% mineral oil, 4.5% Span 80, 0.5% Tween 80 and 0.1% Triton X-100 as described (*163, 167*). This oil solution was stirred on ice for 5 min prior to addition of the aqueous solution. Cell-free translation reaction (10 μL) was set up using NEB pure kit as follows: 4 μL solution A, 3 μL solution B, 0.9 ng of biotinylated IVC-*Mj*YRS(WT)-STA PCR product, 0.8 ng of biotinylated IVC-*Mm*PylRS(BocLys)-STA(2TAG) PCR product and 2 μL nuclease-free water. This 10 μL aqueous phase solution was added slowly into 100 μL spinning oil solution over 1 min. The mixture was continued stirring on ice for 5 min. Then this solution was incubated at 37 ℃ for 3 h without shaking. To the finished reaction was added 100 μL Tris buffer saline (TBS, 100 mM TrisHCl, pH 7.5, 150 mM NaCl) containing 100 μM biotin (Alfa Aesar). The free biotin was provided to quench the excess STA. To break the emulsion, 200 μL ethyl ether was added and the solution was vortexed and centrifuged at 13,000 g for 5 min. The upper ether layer was discarded and the ether extraction was performed for additional three times. After the last ether extraction, the residual ether was removed by vacuum centrifugation. To this aqueous solution was added 100 μL TBST solution (TBS supplemented with 0.1% Tween 20), 0.5 μg rabbit anti-His antibody and 10 μL settled protein A magnetic beads. This mixture was incubated at 4 ℃ for 1 h with shaking. The

176

supernatant (binding flow-through) was removed and the beads were washed with 200 μL TBST for four times. PCR reaction (10 μL) was performed as above on 1 μL of 1:10 diluted binding flow-through, 1 μL each of the four washes and 1 μL of the beads bound with DNA-protein fusion.

Appendix III: Primer Sequences

Removal of Mly I sites from pGFP$_{UV}$ (mutation sites lowercased)

234_FWD (CL866)   TGCATGCCTGCAGGTCGAtCTAGAGGATCC
234_REV (CL876)   TCGACCTGCAGGCATGCAAGCTTGGCGTAA
2334_FWD (CL880)   TTATCTACACGACGGGGAGcCAGGCAACTAT
2334_REV (CL877)   TCCCCGTCGTGTAGATAACTACGATACGGG
2836_FWD (CL870)   GTGTCTTACCGGGTTGGAtCAAGACGATAG
2836_REV (CL878)   TCCAACCCGGTAAGACACGACTTATCGCCA
3322_FWD (CL872)   AACGACCGAGCGCAGCGAtCAGTGAGCGAG
3322_REV (CL879)   TCGCTGCGCTCGGTCGTTCGGCTGCGGCGA

Cloning GFP$_{UV}$ into pET28b(+) vector (restriction sites underlined)

GFP$_{UV}$_FWD (*Nhe* I) (CL959)   CTA<u>GCTAGC</u>ATGAGTAAAGGAGAAGAACTT
GFP$_{UV}$_REV (*Hind* III) (CL960)   CCC<u>AAGCTT</u>TTATTTGTAGAGCTCATC

Incorporation of F64L and S30R folding-mutations (mutation sites lowercased)

F64L_FWD (CL930) GGCCAACACTTGTCACTACTcttTCTTATGGTGTT
F64L_REV (CL931) AGTAGTGACAAGTGTTGGCCATGGAACAGGTA

F64L_FWD for P58Δ[a] (CL1038):
ATGGACACTTGTCACTACTcttTCTTATGGTGTT

F64L_REV for P58Δ[a] (CL1039):
AGTAGTGACAAGTGTCCATGGAACAGGTAGT

S30R_FWD (CL928) ATGGGCACAAATTTTCTGTCcgcGGAGAGGGTGAA
S30R_REV (CL929) GACAGAAAATTTGTGCCCATTAACATCACCAT
C48Δ_FWD (CL916) AAACTTACCCTTAAATTTATCACTACTGGAAA
C48Δ_REV (CL917) ATAAATTTAAGGGTAAGTTTTCCGTATGTTGC

[a] In case that two deletions/mutations are too close, a special primer including both
    deletions/mutations is designed

Generation of mutants with double internal deletions

P75/D76→H_FWD (CL926) TCAATGCTTTTCCCGTTATCATCATATGAAAC
P75/D76→H_REV (CL927) GATAACGGGAAAAGCATTGAACACCATAAGAG
P75Δ_FWD (CL918)   TTCAATGCTTTTCCCGTTATGATCATATGAAA
P75Δ_REV (CL919)   ATAACGGGAAAAGCATTGAACACCATAAGAGA

E172Δ_FWD (CL920)
TTCAAAATTCGCCACAACATAGATGGATCCGT

E172Δ_REV (CL921)

ATGTTGTGGCGAATTTTGAAGTTAGCTTTGAT

S175/V176→F_FWD (CL922)
CCACAACATTGAAGATGGATTTCAACTAGCAG

S175/V176→F_REV (CL923)
ATCCATCTTCAATGTTGTGGCGAATTTTGAAG

S175/V176→F for E172Δ_FWD[a] (CL961)
TCGCCACAACATAGATGGATTTCAACTAGCAG

S175/V176→F for E172Δ_REV[a] (CL962)
ATCCATCTATGTTGTGGCGAATTTTGAAGTTA

[a] In case that two deletions/mutations are too close, a special primer including both deletions/mutations is designed

Removal of restriction sites from pTrcHisA (mutation sites lowercased)

Bsg6_FWD (CL1089)	TTGACAGCTTATCATCGAtTGCACGGTGCAC
Bsg6_REV (CL1090)	TCGATGATAAGCTGTCAAACCAGATCAATT
Bsg100_FWD (CL1091)	GGAAGCTGTGGTATGGCTcTGCAGGTCGTAA
Bsg100_REV (CL1092)	AGCCATACCACAGCTTCCGATGGCTGCCTG
Bsg3494_FWD (CL1093)	CACCTCCAGTCTGGCCCTtCACGCGCCGTCG
Bsg3494_REV (CL1094)	GGCCAGACTGGAGGTGGCAACGCCAATCAG
Bsg3694_FWD (CL1095)	CATTGCTGTGGAAGCTGCgTGCACTAATGTT
Bsg3694_REV (CL1096)	GCAGCTTCCACAGCAATGGCATCCTGGTCA
Mme2366_FWD (CL1122)	TAAGTCGTGTCTTACCGGaTTGGACTCAAGA
Mme2366_REV (CL1123)	CCGGTAAGACACGACTTATCGCCACTGGCA
Mme2550_FWD (CL1124)	TATCCGGTAAGCGGCAGGaTCGGAACAGGAG
Mme2550_REV (CL1125)	CCTGCCGCTTACCGGATACCTGTCCGCCTT
Bpm1802_FWD (CL1177)	GGTTTATTGCTGATAAATCaGGAGCCGGTGAG
Bpm1802_REV (CL1178)	ATTTATCAGCAATAAACCAGCCAGCCGGAA
Bpm3481_FWD (CL1192)	TGATTGGCGTTGCCACCTCaAGTCTGGCCCTT
Bpm3481_REV (CL1180)	GTGGCAACGCCAATCAGCAACGACTGTTTG
Bpm3970_FWD (CL1193)	AGCGGAACGGGAAGGCGAtTGGAGTGCCATGT
Bpm3970_REV (CL1182)	TCGCCTTCCCGTTCCGCTATCGGCTGAATT
Sap_FWD (CL1173)	TCAGTGAGCGAGGAAGCGGttGAGCGCCTGATG
Sap_REV (CL1174)	CGCTTCCTCGCTCACTGACTCGCTGCGCTCGGT

Note that only *Bsg* I and *Bpm* I sites need to be removed for this study. Other restriction sites (*Mme* I and *Sap* I) were removed to accommodate early tests.

M. barkeri *pyrrolysyl tRNA synthetic gene with upstream* proK *promoter*

PylT#1 (CL618)
AAAGCGGCCGCAAAACTAGTGGCAGCGGCTAACTAAGCGGCCTGCTGACTTT
CTCGCCGATCAAAAGGC

PylT#2 (CL619)
CAGATACGCCCTCGTCAATCCCTTAATAGCAAAATGCCTTTTGATCGGCGAGA
AAGTC

PylT#3 (CL620)
AGGGATTGACGAGGGCGTATCTGCGCAGTAAGATGCGCCCCGCATTGGGAAC
CTGA

PylT#4 (CL621)
CCCGGCTGAACGGATTTAGAGTCCATTCGATCTACATGATCAGGTTCCCAATG
CGGGGC

PylT#5 (CL622)
GACTCTAAATCCGTTCAGCCGGGTTAGATTCCCGGGGTTTCCGCCAAATTCGA
AAAGCCTG

PylT#6 (CL623)
TTTGTCGACCAAAAAGCCTGCTCGTTGAGCAGGCTTTTCGAATTTGGCGGAA
AC

Cloning of proK_Mb*PylT gene into pSup vector (restriction sites underlined)*

pSup_PylT_FWD (*ApaL* I) (CL1128)
AACCAA<u>GTGCAG</u>TGATCAAAAGCGGCCGCAAAACTAGTGGCAGCG

pSup_PylT_REV (*Xho* I) (CL1129)
AACAA<u>CTCGAG</u>TTTGTCGACCAAAAAGCCT

Cloning of proK_Mb*PylT gene into pYC vector (restriction sites underlined)*

pYC_PylT_FWD (*Pst* I) (CL1103)
AAAA<u>CTGCAG</u>TTTGTCGACCAAAAAGCCT

pYC_PylT_REV (*Not* I)   (CL1116)
AAA<u>GCGGCCGC</u>AAAGCGGCCGCAAAACTAGTGGCAGCG

Cloning of Mm*PylRS gene into pSup or pBK vector (restriction sites underlined)*

PylRS_FWD (*Nde* I) (CL609)
AAAGCGGCCGCAAA<u>CATATG</u>GATAAAAAACCACTAAACACTCT

PylRS_REV (*Nsi* I) (CL1126)
ACACA<u>ATGCAT</u>TTACAGGTTGGTAGAAATCCCGTTATAGTA

Generation of Mm*PylRS mutants for lysine derivatives (*mutation sites lowercased*)*

PylRS_Y384F_FWD (CL1114)   GCGATTCCTGCATGGTCTtTGGGGATACCCT
PylRS_Y384F_REV (CL1115)   AGACCATGCAGGAATCGCCTACGATCTTGA

PylRS_Y306A_FWD (CL1112)
ATGCTTGCTCCAAACCTTgcgAACTACCTGCGC

PylRS_Y306A_REV (CL1113)   AAGGTTTGGAGCAAGCATGGGTCTCAGGCA

PylRS_Y306M&L309A_FWD (CL1198)
ATGCTTGCTCCAAACCTTatgAACTACgcgCGCAAGCTTGAC

PylRS_C348A_FWD (CL1199)
      TTTACCATGCTGAACTTCgcgCAGATGGGATCG

PylRS_C348A_REV (CL1200)
      GAAGTTCAGCATGGTAAACTCTTCGAGGTGTTC

Generation of Codon-optimized human Ub (restriction sites underlined)

SynUb #1 (*Nde*I) (CL1158)
GGAATTC<u>CATATG</u>CAGATTTTTGTGAAAACCCTGACCG

SynUb #2 (CL1159)
ACTTCCAGGGTAATGGTTTTGCCGGTCAGGGTTTTCACAAA

SynUb #3 (CL1160)
AAAACCATTACCCTGGAAGTGGAACCGAGCGATACCATTGA

SynUb #4 (CL1161) CCTGAATTTTCGCTTTCACATTTTCAATGGTATCGCTCGGT

SynUb #5 (CL1162)
ATGTGAAAGCGAAAATTCAGGATAAAGAAGGCATTCCGCCG

SynUb #6 (CL1163)
CCGCAAAAATCAGACGCTGCTGATCCGGCGGAATGCCTTCT

SynUb #7 (CL1164)
AGCGTCTGATTTTTGCGGGCAAACAGCTGGAAGATGGTCGT

SynUb #8 (CL1165)
CTGAATATTGTAATCGCTCAGGGTACGACCATCTTCCAGCT

SynUb #9 (CL1166)
CCTGAGCGATTACAATATTCAGAAAGAAAGCACCCTGCATC

SynUb #10 (*BamH*I) (CL1167)
CGC<u>GGATCC</u>TTAGCCGCCACGCAGACGCAGCACCAGATGCAGGGTGCTTTCT
T

Cloning Ub into pTXB1 plasmid (restriction sites underlined)

pTXB1_FWD (*Nde*I) (CL1216)   GGAATTC<u>CATATG</u>CAGATTTTTGTGA

pTXB1_REV (*Sap*I) (CL1217)
      GGTGGTT<u>GCTCTT</u>CCGCAGCCGCCACGCAGACGCA

Introduction of TAG stop codons into lysine residues in Ub (mutation site underlined)

K48tag_FWD (CL1231)   CGTCTGATTTTTGCGGGCtagCAGCTGGAAGAT

K48tag_REV (CL1232)	GCCCGCAAAAATCAGACGCTGCTGATCCGG
K11tag_FWD (CL1212)	GTGAAACCCTGACCGGCtagACCATTACCCTG
K11tag_REV (CL1213)	GCCGGTCAGGGTTTTCACAAAAATCTGCAT
K63tag_FWD (CL1214)	AGCGATTACAATATTCAGtagGAAAGCACCCTG
K63tag_REV (CL1215)	CTGAATATTGTAATCGCTCAGGGTACGACC

Generation of codon-optimized sfGFP (restriction sites underlined)

sfGFP #1 (*Nhe* I) (CL966)
CTAGCTAGCATGATGAGCAAAGGCGAAGAACTGTTTACCGGCGTGGTTCCG

sfGFP #2 (CL967)
ATCGCCATCCAGTTCCACCAGAATCGGAACCACGCCGGTAAA

sfGFP #3 (CL968)
GGTGGAACTGGATGGCGATGTAAATGGCCACAAGTTTAGCGT

sfGFP #4 (CL969)
TCGCCCTCCCCTTCCCCACGCACGCTAAACTTGTGGCCATTT

sfGFP #5 (CL970)
GGGAAGGGGAGGGCGATGCGACCAATGGCAAACTGACCCTGA

sfGFP #6 (CL971)
TTCCCGGTCGTGCAAATAAACTTCAGGGTCAGTTTGCCATTG

sfGFP #7 (CL972)
GTTTATTTGCACGACCGGGAAACTGCCGGTTCCTTGGCCCAC

sfGFP #8 (CL973)
GCCATACGTCAGGGTGGTGACAAGGGTGGGCCAAGGAACCGG

sfGFP #9 (CL974)
CACCACCCTGACGTATGGCGTGCAATGCTTTAGCCGTTACCC

sfGFP #10 (CL975)
TCATGCCGCTTCATGTGGTCCGGGTAACGGCTAAAGCATTGC

sfGFP #11 (CL976)
ACCACATGAAGCGGCATGACTTCTTCAAAAGCGCCATGCCTG

sfGFP #12 (CL977)
CGTCCGTTCCTGAACATAGCCTTCAGGCATGGCGCTTTTGAA

sfGFP #13 (CL978)
GCTATGTTCAGGAACGGACGATCTCGTTTAAGGATGACGGCA

sfGFP #14 (CL979)
CCTCCGCACGGGTCTTATAGGTGCCGTCATCCTTAAACGAGA

sfGFP #15 (CL980)
CTATAAGACCCGTGCGGAGGTCAAATTCGAAGGCGATACCCT

sfGFP #16 (CL981)
TCAGCTCAATGCGGTTCACCAGGGTATCGCCTTCGAATTTGA

sfGFP #17 (CL982)
GGTGAACCGCATTGAGCTGAAGGGCATCGACTTCAAAGAGGA

sfGFP #18 (CL983)
TTGTGCCCCAGTATGTTGCCATCCTCTTTGAAGTCGATGCCC

sfGFP #19 (CL984)
GGCAACATACTGGGGCACAAGCTGGAGTACAACTTCAACAGC

sfGFP #20 (CL985)
GCGGTGATGTAGACGTTGTGGCTGTTGAAGTTGTACTCCAGC

sfGFP #21 (CL986)
CACAACGTCTACATCACCGCCGACAAGCAGAAGAACGGCATT

sfGFP #22 (CL987)
CCGAATCTTGAAGTTGGCCTTAATGCCGTTCTTCTGCTTGTC

sfGFP #23 (CL988)
AAGGCCAACTTCAAGATTCGGCACAATGTGGAGGACGGAAGC

sfGFP #24 (CL989)
GATAATGATCCGCCAGCTGAACGCTTCCGTCCTCCACATTGT

sfGFP #25 (CL990)
GTTCAGCTGGCGGATCATTATCAACAGAATACCCCCATTGGC

sfGFP #26 (CL991)
CCGGCAGAAGCACGGGACCGTCGCCAATGGGGGTATTCTGTT

sfGFP #27 (CL992)
CCCGTGCTTCTGCCGGATAATCATTACTTGAGCACCCAGAGC

sfGFP #28 (CL993)
ATTCGGGTCCTTGCTCAGCACGCTCTGGGTGCTCAAGTAATG

sfGFP #29 (CL994)
GCTGAGCAAGGACCCGAATGAGAAACGGGATCACATGGTGCT

sfGFP #30 (CL995)
CGCAGCGGTCACAAATTCCAGCAGCACCATGTGATCCCGTTT

sfGFP #31 (CL996)
GGAATTTGTGACCGCTGCGGGCATTACACATGGCATGGATGA

sfGFP #32 (*Hind* III) (CL997)
CCC<u>AAGCTT</u>TTATTTATACAGTTCATCCATGCCATGTGTAATGCCCG

Introduction of TAG stop codon into sfGFP (mutation site lowercased)

sfGFP_Y66tag_FWD (CL1062)
CCCTTGTCACCACCCTGACGtagGGCGTGCAATGC

sfGFP_Y66tag_REV (CL1063)
CGTCAGGGTGGTGACAAGGGTGGGCCAAGGAA

sfGFP_V150tag_FWD (CL1074)   TTCAACAGCCACAACtagTACATCACCGCC

sfGFP_V150tag_REV (CL1075)   GTTGTGGCTGTTGAAGTTGTACTCCAG
sfGFP_V151tag_FWD (CL1276)   AACAGCCACAACGTCtagATCACCGCCGAC
sfGFP_V151tag_REV (CL1277)   GACGTTGTGGCTGTTGAAGTTGTACTC

Cloning of Ub into pET11a vector (restriction sites underlined)

SynUb_pET11a_REV (*BamH* I) (CL1377)
CGCGGATCCTTAATGATGATGATGATGATGGCCGCCACGCAGACG

Introduction of N346A/C348A double mutations into PylRS (mutation sites lowercased)

PylRS_ N346A/C348A _FWD (CL1422)
GAAGAGTTTACCATGCTGgccTTCgccCAGATGGGATCG

PylRS_ N346A/C348A _REV (CL1423)
CAGCATGGTAAACTCTTCGAGGTGTTCTTT

Introduction of TAG codons into STA gene (mutation sites lowercased)

STA_one TAG_FWD (CL1381)
CATCATCATAGCAGCGGCATCGAAGGCCGCtagCGCTTAATT

STA_two TAG_FWD (CL1382)
CATCATCATAGCAGCtagATCGAAGGCCGCtagCGCTTAATT

STA_TAG_REV (CL1369)
GCTGCTATGATGATGATGATGATGAGAACCAGAC

Cloning of synthetase genes into pIVC construct (restriction sites underlined)

*Mj*YRS_pIVC_FWD (*Nde* I) (CL1031)
AGGAATCCCCATATGGACGAATTTG

*Mj*YRS_pIVC_REV (*Pst* I) (CL1335)
AAACTGCAGTCTAGATTATAATCTCTTTCTAATTGGC

Amplification of IVC constructs

IVC_FWD (CL1353) GATCTCGATCCCGCGAAAT (5' biotinylated)
IVC_REV (CL1334) GCTAGTTATTGCTCAGCGG (5' biotinylated)

Generation of serine suppressor tRNA (SerU)

SerU #1 (CL1385)
tcgagatctcgatcccgcgaaattaatacgactcactataGGAGAGATGCCGGAGCGGCT

SerU #2 (CL1420)

GATGCCGGAGCGGCTGAACGGACCGGTCTCTAAAACCGGAGTAGGGGCAAC
TCTACCGGG

SerU #3 (CL1411)
TGGCGGAGAGAGGGGGATTTGAACCCCCGGTAGAGTTGCC

Appendix IV: DNA Sequences

pGFP~UV~
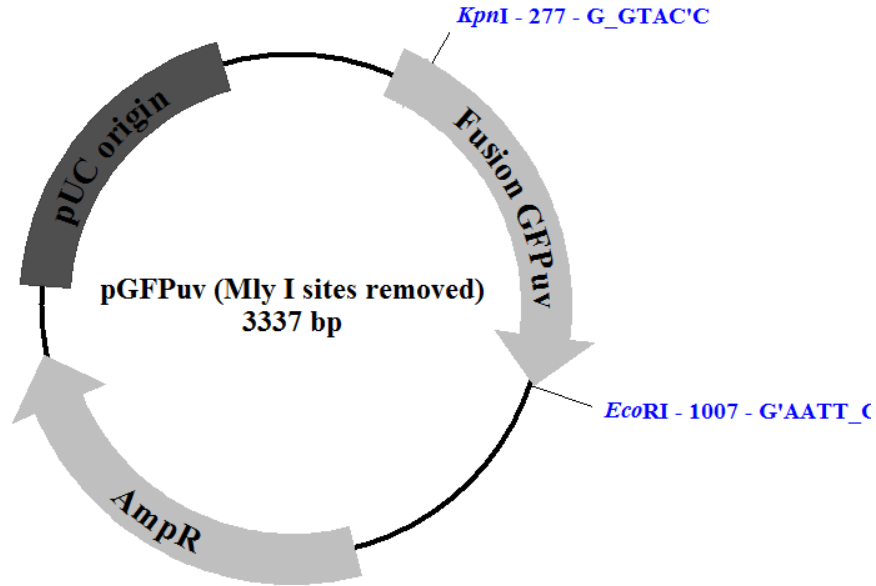

```
AGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATG
CAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGC
AATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGC
TTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGA
AACAGCTATGACCATGATTACGCCAAGCTTGCATGCCTGCAGGTCGATTCTA
GAGGATCCCCGGGTACCGGTAGAAAAAATGAGTAAAGGAGAAGAACTTTTC
ACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAA
ATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACC
CTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGT
CACTACTTTCTCTTATGGTGTTCAATGCTTTTCCCGTTATCCGGATCATATGAA
ACGGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAACGC
ACTATATCTTTCAAAGATGACGGGAACTACAAGACGCGTGCTGAAGTCAAGT
TTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAAGGTATTGATTTTAAA
GAAGATGGAAACATTCTCGGACACAAACTCGAGTACAACTATAACTCACACA
ATGTATACATCACGGCAGACAAACAAAAGAATGGAATCAAAGCTAACTTCAA
AATTCGCCACAACATTGAAGATGGATCCGTTCAACTAGCAGACCATTATCAA
CAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCT
GTCGACACAATCTGCCCTTTCGAAAGATCCCAACGAAAGCGTGACCACATG
GTCCTTCTTGAGTTTGTAACTGCTGCTGGGATTACACATGGCATGGATGAGCT
CTACAAATAATGAATTCCAACTGAGCGCCGGTCGCTACCATTACCAACTTGTC
TGGTGTCAAAAATAATAGGCCTACTAGTCGGCCGTACGGGCCCTTTCGTCTCG
CGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGAC
GGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGC
GCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTATGCGGCATCAG
AGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATG
```

186

```
CGTAAGGAGAAAATACCGCATCAGGCGGCCTTAAGGGCCTCGTGATACGCCT
ATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGGCA
CTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATT
CAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATAT
TGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTT
TTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAG
TAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGA
TCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAA
TGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGAC
GCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGG
TTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAG
AGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTA
CTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACA
TGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGC
CATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAAC
GTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAAT
TAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGC
CCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGT
CTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGT
AGTTATCTACACGACGGGGAGCCAGGCAACTATGGATGAACGAAATAGACA
GATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAA
GTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGG
ATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGA
GTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTT
GAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCG
CTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAA
GGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAG
CCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGC
TCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTA
CCGGGTTGGATTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTG
AACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGA
ACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGG
AGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGC
ACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGT
TTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGG
AGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTG
CTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAA
CCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACC
GAGCGCAGCGAGTTCAGTGAGCGAGGAAGCGGAAG
```

pTTGFP_UV

AAGCTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTT
GATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACG
AGTAGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTC
GGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGA
TAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCT
AATCAGAATTGGTTAATTGGTTGTAACACTGGCAGAGCATTACGCTGACTTG
ACGGGACGGCGGCTTTGTTGAATAAATCGAACTTTTGCTGAGTTGAAGGATC
TCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGAC
TGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATT
AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATT
GTGAGCGGATAACAATTTCAAGGAGATATACATATGAACAATAACGATCTCT
TTCAGGCATCACGTCGGCGTTTTCTGGCACAACTCGGCGGCTTAACCGTCGCC
GGGATGCTGGGGCCGTCATTGTTAACGCCGCGACGTGCGACTGCGCATAATG
GGTGCTTTGCCAAGGGTACCAATGTTTTAATGGCGGATGGGTCTATTGAATGT
ATTGAAAACATTGAGGTTGGTAATAAGGTCATGGGTAAAGATGGCAGACCTC
GTGAGGTAATTAAATTGCCCAGAGGAAGAGAAACTATGTACAGCGTCGTTCA
GAAAAGTCAGCACAGAGCCCACAAAAGTGATTCAAGTCGTGAAGTGCCAGA
ATTACTCAAGTTTACGTGTAATGCGACCCATGAGTTGGTTGTTAGAACACCTC
GTAGTGTCCGCCGTTTGTCTCGTACCATTAAGGGTGTCGAATATTTTGAAGTT
ATTACTTTTGAGATGGGCCAAAAGAAAGCCCCCGACGGTAGAATTGTTGAGC
TTGTCAAGGAAGTTTCAAAGAGCTACCCAATATCTGAGGGGCCTGAGAGAGC
CAACGAATTAGTAGAATCCTATAGAAAGGCTTCAAATAAAGCTTATTTTGAG
TGGACTATTGAGGCCAGAGATCTTTCTCTGTTGGGTTCCCATGTTCGTAAAGC
TACCTACCAGACTTACGCGCCAATTCTTTATGAGAATGACCACTTTTTCGACT
ACATGCAAAAAGTAAGTTTCATCTCACCATTGAAGGTCCAAAAGTACTTGC
TTATTTACTTGGTTTATGGATTGGTGATGGATTGTCTGACAGGGCAACTTTTTC
GGTTGATTCCAGAGATACTTCTTTGATGGAACGTGTTACTGAATATGCTGAAA

AGTTGAATTTGTGCGCCGAGTATAAGGACAGAAAAGAACCACAAGTTGCCAA
AACTGTTAATTTGTACTCTAAAGTTGTCAGAGGTACCATGGCCGGATCTATGA
GTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGAT
GGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATG
CAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCT
GTTCCATGGCCAACACTTGTCACTACTTTCTCTTATGGTGTTCAATGCTTTTCC
CGTTATCCGGATCATATGAAACGGCATGACTTTTTCAAGAGTGCCATGCCCGA
AGGTTATGTACAGGAACGCACTATATCTTTCAAAGATGACGGGAACTACAAG
ACGCGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTT
AAAAGGTATTGATTTTAAAGAAGATGGAAACATTCTCGGACACAAACTCGAG
TACAACTATAACTCACACAATGTATACATCACGGCAGACAAACAAAAGAATG
GAATCAAAGCTAACTTCAAAATTCGCCACAACATTGAAGATGGATCCGTTCA
ACTAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTT
TACCAGACAACCATTACCTGTCGACACAATCTGCCCTTTCGAAAGATCCCAA
CGAAAAGCGTGACCACATGGTCCTTCTTGAGTTTGTAACTGCTGCTGGGATTA
CACATGGCATGGATGAGCTCTACAAATAATGAATTCTGCTTTTCTTCGCGAAT
TAATTCCGCTTCGCAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACC
GTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGA
GCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACT
ATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTC
CGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTG
GCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCG
CTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCC
TTATCCGGTAACTATCGTCTTGAATCCAACCCGGTAAGACACGACTTATCGCC
ACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGT
GCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAG
TATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGT
AGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTG
CAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATC
TTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTT
GGTCATGAGTTGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAAAAAT
ATATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGG
GTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGATTAAA
TTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCG
GGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGA
GTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGATGAG
ATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGC
ATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGG
AAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTG
TTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATT
GTCCTTTTAACAGCGATCGCGTATTTCGTCTAGCTCAGGCGCAATCACGAATG
AATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGC
CTGTTGAACAAGTCTGGAAAGAAATGCAT

pTTsfGFP

AAGCTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTT
GATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACG
AGTAGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTC
GGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGA
TAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCT
AATCAGAATTGGTTAATTGGTTGTAACACTGGCAGAGCATTACGCTGACTTG
ACGGGACGGCGGCTTTGTTGAATAAATCGAACTTTTGCTGAGTTGAAGGATC
TCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGAC
TGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATT
AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATT
GTGAGCGGATAACAATTTCAAGGAGATATACATATGAACAATAACGATCTCT
TTCAGGCATCACGTCGGCGTTTTCTGGCACAACTCGGCGGCTTAACCGTCGCC
GGGATGCTGGGGCCGTCATTGTTAACGCCGCGACGTGCGACTGCGCATAATG
GGTGCTTTGCCAAGGGTACCAATGTTTTAATGGCGGATGGGTCTATTGAATGT
ATTGAAAACATTGAGGTTGGTAATAAGGTCATGGGTAAAGATGGCAGACCTC
GTGAGGTAATTAAATTGCCCAGAGGAAGAGAAACTATGTACAGCGTCGTTCA
GAAAAGTCAGCACAGAGCCCACAAAAGTGATTCAAGTCGTGAAGTGCCAGA
ATTACTCAAGTTTACGTGTAATGCGACCCATGAGTTGGTTGTTAGAACACCTC
GTAGTGTCCGCCGTTTGTCTCGTACCATTAAGGGTGTCGAATATTTTGAAGTT
ATTACTTTTGAGATGGGCCAAAAGAAAGCCCCCGACGGTAGAATTGTTGAGC
TTGTCAAGGAAGTTTCAAAGAGCTACCCAATATCTGAGGGGCCTGAGAGAGC
CAACGAATTAGTAGAATCCTATAGAAAGGCTTCAAATAAAGCTTATTTTGAG
TGGACTATTGAGGCCAGAGATCTTTCTCTGTTGGGTTCCCATGTTCGTAAAGC
TACCTACCAGACTTACGCGCCAATTCTTTATGAGAATGACCACTTTTTCGACT
ACATGCAAAAAGTAAGTTTCATCTCACCATTGAAGGTCCAAAAGTACTTGC
TTATTTACTTGGTTTATGGATTGGTGATGGATTGTCTGACAGGGCAACTTTTTC
GGTTGATTCCAGAGATACTTCTTTGATGGAACGTGTTACTGAATATGCTGAAA

AGTTGAATTTGTGCGCCGAGTATAAGGACAGAAAAGAACCACAAGTTGCCAA
AACTGTTAATTTGTACTCTAAAGTTGTCAGAGGTACCATGGCCGGATCTGCTA
GCATGAGCAAAGGCGAAGAACTGTTTACCGGCGTGGTTCCGATTCTGGTGGA
ACTGGATGGCGATGTAAATGGCCACAAGTTTAGCGTGCGTGGGGAAGGGGA
GGGCGATGCGACCAATGGCAAACTGACCCTGAAGTTTATTTGCACGACCGGG
AAACTGCCGGTTCCTTGGCCCACCCTTGTCACCACCCTGACGTATGGCGTGCA
ATGCTTTAGCCGTTACCCGGACCACATGAAGCGGCATGACTTCTTCAAAAGC
GCCATGCCTGAAGGCTATGTTCAGGAACGGACGATCTCGTTTAAGGATGACG
GCACCTATAAGACCCGTGCGGAGGTCAAATTCGAAGGCGATACCCTGGTGAA
CCGCATTGAGCTGAAGGGCATCGACTTCAAAGAGGATGGCAACATACTGGGG
CACAAGCTGGAGTACAACTTCAACAGCCACAACGTCTACATCACCGCCGACA
AGCAGAAGAACGGCATTAAGGCCAACTTCAAGATTCGGCACAATGTGGAGG
ACGGAAGCGTTCAGCTGGCGGATCATTATCAACAGAATACCCCCATTGGCGA
CGGTCCCGTGCTTCTGCCGGATAATCATTACTTGAGCACCCAGAGCGTGCTGA
GCAAGGACCCGAATGAGAAACGGGATCACATGGTGCTGCTGGAATTTGTGAC
CGCTGCGGGCATTACACATGGCATGGATGAACTGTATAAATAAGAATTCTGC
TTTTCTTCGCGAATTAATTCCGCTTCGCAACATGTGAGCAAAAGGCCAGCAAA
AGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCG
CCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAAC
CCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGC
GCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTT
CGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTG
TAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCG
ACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAATCCAACCCGGTAAGACAC
GACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGT
ATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACAC
TAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAA
AAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGG
TTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAA
GATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACG
TTAAGGGATTTTGGTCATGAGTTGTGTCTCAAAATCTCTGATGTTACATTGCA
CAAGATAAAAATATATCATCATGAACAATAAAACTGTCTGCTTACATAAACA
GTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAG
GCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTC
GCGATAATGTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCC
CGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGAT
GTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCC
GACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTG
CGATCCCCGGGAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGG
TGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTC
CTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTAGCTCAGGCG
CAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGC
GTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCAT

pBK-BocLys

*Nde*I - 74 - CA'TA_TG

PyIRS (Y384F)

KanR

pBKBocLys
3386 bp

*Xba*I - 1442 - T'CTAG_A
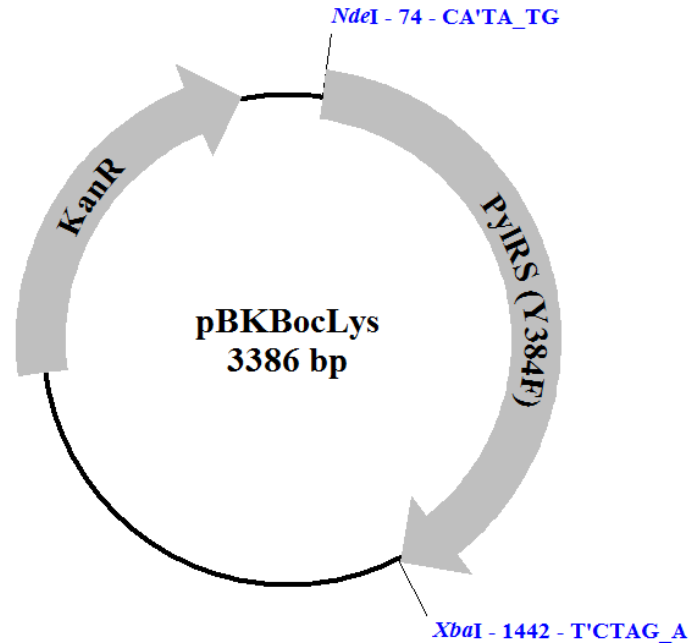
The Y384F mutation is highlighted in red

CTCGGGAGTTGTCAGCCTGTCCCGCTTATAAGATCATACGCCGTTATACGTTG
TTTACGCTTTGAGGAATCCCATATGGATAAAAAACCACTAAACACTCTGATAT
CTGCAACCGGGCTCTGGATGTCCAGGACCGGAACAATTCATAAAATAAAACA
CCACGAAGTCTCTCGAAGCAAAATCTATATTGAAATGGCATGCGGAGACCAC
CTTGTTGTAAACAACTCCAGGAGCAGCAGGACTGCAAGAGCGCTCAGGCACC
ACAAATACAGGAAGACCTGCAAACGCTGCAGGGTTTCGGATGAGGATCTCAA
TAAGTTCCTCACAAAGGCAAACGAAGACCAGACAAGCGTAAAAGTCAAGGT
CGTTTCTGCCCCTACCAGAACGAAAAAGGCAATGCCAAAATCCGTTGCGAGA
GCCCCGAAACCTCTTGAGAATACAGAAGCGGCACAGGCTCAACCTTCTGGAT
CTAAATTTTCACCTGCGATACCGGTTTCCACCCAAGAGTCAGTTTCTGTCCCG
GCATCTGTTTCAACATCAATATCAAGCATTTCTACAGGAGCAACTGCATCCGC
ACTGGTAAAAGGGAATACGAACCCCATTACATCCATGTCTGCCCCTGTTCAG
GCAAGTGCCCCCGCACTTACGAAGAGCCAGACTGACAGGCTTGAAGTCCTGT
TAAACCCAAAAGATGAGATTTCCCTGAATTCCGGCAAGCCTTTCAGGGAGCT
TGAGTCCGAATTGCTCTCTCGCAGAAAAAAAGACCTGCAGCAGATCTACGCG
GAAGAAAGGGAGAATTATCTGGGGAAACTCGAGCGTGAAATTACCAGGTTCT
TTGTGGACAGGGGTTTTCTGGAAATAAAATCCCCGATCCTGATCCCTCTTGAG
TATATCGAAAGGATGGGCATTGATAATGATACCGAACTTTCAAAACAGATCT
TCAGGGTTGACAAGAACTTCTGCCTGAGACCCATGCTTGCTCCAAACCTTTAC
AACTACCTGCGCAAGCTTGACAGGGCCCTGCCTGATCCAATAAAAATTTTTG
AAATAGGCCCATGCTACAGAAAAGAGTCCGACGGCAAAGAACACCTCGAAG
AGTTTACCATGCTGAACTTCTGCCAGATGGGATCGGGATGCACACGGGAAAA
TCTTGAAAGCATAATTACGGACTTCCTGAACCACCTGGGAATTGATTTCAAGA

192

TCGTAGGCGATTCCTGCATGGTCTTTGGGGATACCCTTGATGTAATGCACGGA
GACCTGGAACTTTCCTCTGCAGTAGTCGGACCCATACCGCTTGACCGGGAAT
GGGGTATTGATAAACCCTGGATAGGGGCAGGTTTCGGGCTCGAACGCCTTCT
AAAGGTTAAACACGACTTTAAAAATATCAAGAGAGCTGCAAGGTCCGAGTCT
TACTATAACGGGATTTCTACCAACCTGTAAATCTAGATTTCAAACGCTAAATT
GCCTGATGCGCTACGCTTATCAGGCCTACATGATCTCTGCAACACACCGAGC
CCGCGTGCTTTTGCAGGCCGGATAAGGCGTTCGCGCCGCATCCGGCAAGAAA
CAGCAAACAATCCAAAACGCCGCGTTCGGCGGCGTTTTTCTGCTTTTCTTCG
CGAATTAATTCCGCTTCGCAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGG
AACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGA
CGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGG
ACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTG
TTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGC
GTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGT
TCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGC
GCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATC
GCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGC
GGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGA
CAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTT
GGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGT
TTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTG
ATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGA
TTTTGGTCATGAGTTGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAA
AAATATATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACA
AGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGAT
TAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAAT
GTCGGGCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGC
CAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCAATGATGTTACAGA
TGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCA
AGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCC
GGGAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATA
TTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTA
ATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTCGCTCAGGCGCAATCACGA
ATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCT
GGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAGCTTTTGCCATTCTCACC
GGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACG
AGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCG
ATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCAT
TACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGATATGAATAA
ATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTG
GTTGTAACACTGGCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTT
GAATAAATCGAACTTTTGCTGAGTTGAAGGATC

pYC-PylT-Myo4tag
5862 bp

*Not*I - 4014 - GC'GGCC_GC

*Pst*I - 3778 - C_TGCA'G

PylT

TetR

Myo4tag
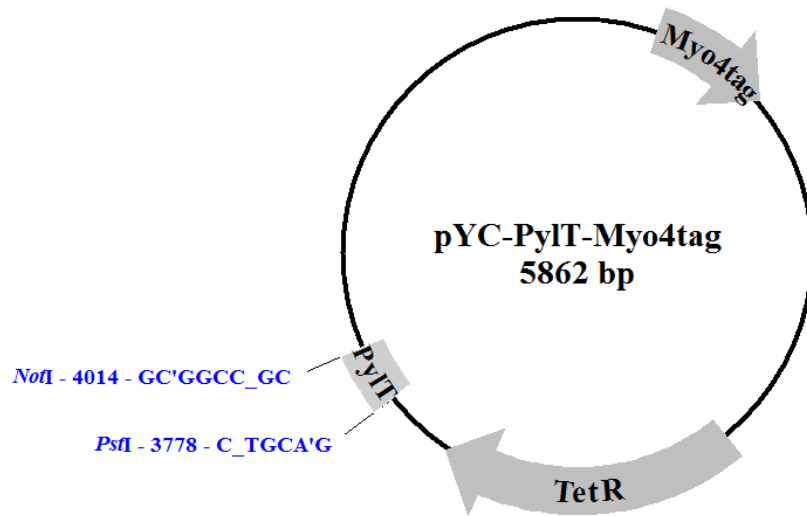
AAGAAACCAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTAC
TGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAGCATTCTGTA
ACAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAA
TCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATG
CCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTTTTATCG
CAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTAACAGGAGGAATTA
ACCATGGTTCTGTCTGAAGGTGAATGGCAGCTGGTTCTGCATGTTTGGGCTAA
AGTTGAAGCTGACGTCGCTGGTCATGGTCAGGACATCTTGATTCGACTGTTCA
AATCTCATCCGGAAACTCTGGAAAAATTCGATCGTTTCAAACATCTGAAAAC
TGAAGCTGAAATGAAAGCTTCTGAAGATCTGAAAAAACATGGTGTTACCGTG
TTAACTGCCCTAGGTGCTATCCTTAAGAAAAAGGGCATCATGAAGCTGAGC
TCAAACCGCTTGCACAATCGCATGCTACTAAACATAAGATCCCGATCAAATA
CCTGGAATTCATCTCTGAAGCGATCATCCATGTTCTGCATTCTAGACATCCAG
GTGACTTCGGTGCTGACGCTCAGGGTGCTATGAACAAAGCTCTCGAGCTGTTC
CGTAAAGATATCGCTGCTAAGTACAAAGAACTGGGTTACCAGGGTGGCTCGG
GACATCATCACCATCACCATTGAGTTTAACGGTACCATATGGGAATTCGAAG
CTTGGGCCCGAACAAAAACTCATCTCAGAAGAGGATCTGAATAGCGCCGTCG
ACCATCATCATCATCATTGAGTTTAAACGGTCTCCAGCTTGGCTGTTTTG
GCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAG
CGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCT
GACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGG
GGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAG
GCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGC
TCTCCTGAGTAGGACAAATCCGCCGGGAGCTGTCCCTCCTGTTCAGCTACTGA
CGGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACATCAGCGCTAGCGGAG
TGTATACTGGCTTACTATGTTGGCACTGATGAGGGTGTCAGTGAAGTGCTTCA
TGTGGCAGGAGAAAAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATAC
AGGATATATTCCGCTTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACT
GCGGCGAGCGGAAATGGCTTACGAACGGGGCGGAGATTTCCTGGAAGATGC

CAGGAAGATACTTAACAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCC
ATAGGCTCCGCCCCCCTGACAAGCATCACGAAATCTGACGCTCAAATCAGTG
GTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCTGGCGGCT
CCCTCGTGCGCTCTCCTGTTCCTGCCTTTCGGTTTACCGGTGTCATTCCGCTGT
TATGGCCGCGTTTGTCTCATTCCACGCCTGACACTCAGTTCCGGGTAGGCAGT
TCGCTCCAAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTGC
GCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGAAAGACATGCAAAAG
CACCACTGGCAGCAGCCACTGGTAATTGATTTAGAGGAGTTAGTCTTGAAGT
CATGCGCCGGTTAAGGCTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCC
TCCAAGCCAGTTACCTCGGTTCAAAGAGTTGGTAGCTCAGAGAACCTTCGAA
AAACCGCCCTGCAAGGCGGTTTTTTCGTTTTCAGAGCAAGAGATTACGCGCA
GACCAAAACGATCTCAAGAAGATCATCTTATTAATCAGATAAAATATTTCTA
GATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAAGTCAGCCCCATAC
GATATAAGTTGTAATTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATG
CGGTAGTTTATCACAGTTAAATTGCTAACGCAGTCAGGCACCGTGTATGAAA
TCTAACAATGCGCTCATCGTCATCCTCGGCACCGTCACCCTGGATGCTGTAGG
CATAGGCTTGGTTATGCCGGTACTGCCGGGCCTCTTGCGGGATATCGTCCATT
CCGACAGCATCGCCAGTCACTATGGCGTGCTGCTAGCGCTATATGCGTTGATG
CAATTTCTATGCGCACCCGTTCTCGGAGCACTGTCCGACCGCTTTGGCCGCCG
CCCAGTCCTGCTCGCTTCGCTACTTGGAGCCACTATCGACTACGCGATCATGG
CGACCACACCCGTCCTGTGGATCCTCTACGCCGGACGCATCGTGGCCGGCAT
CACCGGCGCCACAGGTGCGGTTGCTGGCGCCTATATCGCCGACATCACCGAT
GGGGAAGATCGGGCTCGCCACTTCGGGCTCATGAGCGCTTGTTTCGGCGTGG
GTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTGCA
TGCACCATTCCTTGCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCT
GCTTCCTAATGCAGGAGTCGCATAAGGGAGAGCGTCGACCGATGCCCTTGAG
AGCCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCGGGGCATGACTATCGTC
GCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACAGGTGCCGGC
AGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGATG
ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTT
CGTCACTGGTCCCGCCACCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCC
GGCATGGCGGCCGACGCGCTGGGCTACGTCTTGCTGGCGTTCGCGACGCGAG
GCTGGATGGCCTTCCCCATTATGATTCTTCTCGCTTCCGGCGGCATCGGGATG
CCCGCGTTGCAGGCCATGCTGTCCAGGCAGGTAGATGACGACCATCAGGGAC
AGCTTCAAGGATCGCTCGCGGCTCTTACCAGCCTAACTTCGATCACTGGACCG
CTGATCGTCACGGCGATTTATGCCGCCTCGGCGAGCACATGGAACGGGTTGG
CATGGATTGTAGGCGCCGCCCTATACCTTGTCTGCCTCCCCGCGTTGCGTCGC
GGTGCATGGAGCCGGGCCACCTCGACCTGAATGGAAGCCGGCGGCACCTCGC
TAACGGATTCACCACTCCAAGAATTGGAGCCAATCAATTCTTGCGGAGAACT
GTGAATGCGCAAACCAACCCTTGGCAGAACATATCCATCGCGTCCGCCATCT
CCAGCAGCCGCACGCGGCGCATCTCGGGCTCCTTGCATGCACCATTCCTTGCG
GCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATGCAGG
AGTCGCATAAGGGAGAGCGTAAGCTTAAAAAAAATCCTTAGCTTTCGCTAAG
GATCTGCAGTTTGTCGACCAAAAAGCCTGCTCGTTGAGCAGGCTTTTCGAAT
TTGGCGGAAACCCCGGGAATCTAACCCGGCTGAACGGATTTAGAGTCCATTC
GATCTACATGATCAGGTTCCCAATGCGGGGCGCATCTTACTGCGCAGATACG

```
CCCTCGTCAATCCCTTAATAGCAAAATGCCTTTTGATCGGCGAGAAAGTCAGC
AGGCCGCTTAGTTAGCCGCTGCCACTAGTTTTGCGGCCGCACTTATTTTTGAT
CGTTCGCTCAAAGAAGCGGCGCAACGCCATGAGCGGCCTCATTTCTTATTCTG
AGTTACAACAGTCCGCACCGCTGTCCGGTAGCTCCTTCCGGTGGGCGCGGGG
CATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAG
GACAGGTGCCGGCAGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCAT
ACCCACGCCGAAACAAGCGCCCTGCACCATTATGTTCCGGATCTGCATCGCA
GGATGCTGCTGGCTACCCTGTGGAACACCTACATCTGTATTAACGAAGCGCT
AACCGTTTTTATCAGGCTCTGGGAGGCAGAATAAATGATCATATCGTCAATTA
TTACCTCCACGGGGAGAGCCTGAGCAAACTGGCCTCAGGCATTTGAGAAGCA
CACGGTCACACTGCTTCCGGTAGTCAATAAACCGGTAAACCAGCAATAGACA
TAAGCGGCTATTTAACGACCCTGCCCTGAACCGACGACCGGGTCGAATTTGC
TTTCGAATTTCTGCCATTCATCCGCTTATTATCACTTATTCAGGCGTAGCACCA
GGCGTTTAAGGGCACCAATAACTGCCTTAAAAAAATTACGCCCCGCCCTGCC
ACTCATCGCAGTGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACC
CGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAA
GCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACC
GAAACGCGCGAGGCAGCAGATCAATTCGCGCGCGAAGGCGAAGCGGCATGC
ATAATGTGCCTGTCAAATGGACGAAGCAGGGATTCTGCAAACCCTATGCTAC
TCCGTCAAGCCGTCAATTGTCTGATTCGTTACCAATTATGACAACTTGACGGC
TACATCATTCACTTTTTCTTCACAACCGGCACGGAACTCGCTCGGGCTGGCCC
CGGTGCATTTTTTAAATACCCGCGAGAAATAGAGTTGATCGTCAAAACCAAC
ATTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCTCAAAAGCAGCTTC
GCCTGGCTGATACGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTG
CTGGCGGAAAAGATGTGACAGACGCGACGGCGACAAGCAAACATGCTGTGC
GACGCTGGCGATATCAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGA
CAAGCCTCGCGTACCCGATTATCCATCGGTGGATGGAGCGACTCGTTAATCG
CTTCCATGCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCC
GAATAGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCGCT
GAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGCAAAT
ATTGACGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGACGAAAGTAAA
CCCACTGGTGATACCATTCGCGAGCCTCCGGATGACGACCGTAGTGATGAAT
CTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCGGCAAACAAATTCTCGT
CCCTGATTTTTCACCACCCCCTGACCGCGAATGGTGAGATTGAGAATATAACC
TTTCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGATAACCGTTGGCCTCA
ATCGGCGTTAAACCCGCCACCAGATGGGCATTAAACGAGTATCCCGGCAGCA
GGGGATCATTTTGCGCTTCAGCCATACTTTTCATACTCCCGCCATTCAGAG
```

pSup-BocLys
4150 bp
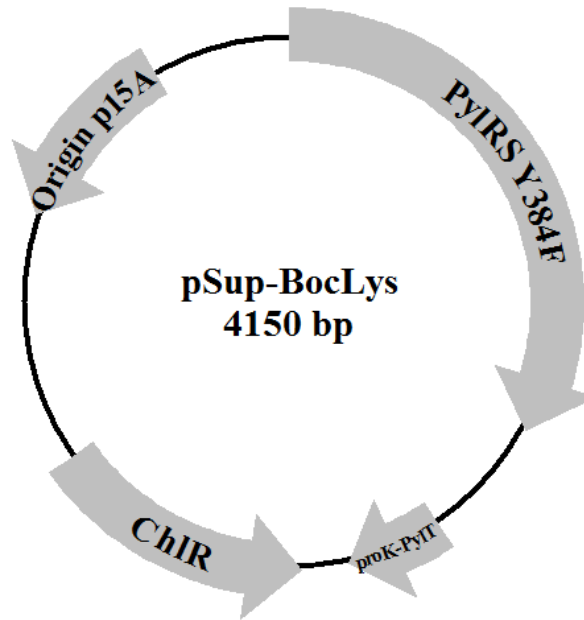
*Mm*PylRS Y384F highlighted in red (Y384F underlined)
*proK-Mb*PylT sequence highlighted in blue (core MbPylT sequence shaded)

ATGGATAAAAAACCACTAAACACTCTGATATCTGCAACCGGGCTCTGGATGT
CCAGGACCGGAACAATTCATAAAATAAAACACCACGAAGTCTCTCGAAGCA
AAATCTATATTGAAATGGCATGCGGAGACCACCTTGTTGTAAACAACTCCAG
GAGCAGCAGGACTGCAAGAGCGCTCAGGCACCACAAATACAGGAAGACCTG
CAAACGCTGCAGGGTTTCGGATGAGGATCTCAATAAGTTCCTCACAAAGGCA
AACGAAGACCAGACAAGCGTAAAAGTCAAGGTCGTTTCTGCCCCTACCAGAA
CGAAAAAGGCAATGCCAAAATCCGTTGCGAGAGCCCCGAAACCTCTTGAGA
ATACAGAAGCGGCACAGGCTCAACCTTCTGGATCTAAATTTTCACCTGCGAT
ACCGGTTTCCACCCAAGAGTCAGTTTCTGTCCCGGCATCTGTTTCAACATCAA
TATCAAGCATTTCTACAGGAGCAACTGCATCCGCACTGGTAAAAGGGAATAC
GAACCCCATTACATCCATGTCTGCCCCTGTTCAGGCAAGTGCCCCCGCACTTA
CGAAGAGCCAGACTGACAGGCTTGAAGTCCTGTTAAACCCAAAAGATGAGAT
TTCCCTGAATTCCGGCAAGCCTTTCAGGGAGCTTGAGTCCGAATTGCTCTCTC
GCAGAAAAAAAGACCTGCAGCAGATCTACGCGGAAGAAAGGGAGAATTATC
TGGGGAAACTCGAGCGTGAAATTACCAGGTTCTTTGTGGACAGGGGTTTTCT
GGAAATAAAATCCCCGATCCTGATCCCTCTTGAGTATATCGAAAGGATGGGC
ATTGATAATGATACCGAACTTTCAAAACAGATCTTCAGGGTTGACAAGAACT
TCTGCCTGAGACCCATGCTTGCTCCAAACCTTTACAACTACCTGCGCAAGCTT
GACAGGGCCCTGCCTGATCCAATAAAAATTTTTGAAATAGGCCCATGCTACA
GAAAAGAGTCCGACGGCAAAGAACACCTCGAAGAGTTTACCATGCTGAACTT
CTGCCAGATGGGATCGGGATGCACACGGGAAAATCTTGAAAGCATAATTACG
GACTTCCTGAACCACCTGGGAATTGATTTCAAGATCGTAGGCGATTCCTGCAT

GGTCTTTGGGGATACCCTTGATGTAATGCACGGAGACCTGGAACTTTCCTCTG
CAGTAGTCGGACCCATACCGCTTGACCGGGAATGGGGTATTGATAAACCCTG
GATAGGGGCAGGTTTCGGGCTCGAACGCCTTCTAAAGGTTAAACACGACTTT
AAAAATATCAAGAGAGCTGCAAGGTCCGAGTCTTACTATAACGGGATTTCTA
CCAACCTGTAAATGCAGTTTCAAACGCTAAATTGCCTGATGCGCTACGCTTAT
CAGGCCTACATGATCTCTGCAATATATTGAGTTTGCGTGCTTTTGTAGGCCGG
ATAAGGCGTTCACGCCGCATCCGGCAAGAAACAGCAAACAATCCAAAACGC
CGCGTTCAGCGGCGTTTTTTCTGCTTTTCTTCGCGAATTAATTCCGCTTCGCAA
CATGTGAGCACCGGTTTATTGACTACCGGAAGCAGTGTGACCGTGTGCTTCTC
AAATGCCTGAGGCCAGTTTGCTCAGGCTCTCCCCGTGGAGGTAATAATTGAC
GATATGATCAGTCAAAAGCGGCCGCAAAACTAGTGGCAGCGGCTAACTAAGC
GGCCTGCTGACTTTCTCGCCGATCAAAAGGCATTTTGCTATTAAGGGATTGAC
GAGGGCGTATCTGCGCAGTAAGATGCGCCCCGCATTGGGAACCTGATCATGT
AGATCGAATGGACTCTAAATCCGTTCAGCCGGGTTAGATTCCCGGGGTTTCCG
CCAAATTCGAAAAGCCTGCTCAACGAGCAGGCTTTTTGGTCGACAAACTCG
AGCAGCTCAGGGTCGAATTTGCTTTCGAATTTCTGCCATTCATCCGCTTATTA
TCACTTATTCAGGCGTAGCAACCAGGCGTTTAAGGGCACCAATAACTGCCTT
AAAAAAATTACGCCCCGCCCTGCCACTCATCGCAGTACTGTTGTAATTCATTA
AGCATTCTGCCGACATGGAAGCCATCACAAACGGCATGATGAACCTGAATCG
CCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAA
ACGGGGGCGAAGAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGA
AACTCACCCAGGGATTGGCTGAGACGAAAAACATATTCTCAATAAACCCTTT
AGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATG
TGTAGAAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACG
TTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGTGAACACTATCCCATATC
ACCAGCTCACCGTCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATCAG
GCGGGCAAGAATGTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTA
CGGTCTTTAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACA
TTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTTACGATGCCATTGGGATA
TATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGCTTCCTTAGCTC
CTGAAAATCTCGATAACTCAAAAAATACGCCCGGTAGTGATCTTATTTCATTA
TGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAAA
GTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTG
CGAAGTGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGAT
GCTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGG
CTTCTGTTTCTATCAGCTGTCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGT
AACGGCAAAAGCACCGCCGGACATCAGCGCTAGCGGAGTGTATACTGGCTTA
CTATGTTGGCACTGATGAGGGTGTCAGTGAAGTGCTTCATGTGGCAGGAGAA
AAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATACAGGATATATTCCGC
TTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAGCGGAAA
TGGCTTACGAACGGGGCGGAGATTTCCTGGAAGATGCCAGGAAGATACTTAA
CAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCCATAGGCTCCGCCCCC
CTGACAAGCATCACGAAATCTGACGCTCAAATCAGTGGTGGCGAAACCCGAC
AGGACTATAAAGATACCAGGCGTTTCCCCCTGGCGGCTCCCTCGTGCGCTCTC
CTGTTCCTGCCTTTCGGTTTACCGGTGTCATTCCGCTGTTATGGCCGCGTTTGT
CTCATTCCACGCCTGACACTCAGTTCCGGGTAGGCAGTTCGCTCCAAGCTGGA

CTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTGCGCCTTATCCGGTAACT
ATCGTCTTGAGTCCAACCCGGAAAGACATGCAAAAGCACCACTGGCAGCAGC
CACTGGTAATTGATTTAGAGGAGTTAGTCTTGAAGTCATGCGCCGGTTAAGG
CTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCCTCCAAGCCAGTTACCTC
GGTTCAAAGAGTTGGTAGCTCAGAGAACCTTCGAAAAACCGCCCTGCAAGGC
GGTTTTTTCGTTTTCAGAGCAAGAGATTACGCGCAGACCAAAACGATCTCAA
GAAGATCATCTTATTAATCAGATAAAATATTTCTAGATTTCAGTGCAATTTAT
CTCTTCAAATGTAGCACCTGAAGTCAGCCCCATACGATATAAGTTGTAATTCT
CATGTTTGACAGCTTATCATCGATAAGCTTGGTACCGAGCTCCCGGTCATCAA
TCATCCCCATAATCCTTGTTAGATTATCAATTTTAAAAAACTAACAGTTGTCA
GCCTGTCCCGCTTTAATATCATACGCCGTTATACGTTGTTTACGCTTTGAGGA
ATCCCAT

pTrcHisA-sfGFP (V150tag)

pTrcHisA-sfGFP150tag
5030 bp
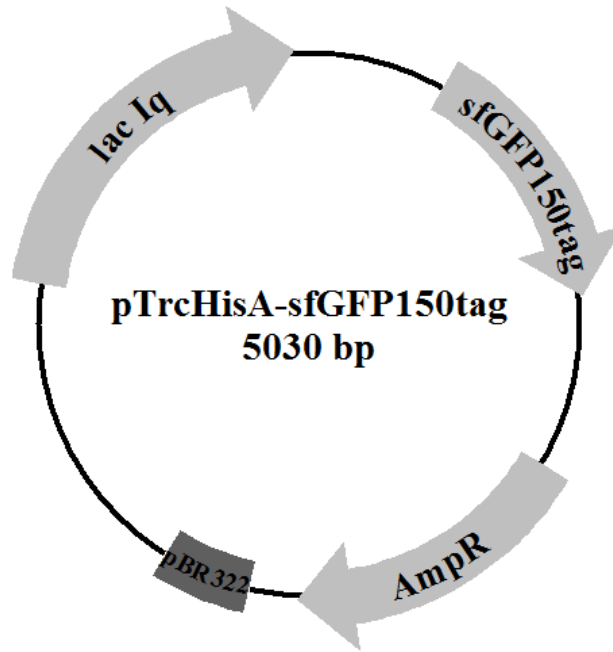
Highlighted in red is the ORF of sfGFP with an upstream fusion peptide. The coding region of sfGFP starts from the shaded "ATG". Highlighted in blue is the V150tag mutation.

GTTTGACAGCTTATCATCGACTGCACGGTGCACCAATGCTTCTGGCGTCAGGC
AGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATT
CGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATC
ATAACGGTTCTGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCCGG
CTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAG
CGCCGCTGAGAAAAAGCGAAGCGGCACTGCTCTTTAACAATTTATCAGACAA
TCTGTGTGGGCACTCGACCGGAATTATCGATTAACTTTATTATTAAAAATTAA
AGAGGTATATATTAATGTATCGATTAAATAAGGAGGAATAAACCATGGGGGG
TTCTCATCATCATCATCATGGTATGGCTAGCATG<mark>ATG</mark>AGCAAAGGCGAA
GAACTGTTTACCGGCGTGGTTCCGATTCTGGTGGAACTGGATGGCGATGTAA
ATGGCCACAAGTTTAGCGTGCGTGGGGAAGGGGAGGGCGATGCGACCAATG
GCAAACTGACCCTGAAGTTTATTTGCACGACCGGGAAACTGCCGGTTCCTTG
GCCCACCCTTGTCACCACCCTGACGTATGGCGTGCAATGCTTTAGCCGTTACC
CGGACCACATGAAGCGGCATGACTTCTTCAAAAGCGCCATGCCTGAAGGCTA
TGTTCAGGAACGGACGATCTCGTTTAAGGATGACGGCACCTATAAGACCCGT
GCGGAGGTCAAATTCGAAGGCGATACCCTGGTGAACCGCATTGAGCTGAAGG
GCATCGACTTCAAAGAGGATGGCAACATACTGGGGCACAAGCTGGAGTACA
ACTTCAACAGCCACAACtagTACATCACCGCCGACAAGCAGAAGAACGGCATT
AAGGCCAACTTCAAGATTCGGCACAATGTGGAGGACGGAAGCGTTCAGCTGG
CGGATCATTATCAACAGAATACCCCCATTGGCGACGGTCCCGTGCTTCTGCCG
GATAATCATTACTTGAGCACCCAGAGCGTGCTGAGCAAGGACCCGAATGAGA

AACGGGATCACATGGTGCTGCTGGAATTTGTGACCGCTGCGGGCATTACACA
TGGCATGGATGAACTGTATAAATAAAAGCTTGGCTGTTTTGGCGGATGAGAG
AAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAA
AACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCC
GAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCAT
GCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAA
AGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTA
GGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAG
GGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGA
AGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTTGTTTATT
TTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAA
TGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTC
GCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAA
ACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTT
ACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGA
AGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTAT
TATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTC
TCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGAT
GGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGC
TTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGG
AGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGC
AATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTT
CCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACT
TCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCG
GTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCC
CTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAA
CGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAC
TGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTT
AATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATC
CCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCA
AAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACA
AAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAA
CTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTC
CTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCC
TACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATA
AGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCA
GCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAAC
GACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACG
CTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGA
ACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATA
GTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGT
CAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTT
CCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGA
TTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCA
GCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCC

TGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGG
TGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACT
CCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACAC
CCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACA
AGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCAC
CGAAACGCGCGAGGCAGCAGATCAATTCGCGCGCGAAGGCGAAGCGGCATG
CATTTACGTTGACACCATCGAATGGTGCAAAACCTTTCGCGGTATGGCATGAT
AGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATGTGAAACCAGTAACGTT
ATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGG
TGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGG
CGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGG
CAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGC
CGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAG
CGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCG
GTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCT
GGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCG
TTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCAT
GAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGC
AAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTG
GCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAAC
GGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCT
GAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCG
CTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATA
TCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCC
GTCAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGAC
CGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGC
CCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGC
CTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCC
GACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCGCGAA
TTGATCTG

pET3a-SynUb
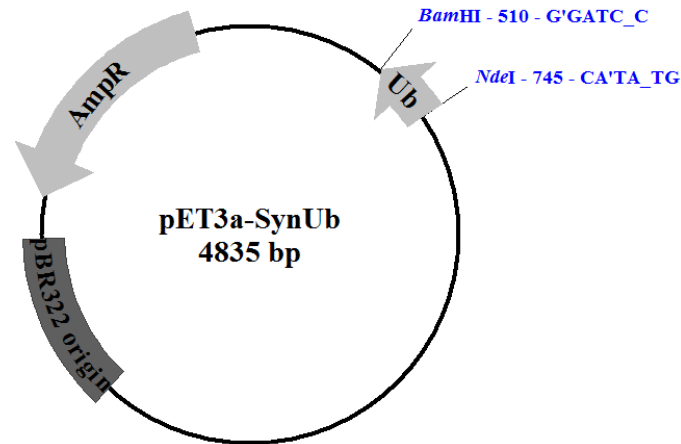
TTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATGCGGTAGTTTATCAC
AGTTAAATTGCTAACGCAGTCAGGCACCGTGTATGAAATCTAACAATGCGCT
CATCGTCATCCTCGGCACCGTCACCCTGGATGCTGTAGGCATAGGCTTGGTTA
TGCCGGTACTGCCGGGCCTCTTGCGGGATATCGTCCATTCCGACAGCATCGCC
AGTCACTATGGCGTGCTGCTAGCGCTATATGCGTTGATGCAATTTCTATGCGC
ACCCGTTCTCGGAGCACTGTCCGACCGCTTTGGCCGCCGCCCAGTCCTGCTCG
CTTCGCTACTTGGAGCCACTATCGACTACGCGATCATGGCGACCACACCCGTC
CTGTGGATATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTCAAGACCC
GTTTAGAGGCCCCAAGGGGTTATGCTAGTTATTGCTCAGCGGTGGCAGCAGC
CAACTCAGCTTCCTTTCGGGCTTTGTTAGCAGCCGGATCCTTAGCCGCCACGC
AGACGCAGCACCAGATGCAGGGTGCTTTCTTTCTGAATATTGTAATCGCTCAG
GGTACGACCATCTTCCAGCTGTTTGCCCGCAAAAATCAGACGCTGCTGATCC
GGCGGAATGCCTTCTTTATCCTGAATTTTCGCTTTCACATTTTCAATGGTATCG
CTCGGTTCCACTTCCAGGGTAATGGTTTTGCCGGTCAGGGTTTTCACAAAAAT
CTGCATATGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGGAA
ACCGTTGTGGTCTCCCTATAGTGAGTCGTATTAATTTCGCGGGATCGAGATCT
CGATCCTCTACGCCGGACGCATCGTGGCCGGCATCACCGGCGCCACAGGTGC
GGTTGCTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGC
CACTTCGGGCTCATGAGCGCTTGTTTCGGCGTGGGTATGGTGGCAGGCCCCGT
GGCCGGGGGACTGTTGGGCGCCATCTCCTTGCATGCACCATTCCTTGCGGCG
GCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATGCAGGAGT
CGCATAAGGGAGAGCGTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAG
CTCCTTCCGGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCT
TCTTTATCATGCAACTCGTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTC
GGCGAGGACCGCTTTCGCTGGAGCGCGACGATGATCGGCCTGTCGCTTGCGG
TATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACTGGTCCCGCCACC
AAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATGGCGGCCGACGCGC
TGGGCTACGTCTTGCTGGCGTTCGCGACGCGAGGCTGGATGGCCTTCCCCATT
ATGATTCTTCTCGCTTCCGGCGGCATCGGGATGCCCGCGTTGCAGGCCATGCT
GTCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAAGGATCGCTCGCG
GCTCTTACCAGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTA
TGCCGCCTCGGCGAGCACATGGAACGGGTTGGCATGGATTGTAGGCGCCGCC

```
CTATACCTTGTCTGCCTCCCCGCGTTGCGTCGCGGTGCATGGAGCCGGGCCAC
CTCGACCTGAATGGAAGCCGGCGGCACCTCGCTAACGGATTCACCACTCCAA
GAATTGGAGCCAATCAATTCTTGCGGAGAACTGTGAATGCGCAAACCAACCC
TTGGCAGAACATATCCATCGCGTCCGCCATCTCCAGCAGCCGCACGCGGCGC
ATCTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCGTGCTCCTGTC
GTTGAGGACCCGGCTAGGCTGGCGGGGTTGCCTTACTGGTTAGCAGAATGAA
TCACCGATACGCGAGCGAACGTGAAGCGACTGCTGCTGCAAAACGTCTGCGA
CCTGAGCAACAACATGAATGGTCTTCGGTTTCCGTGTTTCGTAAAGTCTGGAA
ACGCGGAAGTCAGCGCCCTGCACCATTATGTTCCGGATCTGCATCGCAGGAT
GCTGCTGGCTACCCTGTGGAACACCTACATCTGTATTAACGAAGCGCTGGCAT
TGACCCTGAGTGATTTTTCTCTGGTCCCGCCGCATCCATACCGCCAGTTGTTT
ACCCTCACAACGTTCCAGTAACCGGGCATGTTCATCATCAGTAACCCGTATCG
TGAGCATCCTCTCTCGTTTCATCGGTATCATTACCCCCATGAACAGAAATCCC
CCTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGG
CCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCT
GGACGCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGAT
GAGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTG
ACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGG
AGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGC
GCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTA
TGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATATGCGGTGTGAAA
TACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTC
CTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAG
CTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGG
AAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGC
CGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAA
ATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACC
AGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCG
CTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCA
TAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGG
GCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAAC
TATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAG
CCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTT
CTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATC
TGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATC
CGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTGTTTGCAAGCAGCAG
ATTACGCGCAGAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGG
GGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAG
ATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTA
AATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTA
ATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGC
CTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGC
CCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTAT
CAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAA
CTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGT
AGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTGCAGGCATCGT
```

GGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGAT
CAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTT
CGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGG
TTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTT
TCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGC
GACCGAGTTGCTCTTGCCCGGCGTCAACACGGGATAATACCGCGCCACATAG
CAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTC
TCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACC
CAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAA
CAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTT
GAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATT
GTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGG
GGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATT
ATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCT
TCAAGAA

pTXB1-SynUb

NdeI - 5723 - CA'TA_TG

pTXB1-SynUb
6883 bp
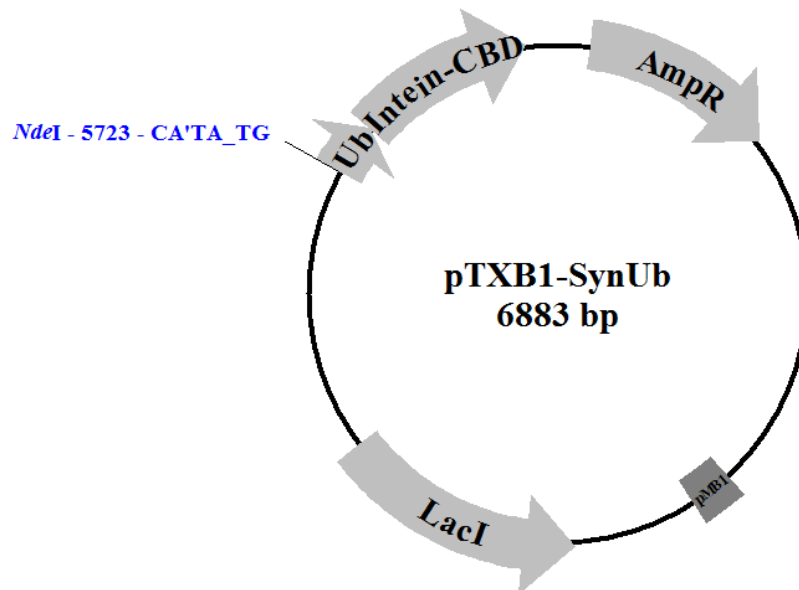
AACTACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTT
TATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGA
TAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCC
GTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCC
AGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGT
GGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
CCCGAAGAACGTTCTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACAC
TATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTAC

205

GGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGAT
AACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAA
CCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAA
CCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCT
GTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTC
TAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGG
ACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTG
GAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGG
TAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATG
GATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATT
GGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTACCCCGG
TTGATAATCAGAAAAGCCCCAAAAACAGGAAGATTGTATAAGCAAATATTTA
AATTGTAAACGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCA
GCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAA
AGAATAGCCCGAGATAGGGTTGAGTGTTGTTCCAGTTTGGAACAAGAGTCCA
CTATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGG
GCGATGGCCCACTACGTGAACCATCACCCAAATCAAGTTTTTTGGGGTCGAG
GTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCT
TGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAA
AGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAAC
CACCACACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCGTAAAAGGATCTA
GGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTT
CGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGA
TCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTAC
CAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTA
ACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGT
AGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTG
CTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGG
GTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACG
GGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGA
GATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAA
AGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGA
GGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGC
CACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCT
ATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGC
CTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTA
TTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCG
CAGCGAGTCAGTGAGCGAGGAAGCTATGGTGCACTCTCAGTACAATCTGCTC
TGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGT
CATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCT
TGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTG
CATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGG
TAAAGCTCATCAGCGTGGTCGTGCAGCGATTCACAGATGTCTGCCTGTTCATC
CGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAA
AGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTG
TAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGA

TGCTCACGATACGGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGT
GAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACT
CAGGGTCAATGCCAGCCGAACGCCAGCAAGACGTAGCCCAGCGCGTCGGCC
GCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGAC
CAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCG
ACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGAC
CCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTC
ATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTG
GGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAG
CTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACC
TGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTT
GCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAG
CTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACG
CTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGA
TATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCA
CCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCT
GATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTG
CATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTA
TCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAG
ACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGA
CCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGA
AAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGG
AACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGA
TAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCG
CTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCA
CCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGT
GCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGC
CAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTT
CCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGG
GAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTA
CTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCC
ATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTC
CCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTT
GAGCACCGCCGCCGCAAGGAATGGTGCATGCCGGCATGCCGCCCTTTCGTCT
TCAAGAATTAATTCCCAATTCCCCAGGCATCAAATAAAACGAAAGGCTCAGT
CGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTG
AGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCC
GGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGAATTAATTCCCCA
GGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTAT
CTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCG
GATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCG
CCATAAACTGCCAGGAATTAATTCCCCAGGCATCAAATAAAACGAAAGGCTC
AGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTC
CTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGC
CCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGAATTAATTCCC
CAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTT

ATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGC
GGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCC
GCCATAAACTGCCAGGAATTAATTCCCCAGGCATCAAATAAAACGAAAGGCT
CAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCT
CCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGG
CCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGAATTGGGGA
TCGGAATTAATTCCCGGTTTAAACCGGGGATCTCGATCCCGCGAAATTAATAC
GACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAAT
TTTGTTTAACTTTAAGAAGGAGATATACATATGCAGATTTTTGTGAAAACCCT
GACCGGCAAAACCATTACCCTGGAAGTGGAACCGAGCGATACCATTGAAAAT
GTGAAAGCGAAAATTCAGGATAAAGAAGGCATTCCGCCGGATCAGCAGCGT
CTGATTTTTGCGGGCAAACAGCTGGAAGATGGTCGTACCCTGAGCGATTACA
ATATTCAGAAAGAAAGCACCCTGCATCTGGTGCTGCGTCTGCGTGGCGGCTG
CATCACGGGAGATGCACTAGTTGCCCTACCCGAGGGCGAGTCGGTACGCATC
GCCGACATCGTGCCGGGTGCGCGGCCCAACAGTGACAACGCCATCGACCTGA
AAGTCCTTGACCGGCATGGCAATCCCGTGCTCGCCGACCGGCTGTTCCACTCC
GGCGAGCATCCGGTGTACACGGTGCGTACGGTCGAAGGTCTGCGTGTGACGG
GCACCGCGAACCACCCGTTGTTGTGTTTGGTCGACGTCGCCGGGGTGCCGAC
CCTGCTGTGGAAGCTGATCGACGAAATCAAGCCGGGCGATTACGCGGTGATT
CAACGCAGCGCATTCAGCGTCGACTGTGCAGGTTTTGCCCGCGGAAAACCCG
AATTTGCGCCCACAACCTACACAGTCGGCGTCCCTGGACTGGTGCGTTTCTTG
GAAGCACACCACCGAGACCCGGACGCCCAAGCTATCGCCGACGAGCTGACC
GACGGGCGGTTCTACTACGCGAAAGTCGCCAGTGTCACCGACGCCGGCGTGC
AGCCGGTGTATAGCCTTCGTGTCGACACGGCAGACCACGCGTTTATCACGAA
CGGGTTCGTCAGCCACGCTACTGGCCTCACCGGTCTGAACTCAGGCCTCACG
ACAAATCCTGGTGTATCCGCTTGGCAGGTCAACACAGCTTATACTGCGGGAC
AATTGGTCACATATAACGGCAAGACGTATAAATGTTTGCAGCCCCACACCTC
CTTGGCAGGATGGGAACCATCCAACGTTCCTGCCTTGTGGCAGCTTCAATGAC
TGCAGGAAGGGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGG
CTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG
GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

pET11a-Ub-His

BamHI - 5567 - G'GATC_C
NdeI - 5316 - CA'TA_TG

Ub
Amp R
pBR322
LacI

pET11a-Ub His WT
5890 bp
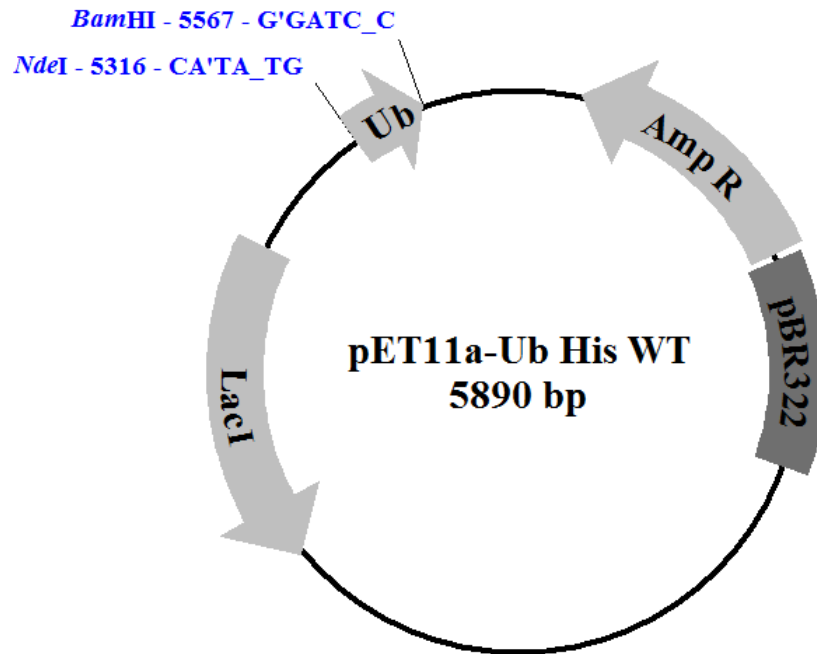
```
TTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCA
TGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGC
GGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATG
AGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGA
GTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTC
CTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCA
GTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATC
CTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGT
TCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTC
GGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCAC
AGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCC
ATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAG
GACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCG
CCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGT
GACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTG
GCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGC
GGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTA
TTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGC
ACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGG
AGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCT
CACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAG
ATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTT
TGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGT
CAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGC
GTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTT
GCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGA
```

GCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTT
CAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAG
TGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACG
ATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACA
CAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTG
AGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATC
CGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGG
GAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAG
CGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCA
GCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGT
TCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGT
GAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGA
GCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTG
CGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGC
CGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGC
TGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTG
CTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGT
GTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAG
CTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGT
CCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGG
GCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGG
GGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTC
ACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGG
GTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGG
GTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCA
GCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTC
CGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTG
CTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATC
GGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCC
TCAACGACAGGAGCACGATCATGCGCACCCGTGGCCAGGACCCAACGCTGCC
CGAGATGCGCCGCGTGCGGCTGCTGGAGATGGCGGACGCGATGGATATGTTC
TGCCAAGGGTTGGTTTGCGCATTCACAGTTCTCCGCAAGAATTGATTGGCTCC
AATTCTTGGAGTGGTGAATCCGTTAGCGAGGTGCCGCCGGCTTCCATTCAGGT
CGAGGTGGCCCGGCTCCATGCACCGCGACGCAACGCGGGGAGGCAGACAAG
GTATAGGGCGGCGCCTACAATCCATGCCAACCCGTTCCATGTGCTCGCCGAG
GCGGCATAAATCGCCGTGACGATCAGCGGTCCAGTGATCGAAGTTAGGCTGG
TAAGAGCCGCGAGCGATCCTTGAAGCTGTCCCTGATGGTCGTCATCTACCTGC
CTGGACAGCATGGCCTGCAACGCGGGCATCCCGATGCCGCCGGAAGCGAGA
AGAATCATAATGGGGAAGGCCATCCAGCCTCGCGTCGCGAACGCCAGCAAG
ACGTAGCCCAGCGCGTCGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGC
CGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCA
AGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAA
GCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGT
TGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCG
CCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGA
TCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGC

CCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCA
ACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTC
ACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGA
GTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTT
GATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATC
CCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCG
CATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACG
ATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCA
GTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTAT
GCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTA
ACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCG
CGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAG
ACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGG
CATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGC
GAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCA
TCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGC
GACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAAT
CAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAAT
TCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGG
CTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACT
CTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTC
TCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGT
GTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCC
AGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCA
AGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCAC
GCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCG
GTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGC
CGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTA
ATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAA
TAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCAGATTTTTGTGAAAA
CCCTGACCGGCAAAACCATTACCCTGGAAGTGGAACCGAGCGATACCATTGA
AAATGTGAAAGCGAAAATTCAGGATAAAGAAGGCATTCCGCCGGATCAGCA
GCGTCTGATTTTTGCGGGCAAACAGCTGGAAGATGGTCGTACCCTGAGCGAT
TACAATATTCAGAAAGAAAGCACCCTGCATCTGGTGCTGCGTCTGCGTGGCG
GCCATCATCATCATCATCATTAAGGATCCGGCTGCTAACAAAGCCCGAAAGG
AAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGG
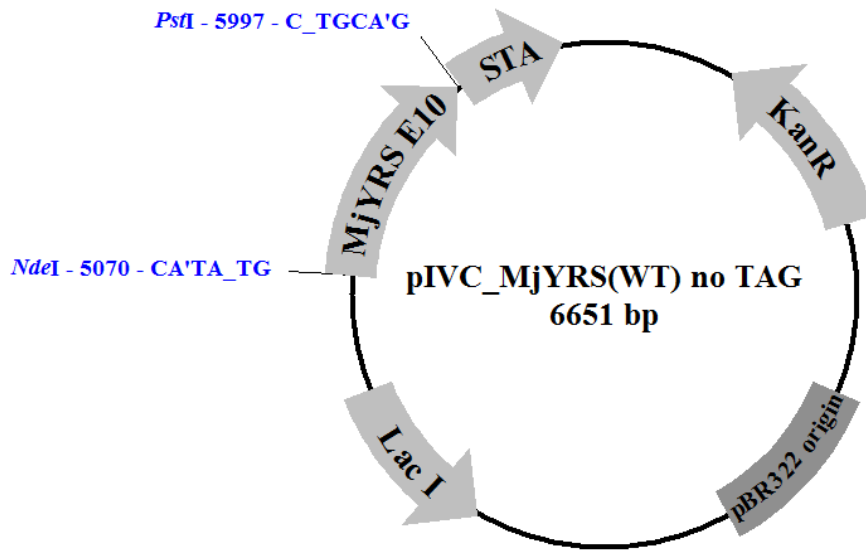GGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCC
GGATATCCCGCAAGAGGCCCGGCAGTACCGGCATAACCAAGCCTATGCCTAC
AGCATCCAGGGTGACGGTGCCGAGGATGACGATGAGCGCATTGTTAGATTTC
ATACACGGTGCCTGACTGCGTTAGCAATTTAACTGTGATAAACTACCGCATTA
AAGCTTATCGATGATAAGCTGTCAAACATGAGAA

pIVC-MjYRS(WT)-STA(no TAG)

*Pst*I - 5997 - C_TGCA'G

*Nde*I - 5070 - CA'TA_TG

STA

MjYRS E10

KanR

pIVC_MjYRS(WT) no TAG
6651 bp

Lac I

pBR322 origin

Synthetase (shown in yellow) and STA (shown in blue) gene were cloned into pET24a plasmid to create pIVC construct. RBSs are indicated as lowcased letters. T7 promoter and terminator sequences are shown in green.

The synthetase gene shown in the following is WT *Mj*YRS and can be replaced by other synthetases as well. The STA ORF contains an N-terminal fusion peptide (including a His$_6$ tag) and the core STA sequence from residues 13 to 140. One or two amber stop codons (TAG) are introduced into STA. The positions of TAG codons are shown by underlined letters (for construct containing one stop codon, TAG is introduced to the second position)

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGT
GGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCC
CGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTC
CCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGC
TTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACG
TAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAG
TCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCA
ACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCG
GCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAAT
TTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAA
ATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATAT
GTATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAAT
GAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAA
AAGCCGTTTCTGTAATGAAGGAGAAAACTCACCGAGGCAGTTCCATAG
GATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATC
AATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGA

212

GAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTT
ATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCA
TCAAAATCACTCGCATCAACCAAACCGTTATTCATTCGTGATTGCGCCT
GAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACA
GGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAAT
ATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTC
CCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTACGGATA
AAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGT
CTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTT
TCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTG
TCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAA
ATCAGCATCCATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTC
CCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCA
GACAGTTTTATTGTTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTC
CACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGAT
CCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGC
TACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTC
CGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTC
TAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGC
CTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGG
CGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGA
TAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCA
GCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGC
TATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTAT
CCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCC
AGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTC
TGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTAT
GGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTG
GCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATA
ACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAA
CGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTG
ATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATAT
ATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAG
TATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACA
CCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGC
ATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCA
GAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAG
CTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCC
GCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGA
TAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGC
CTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAA
ACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCG
GTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCG

```
GGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATAC
AGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGAT
CCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACTTTA
CGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGA
CGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCA
TTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAAC
GACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATA
ATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAG
GCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCC
GATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCA
GAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGT
CATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCT
GACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTA
ATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTC
CAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGC
GCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTC
ACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGA
GAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAA
TCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGG
TATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGG
ACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAA
CCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTG
TTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGC
TGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGA
CGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGG
TGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCAT
GGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAA
ATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCT
GGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGA
GAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTAC
CATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAAT
CGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGC
AACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCG
GTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGC
GTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTC
TGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGT
TTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCA
TACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCT
CTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGA
GGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATG
GCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCG
AAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCG
GTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTG
```

ATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCG
CGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATT
CCCCTCTAGAAATAATTTTGTTTAACTTTAAGaaggagATATACATATGGACG
AATTTGAAATGATAAAGAGAAACACATCTGAAATTATCAGCGAGGAAGAGTT
AAGAGAGGTTTTAAAAAAGATGAAAAATCTGCTTACATAGGTTTTGAAC
CAAGTGGTAAAATACATTTAGGGCATTATCTCCAAATAAAAAAGATGA
TTGATTTACAAAATGCTGGATTTGATATAATTATATTGTTGGCTGATTT
ACACGCCTATTTAAACCAGAAAGGAGAGTTGGATGAGATTAGAAAAAT
AGGAGATTATAACAAAAAAGTTTTTGAAGCAATGGGGGTTAAAGGCAA
AATATGTTTATGGAAGTGAATTCCAGCTTGATAAGGATTATACACTGA
ATGTCTATAGATTGGCTTTAAAAACTACCTTAAAAAGAGCAAGAAGGA
GTATGGAACTTATAGCAAGAGAGGATGAAAATCCAAAGGTTGCTGAA
GTTATCTATCCAATAATGCAGGTTAATGATATTCATTATTTAGGCGTTG
ATGTTGCAGTTGGAGGGATGGAGCAGAGAAAAATACACATGTTAGCA
AGGGAGCTTTTACCAAAAAAGGTTGTTTGTATTCACAACCCTGTCTTAA
CGGGTTTGGATGGAGAAGGAAAGATGAGTTCTTCAAAAGGGAATTTTA
TAGCTGTTGATGACTCTCCAGAAGAGATTAGGGCTAAGATAAAGAAAG
CATACTGCCCAGCTGGAGTTGTTGAAGGAAATCCAATAATGGAGATAG
CTAAATACTTCCTTGAATATCCTTTAACCATAAAAAGGCCAGAAAAAT
TTGGTGGAGATTTGACAGTTAATAGCTATGAGGAGTTAGAGAGTTTAT
TTAAAAATAAGGAATTGCATCCAATGGATTTAAAAAATGCTGTAGCTG
AAGAACTTATAAAGATTTTAGAGCCAATTAGAAAGAGATTATAACTGCagagga
GAAATTTTATGTCTGGTTCTCATCATCATCATCATCATAGCAGC<u>GGC</u>ATCGAA
GGCCGC<u>GGC</u>CGCTTAATTAAACATATGACCGCGGAAGCGGGCATTACAGGCA
CCTGGTATAATCAGCTGGGCAGCACCTTTATAGTGACCGCGGGAGCGGATGG
GGCATTGACGGGAACCTATGAAAGCGCGGTGGGGAATGCGGAGAGCCGTTA
TGTGCTGACCGGCCGTTACGATAGCGCCCCTGCGACCGATGGCTCTGGCACT
GCGCTGGGCTGGACCGTGGCATGGAAAAACAATTACCGTAATGCGCACAGCG
CGACCACCTGGAGCGGCCAATATGTGGGCGGAGCAGAGGCGCGTATTAACA
CCCAATGGCTGTTAACCAGCGGCACCACCGAAGCAAATGCGTGGAAGAGCA
CCCTGGTGGGCCATGATACCTTTACCAAGGTGAAACCGTCAGCGGCGAGCAT
TTAATAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAG
ATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCA
CCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCT
TGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

pRSET-sfGFP
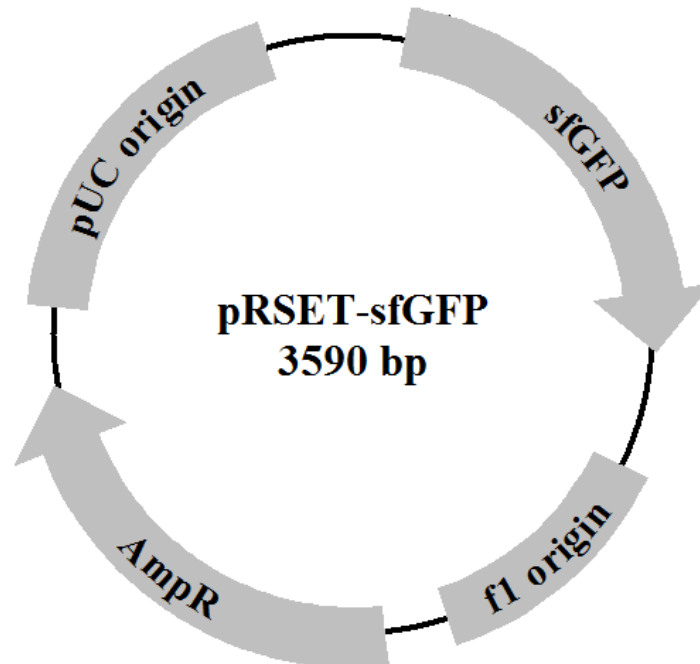
The ORF for sfGFP is highlighted in red (including an N-terminal His tag and upstream fusion peptide). The coding region of sfGFP starts from the shaded ATG codon. Ribosome binding site is shown in green. *Xba* I, *Nhe* I and *Hind* III sites are underlined.

GATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGAGACCACAACGGT
TTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCA
GCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCA
TATGGCTAGCATGATGAGCAAAGGCGAAGAACTGTTTACCGGCGTGGTTCCG
ATTCTGGTGGAACTGGATGGCGATGTAAATGGCCACAAGTTTAGCGTGCGTG
GGGAAGGGGAGGGCGATGCGACCAATGGCAAACTGACCCTGAAGTTTATTTG
CACGACCGGGAAACTGCCGGTTCCTTGGCCCACCCTTGTCACCACCCTGACGT
ATGGCGTGCAATGCTTTAGCCGTTACCCGGACCACATGAAGCGGCATGACTT
CTTCAAAAGCGCCATGCCTGAAGGCTATGTTCAGGAACGGACGATCTCGTTT
AAGGATGACGGCACCTATAAGACCCGTGCGGAGGTCAAATTCGAAGGCGAT
ACCCTGGTGAACCGCATTGAGCTGAAGGGCATCGACTTCAAAGAGGATGGCA
ACATACTGGGGCACAAGCTGGAGTACAACTTCAACAGCCACAACGTCTACAT
CACCGCCGACAAGCAGAAGAACGGCATTAAGGCCAACTTCAAGATTCGGCA
CAATGTGGAGGACGGAAGCGTTCAGCTGGCGGATCATTATCAACAGAATACC
CCCATTGGCGACGGTCCCGTGCTTCTGCCGGATAATCATTACTTGAGCACCCA
GAGCGTGCTGAGCAAGGACCCGAATGAGAAACGGGATCACATGGTGCTGCT
GGAATTTGTGACCGCTGCGGGCATTACACATGGCATGGATGAACTGTATAAA
TAAAAGCTTGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTG
CTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT
CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATCTGGCGTAATA
GCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGG

CGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTT
ACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGC
TTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAA
TCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCA
AAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGAC
GGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTT
CCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAG
GGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAA
TTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTT
CGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAA
TATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAA
AAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTT
GCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAA
AGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTC
AACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGA
TGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCC
GGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTG
AGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGA
ATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTC
TGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGG
GGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATA
CCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGC
GCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATA
GACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTC
CGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGC
GGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTAT
CTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGC
TGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACT
CATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAG
GTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTC
GTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGAT
CCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACC
AGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAA
CTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAG
TTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCT
AATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGT
TGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGG
GGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAG
ATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAG
GGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCC
ACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTA
TGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCC
TTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTAT
TACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGC

217

AGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCT
CTCCCCGCGCGTTGGCCGATTCATTAATGCAG

MuCDM transposon (asymmetric end sequences highlighted in red)

AAAAGATCTTTAGCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGC
GAAAACGGAAGGTATTCGCAATAATCTTAATACTGAGAATCCATTATGGGAC
GCTATTGTTGGCTTAGGATTCTTGAAGGACGGTGTCAAAAATATTCCTTCTTT
CTTGTCTACGGACAATATCGGTACTCGTGAAACATTTCTTGCTGGTCTAATTG
ATTCTGATGGCTATGTTACTGATGAGCATGGTATTAAAGCAACAATAAAGAC
AATTCATACTTCTGTCAGAGATGGTTTGGTTTCCCTTGCTCGTTCTTTAGGCTT
AGTAGTCTCGGTTAACGCAGAACCTGCTAAGGTTGACATGAATGGCACCAAA
CATAAAATTAGTTATGCTATTTATATGTCTGGTGGAGATGTTTTGCTTAACGT
TCTTTCGAAGTGTGCCGGCTCTAAAAAATTCAGGCCTGCTCCCGCCGCTGCTT
TTGCACGTGAGTGCCGCGGATTTTATTTCGAGTTACAAGAATTGAAGGAAGA
CGATTATTATGGGATTACTTTATCTGATGATTCTGATCATCAGTTTTTGCTTGC
CAACCAGGTTGTCGTCCATAATTGCGCTAGTCACCCAGAAACGCTGGTGAAA
GTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGG
ATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCA
ATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGA
CGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG
GTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAA
GAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACCGCGGCCAACTT
ACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAAC
ATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAG
CCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAA
CGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAA
TTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGG
CCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGG
TCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGT
AGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACA
GATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGAGTCAGACCGC
GTTTACTCATTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTCGTGCGCCGT
CGGAGATCTTTT

MuCSM transposon (asymmetric end sequences highlighted in red)

AAAGGATCTTTAGCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGC
GAAAACGGAAGGTATTCGCAATAATCTTAATACTGAGAATCCATTATGGGAC
GCTATTGTTGGCTTAGGATTCTTGAAGGACGGTGTCAAAAATATTCCTTCTTT
CTTGTCTACGGACAATATCGGTACTCGTGAAACATTTCTTGCTGGTCTAATTG
ATTCTGATGGCTATGTTACTGATGAGCATGGTATTAAAGCAACAATAAAGAC

AATTCATACTTCTGTCAGAGATGGTTTGGTTTCCCTTGCTCGTTCTTTAGGCTT
AGTAGTCTCGGTTAACGCAGAACCTGCTAAGGTTGACATGAATGGCACCAAA
CATAAAATTAGTTATGCTATTTATATGTCTGGTGGAGATGTTTTGCTTAACGT
TCTTTCGAAGTGTGCCGGCTCTAAAAAATTCAGGCCTGCTCCCGCCGCTGCTT
TTGCACGTGAGTGCCGCGGATTTTATTTCGAGTTACAAGAATTGAAGGAAGA
CGATTATTATGGGATTACTTTATCTGATGATTCTGATCATCAGTTTTTGCTTGC
CAACCAGGTTGTCGTCCATAATTGCGCTAGTCACCCAGAAACGCTGGTGAAA
GTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGG
ATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCA
ATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGA
CGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG
GTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAA
GAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACCGCGGCCAACTT
ACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAAC
ATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAG
CCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAA
CGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAA
TTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGG
CCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGG
TCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGT
AGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACA
GATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGAGTCAGACCGC
GTTTACTCATTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTCGTGCGCCGT
CGGAGATCTTTT

M. barkeri PylT

GGGAACCTGATCATGTAGATCGAATGGACTCTAAATCCGTTCAGCCGGGTTA
GATTCCCGGGGTTTCCGCCA

Note: *M barkeri* PylT (*Mb*PylT) has one nucleotide difference from *M. mazei* PylT (*Mm*PylT) (highlighted in red). In *Mm*PylT, position 42 is a T instead of C. Both *Mb*PylT and *Mm*PylT can be readily aminoacylated by *M. mazei* tRNA synthetase.

M. mazei pyrrolysyl-tRNA synthetase

ATGGATAAAAAACCACTAAACACTCTGATATCTGCAACCGGGCTCTGGATGT
CCAGGACCGGAACAATTCATAAAATAAAACACCACGAAGTCTCTCGAAGCA
AAATCTATATTGAAATGGCATGCGGAGACCACCTTGTTGTAAACAACTCCAG
GAGCAGCAGGACTGCAAGAGCGCTCAGGCACCACAAATACAGGAAGACCTG
CAAACGCTGCAGGGTTTCGGATGAGGATCTCAATAAGTTCCTCACAAAGGCA
AACGAAGACCAGACAAGCGTAAAGTCAAGGTCGTTTCTGCCCCTACCAGAA
CGAAAAAGGCAATGCCAAAATCCGTTGCGAGAGCCCCGAAACCTCTTGAGA
ATACAGAAGCGGCACAGGCTCAACCTTCTGGATCTAAATTTTCACCTGCGAT
ACCGGTTTCCACCCAAGAGTCAGTTTCTGTCCCGGCATCTGTTTCAACATCAA
TATCAAGCATTTCTACAGGAGCAACTGCATCCGCACTGGTAAAAGGGAATAC

GAACCCCATTACATCCATGTCTGCCCCTGTTCAGGCAAGTGCCCCCGCACTTA
CGAAGAGCCAGACTGACAGGCTTGAAGTCCTGTTAAACCCAAAAGATGAGAT
TTCCCTGAATTCCGGCAAGCCTTTCAGGGAGCTTGAGTCCGAATTGCTCTCTC
GCAGAAAAAAGACCTGCAGCAGATCTACGCGGAAGAAAGGGAGAATTATC
TGGGGAAACTCGAGCGTGAAATTACCAGGTTCTTTGTGGACAGGGGTTTTCT
GGAAATAAAATCCCCGATCCTGATCCCTCTTGAGTATATCGAAAGGATGGGC
ATTGATAATGATACCGAACTTTCAAAACAGATCTTCAGGGTTGACAAGAACT
TCTGCCTGAGACCCATGCTTGCTCCAAACCTTTACAACTACCTGCGCAAGCTT
GACAGGGCCCTGCCTGATCCAATAAAAATTTTTGAAATAGGCCCATGCTACA
GAAAAGAGTCCGACGGCAAAGAACACCTCGAAGAGTTTACCATGCTGAACTT
CTGCCAGATGGGATCGGGATGCACACGGGAAAATCTTGAAAGCATAATTACG
GACTTCCTGAACCACCTGGGAATTGATTTCAAGATCGTAGGCGATTCCTGCAT
GGTCTATGGGGATACCCTTGATGTAATGCACGGAGACCTGGAACTTTCCTCTG
CAGTAGTCGGACCCATACCGCTTGACCGGGAATGGGGTATTGATAAACCCTG
GATAGGGGCAGGTTTCGGGCTCGAACGCCTTCTAAAGGTTAAACACGACTTT
AAAAATATCAAGAGAGCTGCAAGGTCCGAGTCTTACTATAACGGGATTTCTA
CCAACCTG

Codon-optimized Ub (SynUb) for expression in E. coli

All the seven lysines are highlighted in red (K6, K11, K27, K29, K33, K48 and K63)

ATGCAGATTTTTGTGAAAACCCTGACCGGCAAAACCATTACCCTGGAAGTGG
AACCGAGCGATACCATTGAAAATGTGAAAGCGAAAATTCAGGATAAAGAAG
GCATTCCGCCGGATCAGCAGCGTCTGATTTTTGCGGGCAAACAGCTGGAAGA
TGGTCGTACCCTGAGCGATTACAATATTCAGAAAGAAAGCACCCTGCATCTG
GTGCTGCGTCTGCGTGGCGGCTAA

Codon-optimized sfGFP for expression in E. coli

ATGAGCAAAGGCGAAGAACTGTTTACCGGCGTGGTTCCGATTCTGGTGGAAC
TGGATGGCGATGTAAATGGCCACAAGTTTAGCGTGCGTGGGGAAGGGGAGG
GCGATGCGACCAATGGCAAACTGACCCTGAAGTTTATTTGCACGACCGGGAA
ACTGCCGGTTCCTTGGCCCACCCTTGTCACCACCCTGACGTATGGCGTGCAAT
GCTTTAGCCGTTACCCGGACCACATGAAGCGGCATGACTTCTTCAAAAGCGC
CATGCCTGAAGGCTATGTTCAGGAACGGACGATCTCGTTTAAGGATGACGGC
ACCTATAAGACCCGTGCGGAGGTCAAATTCGAAGGCGATACCCTGGTGAACC
GCATTGAGCTGAAGGGCATCGACTTCAAAGAGGATGGCAACATACTGGGGCA
CAAGCTGGAGTACAACTTCAACAGCCACAACGTCTACATCACCGCCGACAAG
CAGAAGAACGGCATTAAGGCCAACTTCAAGATTCGGCACAATGTGGAGGAC
GGAAGCGTTCAGCTGGCGGATCATTATCAACAGAATACCCCCATTGGCGACG
GTCCCGTGCTTCTGCCGGATAATCATTACTTGAGCACCCAGAGCGTGCTGAGC
AAGGACCCGAATGAGAAACGGGATCACATGGTGCTGCTGGAATTTGTGACCG
CTGCGGGCATTACACATGGCATGGATGAACTGTATAAATAA

Appendix V: Sequencing Results from Transposon Mutagenesis

Sequence analysis of the deletion mutants identified from MuDel deletion experiments

Table S-1. Sequencing results of fluorescent and non-fluorescent clones

	Deletions/mutations[a][b]	Frequency
Fluorescent mutants	M1	1
	S2/K3→R	4
	G4Δ	11
	E5/E6→E	3
	F8/T9→S	1
	C48Δ	1
	P75/D76→H	1
	P75Δ	1
	E172Δ	1
	S175/V176→F	1
	A226/A227→A	2
	G228Δ	4
	G228/I229→V	2
Non-fluorescent mutants	G10Δ	1
	D21Δ	2
	G24Δ	1
	E34Δ	3
	P58Δ	1
	P89/E90→Q	1
	T203Δ	1
Unexpected	No deletions stop codon mutations 4 bp deletions	21

Note: [a] If occurred in two neighboring codons, the 3-bp deletions can result in unexpected mutations

[b] Many mutants have more than one possibility for the deletion positions. The numbers indicate the positions in the DNA sequence of GFP$_{UV}$ gene.

Library characterization of CDM

Table S-2. Deletions in the 104 sequenced clones

	Total	Incorrect	Correct in VMAI-N[bc]	Correct in GFP$_{UV}$[c]
One codon	20	2	-27(3), -17	D21 (2), G33, G35, T49, N105 (2), L125, I136, N170, H181, H217 (2), I229
Two codon[a]	21	1	-35	G10 (2), D19 (3), D21 (2), G24 (2), G33 (3), R73, K85, E111, E124, G138 (2), E142
Three codon[a]	21	2	-	T9 (4), D19 (2), N23 (1), G33 (2), P89, G103 (2), E115, I123 (3), L137 (2), S147
Four codon[a]	20	1	-59	F8 (2), L18 (3), F71 (2), M88, F114 (2), D117, R122 (3), K140, N146, I188, R215
Five codon[a]	22	2	-59, -56, -40(2), -27	S2, F8, V12, L18, V22 (2), E34, F71, R80, K101, D102, R122, I136, H181, I229

Note:

[a] In case of multiple codon deletions, only the first codons are shown

[b] The C-terminal residue of the leading peptide, which is adjacent to GFP$_{UV}$, is numbered as -1.

[c] The number of identical clones is indicated in parentheses.

Library characterization of CSM

Table S-3. Correct mutations in the 104 sequenced clones

Codon libraries	Correct mutations
One	D21L, D21G, E32S, Q69L, N149H, N149D, N149R, N149G, Y151R, S205H
Two	V11H/V12N, V11H/V12H, V22H/N23R, S28N/V29S, Y74R/P75L, F83N/F84H, G91L/Y92N, T105H/Y106R, N135H/I136R, I161C/K162L, S175L/V176S, S205H/V206L, G232S/M233D
Three	G10H/V11R/V12H, G10H/V11R/V12H G10Y/V11H/V12R, G10G/V11D/V12C, G10Y/V11C/V12V, G10H/V11V/V12Y, G10Y/V11H/V12H, G20S/D21Y/V22R, V68R/Q69Y/C70H, V68R/Q69G/C70N, H148R/N149F/V150Y, G160H/I161V/K162N, H169H/N170R/V171D, P187H/I188N/G189Y

Appendix VI: Synthesis of UAAs

Synthesis of OnbLys

Step 1: synthesis of *N-tert*-butoxycarbonyl-$^{\varepsilon}$*N-o*-nitrobenzyloxycarbonyl-L-lysine

First, 1g (4.060 mmol) of *N*-Boc-L-lysine was dissolved in 6 mL of tetrahydrofuran (THF) (0.67 M). The solution was cooled to 0 ℃ in an ice bath and then was added 13 mL of 1 M NaOH. The solution was left stirring for 5 min at the same temperature and then was added slowly 900 mg (4.18 mmol; 1.03 eq) of *o*-nitrobenzyloxychloride over a period of 10 min. The reaction mixture was left stirring 12 h at room temperature and then cooled down to 0 ℃ with an ice bath. The prechilled solution was acidified by the addition of 10 mL of 1 M HCl. The acidified reaction mixture was extracted with 5 mL ether three times. The organic phase was discarded. The aqueous phase was extracted with 10 mL ethyl acetate for three times, and dried by anhydrous Na_2SO_4 and finally purified by column chromatography (60:30:10 for hexane/ethyl acetate/acetic acid). In total, 0.740 g (43% yield) of *N*-Boc-$^{\varepsilon}$*N-o*-nitrobenzyl-L-lysine (*N*-Boc-$^{\varepsilon}$*N*-OnbLys) was isolated as white crystals (Rf: 0.60). The purified product was characterized by NMR and MS. The compound signals were consistent with previous study (*121*).

Step 2: synthesis of $^{\varepsilon}$*N*-Onb-L-lysine (OnbLys)

The *N*-Boc-$^{\varepsilon}$*N*-OnbLys from last step was dissolved in 2 mL of dioxane (0.87M). Then was added, under argon atmosphere, 10 mL of 4 M HCl-dioxane (0.04 mol; 23 eq) by syringe. The reaction mixture was left stirring under argon and at room temperature over 2 h. Finally, the solvent was removed by reducing pressure and triturating with ether for 4 h. Finally, 0.400 g (57% yield) of OnbLys was isolated as the HCl salt and characterized by NMR and MS. The compound signals were consistent with previous study (*121*).

Synthesis of OnbSTyr and OabSTyr

Scheme 1: Structure and synthesis of *o*-nitrobenzyl thiotyrosine (OnbSTyr) and *o*-amino benzyl thiotyrosine (OabSTyr)
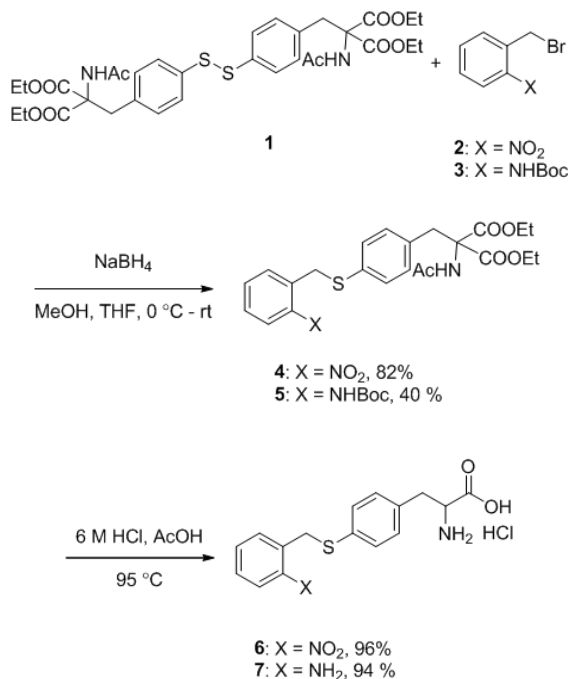
The synthesis of *o*-nitrobenzyl thiotyrosine hydrochloride **6** and its reduced analogue *o*-amino benzyl tyrosine hydrochloride **7** was started from the known disulfide ester **1** which was synthesized in two steps starting from toluene disulfide as reported in literature (*168*). Reduction of disulfide bond of **1** in the presence of NaBH$_4$ followed by alkylation with *o*-nitrobenzylbromide **2** or Boc protected aminobenzyl bromide **3** in a THF/EtOH mixture (1:1) at room temperature furnished corresponding sulfides **4** (yield 82%) or **5** (yield 40%). Boc protected aminobenzyl bromide **3** was synthesized from commercial product aminobenzyl alcohol in two steps according to the reported procedure (*169*). Finally, acetamidomalonate **4** or **5** was refluxed in a mixture of aqueous 6 M HCl and acetic acid (1:1) to obtain corresponding amino acid hydrochlorides **6** (yield 96%) or **7** (yield 94%) respectively in pure crystalline form.

Bibliography

1.   Tsien, R. Y. (2009) Constructing and exploiting the fluorescent protein paintbox (Nobel Lecture). *Angew Chem Int Ed Engl 48*, 5612-26.

2.   Arnold, F. H., and Georgiou, G. (2003) Directed evolution library creation: methods and protocols, in *Methods in Molecular Biology* (Arnold, F. H., and Georgiou, G., Eds.), Humana Press, Totowa.

3.   Arnold, F. H., and Georgiou, G. (2003) Directed enzyme evolution: screening and selection methods, in *Methods in Molecular Biology* (Arnold, F. H., and Georgiou, G., Eds.), Humana Press, Totowa.

4.   Hutchison, C. A., 3rd, Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., and Smith, M. (1978) Mutagenesis at a specific position in a DNA sequence. *J Biol Chem 253*, 6551-60.

5.   Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. (1996) Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science 273*, 1392-5.

6.   Cirino, P. C., Mayer, K. M., and Umeno, D. (2003) Generating mutant libraries using error-prone PCR, in *Methods Mol Biol* (Arnold, F. H., and Georgiou, G., Eds.) pp 3-9.

7.   Stemmer, W. P. (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature 370*, 389-91.

8.   Fromant, M., Blanquet, S., and Plateau, P. (1995) Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal Biochem 224*, 347-53.

9.   Lin-Goerke, J. L., Robbins, D. J., and Burczak, J. D. (1997) PCR-based random mutagenesis using manganese and reduced dNTP concentration. *Biotechniques 23*, 409-12.

10.   Vanhercke, T., Ampe, C., Tirry, L., and Denolf, P. (2005) Reducing mutational bias in random protein libraries. *Anal Biochem 339*, 9-14.

11.   Nagai, T., Ibata, K., Park, E. S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol 20*, 87-90.

12.   Tsien, R. Y. (1998) The green fluorescent protein. *Annu Rev Biochem 67*, 509-44.

13. Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. P. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol 14*, 315-9.

14. Georgescu, R., Bandara, G., and Sun, L. (2003) Saturation mutagenesis, in *Methods Mol Biol* (Arnold, F. H., and Georgiou, G., Eds.) pp 75-83.

15. Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N., Palmer, A. E., and Tsien, R. Y. (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotechnol 22*, 1567-72.

16. Garfinkel, D. J., and Strathern, J. N. (1991) Ty mutagenesis in *Saccharomyces cerevisiae*. *Methods Enzymol 194*, 342-61.

17. Groisman, E. A. (1991) *In vivo* genetic engineering with bacteriophage Mu. *Methods Enzymol 204*, 180-212.

18. Berg, C. M., Wang, G., Strausbaugh, L. D., and Berg, D. E. (1993) Transposon-facilitated sequencing of DNAs cloned in plasmids. *Methods Enzymol 218*, 279-306.

19. Mizuuchi, K. (1983) *In vitro* transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction. *Cell 35*, 785-94.

20. Mizuuchi, K. (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annu Rev Biochem 61*, 1011-51.

21. Craigie, R., and Mizuuchi, K. (1987) Transposition of Mu DNA: joining of Mu to target DNA can be uncoupled from cleavage at the ends of Mu. *Cell 51*, 493-501.

22. Savilahti, H., Rice, P. A., and Mizuuchi, K. (1995) The phage Mu transpososome core: DNA requirements for assembly and function. *Embo J 14*, 4893-903.

23. Haapa, S., Taira, S., Heikkinen, E., and Savilahti, H. (1999) An efficient and accurate integration of mini-Mu transposons *in vitro*: a general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res 27*, 2777-84.

24. Poussu, E., Vihinen, M., Paulin, L., and Savilahti, H. (2004) Probing the alpha-complementing domain of *E. coli* beta-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu *in vitro* DNA transposition. *Proteins 54*, 681-92.

25. Jones, D. D. (2005) Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. *Nucleic Acids Res 33*, e80.

26. Allet, B. (1979) Mu insertion duplicates a 5 base pair sequence at the host inserted site. *Cell 16*, 123-9.

27. Kahmann, R., and Kamp, D. (1979) Nucleotide sequences of the attachment sites of bacteriophage Mu DNA. *Nature 280*, 247-50.

28. Bock, A., Forchhammer, K. c., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., and Zinoni, F. (1991) Selenocysteine: the 21st amino acid. *Mol Microbiol 5*, 515-20.

29. Srinivasan, G., James, C. M., and Krzycki, J. A. (2002) Pyrrolysine encoded by UAG in archaea: charging of a UAG-decoding specialized tRNA. *Science 296*, 1459-62.

30. Merrifield, R. B. (1969) Solid-phase peptide synthesis. *Adv Enzymol Relat Areas Mol Biol 32*, 221-96.

31. Dawson, P. E., and Kent, S. B. (2000) Synthesis of native proteins by chemical ligation. *Annu Rev Biochem 69*, 923-60.

32. Muir, T. W. (2003) Semisynthesis of proteins by expressed protein ligation. *Annu Rev Biochem 72*, 249-89.

33. Hecht, S. M., Alford, B. L., Kuroda, Y., and Kitano, S. (1978) "Chemical aminoacylation" of tRNA's. *J Biol Chem 253*, 4517-20.

34. Heckler, T. G., Chang, L. H., Zama, Y., Naka, T., Chorghade, M. S., and Hecht, S. M. (1984) T4 RNA ligase mediated preparation of novel "chemically misacylated" tRNA^{Phe}S. *Biochemistry 23*, 1468-73.

35. 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*, 182-8.

36. Kohrer, C., Yoo, J. H., Bennett, M., Schaack, J., and RajBhandary, U. L. (2003) A possible approach to site-specific insertion of two different unnatural amino acids into proteins in mammalian cells *via* nonsense suppression. *Chem Biol 10*, 1095-102.

37. Rodriguez, E. A., Lester, H. A., and Dougherty, D. A. (2006) *In vivo* incorporation of multiple unnatural amino acids through nonsense and frameshift suppression. *Proc Natl Acad Sci U S A 103*, 8650-5.

38. Hendrickson, T. L., de Crecy-Lagard, V., and Schimmel, P. (2004) Incorporation of nonnatural amino acids into proteins. *Annu Rev Biochem 73*, 147-76.

39. Link, A. J., Mock, M. L., and Tirrell, D. A. (2003) Non-canonical amino acids in protein engineering. *Curr Opin Biotechnol 14*, 603-9.

40. 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*, 5652-3.

41. Sharma, N., Furter, R., Kast, P., and Tirrell, D. A. (2000) Efficient introduction of aryl bromide functionality into proteins *in vivo*. *FEBS Lett 467*, 37-40.

42. Chin, J. W., Cropp, T. A., Anderson, J. C., Mukherji, M., Zhang, Z., and Schultz, P. G. (2003) An expanded eukaryotic genetic code. *Science 301*, 964-7.

43. Wang, L., Brock, A., Herberich, B., and Schultz, P. G. (2001) Expanding the genetic code of *Escherichia coli*. *Science 292*, 498-500.

44. Liu, C. C., and Schultz, P. G. (2010) Adding new chemistries to the genetic code. *Annu Rev Biochem 79*, 413-44.

45. Greiss, S., and Chin, J. W. (2011) Expanding the genetic code of an animal. *J Am Chem Soc 133*, 14196-9.

46. Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., Blumenthal, R. M., Degtyarev, S., Dryden, D. T., Dybvig, K., Firman, K., Gromova, E. S., Gumport, R. I., Halford, S. E., Hattman, S., Heitman, J., Hornby, D. P., Janulaitis, A., Jeltsch, A., Josephsen, J., Kiss, A., Klaenhammer, T. R., Kobayashi, I., Kong, H., Kruger, D. H., Lacks, S., Marinus, M. G., Miyahara, M., Morgan, R. D., Murray, N. E., Nagaraja, V., Piekarowicz, A., Pingoud, A., Raleigh, E., Rao, D. N., Reich, N., Repin, V. E., Selker, E. U., Shaw, P. C., Stein, D. C., Stoddard, B. L., Szybalski, W., Trautner, T. A., Van Etten, J. L., Vitor, J. M., Wilson, G. G., and Xu, S. Y. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res 31*, 1805-12.

47. Ochman, H., Gerber, A. S., and Hartl, D. L. (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics 120*, 621-3.

48. Gerth, M. L., Patrick, W. M., and Lutz, S. (2004) A second-generation system for unbiased reading frame selection. *Protein Eng Des Sel 17*, 595-602.

49. Daggett, K. A., Layer, M., and Cropp, T. A. (2009) A general method for scanning unnatural amino acid mutagenesis. *ACS Chem. Biol. 4*, 109-13.

50. Dopf, J., and Horiagon, T. M. (1996) Deletion mapping of the *Aequorea victoria* green fluorescent protein. *Gene 173*, 39-44.

51. Li, X., Zhang, G., Ngo, N., Zhao, X., Kain, S. R., and Huang, C. C. (1997) Deletions of the *Aequorea victoria* green fluorescent protein define the minimal domain required for fluorescence. *J Biol Chem 272*, 28545-9.

52. Flores-Ramirez, G., Rivera, M., Morales-Pablos, A., Osuna, J., Soberon, X., and Gaytan, P. (2007) The effect of amino acid deletions and substitutions in the longest loop of GFP. *BMC Chem Biol 7*, 1.

53. Shaner, N. C., Patterson, G. H., and Davidson, M. W. (2007) Advances in fluorescent protein technology. *J Cell Sci 120*, 4247-60.

54. Yang, F., Moss, L. G., and Phillips, G. N., Jr. (1996) The molecular structure of green fluorescent protein. *Nat Biotechnol 14*, 1246-51.

55. Heim, R., Prasher, D. C., and Tsien, R. Y. (1994) Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc Natl Acad Sci U S A 91*, 12501-4.

56. Haseloff, J., Siemering, K. R., Prasher, D. C., and Hodge, S. (1997) Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic *Arabidopsis* plants brightly. *Proc Natl Acad Sci U S A 94*, 2122-7.

57. Hanazono, Y., Yu, J. M., Dunbar, C. E., and Emmons, R. V. (1997) Green fluorescent protein retroviral vectors: low titer and high recombination frequency suggest a selective disadvantage. *Hum Gene Ther 8*, 1313-9.

58. Liu, H. S., Jan, M. S., Chou, C. K., Chen, P. H., and Ke, N. J. (1999) Is green fluorescent protein toxic to the living cells? *Biochem Biophys Res Commun 260*, 712-7.

59. Kimura, T., Takeda, S., Kyozuka, J., Asahi, T., Shimamoto, K., and Nakamura, K. (1993) The presequence of a precursor to the delta-subunit of sweet potato mitochondrial F1ATPase is not sufficient for the transport of beta-glucuronidase (GUS) into mitochondria of tobacco, rice and yeast cells. *Plant Cell Physiol 34*, 345-55.

60. Lisenbee, C. S., Karnik, S. K., and Trelease, R. N. (2003) Overexpression and mislocalization of a tail-anchored GFP redefines the identity of peroxisomal ER. *Traffic 4*, 491-501.

61. Hanson, M. R., and Kohler, R. H. (2001) GFP imaging: methodology and application to investigate cellular compartmentation in plants. *J Exp Bot 52*, 529-39.

62. Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J., and King, J. (1991) Global suppression of protein folding defects and inclusion body formation. *Science 253*, 54-8.

63. Sano, T., Pandori, M. W., Chen, X., Smith, C. L., and Cantor, C. R. (1995) Recombinant core streptavidins. A minimum-sized core streptavidin has enhanced

structural stability and higher accessibility to biotinylated macromolecules. *J Biol Chem 270*, 28204-9.

64. Kim, Y. C., and Morrison, S. L. (2009) N-terminal domain-deleted mu transposase exhibits increased transposition activity with low target site preference in modified buffers. *J Mol Microbiol Biotechnol 17*, 30-40.

65. de Wildt, R. M., van Venrooij, W. J., Winter, G., Hoet, R. M., and Tomlinson, I. M. (1999) Somatic insertions and deletions shape the human antibody repertoire. *J Mol Biol 294*, 701-10.

66. Wong, T. S., Roccatano, D., Zacharias, M., and Schwaneberg, U. (2006) A statistical analysis of random mutagenesis methods used for directed protein evolution. *J Mol Biol 355*, 858-71.

67. Murakami, H., Hohsaka, T., and Sisido, M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. *Nat Biotechnol 20*, 76-81.

68. Baldwin, A. J., Busse, K., Simm, A. M., and Jones, D. D. (2008) Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). *Nucleic Acids Res 36*, e77.

69. Edwards, W. R., Busse, K., Allemann, R. K., and Jones, D. D. (2008) Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Res 36*, e78.

70. Liu, J., and Cropp, T. A. (2012) A method for multi-codon scanning mutagenesis of proteins based on asymmetric transposons. *Protein Eng Des Sel 25*, 67-72.

71. Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008) Addressing the numbers problem in directed evolution. *Chembiochem 9*, 1797-804.

72. Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol 24*, 79-88.

73. Pascarella, S., and Argos, P. (1992) Analysis of insertions/deletions in protein structures. *J Mol Biol 224*, 461-71.

74. Zacharias, D. A., and Tsien, R. Y. (2005) *Green fluorescent protein: properties, applications and protocols*, John Wiley & Sons, Inc., Hoboken.

75. Ward, W. W. (1981) *Bioluminescence and Chemiluminescence*, Academic Press, New York.

76. Patterson, G. H., Knobel, S. M., Sharif, W. D., Kain, S. R., and Piston, D. W. (1997) Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophys J 73*, 2782-90.

77. Cormack, B. P., Valdivia, R. H., and Falkow, S. (1996) FACS-optimized mutants of the green fluorescent protein (GFP). *Gene 173*, 33-8.

78. Patrick, W. M., Firth, A. E., and Blackburn, J. M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng 16*, 451-7.

79. Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., and Griffiths, A. D. (2006) Amplification of complex gene libraries by emulsion PCR. *Nat Methods 3*, 545-50.

80. Brejc, K., Sixma, T. K., Kitts, P. A., Kain, S. R., Tsien, R. Y., Ormo, M., and Remington, S. J. (1997) Structural basis for dual excitation and photoisomerization of the *Aequorea victoria* green fluorescent protein. *Proc Natl Acad Sci U S A 94*, 2306-11.

81. Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A 103*, 5869-74.

82. Cunningham, B. C., and Wells, J. A. (1989) High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science 244*, 1081-5.

83. Depuydt, M., Messens, J., and Collet, J. F. (2011) How proteins form disulfide bonds. *Antioxid Redox Signal 15*, 49-66.

84. Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990) DNA recombination during PCR. *Nucleic Acids Res 18*, 1687-91.

85. Liu, J., Castaneda, 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*, 5613-6.

86. Ikeda, F., and Dikic, I. (2008) Atypical ubiquitin chains: new molecular signals. 'Protein Modifications: Beyond the Usual Suspects' review series. *EMBO Rep 9*, 536-42.

87. Hochstrasser, M. (2006) Lingering mysteries of ubiquitin-chain assembly. *Cell 124*, 27-34.

88. Pickart, C. M. (2004) Back to the future with ubiquitin. *Cell 116*, 181-90.

89. Dikic, I., Wakatsuki, S., and Walters, K. J. (2009) Ubiquitin-binding domains - from structures to functions. *Nat Rev Mol Cell Biol 10*, 659-71.

90. Kirkin, V., and Dikic, I. (2007) Role of ubiquitin- and Ubl-binding proteins in cell signaling. *Curr Opin Cell Biol 19*, 199-205.

91. Sun, L., and Chen, Z. J. (2004) The novel functions of ubiquitination in signaling. *Curr Opin Cell Biol 16*, 119-26.

92. Pickart, C. M., and Eddins, M. J. (2004) Ubiquitin: structures, functions, mechanisms. *Biochim Biophys Acta 1695*, 55-72.

93. Behrends, C., and Harper, J. W. (2011) Constructing and decoding unconventional ubiquitin chains. *Nat Struct Mol Biol 18*, 520-8.

94. Piotrowski, J., Beal, R., Hoffman, L., Wilkinson, K. D., Cohen, R. E., and Pickart, C. M. (1997) Inhibition of the 26 S proteasome by polyubiquitin chains synthesized to have defined lengths. *J Biol Chem 272*, 23712-21.

95. Pickart, C. M., and Raasi, S. (2005) Controlled synthesis of polyubiquitin chains. *Methods Enzymol 399*, 21-36.

96. Chatterjee, C., McGinty, R. K., Pellois, J. P., and Muir, T. W. (2007) Auxiliary-mediated site-specific peptide ubiquitylation. *Angew Chem Int Ed Engl 46*, 2814-8.

97. Yang, R., Pasunooti, K. K., Li, F., Liu, X. W., and Liu, C. F. (2009) Dual native chemical ligation at lysine. *J Am Chem Soc 131*, 13592-3.

98. Nguyen, D. P., Garcia Alai, M. M., Virdee, S., and Chin, J. W. Genetically directing varepsilon-N, N-dimethyl-L-lysine in recombinant histones. *Chem Biol 17*, 1072-6.

99. Virdee, S., Ye, Y., Nguyen, D. P., Komander, D., and Chin, J. W. (2010) Engineered diubiquitin synthesis reveals Lys29-isopeptide specificity of an OTU deubiquitinase. *Nat Chem Biol 6*, 750-7.

100. Castaneda, C. A., Liu, J., Kashyap, T. R., Singh, R. K., Fushman, D., and Cropp, T. A. (2011) Controlled enzymatic synthesis of natural-linkage, defined-length polyubiquitin chains using lysines with removable protecting groups. *Chem Commun (Camb) 47*, 2026-8.

101. Castaneda, C., Liu, J., Chaturvedi, A., Nowicka, U., Cropp, T. A., and Fushman, D. Nonenzymatic assembly of natural polyubiquitin chains of any linkage composition and isotopic labeling scheme. *J Am Chem Soc 133*, 17855-68.

102. Williamson, A., Wickliffe, K. E., Mellone, B. G., Song, L., Karpen, G. H., and Rape, M. (2009) Identification of a physiological E2 module for the human anaphase-promoting complex. *Proc Natl Acad Sci U S A 106*, 18213-8.

103. Hofmann, R. M., and Pickart, C. M. (2001) *In vitro* assembly and recognition of Lys-63 polyubiquitin chains. *J Biol Chem 276*, 27936-43.

104. Bang, D., Makhatadze, G. I., Tereshko, V., Kossiakoff, A. A., and Kent, S. B. (2005) Total chemical synthesis and X-ray crystal structure of a protein diastereomer: [D-Gln 35]ubiquitin. *Angew Chem Int Ed Engl 44*, 3852-6.

105. Yang, R., Pasunooti, K. K., Li, F., Liu, X. W., and Liu, C. F. (2010) Synthesis of K48-linked diubiquitin using dual native chemical ligation at lysine. *Chem Commun (Camb) 46*, 7199-201.

106. Kumar, K. S., Spasser, L., Erlich, L. A., Bavikar, S. N., and Brik, A. (2010) Total chemical synthesis of di-ubiquitin chains. *Angew Chem Int Ed Engl 49*, 9126-31.

107. Kumar, K. S., Bavikar, S. N., Spasser, L., Moyal, T., Ohayon, S., and Brik, A. (2011) Total chemical synthesis of a 304 amino acid K48-linked tetraubiquitin protein. *Angew Chem Int Ed Engl 50*, 6137-41.

108. Castaneda, C., Liu, J., Chaturvedi, A., Nowicka, U., Cropp, T. A., and Fushman, D. (2011) Nonenzymatic assembly of natural polyubiquitin chains of any linkage composition and isotopic labeling scheme. *J Am Chem Soc 133*, 17855-68.

109. Virdee, S., Kapadnis, P. B., Elliott, T., Lang, K., Madrzak, J., Nguyen, D. P., Riechmann, L., and Chin, J. W. (2011) Traceless and site-specific ubiquitination of recombinant proteins. *J Am Chem Soc 133*, 10708-11.

110. 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*, 1462-6.

111. Kavran, J. M., Gundllapalli, S., O'Donoghue, P., Englert, M., Soll, D., and Steitz, T. A. (2007) Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. *Proc Natl Acad Sci U S A 104*, 11268-73.

112. Blight, S. K., Larue, R. C., Mahapatra, A., Longstaff, D. G., Chang, E., Zhao, G., Kang, P. T., Green-Church, K. B., Chan, M. K., and Krzycki, J. A. (2004) Direct charging of tRNA$_{CUA}$ with pyrrolysine *in vitro* and *in vivo*. *Nature 431*, 333-5.

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

114. Polycarpo, C. R., Herring, S., Berube, A., Wood, J. L., Soll, D., and Ambrogelly, A. (2006) Pyrrolysine analogues as substrates for pyrrolysyl-tRNA synthetase. *FEBS Lett 580*, 6695-700.

115. Mukai, T., Kobayashi, T., Hino, N., Yanagisawa, T., Sakamoto, K., and Yokoyama, S. (2008) Adding L-lysine derivatives to the genetic code of

mammalian cells with engineered pyrrolysyl-tRNA synthetases. *Biochem Biophys Res Commun 371*, 818-22.

116. Yanagisawa, T., Ishii, R., Fukunaga, R., Kobayashi, T., Sakamoto, K., and Yokoyama, S. (2008) Multistep engineering of pyrrolysyl-tRNA synthetase to genetically encode N(epsilon)-(o-azidobenzyloxycarbonyl) lysine for site-specific protein modification. *Chem Biol 15*, 1187-97.

117. 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*, 2950-3.

118. Aimoto, S. (1999) Polypeptide synthesis by the thioester method. *Biopolymers 51*, 247-65.

119. El Oualid, F., Merkx, R., Ekkebus, R., Hameed, D. S., Smit, J. J., de Jong, A., Hilkmann, H., Sixma, T. K., and Ovaa, H. (2010) Chemical synthesis of ubiquitin, ubiquitin-based probes, and diubiquitin. *Angew Chem Int Ed Engl 49*, 10149-53.

120. Ryu, Y., and Schultz, P. G. (2006) Efficient incorporation of unnatural amino acids into proteins in *Escherichia coli*. *Nat Methods 3*, 263-5.

121. Chen, P. R., Groff, D., Guo, J., Ou, W., Cellitti, S., Geierstanger, B. H., and Schultz, P. G. (2009) A facile system for encoding unnatural amino acids in mammalian cells. *Angew Chem Int Ed Engl 48*, 4052-5.

122. Varadan, R., Walker, O., Pickart, C., and Fushman, D. (2002) Structural properties of polyubiquitin chains in solution. *J Mol Biol 324*, 637-47.

123. Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z., and Pickart, C. M. (1992) Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J Biol Chem 267*, 16467-71.

124. Ryabov, Y., and Fushman, D. (2006) Analysis of interdomain dynamics in a two-domain protein using residual dipolar couplings together with ^{15}N relaxation data. *Magn Reson Chem 44 Spec No*, S143-51.

125. Ryabov, Y., and Fushman, D. (2006) Interdomain mobility in di-ubiquitin revealed by NMR. *Proteins: Struct., Funct., Bioinf. 63*, 787-96.

126. Lam, Y. A., DeMartino, G. N., Pickart, C. M., and Cohen, R. E. (1997) Specificity of the ubiquitin isopeptidase in the PA700 regulatory complex of 26 S proteasomes. *J Biol Chem 272*, 28438-46.

127. Streu, C., and Meggers, E. (2006) Ruthenium-induced allylcarbamate cleavage in living cells. *Angew Chem Int Ed Engl 45*, 5645-8.

128. Xu, P., Duong, D. M., Seyfried, N. T., Cheng, D., Xie, Y., Robert, J., Rush, J., Hochstrasser, M., Finley, D., and Peng, J. (2009) Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell 137*, 133-45.

129. Wickliffe, K., Williamson, A., Jin, L., and Rape, M. (2009) The multiple layers of ubiquitin-dependent cell cycle control. *Chem Rev 109*, 1537-48.

130. Pervushin, K., Riek, R., Wider, G., and Wuthrich, K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A 94*, 12366-71.

131. Varadan, R., Assfalg, M., Haririnia, A., Raasi, S., Pickart, C., and Fushman, D. (2004) Solution conformation of Lys63-linked di-ubiquitin chain provides clues to functional diversity of polyubiquitin signaling. *J Biol Chem 279*, 7055-63.

132. Mao, L., Tang, Y., Vaiphei, S. T., Shimazu, T., Kim, S. G., Mani, R., Fakhoury, E., White, E., Montelione, G. T., and Inouye, M. (2009) Production of membrane proteins for NMR studies using the condensed single protein (cSPP) production system. *J Struct Funct Genomics 10*, 281-9.

133. Suzuki, M., Roy, R., Zheng, H., Woychik, N., and Inouye, M. (2006) Bacterial bioreactors for high yield production of recombinant protein. *J Biol Chem 281*, 37559-65.

134. Chin, J. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002) Addition of a photocrosslinking amino acid to the genetic code of Escherichiacoli. *Proc Natl Acad Sci U S A 99*, 11020-4.

135. David, Y., Ziv, T., Admon, A., and Navon, A. (2010) The E2 ubiquitin-conjugating enzymes direct polyubiquitination to preferred lysines. *J Biol Chem 285*, 8595-604.

136. Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science 228*, 1315-7.

137. Nakamura, Y., Gojobori, T., and Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res 28*, 292.

138. Wilkins, B. J., Marionni, S., Young, D. D., Liu, J., Wang, Y., Di Salvo, M. L., Deiters, A., and Cropp, T. A. (2010) Site-specific incorporation of fluorotyrosines into proteins in *Escherichia coli* by photochemical disguise. *Biochemistry 49*, 1557-9.

139. Zhang, Z., and Marshall, A. G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J Am Soc Mass Spectrom 9*, 225-33.

140. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR 6*, 277-93.

141. Lee, W., Westler, W. M., Bahrami, A., Eghbalnia, H. R., and Markley, J. L. (2009) PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. *Bioinformatics 25*, 2085-7.

142. Hanks, S. K., Quinn, A. M., and Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science 241*, 42-52.

143. Huttner, W. B. (1982) Sulphation of tyrosine residues-a widespread modification of proteins. *Nature 299*, 273-6.

144. Nara, S. J., Valgimigli, L., Pedulli, G. F., and Pratt, D. A. (2009) Tyrosine analogues for probing proton-coupled electron transfer processes in peptides and proteins. *J Am Chem Soc 132*, 863-72.

145. Pal, P. P., Bae, J. H., Azim, M. K., Hess, P., Friedrich, R., Huber, R., Moroder, L., and Budisa, N. (2005) Structural and spectral response of *Aequorea victoria* green fluorescent proteins to chromophore fluorination. *Biochemistry 44*, 3663-72.

146. Seyedsayamdost, M. R., Yee, C. S., and Stubbe, J. (2007) Site-specific incorporation of fluorotyrosines into the R2 subunit of *E. coli* ribonucleotide reductase by expressed protein ligation. *Nat Protoc 2*, 1225-35.

147. Adams, S. R., and Tsien, R. Y. (1993) Controlling cell chemistry with caged compounds. *Annu Rev Physiol 55*, 755-84.

148. Deiters, A., Groff, D., Ryu, Y., Xie, J., and Schultz, P. G. (2006) A genetically encoded photocaged tyrosine. *Angew Chem Int Ed Engl 45*, 2728-31.

149. Zhu, Y., Pavlos, C. M., Toscano, J. P., and Dore, T. M. (2006) 8-Bromo-7-hydroxyquinoline as a photoremovable protecting group for physiological use: mechanism and scope. *J Am Chem Soc 128*, 4267-76.

150. Momotake, A., Lindegger, N., Niggli, E., Barsotti, R. J., and Ellis-Davies, G. C. (2006) The nitrodibenzofuran chromophore: a new caging group for ultra-efficient photolysis in living cells. *Nat Methods 3*, 35-40.

151. Furuta, T., Wang, S. S., Dantzker, J. L., Dore, T. M., Bybee, W. J., Callaway, E. M., Denk, W., and Tsien, R. Y. (1999) Brominated 7-hydroxycoumarin-4-

ylmethyls: photolabile protecting groups with biologically useful cross-sections for two photon photolysis. *Proc Natl Acad Sci U S A 96*, 1193-200.

152.   Edwards, W. F., Young, D. D., and Deiters, A. (2009) Light-activated Cre recombinase as a tool for the spatial and temporal control of gene function in mammalian cells. *ACS Chem Biol 4*, 441-5.

153.   Nguyen, D. P., Lusic, H., Neumann, H., Kapadnis, P. B., Deiters, A., and Chin, J. W. (2009) Genetic encoding and labeling of aliphatic azides and alkynes in recombinant proteins via a pyrrolysyl-tRNA Synthetase/tRNA$_{CUA}$ pair and click chemistry. *J Am Chem Soc 131*, 8720-1.

154.   Chou, C., Young, D. D., and Deiters, A. (2009) A light-activated DNA polymerase. *Angew Chem Int Ed Engl 48*, 5950-3.

155.   Chou, C., Young, D. D., and Deiters, A. (2010) Photocaged T7 RNA polymerase for the light activation of transcription and gene function in pro- and eukaryotic cells. *Chembiochem 11*, 972-7.

156.   Gautier, A., Nguyen, D. P., Lusic, H., An, W., Deiters, A., and Chin, J. W. (2010) Genetically encoded photocontrol of protein localization in mammalian cells. *J Am Chem Soc 132*, 4086-8.

157.   Gautier, A., Deiters, A., and Chin, J. W. (2011) Light-activated kinases enable temporal dissection of signaling networks in living cells. *J Am Chem Soc 133*, 2124-7.

158.   Chou, C., and Deiters, A. (2011) Light-activated gene editing with a photocaged zinc-finger nuclease. *Angew Chem Int Ed Engl 50*, 6839-42.

159.   Lang, K., Davis, L., Torres-Kolbus, J., Chou, C., Deiters, A., and Chin, J. W. (2012) Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. *Nat Chem 4*, 298-304.

160.   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 Biol 10*, 425-32.

161.   Johansson, E., Parkinson, G. N., Denny, W. A., and Neidle, S. (2003) Studies on the nitroreductase prodrug-activating system. Crystal structures of complexes with the inhibitor dicoumarol and dinitrobenzamide prodrugs and of the enzyme active form. *J Med Chem 46*, 4009-20.

162.   Parkinson, G. N., Skelly, J. V., and Neidle, S. (2000) Crystal structure of FMN-dependent nitroreductase from *Escherichia coli* B: a prodrug-activating enzyme. *J Med Chem 43*, 3624-31.

163. Levy, M., and Ellington, A. D. (2008) Directed evolution of streptavidin variants using *in vitro* compartmentalization. *Chem Biol 15*, 979-89.

164. Liu, D. V., Zawada, J. F., and Swartz, J. R. (2005) Streamlining *Escherichia coli* S30 extract preparation for economical cell-free protein synthesis. *Biotechnol Prog 21*, 460-5.

165. Goerke, A. R., and Swartz, J. R. (2009) High-level cell-free synthesis yields of proteins containing site-specific non-natural amino acids. *Biotechnol Bioeng 102*, 400-16.

166. Deutscher, M. P., and Hilderman, R. H. (1974) Isolation and partial characterization of *Escherichia coli* mutants with low levels of transfer ribonucleic acid nucleotidyltransferase. *J Bacteriol 118*, 621-7.

167. Davidson, E. A., Dlugosz, P. J., Levy, M., and Ellington, A. D. (2009) Directed evolution of proteins *in vitro* using compartmentalization in emulsions. *Curr Protoc Mol Biol Chapter 24*, Unit 24 6.

168. Miao, Z., Liu, J., Norman, T., and Driver, R. (2006) New non-natural amino acids useful for preparation of modified or non-modified, non-natural amino acid polypeptides, useful in e.g. treatment of a disorder, condition or disease. *PCT Int Appl* WO2006069246 Jun 29, 2006.

169. Mukai, C., Kobayashi, M., Kubota, S., Takahashi, Y., and Kitagaki, S. (2004) Construction of azacycles based on endo-mode cyclization of allenes. *J Org Chem 69*, 2128-36.

Curriculum Vitae
Jia Liu

Education

- Ph.D candidate in Biochemistry, University of Maryland, College Park
  Expected graduation date: May 2012
- B.S. in Biological Sciences, July 2006, Nankai University, China

Research experience

- September 2007–present, unnatural amino acid technologies, DNA mutagenesis methods, with Dr. Ashton Cropp
- May 2005–August 2007, DNA mutagenesis methods, mechanistic studies of *Arabidopsis thaliana* glutathione transferase zeta class, with Dr. Defu Chen

Publications

Liu, J. and Cropp, T. A. (2012) Experimental methods for scanning unnatural amino acid mutagenesis, in *Methods in Molecular Biology* (Pollegioni, L. and Servi, S., Eds.), *794,* pp. 187-197, Humana Press, Totowa.

Liu, J. and Cropp, T. A. (2012). A method for multi-codon scanning mutagenesis of proteins based on asymmetric transposons. *Protein Eng. Des. Sel. 25*, 67-72.

Castaneda, C., **Liu, J.**, Chaturvedi, A., Urszula, N., Cropp, T. A. and Fushman, D. (2011) Nonenzymatic assembly of natural polyubiquitin chains of any linkage composition and isotopic labeling scheme. *J. Am. Chem. Soc.*, *133,* 17855-17868.

Castaneda, C. A., **Liu, J.**, Kashyap, T. R., Singh, R. K., Fushman, D. and Cropp, T. A. (2011) Controlled enzymatic synthesis of natural-linkage, defined-length polyubiquitin chains using lysines with removable protecting groups. *Chem. Commun. 47*, 2026-2028.

Liu, J., Castaneda, 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.

Wilkins, B. J., Marionni, S., Young, D. D., **Liu, J.**, Wang, Y., Di Salvo, M. L., Deiters, A. and Cropp, T. A. (2010) Site-specific incorporation of fluorotyrosines into proteins in *Escherichia coli* by photochemical disguise. *Biochemistry 49*, 1557-1559.

Chen, D., **Liu, J.**, Liu, J. and Chen, X. (2010) Evolution of the catalytic activity of *Arabidopsis thaliana* glutathione transferase zeta class-1 by saturation mutagenesis. *Biosci. Biotechnol. Biochem. 74*, 1458-1461.

Chen, X., **Liu, J.**, Yang, P. and Chen, D. (2010) Identifying functional residues in *Arabidopsis thaliana* zeta class glutathione *S*-transferase through screening inactive point mutants. *Biochemistry (Moscow) 75*, 110-114.

Tao, S., Chen, X., **Liu, J.**, Ming, M., Chong, N. and Chen, D. (2008) Characterization of Ser73 in *Arabidopsis thaliana* glutathione *S*-transferase zeta class. *J. Genet. Genomics 35*, 507-512.

Meeting Presentations

Liu, J. and Cropp, T. A. (2010) Codon scanning mutagenesis by asymmetric transposons. In *240th ACS National Meeting, Boston, MA, August 22-26, 2010*.

Liu, J. and Cropp, T. A. (2010) Facile codon deletion mutagenesis using asymmetric transposons. In *3rd Frontiers at the Chemistry and Biology Interface, College Park, MD, May 1, 2010.*

Liu, J. and Cropp, T. A. (2009) Identification and characterization of fluorescent GFP mutants with internal deletions. In *238th ACS National Meeting, Washington, DC, August 16-20, 2009.*