Proteome-wide amino acid substitution and biochemical selection to understand protein biology

Bianca Ynez Ruiz

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Judit Villén, Chair
Stanley Fields, Chair
James Carothers

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Proteome-wide amino acid substitution and biochemical selection to understand protein biology

Bianca Ynez Ruiz

Chairs of the Supervisory Committee:
Judit Villén and Stanley Fields
Department of Genome Sciences

Proteins are important macromolecules that carry out most biological functions and are enabled to do so by the diverse chemistries of the twenty amino acids. Changes in the amino acid sequence of a protein can change structure, function, and other biochemical properties, sometimes resulting in disease. While genome sequencing has identified numerous mutations that cause changes to the encoded amino acid sequence of proteins, our understanding of the functional consequences is limited. During my thesis work I have contributed to the development and application of mass spectrometry-based methods for understanding the effects of amino acid substitutions at the proteome scale. The work I describe here includes: 1) a screen of noncanonical amino acids to generate random amino acid substitutions across the *E. coli* proteome for use in functional studies; 2) a high throughput selection method for understanding the impact of pH on protein solubility; and 3) a proteome-wide study of the impact of amino acid substitutions on the critical post-

translational modification of phosphorylation. Taken together, these projects provide high throughput methods for directly measuring proteins and understanding how amino acid substitutions impact their own biochemical properties and more broadly, the proteome.

TABLE OF CONTENTS

Chapter 1. Introduction 1

1.1 Proteins and the consequences of amino acid substitutions 1

1.2 Advances in DNA sequencing and detection of missense mutations 4

1.3 Functional assays available for proteins 6

1.4 Advances in mass spectrometry to study proteins 8

Chapter 2. High throughput screening of noncanonical amino acids in *E. coli* 10

2.1 Abstract 10

2.2 Introduction 12

2.3 Results 17

2.3.1 Toxicity of noncanonical amino acids 18

2.3.2 Incorporation of noncanonical amino acids 19

2.4 Discussion 22

2.5 Methods 24

2.5.1 Toxicity assays 24

2.5.2 MS sample preparation 24

2.5.3 MS methods 25

2.5.4 MS data analysis 25

2.6 Contributions 27

Chapter 3. Chemical selection to understand protein solubility 28

3.1 Abstract 28
3.2 Introduction 29
3.3 Results 31
3.3.1 Construction of solubility profiles by spline fitting 33
3.3.2 Proteins from different subcellular localizations tolerate different pH ranges 35
3.3.3 Principal component analysis of pH selection data 38
3.4 Discussion 40
3.5 Methods 42
3.5.1 Yeast strains and growth 42
3.5.2 Polybuffer system 42
3.5.3 Solubility selection across pH gradient 42
3.5.4 MS sample preparation 43
3.5.5 MS methods 44
3.5.6 MS data analysis 45
3.6 Contributions 46
Chapter 4. Interactions between amino acid substitutions and the phosphoproteome 47
4.1 Abstract 47
4.2 Introduction 48
4.3 Results 51
4.3.1 Growth and mistranslation rates 52
4.3.2 Phosphoproteome data quality and reproducibility 56
4.3.3 Mistranslation-driven changes in abundance of phosphopeptides 58
4.3.4 GO analysis of regulated phosphoproteins 61

4.3.5 Changes in phosphorylation of known motifs 64
4.4 Discussion 68
4.5 Methods 71
4.5.1 Yeast strains and growth 71
4.5.2 DNA constructs 71
4.5.3 MS sample preparation 71
4.5.4 MS methods 72
4.5.5 MS data analysis 73
4.6 Contributions 74
Chapter 5. Next steps for high-throughput studies of proteins 75
5.1 Generating protein variants with noncanonical amino acids 75
5.2 Functional selections for high-throughput studies of proteins 76
5.3 Developing biological selections to understand amino acid substitutions in proteins. 79
5.4 Closing remarks 81
Funding 82
Bibliography 83
Appendix A 86
Appendix B 89
Appendix C 94

ACKNOWLEDGEMENTS

The completion of my dissertation would not have been possible without the support of many people in my life.

I would first like to thank my thesis committee for dedicating their time and expertise to advise me on my science, as well as my career goals. Thanks to my advisors, Judit Villén and Stanley Fields, for forming a co-mentorship in order to support my research in both of their labs. Both of my advisors are creative and profoundly knowledgeable scientists, and I feel very fortunate to have had their guidance for my research goals. I am grateful to Judit for reminding me to pursue the science that made me feel the most excited above all else, and I hope to take this reminder with me wherever I do research.

I want to express my deepest gratitude to Stan, for his unwavering support throughout all of the changes I underwent both professionally and personally during graduate school. In addition to providing invaluable scientific insights to my research, Stan has been an incredibly caring and open-minded mentor. He always made the time to talk with me about anything that was on my mind, boldly advocated for my best interests, and challenged me to be the best scientist I could be. I feel extremely fortunate that life's timing worked out so that I had the chance to be Stan's last graduate student.

Throughout graduate school, I have had the opportunity to work alongside many kind and brilliant people in the Genome Sciences Department. A special thanks to members of the Villén and Fields Labs, particularly Stephanie Zimmerman and Ricard Rodriguez, who both provided guidance and support in the early days when I was still finding my footing. I would also like to thank Ian Smith, Matt Berg, and Anthony Barente for their collaboration and support for my research projects.

I am also deeply grateful to my family for being by my side at every step of the way. To my mother and father, thank you for unconditionally loving and supporting me throughout my entire life, and for everything you have done to make my dreams and my brother's dreams possible. To my partner, Ian, thank you for holding my hand and for having the kindest heart I have ever known. Thanks also, to our puppy Clifford, who makes us smile every single day.

Chapter 1. INTRODUCTION

1.1 PROTEINS AND THE CONSEQUENCES OF AMINO ACID SUBSTITUTIONS

Proteins are diverse molecular machines that are involved in virtually all biological structures and processes. They execute this almost infinite array of functions through the chemistry of the twenty amino acids. Given that the sequence of amino acids dictates the capabilities of a protein, alterations to this sequence can have major implications that manifest at the molecular, cellular, and organismal level. Changes in the amino acid sequence of a protein can arise by many mechanisms, ranging from missense mutations in protein-coding DNA to errors during transcription or translation. The insertions, truncations, or substitutions in amino acid sequence subsequent to such errors can damage, enhance, or have no effect on the activity of a protein. The ability to predict these consequences remains one of the most important goals in molecular biology and could revolutionize our understanding and treatment of numerous human diseases.

The causal relationship between changes in amino acid sequence and disease has long been established, yet new examples continually surface today. One of the best studied examples of this relationship is sickle cell disease. Sickle cell disease was first described in the literature in 1910 by physician and professor James Bryan Herrick, who reported abnormally shaped red blood cells in a patient suffering from severe anemia (Herrick, 1910). In the next few decades, scientists would make a series of discoveries based on observation of patient symptoms as well as the biochemical properties of the hemoglobin protein. This trail of findings would ultimately lead to a profound new concept by Vernon Ingram, who demonstrated that the difference between sickle hemoglobin and normal hemoglobin came down to the substitution of a single amino acid (Ingram, 1958).

Another well-known protein that illustrates disease-causing amino acid substitutions is breast cancer type 1 susceptibility protein (BRCA1). Since the breakthrough discovery of the

linkage between BRCA1 and familial breast cancer by the laboratory of Mary-Claire King in 1990 (Hall *et al.*, 1990), the scientific community has launched a concerted effort to catalogue pathogenic mutations within the protein. Many of the disease-causing mutations identified have been found to result in amino acid substitutions that fundamentally change the structural and functional properties of this protein (Tischkowitz *et al.*, 2008; Starita *et al.*, 2015; Fernandes *et al.*, 2019). Today, the link between BRCA1 mutation and development of breast and ovarian cancers is so well-established that doctors often recommend prophylactic surgery to women who are carriers as confirmed by genetic screening. While these surgeries are a relatively drastic measure, they are truly lifesaving and can significantly lower the risk of cancer and mortality (Li *et al.*, 2016). BRCA1 provides a powerful example of how the prevention, diagnosis, and treatment of life-threatening diseases is contingent upon our deep understanding of amino acid substitutions and their consequences.

A current example of the importance of amino acid substitutions in disease is the COVID-19 pandemic. In response to the pandemic, scientists all over the world have come together in an unprecedented effort to understand transmission and symptoms of the virus, and ultimately to develop vaccines to prevent the spread of this highly contagious infection. Most of the vaccines being globally distributed today initiate an immune response that targets the SARS-CoV-2 spike protein, which binds receptors on the host cell to mediate entry and infection (Dai and Gao, 2021). In 2020, researchers identified the spike protein substitution D614G, a variant that has been of great concern, as it significantly increased viral transmission and quickly dominated the pandemic (Shi *et al.*, 2020). Fortunately, this substitution arose early enough in the vaccine development process that it was accounted for, but researchers remain vigilant of amino acid substitutions that increase infectivity, worsen symptoms, or escape existing vaccines. Given the importance of amino

acid substitutions in our understanding of proteins and disease, it is more urgent than ever that we quickly identify and screen function of the missense mutations that cause these substitutions.

1.2 Advances in DNA sequencing and detection of missense mutations

Recent advances in DNA sequencing have revolutionized our ability to sequence genes and genomes and amassed unprecedented volumes of biological data. Next generation sequencing (NGS), in short, encompasses a set of technologies that enable the parallel sequencing of millions of DNA fragments in a single experiment. Today, NGS technologies can sequence the entire human genome in a single day, a feat that originally took thirteen years to complete by Sanger Sequencing in the Human Genome Project. This exponential increase in the speed of DNA sequencing has identified millions of mutations in humans, many of which alter amino acids in proteins. However, the potential for clinical applications of this information is limited by current methods of determining the functional consequences of missense mutations at the protein level.

To address this bottleneck, a technology called Deep Mutational Scanning (DMS) was developed by Douglas Fowler in the laboratory of Stanley Fields (Fowler *et al.*, 2010). The DMS approach harnesses NGS capabilities to annotate the functional consequences of all amino acid substitutions in a protein in parallel. The experiments are typically carried out by synthesizing a DNA library encoding the protein of interest with all possible single amino acid substitutions. The library is then used to express the protein variants, either in cell-based or *in vitro* systems, and a functional selection is applied. The genetic library is sequenced in parallel before selection and after selection, and these data are used to calculate the comparative frequency of each variant. These values are then used to determine whether each variant was enriched, depleted or unchanged in frequency after the given functional selection. An enrichment would indicate that the amino acid substitution enhanced function, no change would indicate that the substitution was neutral, and a depletion would indicate that the substitution caused a loss of function. DMS technologies have dramatically increased the speed with which we can annotate the function of amino acid

substitutions, by collapsing the experimental time required by previous approaches with the same goal, where variants were assayed one by one, such as site-directed mutagenesis and alanine-scanning.

1.3 FUNCTIONAL ASSAYS AVAILABLE FOR PROTEINS

The diversity of structures and functions of proteins has driven the development of equally diverse assays to probe the biological and biochemical properties of these molecules. Cell-based methods for reading out protein properties include viability assays, fluorescence-based assays, and two-hybrid screening. In a cell viability assay, the function of a protein can be indirectly read out by a change in cell growth, sometimes in the presence of a drug or stress condition (Handford *et al.*, 2009). Fluorescence-based assays can be used to read out subcellular localization of a protein by microscopy, or to quantify protein abundance by flow sorting (Gohar *et al.*, 2013). Yeast two-hybrid assays have been used to elucidate protein interactions with other proteins, as well as RNA and DNA (Fields and Song, 1989). Immunological methods are yet another option for studying the properties of proteins and have many unique applications, such as antibody-based assays to define kinase-substrate interactions (Edbauer *et al.*, 2009). Additionally, NMR and X-ray crystallography have revealed the structure of proteins and their complexes at atomic resolution (Saunders, Wishnia and Kirkwood, 1957). These methods represent just a handful of examples in a sea of versatile options for studying protein function.

While most applications of these methods for studying protein variants were initially low throughput, many have been used in deep mutational scanning to screen thousands of variants in a single experiment. One example is the VAMP-seq method, which was developed to quantify cellular abundance of thousands of protein variants by coupling to flow cytometry (Matreyek *et al.*, 2018). However, not all selections of protein properties are compatible with deep mutational scanning technologies. While DMS is applicable to a wide range of cell-based assays, these experiments are not as conducive to identifying substitutions that may result in a change in the localization of a protein. DMS experiments are also limited in the context of proteins that do not

have a well-established molecular function, or proteins with extracellular function (Fowler and Fields, 2014). Outside of the limitations with functional selections, one of the most prominent disadvantages to DMS is the indirectness of readout. This technology extrapolates DNA sequencing data to understand changes in the molecular function of a protein, creating the need for more direct measurements of proteins.

1.4 ADVANCES IN MASS SPECTROMETRY TO STUDY PROTEINS

Mass spectrometry (MS) has been an excellent tool for measuring proteins, and the field is advancing all the time. While mass spectrometry itself has been around since the early nineteenth century, the technology was not applied to proteins until late in the twentieth century, when electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) technologies were developed. In 1984, the introduction of these ionization techniques coupled with MS allowed for the analysis of biological macromolecules (Yamashita, 1984). These techniques were so powerful to our understanding of biology, that John Bennett Fenn and Koichi Tanaka were awarded the 2002 Nobel Prize in Chemistry for developing the methods.

In the following years, these same methods were applied to analyzing purified proteins or entire proteomes in a single experiment. Top-down proteomics methods encompass experiments that use mass spectrometry to analyze intact proteins, and to capture thier various proteoforms with post-translational modifications or other amino acid sequence variations (Sze *et al.*, 2002). In addition to top-down approaches, bottom-up approaches enable the analysis of entire proteomes, by digesting proteins into small peptide fragments that are more amenable to MS-based sequencing (Chait, 2006). Bottom-up approaches have been greatly aided by the availability of genome sequencing, which preemptively tells researchers what protein sequences will possibly be identified in various organisms.

Over the last few decades, MS-based proteomics technologies have significantly improved in throughput and quantitative capabilities. Stable isotope labeling by amino acids in cell culture (SILAC) and isobaric tags such as tandem mass tags (TMT) have enabled the quantitative measurement of samples representing multiple replicates or biological conditions in a single MS run (Ong *et al.*, 2002; Werner *et al.*, 2014). Many MS-based approaches also enable the probing

of protein functions at the proteome scale, such as chemical cross-linking with mass spectrometry (XL-MS) and thermal proteome profiling (TPP). XL-MS technologies have enabled proteome-wide identification of protein-protein interactions as well as protein structures and conformations (Mintseris and Gygi, 2020; Chavez *et al.*, 2021). TPP is another powerful MS-based approach that uses the property of thermal stability to characterize proteins and identify drug targets in an un-biased fashion, at the proteome scale (Savitski *et al.*, 2014; Franken *et al.*, 2015).

We can also use MS to measure proteoforms that deviate from existing protein sequence databases, such as proteins with post-translational modifications (PTMs) or amino acid substitutions, at the proteome scale. MS-based studies on post-translational modification of proteins have enabled the mapping of specific PTM sites across the proteome, and deepened our understanding of the molecular roles PTMs play (Sharma *et al.*, 2014; Smith *et al.*, 2021). There have also been applications to studying amino acid substitutions across the proteome, such as with noncanonical amino acids (ncAAs) (Saleh *et al.*, 2019). While these approaches have demonstrated MS-based capabilities to analyze and quantify protein variants across the proteome, to understand the functions of these protein variants a selection must be applied.

The focus of my thesis was to develop MS-based methods for the high throughput study of protein variants. My projects contributed to the development of these methods by designing and characterizing cell-based generation of protein variants with amino acid substitution via ncAA incorporation, as discussed in Chapter 2. In Chapter 3, I describe a project that combines TMT labeling with a chemical selection applicable to the entire proteome, to understand the behavior of proteins in various pH conditions. In Chapter 4, I explore the effects of amino acid substitutions on the cellular phosphoproteome, in order to learn how substitutions change signaling and recognition by kinases.

Chapter 2. HIGH THROUGHPUT SCREENING OF NONCANONICAL AMINO ACIDS IN *E. COLI*

2.1 ABSTRACT

Amino acid substitutions can change the structure and function of a protein and are the basis of many human diseases. Artificially introduced substitutions facilitated by incorporation of noncanonical amino acids (ncAAs) provide an avenue for understanding the impact of amino acid substitution on proteins and expanding protein biochemistry. Here I describe a pilot study using cell growth and mass spectrometry (MS) assays to screen twenty-two ncAAs for their capacity to produce protein variants by random incorporation across the *E. coli* proteome. I observed a wide range of inhibitory effects on growth, with most compounds causing a clear dose-dependent toxicity, and five compounds (A1, A2, L3, M1 and M2) that did not cause an observable growth defect. Data from my MS assays identified eighteen compounds that incorporate across the proteome at levels above background, facilitating substitutions that span the replacement of twelve of the canonical amino acids. Interestingly, I observed that four compounds did not incorporate into proteins, and that three of these compounds were characterized by the addition of a ring structure to a small hydrophobic side chain. Another trend from the proteome data was that the compounds that were fluorinated derivatives of the aromatic side chains incorporated at exceedingly high levels, replacing up to 40% to 60% of the target amino acid. These trends in the chemical makeup of the ncAAs can inform interesting hypotheses about why certain compounds are recognized and incorporated by endogenous translational machinery, while others are not. The dataset also confirmed the applicability of seven compounds (E1, F3, L2, M3, R1, R2, and V3) to high throughput studies of protein variants, with ideal incorporation rates between approximately 5% and 10%. These compounds targeted the replacement of a diverse group of canonical amino

acids, including charged, aromatic, and hydrophobic residues. Outside of my own research, the combined dataset resulting from this project provides a comprehensive resource to inform other experiments utilizing these ncAAs.

2.2 INTRODUCTION

Proteins are involved in all cellular processes and structures and carry out a broad array of functions through complex chemistries provided by their building blocks, the amino acids. When amino acid substitutions occur, protein structure and function can undergo significant changes. For decades, the functional consequences of substituting protein residues have largely been explored via tools in DNA mutagenesis and recombinant protein expression in *E. coli* (Nicaud *et al.*, 1986; Feng *et al.*, 2000; Rosano *et al.*, 2014). While these approaches are still highly useful today, they are significantly limited in throughput, with many experiments spanning a single amino acid substitution within one protein of interest (Kim *et al.*, 2020). In recent years, deep mutational scanning (DMS) methods addressed this limitation by enabling the functional characterization of all possible amino acid substitutions in a protein within a single experiment (Fowler and Fields, 2014; Fowler *et al.*, 2014). However, the DMS approach typically focuses on only one protein per experiment and relies on an indirect readout by extrapolating DNA sequencing data to protein function.

As an alternative approach, the Villén Lab has developed a technology called Miró (Villén *et al.*, 2017), which rather than examining all types of amino acid substitution within a single protein, assays one type of amino acid substitution across many proteins at the proteome scale. A Miró experiment is performed by first facilitating random proteome-wide amino acid substitutions to generate protein variants, and then applying a functional selection to the substituted proteome. The selection is read out by MS-based quantification of peptides containing the substitution and their counterpart wildtype peptides. The relative abundance of these peptide pairs is used to calculate a ratio, a value that can then be compared before and after the given selection to quantify any changes in function caused by the amino acid substitution (Figure 2.1).

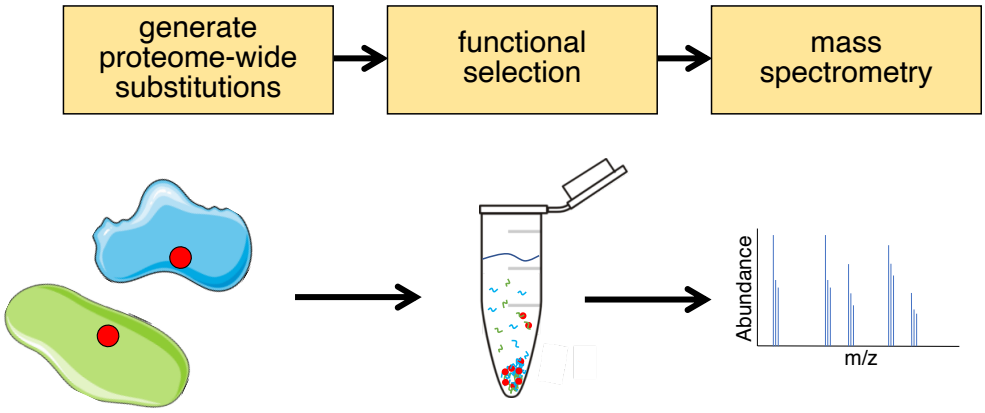


Figure 2.1 Workflow of the Miró technology for carrying out high throughput functional selections on protein variants. The general protocol for applying this technology is shown. First, amino acid substitutions are randomly introduced across the proteome. The substituted proteome is then subjected to a functional selection to identify changes in protein properties that occurred as a result of the amino acid substitutions. Finally, the functional selection is read out by MS analysis, wherein substituted peptides and wildtype peptides can be compared.

In an effort to develop the Miró technology, we first needed to design methods for generating one type of amino acid substitution randomly across the proteome by perturbing translation, leading us to explore noncanonical amino acids as a tool. Noncanonical amino acids are chemically divergent from the twenty canonical amino acids, with unique chemical properties and functional groups (Figure 2.2). Often, due in part to a lack of selectivity by tRNA synthetases, ncAAs can incorporate into proteins via endogenous translational machinery when cells are cultured in a medium containing the ncAA, without any genetic perturbations made to the system. For our purposes, ncAAs provided a straightforward way to produce amino acid substitutions across the proteome to examine the impact of substitutions in proteins at scale.

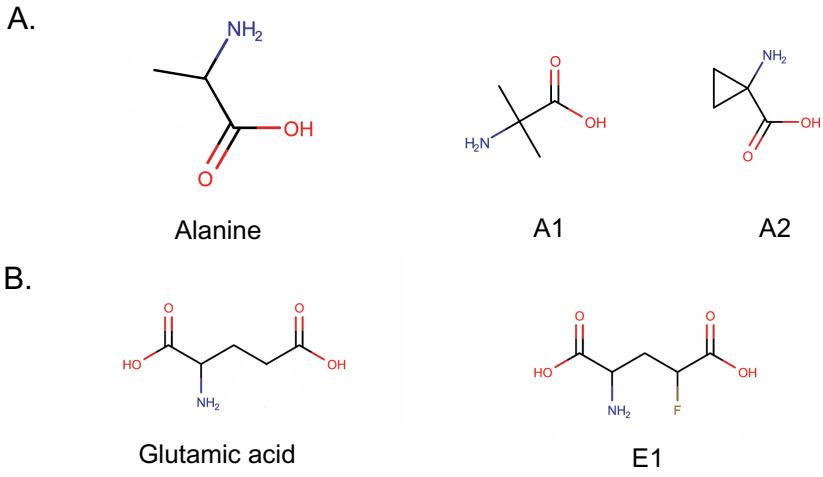


Figure 2.2 Noncanonical amino acids are chemically modified derivatives of the twenty canonical amino acids. Canonical amino acids are shown with the ncAAs that replace them. Our naming system for the ncAAs is the single letter abbreviation of the canonical amino acid followed by a number. A. A1 and A2 are two ncAAs that both replace alanine. Canonical alanine structure is shown on the left. A methylated derivative of alanine, A1 (2-amino-2-methylpropanoic acid) and a cyclic derivative, A2 (1-aminocyclopropane-1-carboxylic acid) are shown on the right. B. Canonical glutamic acid structure is shown on the left, a fluorinated derivative of glutamic acid, E1 (2-amino-4-fluoropentanedioic acid) is show on the right. See Supplementary Figure 2.1 for the structures of all ncAAs used in this study.

In many cases, ncAAs are synthesized for use in chemical biology to amplify our understanding of proteins. However, we have also observed many ncAAs to occur in nature, at times as precursors to the twenty canonical amino acids or as products of related biosynthetic pathways (Reitz *et al.*, 2018). Additionally, the presence of ncAAs in nature has been observed as a protective mechanism in certain plant species. Canavanine, a non-proteinogenic amino acid

derivative of arginine (see ncAA R1 in Supplementary Figure 2.1), has been detected in the seeds of the flowering plant *Hedysarum alpinum* (Krakauer *et al.*, 2015), and acts as both an antimicrobial and insecticidal agent (Dahlman *et al.*, 1975). Similarly, a non-proteinogenic derivative of proline called azetidine (see ncAA P1 in Supplementary Figure 2.1) is produced by the poisonous woodland plant, *Convallaria majalis* (Lily of the valley) (Fowden, 1956), as well as the garden beet, *Beta vulgaris* (Rubenstein *et al.*, 2006). This well-studied compound is thought to protect these plants by poisoning predatory herbivores (Samardzic *et al.*, 2019). Both azetidine and canavanine confer this protective toxicity in plants by mimicking proline or arginine, respectively, incorporating into the proteins of predatory organisms, and disrupting protein structure or function.

Several ncAAs have been demonstrated in the literature to incorporate into *E. coli* proteins. Norleucine, an analog of methionine that does not contain sulfur (also called 2-aminohexanoic acid; see ncAA M2 in Supplementary Figure 2.1), has been shown to replace 30% to 40% of naturally occurring methionine residues in *E. coli* proteins. The same study also showed that a fluorinated derivative of phenylalanine, *p*-fluorophenylalanine (also called (2S)-amino-3-(4-fluorophenyl)propanoic acid; see ncAA F2 in Supplementary Figure 2.1), replaces approximately 50% of naturally occurring phenylalanine residues (Cowei *et al.*, 1959). Another example of a fluorinated aromatic amino acid, fluorotryptophan (also called 2-amino-3-(5-fluoro-1H-indol-3-yl)propanoic acid; see ncAA W1 in Supplementary Figure 2.1), has been shown to replace up to 50% to 60% of tryptophan residues in *E. coli* (Browne *et al.*, 1970; Pratt *et al.*, 1975).

These early studies applying ncAAs to examine amino acid substitution, and many others, comprise a fascinating section of literature in protein biochemistry. Our deeper understanding of fundamental cellular processes, such as translation, has been developed using ncAAs. Many facets

of protein synthesis in the cell have been elucidated through ncAA perturbation, such as the molecular mechanisms of amino acid recognition by tRNA-synthetases and the general selectivity of the tRNA aminoacylation reaction (Old *et al.*, 1977). Historically, ncAA incorporation has also been used to explore the biochemical properties of enzymes, through the quantification of both deleterious and enhancing impacts of ncAA substitution on structure and catalytic activity (Gilles *et al.*, 1988).

With the Miró technology, we aimed to expand the use of ncAAs to probe protein structure or function to the proteome scale, and first needed to identify a subset of compounds that would be ideal for this application. To this end, I screened a panel of twenty-two ncAAs in *E. coli* for growth inhibition and proteome-wide incorporation rates. Generally, I was interested in identifying compounds that caused a moderate dose-dependent toxicity during growth, as this result would be a preliminary indicator that proteins were being disrupted by ncAA substitution. With regard to proteome-wide incorporation levels, I was seeking to identify compounds that replaced approximately 5% to 10% of their target residues. We estimated that this range of incorporation would be high enough to facilitate substitution in a wide array of proteins across the proteome, and low enough to not cause more than one substitution per protein, as this would complicate downstream analyses of functional selections. Here, I summarize my findings on the panel of ncAAs and their impact on both toxicity and proteome-wide incorporation in *E. coli*.

2.3 RESULTS

Within this project, I characterized a set of twenty-two ncAAs (Supplementary Figure 2.1) with regard to their utility in the first step of Miró: creating proteome-wide amino acid substitutions. I applied a two-tier screen to assess how each ncAA impacted growth and quantified the capacity of each compound to generate proteome-wide substitutions in *E. coli* (Figure 2.3). I hypothesized that this set of ncAAs would incorporate via endogenous translational machinery, replacing the natural amino acid the given ncAA most resembled structurally, and that we would be able to detect and quantify these substitutions by mass spectrometry. The naming system I used specified the natural amino acid being replaced along with a numeric identifier. For example, the set of ncAAs I tested included two ncAAs that were closely related to alanine and were thus expected to replace alanine residues; these compounds were denoted as "A1" and "A2" (Figure 2.2).

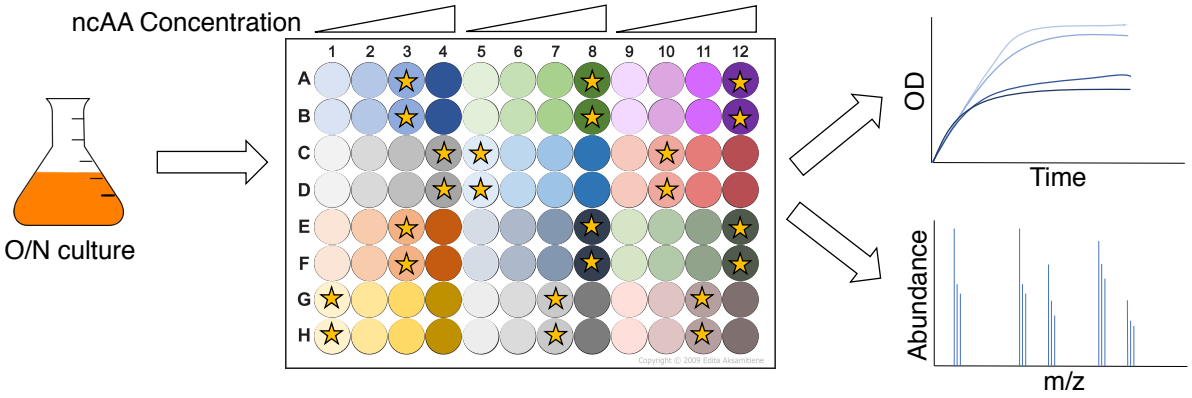


Figure 2.3 Workflow schematic of two-tier screen of ncAAs in *E. coli*. An overnight culture of *E. coli* was diluted into a 96-well plate containing minimal media with varying concentrations of ncAA, and two replicates per concentration. Colors on the 96-well plate indicate different ncAAs, and intensity of the colors represents increasing concentration. Based on results of growth curves, one concentration of each ncAA was chosen for incorporation analysis by MS, denoted by a star.

For the incorporation assay, *E. coli* was grown in the chosen concentrations, and cultures were harvested and prepared for MS analysis.

2.3.1 *Toxicity of noncanonical amino acids*

Among the set of ncAAs tested, I observed a wide range of effects on growth. Certain ncAAs such as F2 ((2S)-2-amino-3-(4-fluorophenyl)propanoic acid), R1 ((2S)-2-amino-4-([(diaminomethylidene)amino]oxy)butanoic acid), and P1 ((2S)-azetidine-2-carboxylic acid), displayed a strong dose-dependent inhibition of growth (Figure 2.4), and were considered strong candidates for proteome incorporation. Proline analog P1 was one of the most toxic compounds, along with leucine analog L1 (2-amino-3-(dimethylamino)propanoic acid) and tryptophan analog W1 (2-amino-3-(5-fluoro-1H-indol-3-yl)propanoic acid) (Supplementary Figure 2.2). Some examples of the least toxic ncAAs were the methionine analog M2 (2-aminohexanoic acid) (Figure 2.4), the alanine analog A2 (1-aminocyclopropane-1-carboxylic acid), and the leucine analog L3 ((2S)-2-amino-3-cyclopentylporpanoic acid). In the cases of A2 and L3, our MS data (as detailed below) showed that these ncAAs were not incorporating into the proteome, explaining their lack of toxicity. However, M2 was found to incorporate at significant levels while showing minimal toxicity, perhaps because the ncAA closely resembles methionine and is not causing severe damage to proteins. There was also a subset of ncAAs, such as isoleucine analog I1 ((2R)-2-amino-2-cyclohexylacetic acid) and proline analog P2 (1,3-thiazolidine-2-carboxylic acid), that slightly impaired growth, but not in a dose-dependent manner (Supplementary Figure 2.2).

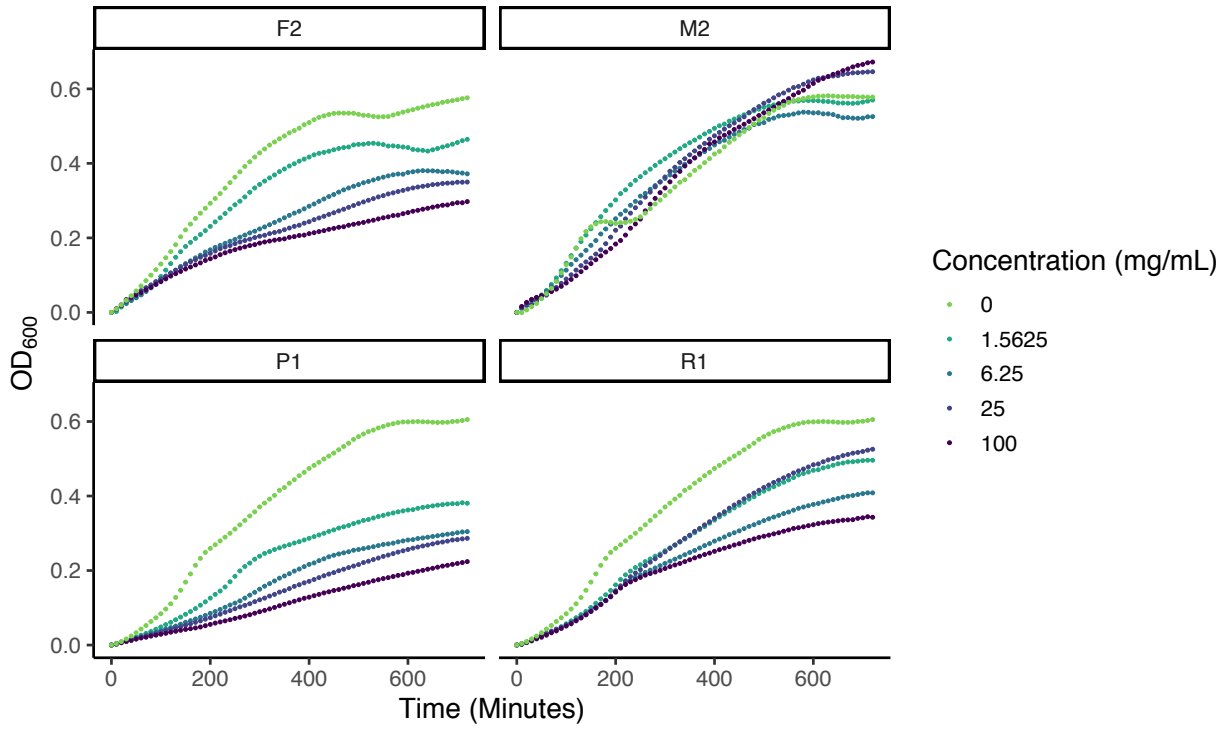


Figure 2.4 Selected growth curves from *E. coli* toxicity assays with ncAA treatment. Overnight culture of *E. coli* was diluted into a 96-well plate containing minimal media with various concentrations of ncAAs, and two replicate wells per condition. Cultures were incubated at 37ºC with agitation on a BioTek plate reader, and OD_{600} was measured every 15 minutes. OD_{600} measurements of two replicates per ncAA concentration were averaged and plotted against time. Plot title indicates ncAA, color of curves indicate concentration. See all growth curves in Supplementary Figure 2.2.

2.3.2 *Incorporation of noncanonical amino acids*

Several of the ncAAs from our panel were able to incorporate into the *E. coli* proteome at levels detectable by mass spectrometry. Comparatively, fluorinated derivatives targeting the aromatic amino acids (F2, W1, and Y1) incorporated at the highest levels, with Y1 ((2S)-2-amino-

3-(3-fluoro-4-hydroxyphenyl)propanoic acid) incorporating at the highest rate of over 60% (Figure 2.5). Compounds that incorporated at moderately high rates included M1 (2-amino-4-ethylsulfanylbutanoic acid), a methylated derivative of methionine, and P1 ((2S)-azetidine-2-carboxylic acid), which replaces the 5-membered ring with a 4-membered ring in proline. The amino acids that did not incorporate well, with very low rates near background levels, included alanine analog A2 (1-aminocyclopropane-1-carboxylic acid), isoleucine analog I1 ((2R)-2-amino-2-cyclohexylacetic acid), leucine analog L3 ((2S)-2-amino-3-cyclopentylpropanoic acid) and glutamine analog Q1 ((2S)-2-amino-5-(ethylamino)-5-oxopentanoic acid).

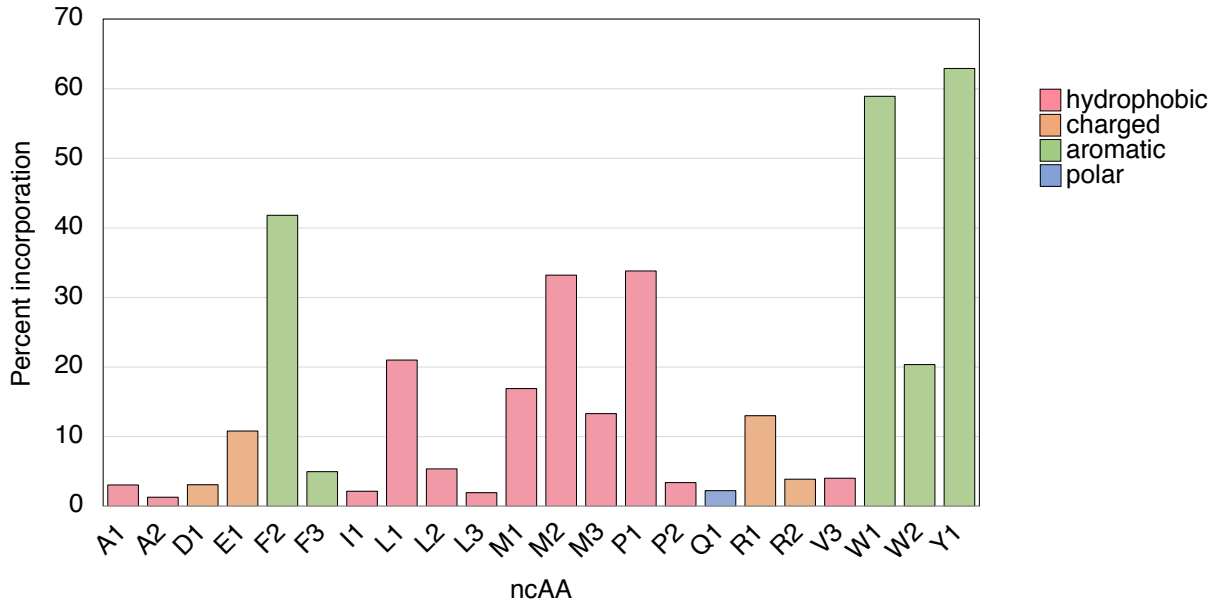


Figure 2.5 Proteome-wide incorporation of noncanonical amino acids in *E. coli*. Percent incorporation was determined for each ncAA, by filtering whole proteome data to unique peptide identifications containing the target amino acid and calculating a percentage of identifications containing the ncAA substitution. Bars are colored according to the chemical properties of the canonical amino acid being targeted.

In total, my screen identified eighteen ncAAs targeting substitution of twelve of the canonical amino acids that incorporated into the *E. coli* proteome above background levels, making them applicable to Miró. The ideal incorporation rate for Miró is between 5% and 10%, as this rate would generate approximately one substitution per protein, thus simplifying downstream analysis and interpretation of biochemical selections. Among the conditions I tested, seven of the ncAAs incorporated at this ideal rate: E1, F3, L2, M3, R1, R2, and V3. These compounds target the replacement of a diverse group of amino acids and are readily applicable to the Miró method. For the ncAAs that incorporated well above or below this level, additional concentrations can be tested in order to tune incorporation rates before application in a Miró protocol.

2.4 DISCUSSION

Collectively, my results demonstrated that ncAAs with diverse chemical properties have a broad range of inhibitory effects on bacterial growth and capacity to incorporate into proteins. In some cases, structural trends could be observed playing out in the data. For instance, of the four ncAAs that did not incorporate above background, three (A2, I1, and L3) were compounds characterized by the addition of a ring structure to a hydrophobic side chain. In terms of growth inhibition, A2 and L3 were some of the least toxic compounds, supporting the idea that these ncAAs are likely not disrupting proteins. Interestingly, I1 displayed a modest inhibition of growth, in a manner seemingly independent of dosage (Supplementary Figure 2.2). This type of growth inhibition combined with the observed lack of incorporation suggests that perhaps this compound is toxic to *E. coli* cells by another mechanism, such as binding as a metabolite and/or disrupting important interactions between proteins.

Another trend that stood out was within the ncAAs that incorporated at the highest levels: F2, W1, and Y1. These compounds are all fluorinated derivatives of aromatic amino acids, and are relatively conservative changes, making them more likely to be recognized by their respective synthetases and incorporated into proteins. However, this trend seemed only to apply to certain fluorinated amino acid derivatives. E1, a fluorinated derivative of the charged residue glutamic acid, incorporated into proteins, but at a much lower rate than the fluorinated aromatic compounds. This result suggests that perhaps the glutaminyl-tRNA synthetase is more discriminatory to the addition of a fluorine atom than the synthetases for the aromatic residues. Though the addition of a fluorine atom is a relatively conservative change, the above ncAAs all still significantly inhibited growth. This combination of observed toxicity and incorporation could indicate that while the addition of fluorine is not severe enough to prevent the ncAA from participating in translation,

these compounds are still causing significant damage at the protein level. In addition to informing ncAA application in the Miró protocol, data from this pilot study also provide a comprehensive resource characterizing effects of growth and incorporation levels across a diverse set of twenty-two compounds, some of which were not previously confirmed to incorporate into proteins in the literature. This resource can provide information to any experiment or study involving the ncAAs covered here in *E. coli* and can additionally be tested for incorporation in other organisms.

2.5 METHODS

2.5.1 *Toxicity assays*

E. coli BL21 cells were grown overnight in M9 minimal media at 37ºC and inoculated to an OD_{600} of 0.1 in a 96-well plate containing the same media with the ncAAs at four different concentrations (1.6 mg/mL, 6.3 mg/mL, 25 mg/mL, and 100 mg/mL), and without ncAAs (0 mg/mL) as a control (two replicate wells per condition) (Supplementary Table 2.1). Plate was incubated at 37ºC with agitation in a BioTek plate reader, and OD_{600} was measured at 15-minute intervals for a total of 12 hours. Growth data were used to construct curves for each ncAA at each concentration and the control with no ncAA added, by calculating the mean OD_{600} of the two replicate wells and plotting over time.

2.5.2 *MS sample preparation*

In order to determine the concentration of each compound to test for incorporation, growth curves from toxicity assays were used to estimate an IC_{50} for ncAAs that caused a clear dose-dependent toxicity. For ncAAs that did not cause a clear dose-dependent toxicity, I treated *E. coli* cultures with higher concentrations to maximize the chance that these compounds would incorporate at detectable levels. *E. coli* cultures were grown in 96-well plate format (as described above), treated with the determined concentration of each ncAA (Supplementary Table 2.1), and harvested at an OD_{600} of 1.0.

Cell pellets were resuspended in a urea buffer (8M urea, 75mM NaCl, and 50mM Tris, pH 8.2) and lysed by four 1-minute cycles of bead-beating at 4ºC. Protein concentration of lysates was measured by a BCA assay. Proteins were reduced and alkylated as described in Leutert *et al.*, 2019. Whole proteome samples were prepared on a KingFisher™ Flex up to quenching of the LysC digestion, as described in steps 1 – 12 of the R2-P1 protocol in (Leutert *et al.*, 2019). Samples were

subsequently desalted by the StageTip protocol (Rappsilber *et al.*, 2007), dried by vacuum centrifugation, and stored at -20ºC until MS analysis.

2.5.3 *MS methods*

Proteome samples were subjected to a 60-minute LC-MS/MS run on a Thermo Easy-nLC online with a QExactive hybrid quadrupole-orbitrap mass spectrometer. Samples were loaded onto a 100-µm internal diameter (ID) x 3-cm trap column filled with 3-µm Reprosil C18 beads (Dr. Maisch GmbH). Peptides were separated via a 45-minute gradient of increasing concentration of acetonitrile in 0.125% formic acid on a 100-µm ID x 30-cm analytical column filled with 1.9-µm Reprosil C18 beads (Dr. Maisch). An MS1 survey scan was used to perform data-dependent acquisition for the top 20 most abundant precursors. We used an MS1 scan over the range of 300 – 1500 m/z with the following parameters: 70,000 resolution, 3e6 automated gain control (AGC), and 120 ms injection time (IT). Top 20 precursors were isolated within a 2.0 m/z window and subjected to 26 normalized collision energy (NCE). MS/MS scans were performed at 17,500 resolution, 5e4 AGC, and 100 ms IT. Targeted precursors were placed on dynamic exclusion for 40 seconds, to prevent redundant sampling.

2.5.4 *MS data analysis*

MS data was searched with Comet (2015.02.rev.5; (Eng *et al.*, 2013) against the *Escherichia coli* K-12 protein sequence database (downloaded from Uniprot in April 2017) using LysC enzyme specificity and allowing for two missed cleavages. I used a Comet configuration file that allowed for a variable modification on the target amino acid with the mass shift corresponding to the ncAA being tested for incorporation. For example, to check for P1 (azetidine) incorporation, we allowed for a -14.015 Dalton mass shift as a variable modification on the target amino acid, proline. Other parameters included oxidation of methionine and acetylation of the protein N-

terminus as variable modifications, carbamidomethylation of cysteine as a static modification, 20 ppm tolerance for precursor mass and 0.02 Dalton tolerance for fragment ions. Percolator (Käll *et al.*, 2007) was used to filter results to a PSM-level FDR of 1% estimated via the target-decoy method (Elias *et al.*, 2007). Percent incorporation was calculated using the number of unique peptide identifications containing ncAA out of all unique peptides containing the target amino acid. For example, P1 (azetidine) incorporation was calculated by taking a percentage of the number of peptides containing P1 out of all peptides containing proline.

2.6 CONTRIBUTIONS

Bianca Ruiz executed toxicity and incorporation assays, with guidance in experimental design by Ricard Rodriguez; ncAA structures illustrated by Ricard Rodriguez.

Chapter 3. CHEMICAL SELECTION TO UNDERSTAND PROTEIN SOLUBILITY

3.1 ABSTRACT

In order for proteins to carry out their many biological functions, they must be in an environment that chemically permits folding into a stable and soluble state. Many critical protein functions rely heavily on environmental pH, such as interacting with other proteins, maintaining or changing conformation, and catalyzing critical biochemical reactions. Though the relationship between proteins and pH is clear, we have a limited understanding how protein stability and solubility behave across different conditions. Here, I develop a method for measuring the solubility of proteins across a wide range of pH conditions at the proteome scale. A single experiment applying my method generates rich proteome-scale data, which enable the construction of pH-dependent solubility profiles for thousands of proteins. My results show that proteins from various subcellular compartments are differentially tolerant of pH ranges, which sometimes coincide with the previously established pH of these respective localizations. Further analyses of this dataset will aid our understanding of how biochemical properties such as amino acid composition, known interactions and complexes, or conformation and structure, enable proteins to tolerate different pH conditions. Additionally, I anticipate that my method will be applicable as a selection to explore how amino acid substitutions impact the solubility of proteins in their microenvironments.

3.2 INTRODUCTION

Across the various organelles and subcellular compartments in a yeast cell, pH can range from as acidic as 5.0 in the vacuole to as alkaline as 8.0 in the mitochondria (Figure 3.1). These conditions are important to the proper functioning of subcellular compartments and the proteins they enclose, and cells maintain internal pH conditions by various regulatory mechanisms, such as ATPases (Eraso *et al.*, 1987; Orij *et al.*, 2011). Proteins fold into intricate structures and carry out complex biological processes within this diversity of intracellular microenvironments, yet the interplay between proteins and pH in biological environments is not fully understood. Some approaches toward this end include algorithmic estimation of the isoelectric point of proteins, which is the pH at which the collective charge of a protein is neutral (Audain *et al.*, 2016). The calculation of isoelectric point has been used in fractionation approaches to purify proteins and in efforts to understand their underlying chemical properties, and provides some understanding of how proteins behave in environments with different pH conditions (Ramos *et al.*, 2008). Bioinformatics studies have even found a proteome-wide correlation between isoelectric point and organellar pH, suggesting that these two properties have co-evolved to maintain and optimize proper protein function (Brett *et al.*, 2006).

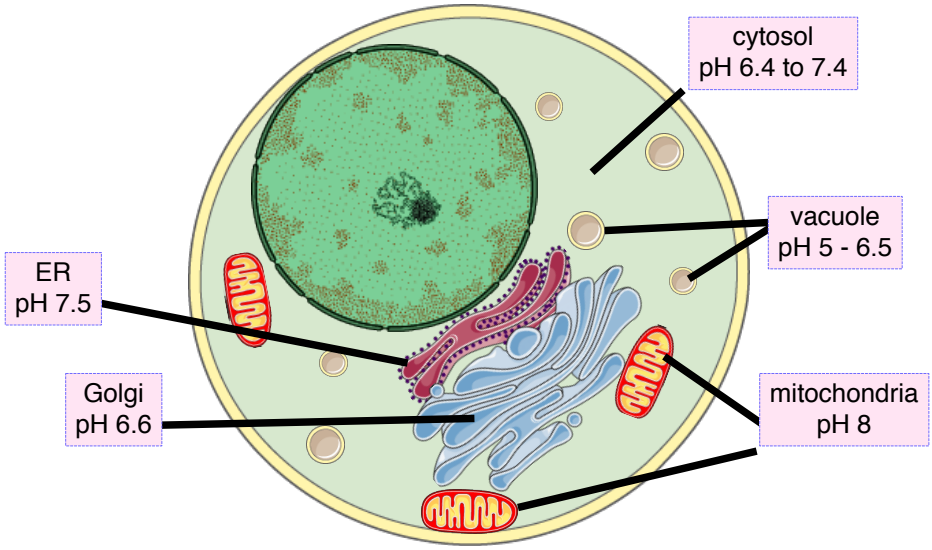


Figure 3.1 Cellular pH varies widely throughout the cell and plays an important role in protein biology. Approximate pH ranges are indicated for selected organelles in a yeast cell, as reported in the literature. Conditions can range from acidic to alkaline, while more generalized regions such as the cytosol are relatively neutral (organelle illustrations from smart.servier.com).

Protein properties in the context of pH have also been characterized *in vitro*, via measurement of protein activity or solubility across several pH conditions. These titration assays often reveal that protein activities take on a bell-shaped curve, enabling identification of a pH optimum at which the given activity is at its peak (Kumar *et al.*, 2004; Bearne, 2014). While these types of approaches have expanded our understanding of protein interactions with pH, *in vitro* studies of purified proteins are far removed from the cellular environments proteins experience. Within the cell, proteins are surrounded by a mixture of other proteins, nucleic acids, and metabolites that all play a role in facilitating protein structure and function.

3.3 RESULTS

To gain a deeper understanding of how proteins tolerate a wide range of pH, I developed a proteomic method to determine the solubility of proteins subjected to a range of pH conditions in the context of a complex lysate. In an effort to develop and optimize this method, I performed several pilot experiments treating proteins with three pH conditions. The goal of the first pilot experiment was to determine whether or not the pH selection would require a heat treatment. In order to assess this question, I treated proteome samples with three different pH conditions representing the range to be assayed (pH 5, pH 7, and pH 9). I then incubated these samples at either 30°C or 52°C and ran the soluble fraction of the proteome on an SDS gel. The gel results revealed that the differences in the amount of protein between conditions was more pronounced when heat was applied, and thus this step was implemented in the selection protocol (Supplementary Figure 3.1).

An additional pilot experiment was performed to assess the importance of the pH during cell lysis, versus the final pH during the selection. Given the protocol design, proteins in the complex lysate were exposed to two different pH conditions – one uniform lysis pH and the final selection pH. I wanted to ensure that the final selection pH, not the lysis pH, was driving any quantitative differences we observed in the MS data. In order to address this question, I split a sample of cells into three batches, and lysed each batch at a pH of 5, 7 or 9. I then split these three samples into three more aliquots each, for a total of nine samples. Each set of three samples was then adjusted to a selection pH of 5, 7, or 9, and heat was applied at 52°C. The soluble fraction from each of these samples was then separated and prepared for MS analysis. After MS analysis, I used peptide-level quantifications to calculate a Pearson correlation value across each crosswise pair of these samples. The Pearson correlation values were much higher when comparing samples

that were adjusted to the same final selection pH, than they were between samples that were lysed in the same pH, demonstrating that the final selection pH was more deterministic of any differences measured by MS (Supplementary Figure 3.1). Taken together, these pilot experiments demonstrated that heat should be applied to achieve the most pronounced selection, and that the selection pH was driving differences in the MS data, regardless of lysis pH.

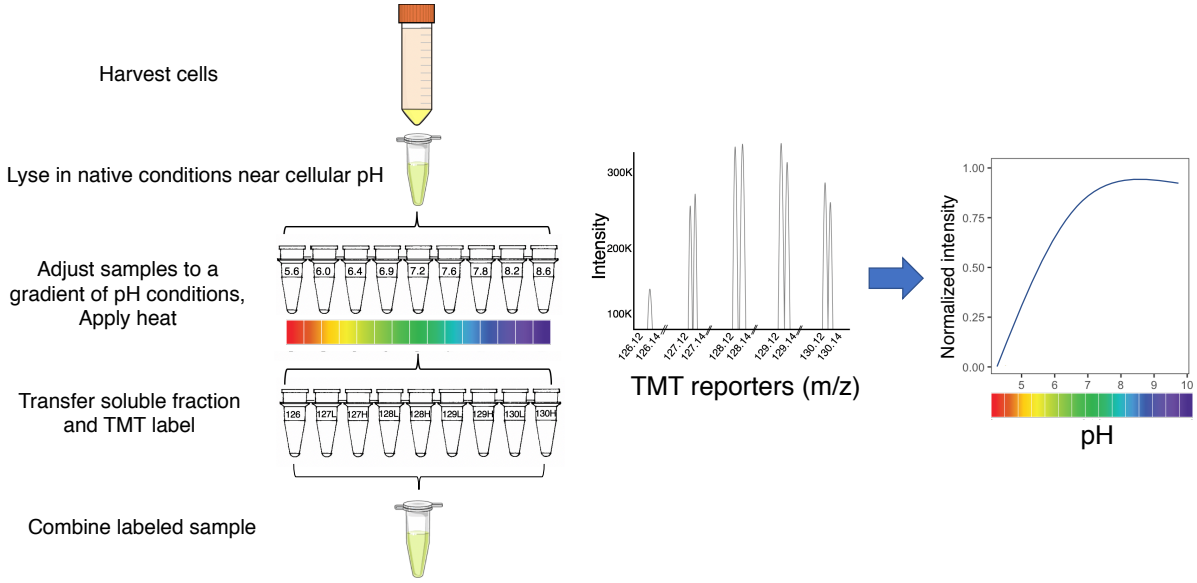


Figure 3.2 Workflow of the proteome-wide selection for interrogating protein solubility in a range of pH conditions. A. Experimental steps for harvesting and lysing cells in native conditions, followed by treatment with the pH gradient. The fractions of each proteome sample remaining soluble underwent TMT labeling and were subsequently combined into a single sample. B. Combined sample was analyzed by MS to identify and quantify peptides. C. MS data was used to construct solubility profiles of peptides and proteins.

In brief, the final experimental design involved incubating a complex yeast lysate in ten pH conditions at 52°C, to differentially precipitate proteins out of solution (Figure 3.2). The

soluble fractions from each condition were collected and labeled with unique TMT isobaric mass tags, combined into a single tube, and analyzed by MS. In two biological replicates, my MS experiments attained high coverage of the yeast proteome, with approximately 4000 to 6000 unique peptide identifications, mapping to approximately 1500 to 2000 unique proteins. Data from MS analysis was used to construct solubility profiles for individual peptides and proteins across the treatment pH range from approximately 4.0 to 10.0.

3.3.1 *Construction of solubility profiles by spline fitting*

I first used data from the pH selection to show that solubility profiles could be constructed at multiple levels – peptide-spectral match (PSM), peptide, protein, and proteome. Similarly to the Thermal Proteome Profiling (TPP) method (Savitski *et al.*, 2014), I first normalized intensity measurements across the pH gradient to the least selective condition – pH 7.2. This pH was chosen because it is the approximate cytosolic pH in yeast, and I predicted that this condition would be the most conducive to maintaining solubility. I ultimately determined that this normalization method was too crude to capture solubility profiles, because for many peptides, this pH was not the optimum. The solubility profiles resulting from this method sometimes displayed noise from quantification and were thus not as informative, and would indicate sharp increases and decreases in abundance across the pH gradient (Supplementary Figure 3.2).

There are many studies in the literature that characterize a purified protein by constructing a curve of stability or enzymatic activity across a pH gradient (Kumar *et al.*, 2004). While this area of literature provided a reference for how solubility profiles would be shaped in our proteome-wide experiment, I acknowledged major differences in my method could result in many unexpected curve patterns. Some of the main differences were that 1) my method strictly probed solubility, and not other commonly assayed protein properties, such as activity and 2) my method

treated proteins in the context of a complex lysate containing other proteins and metabolites, rather than in a purified state. Given these differences, I decided to fit solubility data using spline fitting, because this model is permissive of many different shapes, which would include the classic bell-shaped curve, among many others (Figure 3.3) (Ferguson, 1964). Overall, this method of curve fitting performed well, as demonstrated by a distribution of low sum residual error values across the dataset, indicating that the spline-fit solubility curves closely reflected the raw data (Supplementary Figure 3.2).

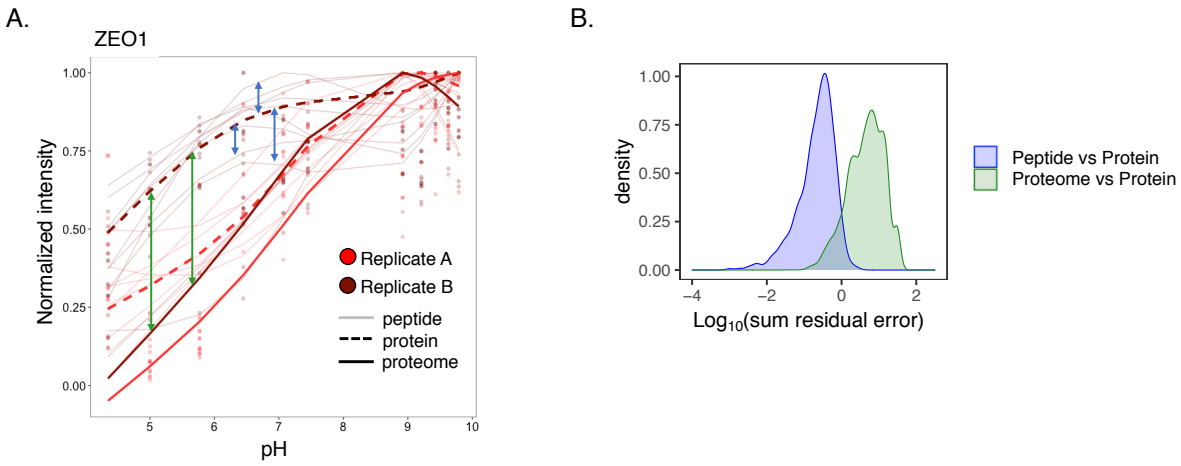


Figure 3.3 Solubility profiles were generated at the peptide-, protein- and proteome-level, and true differences between proteins outweighed technical variance. A. Solubility curves for protein ZEO1 (YOL109W, a plasma membrane protein), generated by spline fitting raw data. Points represent normalized ZEO1 peptide-level abundance measurements, fine lines represent ZEO1 peptide spline curves, bold dashed lines represent aggregated ZEO1 protein spline curves, and bold solid lines represent aggregated proteome spline curves. Colors represent biological replicates. Arrows indicate curves between which sum residual error was calculated. B. Distribution of sum residual error between spline fit curves, across entire dataset. Blue distribution represents error between peptides and their respective proteins (median = -1.43, n = 6305), and

green distribution represents error between the aggregated proteome curve and individual protein curves (median = 0.69, n = 1138).

In order to determine whether the observed differences between solubility profiles were reflective of true differences between proteins rather than technical error, I calculated a sum residual error for several comparisons across the proteome. Higher values of sum residual error indicate large differences between curves, while lower values indicate that curves are similar to each other. First, this statistic was calculated to describe the difference between peptide solubility profiles and that of their respective proteins. The distribution of this sum residual error is representative of technical variation, as curves for peptides of the same protein are effectively replicate measurements of that protein. Second, I calculated the same statistic describing the difference between protein level curves and the aggregated curve representing the entire proteome. These differences are be representative of true differences in protein solubility, as proteins across the proteome would be expected to differentially tolerate pH conditions. The median of this distribution of error values was two orders of magnitude greater than that of peptides and their respective proteins. This result indicated that my method was able to capture differences in pH solubility between proteins (Figure 3.3).

3.3.2 *Proteins from different subcellular localizations tolerate different pH ranges*

I next sought to use my data to compare solubility profiles between proteins localized to various subcellular compartments. Though the pH of many yeast organelles has been experimentally estimated in the literature, differences in solubility of proteins localized to their respective organelles has not been explored at the proteome scale. To address this question, I

aggregated solubility profiles of all proteins from specific subcellular compartments and evaluated the differences between these aggregated curves (Figure 3.4). Interestingly, I observed that the solubility profiles of proteins localized to the extracellular region and cell wall displayed higher solubility in the low pH range compared to proteins in other localizations. This tolerance to acidic pH coincides with the well-established observation that yeast cells grow well in acidic conditions, and that extracellular proteins and cell wall proteins would be in contact with this acidic environment.

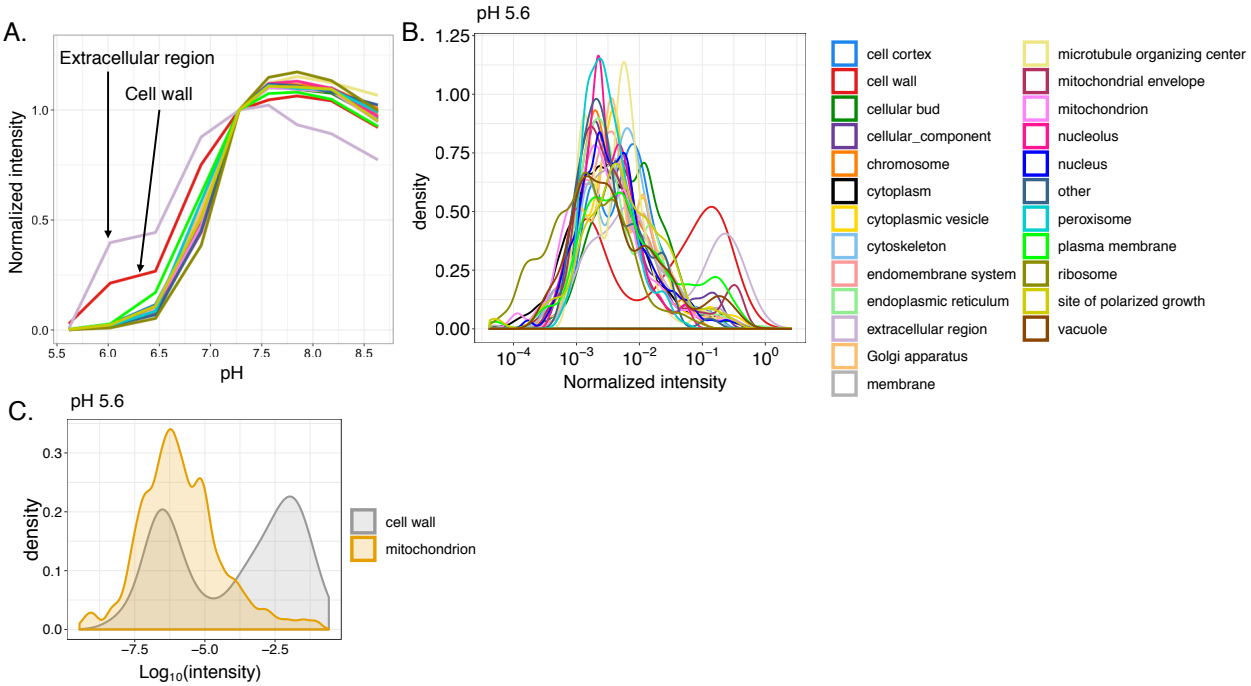


Figure 3.4 Solubility profiles vary depending on subcellular compartment. A. Aggregated solubility profiles for various subcellular compartments, represented by the different colors. B. Distribution of abundance of proteins from different subcellular compartments plotted for data from the pH 5.6 condition. C. Distribution of Log_{10}(intensity) of proteins from organelles with

very different pH, with the cell wall being an acidic environment and the mitochondrion being alkaline.

In order to compare the selectivity of each pH condition, I plotted the distribution of protein abundance in each subcellular compartment across all conditions (Supplementary Figure 3.3). This analysis allowed me to visualize which pH conditions were the most selective for separating protein solubility profiles in the various compartments, which would appear as broader distributions, while distributions would be narrower for less selective pH conditions. In general, I expected that the further away from the approximate cellular pH of 7.2, the more selective the condition would be, and that selectivity would be relatively low in the conditions closest to 7.2. This expected trend was confirmed by the data, with some exceptions. The least selective pH conditions tended to be close to the approximate cellular pH of 7.2, or more basic. In fact, the only basic pH that displayed selectivity between subcellular compartments that was comparable to the acidic conditions was the highest pH at 8.6. This result indicated that proteins are generally more tolerant of the basic pH range tested, but start to change in solubility at a threshold of basicity.

Overall the acidic conditions were markedly more selective, with the two lowest pH conditions, 5.6 and 6.0, resulting in broader distributions across subcellular localizations. While cell wall and extracellular region proteins again display partial solubility even at the lowest pH assayed, ribosomal proteins show a relative sensitivity to this condition and have a distribution of abundance on the lower end. I also used the lowest pH condition to compare solubility between proteins from some of the more extreme conditions that occur in yeast cells – the acidic cell wall and basic mitochondria (Figure 3.4). As expected, proteins localized to the cell wall were markedly

more soluble at pH 5.6 than mitochondrial proteins, which may not tolerate such a low pH because they are likely optimized for the mitochondrial pH of 8.0.

3.3.3 *Principal component analysis of pH selection data*

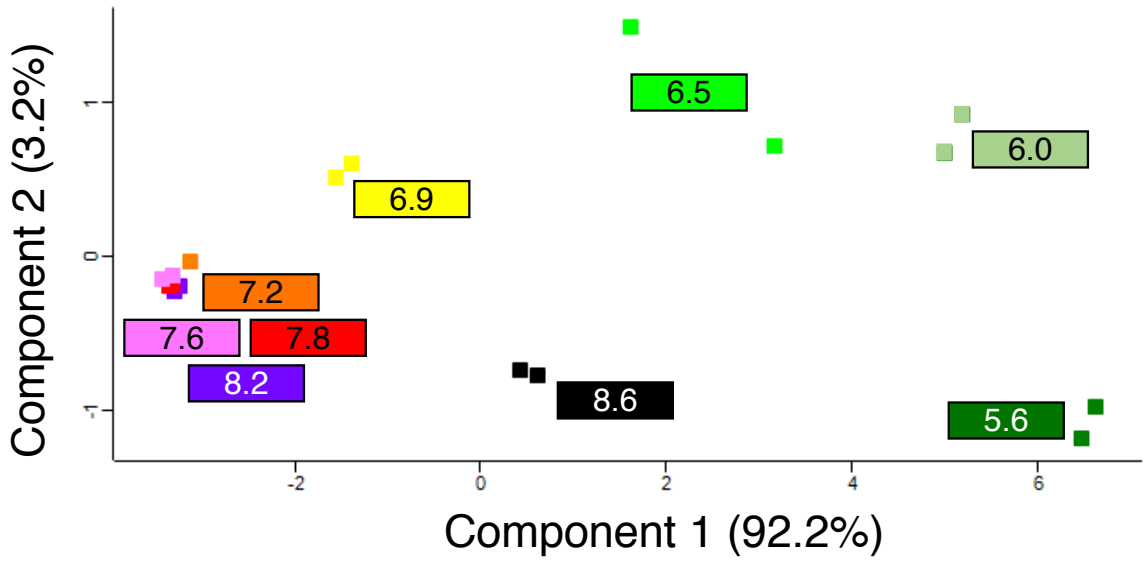


Figure 3.5 Principal component analysis of data from pH selection. Principal component analysis is shown for the top two components. Component 1 accounted for variance in 92.2% of the data, and Component 2 accounted for 3.2%. Points represent dimensionality-reduced data for each pH, with two biological replicates per pH.

In order to determine whether the pH selection was truly driving the differences I was observing in the data, I applied principal component analysis (PCA) as a more exploratory approach. I found that most of the variance in my data was explained by two components, with 92.2% of the separation explained by the first component. I also found that the PCA separated my data by pH conditions, and that data from the two replicates of each condition segregated together

(Figure 3.5). This finding confirmed that the pH selection was driving the global results and any differences in the data I observed, and that the method was reproducible across biological replicates.

3.4 DISCUSSION

Overall, my pH selection successfully captured solubility profiles across the entire proteome in several different pH conditions. There were some data in the literature to suggest what a solubility profile of a purified protein may look like, but one major distinction in my experiment was that proteins would undergo selection in the context of other proteins and metabolites in the proteome samples. Given this difference, I wanted to ensure my analysis and curve fitting did not make assumptions about the shape of the data. I found that spline fitting worked well for this data, as demonstrated by the differences between proteins being greater than that between peptides and their respective proteins.

One of the main goals I had for this dataset was to explore differences in protein solubility in various subcellular localizations. I found that localization was a strong indicator of protein solubility at different ranges of pH, which at times coincided with the known pH of the given organelle. This relationship between localization and pH solubility was supported by several analytical approaches, including those focused on proteins from different organelles, as well as exploratory analyses applying dimensionality reduction via PCA. An interesting direction for this data would be to explore the solubility profiles of proteins that are known localize to two or more organelles, and identify the properties that enable solubility in multiple subcellular compartments.

Next steps for analyses outside of subcellular localization would be to examine the proteins with solubility profiles that deviate significantly from the average proteome profile, to identify other protein properties that drive these differences. For example, one analysis could be to compare amino acid composition in proteins that are tolerant of acidic conditions, versus proteins that are more tolerant of basic conditions, to identify any associated properties (i.e. charge, hydrophobicity) that may facilitate these differences. This approach to analysis, along with others

can utilize my data to decipher the interplay between pH and protein properties such as amino acid composition, charge, structure, and interaction with other proteins, nucleic acids, or small molecules. An experimental application I anticipate for my selection method is to probe changes in solubility caused by amino acid substitutions in proteins through Miró. Thus far, Miró has largely been used to understand the impact of amino acid substitution on protein solubility, protein thermal stability, protein-protein interactions, and protein modifications. My pH selection could be used to expand beyond these selections to identify how amino acid substitutions affect the solubility of proteins in various pH conditions. Amino acid substitutions that alter charge or hydrophobicity of the residue would be particularly interesting for a pH selection in the context of Miró.

In summary, my method will enable a deeper understanding of how protein properties determine and facilitate the folding and function of proteins across diverse microenvironments. This approach generates rich data capturing the solubility of proteins and peptides at an unprecedented throughput, and the dataset itself can be subjected to a wide range of further analyses and be used identify interesting targets for validation.

3.5 METHODS

3.5.1 *Yeast strains and growth*

Wildtype diploid *S. cerevisiae* strain DBY10144 (MAT a/α) is a derivative of S288C and is prototrophic for lysine, with FY3G and FY4H as parental strains (provided by Dunham Lab, Department of Genome Sciences, University of Washington). Two replicate cultures were grown overnight in YEPD at 30°C and inoculated to OD 0.1 into 50 mL harvest cultures.

3.5.2 *Polybuffer system*

A polybuffer system was generated as described in (Newman, 2004) by first preparing the following three buffers: Citric acid (42.8 mg/mL), HEPES (79.4 mg/mL), and CHES (92.1 mg/mL). These buffers were then combined in a ratio of 2 parts Citric acid, 3 parts HEPES, and 4 parts CHES. Half of this buffer mixture became the high pH polybuffer, with the addition 13.5 mL 10M NaOH for a final pH of 10. The other half of the buffer mixture became the low pH polybuffer, with the addition of 2 mL 10M NaOH for a final pH of 4. The high pH and low pH buffers were then combined at various ratios (ranging between 10% to 90% high pH buffer with 90% to 10% low pH buffer) in order to yield the pH ranges specified in figures.

3.5.3 *Solubility selection across pH gradient*

Whole proteome samples were prepared from harvested cultures by lysing cells and adjusting to a protein concentration of 2 mg/mL in a native lysis polybuffer at pH 7.2 (near cellular/cytosolic pH). Lysate was aliquoted into 10 PCR tubes containing the appropriate amount of polybuffer to adjust samples to each pH across the desired gradient. Previous pilot experiments demonstrated that method was more selective, and more likely to generate informative solubility profiles, when heat was applied in addition to adjusting pH (Supplementary Figure 3.1). Thus, in

the final implementation of the protocol I applied heat during the pH incubation, and chose 52°C, the approximate median melting temperature of the yeast proteome. Samples were incubated at 30°C for 5 minutes, and then 52°C for an additional 5 minutes. Insoluble fractions were removed by centrifugation at 4°C and 14,800 xg for one hour (Franken *et al.*, 2015). The soluble fraction was isolated by transferring supernatant to new set of tubes containing the appropriate amount of polybuffer to equilibrate all samples to a uniform pH.

3.5.4 *MS sample preparation*

After selection, samples were denatured by addition of urea to a final concentration of 4M. Proteins were reduced by addition of dithiothreitol (DTT) to a final concentration of 5mM, and incubation at 55°C for 30 minutes with agitation. Reduced samples were then alkylated by the addition of iodoacetamide (IAA) to a final concentration of 15mM and incubation at room temperature for 30 minutes. Alkylation reaction was quenched by addition of DTT to a final concentration of 5mM.

Reduced and alkylated protein samples were then adjusted to a pH of approximately 8 – 9 by the addition of 2M Tris, and prepared on a KingFisher™ Flex up to quenching of the LysC digestion, as described in steps 1 – 12 of the R2-P1 protocol in Leutert *et al.*, 2019. 200 µg peptide aliquots were desalted using HLB stationary phase (Waters), with a standard clean-up protocol, and dried by vacuum centrifugation. 25 µg peptides from each sample were reconstituted 100mM HEPES, pH 8.5 with 30% acetonitrile, and labeled with 100 µg TMT10plex™ (Thermo) for 1 hour at room temperature. Labeling reaction was quenched by addition of hydroxylamine to a final concentration of 0.5% and incubation at room temperature for 30 minutes, followed by acidification with concentrated HCl to a pH of 2 – 3. Labeled samples were then collapsed into a single tube, and approximately 40 µg was subjected to reverse phase basic fractionation on

StageTips containing Empore SDS-RPS C18 material (3M). Samples were washed and eluted into four separate fractions by passing 20mM ammonium hydroxide with increasing amounts of acetonitrile at 10% (fraction 1), 15% (fraction 2), 20% (fraction 3), and 80% (fraction 4).

3.5.5 *MS methods*

Each fractionated proteome sample was subjected to a 120-minute LC-MS/MS run on a NanoAcquity U-HPLC (Waters) online with an Orbitrap Tribrid Fusion Lumos (Thermo) mass spectrometer. Samples were loaded onto a 100-µm internal diameter (ID) x 3-cm trap column filled with 3-µm Reprosil C18 beads (Dr. Maisch GmbH). Peptides were separated via a 90-minute gradient of increasing concentration of acetonitrile in 0.125% formic acid on a 100-µm ID x 30-cm analytical column filled with 1.9-µm Reprosil C18 beads (Dr. Maisch). Different gradients were used to accommodate the complexity of the reverse phase basic fractions (fraction 1: 8% to 20%, fraction 2: 9% to 25%, fraction 3: 11% to 29%, fraction 4: 13% to 32%).

An MS1 survey scan was used to perform data-dependent acquisition for the top N precursors within a 5-second duty cycle. We used an MS1 scan over the range of 500 – 1200 m/z with the following parameters: 60,000 resolution, 5e5 automated gain control (AGC), and 100 ms injection time (IT). Targeted precursors were placed on dynamic exclusion for 45 seconds, to prevent redundant sampling, and top N precursors were subjected to MS/MS and SPS-MS3. Top N precursors were isolated within a 0.5 m/z window and subjected to 30% CID. Ion trap MS/MS scans were acquired at a rapid scan rate, 1e4 AGC, 35 ms IT, and 10 ms activation time. For the SPS-MS3, the top 10 most abundant MS/MS peaks were isolated in parallel (2.5 m/z isolation width each) and fragmented via HCD fragmentation at 55 normalized collision energy (NCE). The resultant fragment ions were then analyzed in the orbitrap using 50,000 resolution, 5e4 AGC, and 86 ms IT, over a 100 – 500 m/z range.

3.5.6 *MS data analysis*

MS data were searched with Comet (2015.02.rev.5; (Eng *et al.*, 2013) against the *Saccharomyces cerevisiae* protein sequence database (downloaded from Uniprot in July 2014) using LysC enzyme specificity and allowing for two missed cleavages. I used a Comet configuration file that allowed for the variable modifications of methionine oxidation and protein N-terminal acetylation. Carbamidomethylation of cysteine and TMT10plex™ (Thermo) labeling on both the peptide N-terminus and lysine were included as static modifications. Other parameters included 20 ppm tolerance for precursor mass and 0.02 Dalton tolerance for fragment ions. Percolator (Käll *et al.*, 2007) was used to filter results to a PSM-level FDR of 1% estimated via the target-decoy method (Elias *et al.*, 2007). MS3 intensity of TMT reporter ions was used to generate solubility profiles at the PSM-, peptide-, protein-, and proteome-level. Intensities were normalized either to the intensity at the lysis pH of 7.2, or to the maximum intensity across the pH gradient, curves were then fit using spline modeling.

3.6 CONTRIBUTIONS

Bianca Ruiz designed, optimized, and executed selection experiments, and analyzed MS data. MS analysis was conducted at the University of Washington's Proteomics Resource (UWPR), with assistance on instrumentation from Priska von Haller and Ian Smith. Anthony Barente provided support for statistical modeling methods.

Chapter 4. INTERACTIONS BETWEEN AMINO ACID SUBSTITUTIONS AND THE PHOSPHOPROTEOME

4.1 ABSTRACT

Amino acid substitutions in proteins can impact their function, disrupt proteostasis, and dysregulate cellular processes. Phosphorylation plays a critical role within these processes, by regulating the functions of individual proteins and complexes, as well as launching global responses to proteotoxicity and a variety of other stressors. In order to deepen our understanding phosphorylation-mediated responses to amino acid substitutions across the proteome, I designed a model system to characterize the interplay between phosphorylation and substitutions in proteins. Using mass spectrometry, I measured the phosphoproteome of three yeast strains that facilitate either proline to alanine, proline to serine, or alanine to serine substitutions randomly across the proteome, as well as a control strain with no substitutions. I found that the abundance of hundreds of phosphopeptides was regulated in the mistranslating strains relative to the control strain, and that aspects of this regulation were both specific and non-specific to the type of amino acid substitution. At the phosphosite level, I identified regulated phosphopeptides and phosphomotifs. Regulated phosphopeptides were found to occur within proteins enriched in specific functions, subcellular compartments, and biological processes. Collectively, these findings expand upon our understanding of the relationship between amino acid substitutions and phosphorylation across the proteome and generate exciting hypotheses for future studies.

4.2 INTRODUCTION

Proteins are incredibly diverse molecules that carry out the biological processes that comprise all of life. In order for proteins to properly execute their vast array of functional and structural duties, the cell must sustain a stable and resilient proteome through proteostasis. Avenues for protein quality control including synthesis, folding, localization, activation and deactivation, and degradation all play an important part in the orchestration of the proteostasis network. The dysregulation of any one of these processes can throw the entire proteome out of balance. Defining the molecular mechanisms by which proteostasis is disrupted remains an important ambition of biomedical research, as the loss of this delicate equilibrium is an acute factor in the aging process as well as the pathogenesis of disease (Balch *et al.*, 2008; Powers *et al.*, 2009; Bouchecareilh *et al.*, 2011; Kaushik *et al.*, 2015).

Within the myriad of dysfunctions that can lead to a loss of proteostasis, protein misfolding and aggregation are particularly well-studied because of their implications for many human diseases. The molecular phenotypes of misfolding and aggregation play instrumental roles in the pathogenesis of progressive neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis (ALS) (Irvine *et al.*, 2008; Blokhuis *et al.*, 2013). Our understanding of the mechanisms underlying the destabilization of protein structure is crucial to the development of therapeutic approaches for these incurable and fatal diseases.

Several factors can trigger misfolding and aggregation of proteins, and many involve changes in the amino acid sequence. Truncations, insertions, and substitutions in the amino acid sequence of a protein can destabilize structure and ultimately result in the formation of insoluble aggregates and proteotoxic stress. Cells manage this stress and maintain proteostasis through the action of many intersecting pathways, including the unfolded protein response and heat shock

response pathways. These response pathways rely on the coordination between a multitude of proteins with diverse functions to detect misfolded aggregates and mediate proper folding or targeted degradation. The protein functions and interactions within these networks are commonly facilitated by signaling through phosphorylation, a sophisticated and reversible post-translational modification that is involved in virtually all cellular processes.

The relationship between proteostasis and phosphorylation has been demonstrated by a number of examples in the literature. One study illustrates this dynamic relationship through protein disulfide isomerase (PDI), a multifunctional enzyme that is central to proper protein folding in the endoplasmic reticulum (ER) of eukaryotic cells. Upon triggering of the unfolded protein response in the ER, PDI is phosphorylated at serine 357, and subsequently undergoes a conformational change. This phosphorylation-driven change in protein structure toggles enzyme function from supporting normal protein folding to actively preventing aggregation (Yu *et al.*, 2020). The link between phosphorylation and mechanisms of maintaining proteostasis is also exemplified by the heat shock transcription factor in yeast. During heat stress, this protein binds to and activates transcription of genes within specific heat shock elements. The initiation of this response by the protein has been established to be contingent upon proper phosphorylation (Hashikawa *et al.*, 2004). While these reports and many others have broadened our understanding of the connections between proteostasis and phosphorylation, the methods applied are limited in scope and throughput, often focusing on the dynamics of one or two proteins per study. Given the complexity of each individual system, proteome-wide approaches are needed to capture a more complete picture of how they interface.

To find relationships between amino acid substitutions and phosphorylation, I used mistranslation in the proteome as a model system in yeast. I used three strains that randomly

introduced different amino acid substitutions relevant to phosphorylation, through the expression of mistranslating tRNAs. In addition to their relevance to phosphorylation, introducing amino acid substitutions across the proteome was expected to cause protein misfolding and aggregation. These effects also make proteome-wide mistranslation a model for disruption of proteostasis. After facilitating random amino acid substitutions across the proteome, I quantified changes in the phosphoproteome to explore these effects.

4.3 RESULTS

To understand the relationship between mistranslation and phosphorylation at the proteome scale, I employed four strains of yeast with six biological replicates per strain. Three of the yeast strains expressed mistranslating tRNA variants engineered to introduce different amino acid substitutions randomly across the proteome, and a fourth control strain expressed a wildtype tRNA. The first strain introduced proline to alanine substitutions, via expression of a proline tRNA with a mutation in the acceptor stem that facilitated charging of an alanine residue rather than a proline residue (Figure 4.1) (Hoffman *et al.*, 2017; Berg *et al.*, 2017). The second strain introduced proline to serine substitutions, via expression of a serine tRNA with a mutation to the anticodon that facilitated reading of proline codons instead of serine codons (Berg *et al.*, 2019). The third strain introduced alanine to serine substitutions, via expression of the same serine tRNA with a mutation to the anticodon that facilitated reading of alanine codons instead of serine codons (see Methods). The fourth strain expressed the wildtype version of the same serine tRNA and thus did not introduce substitutions. I hypothesized that each of these substitutions would perturb phosphorylation in different ways. The two substitutions that remove proline were anticipated to change phosphorylation because this residue is an important amino acid in phosphomotifs and directs recognition and phosphorylation by many kinases. Additionally, for the two substitutions that introduce serine, I hypothesized that these residues could potentially be recognized by kinases and phosphorylated.

4.3.1 *Growth and mistranslation rates*

We first aimed to characterize the yeast strains for relative growth rates, as a growth defect was a preliminary indication that mistranslation was occurring and causing stress to the cell. To measure growth, we cultured six biological replicates of each strain in parallel and monitored optical density. Growth curves were generated and used to calculate doubling time for each culture (Figure 4.1). All three mistranslating strains grew more slowly than the control strain with no substitutions (annotated as "No Sub"), and thus had relatively longer doubling times. The No Sub strain grew the fastest out of all the strains and had an average doubling time of 67 minutes. The proline to alanine (notated as "Pro to Ala") mistranslating strain had an increased average doubling time of 75 minutes. The alanine to serine (notated as "Ala to Ser") mistranslating strain showed a similar increase in doubling time at an average of 74 minutes. The slowest growing strain was the proline to serine ("Pro to Ser") mistranslating strain, at an average doubling time of 83 minutes (Figure 4.1). The increase in doubling time for the mistranslating strains relative to the control was expected and a preliminary indication of proteotoxicity due to mistranslation.

After characterizing growth of the yeast strains, we measured the levels of proteome-wide mistranslation by MS analysis. Cells from six replicates of each strain were harvested and processed as whole proteome samples for MS. Whole proteome data was used to determine mistranslation rate by calculating the percentage of unique peptides containing an amino acid substitution out of all unique peptides containing the target site. For example, the Pro to Ala mistranslation rate was calculated by calculating the percentage of unique peptides containing a proline to alanine substitution, out of all unique peptides containing proline. To calculate a background rate, the No Sub strain was subjected to the same analysis for each substitution type.

In our whole proteome data from the Pro to Ala strain, we detected proline to alanine substitutions in 4.2% of unique proline-containing peptides. When subjected to the same analysis, the whole proteome data from the No Sub strain showed a background mistranslation rate of 0.2%. In the Pro to Ser strain, we detected proline to serine substitutions in 4.7% of unique proline-containing peptides across the proteome, and the background search of the No Sub strain showed a rate of 0.3%. In the Ala to Ser strain, 10. 9% of unique alanine-containing peptides had an alanine to serine substitution, and the background rate calculate in the No Sub strain for this substitution was 3.2%.

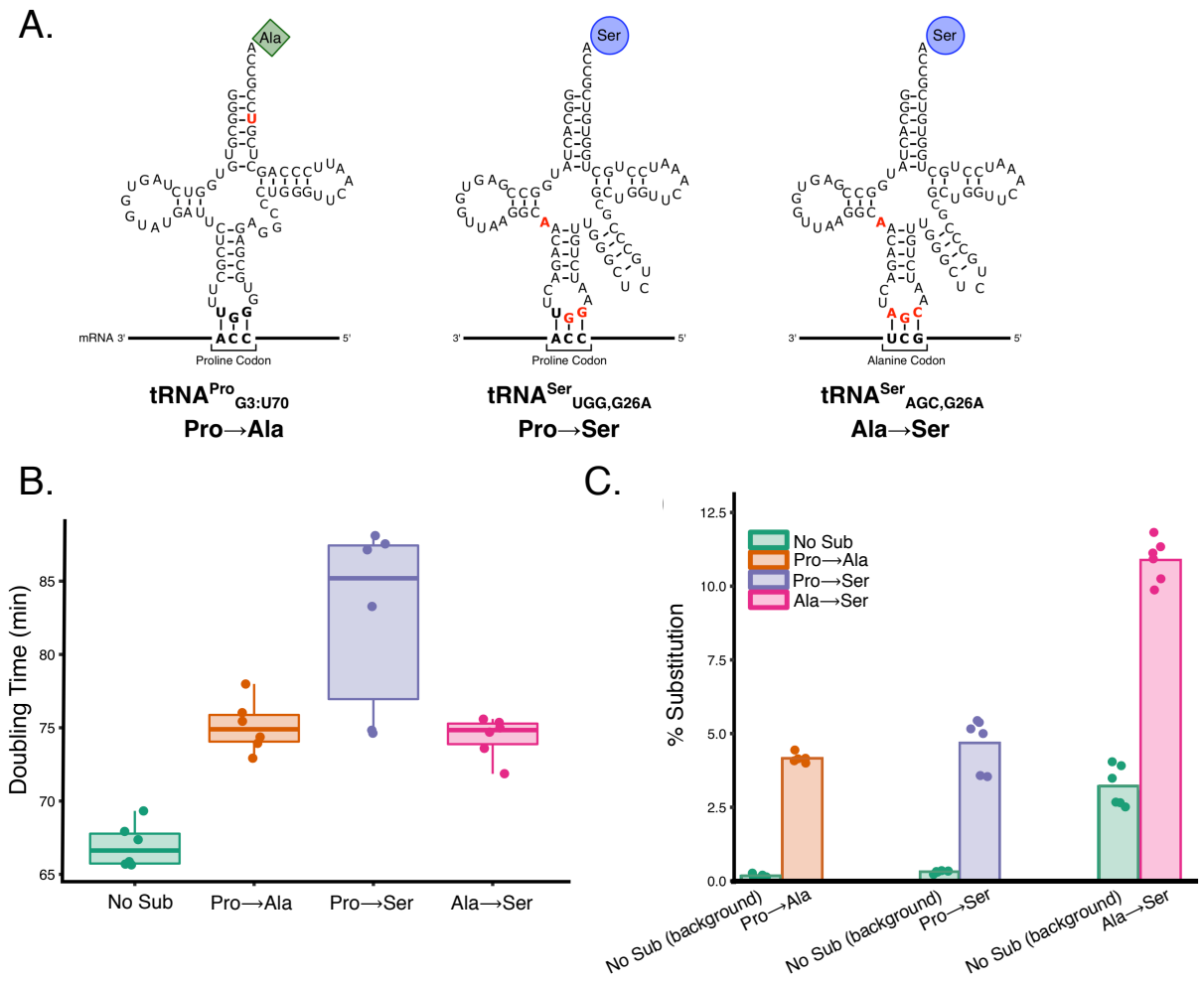


Figure 4.1 Yeast expressing mistranslating tRNAs generate proteome-wide mistranslation. A. Mutations in proline and serine tRNAs that facilitate mistranslation across the proteome. The first diagram shows the sequence of a proline tRNA variant with a mutation to the acceptor stem. This mutation introduces the identity element G3:U70 that enables mischarging of the tRNA with an alanine residue instead of proline. The second diagram shows the sequence of a serine tRNA variant with a destabilizing mutation G26A, as well as mutations to the anticodon. The destabilizing mutation decreases toxicity of this tRNA, and the anticodon mutations introduce a proline anticodon allowing for translation of proline codons as serine. The third diagram shows the sequence of the same serine tRNA variant with the destabilizing G26A mutation, as well as

mutations to the anticodon. The anticodon mutations introduce an alanine anticodon allowing for translation of alanine codons as serine. B. Growth curves were collected for six replicate cultures of each strain and used to calculate doubling time. Doubling time is shown on the y-axis in minutes, and the strain is indicated on the x-axis by the type of amino acid substitution introduced. Each point represents the doubling time of one replicate culture, and the center line within the boxplot indicates mean doubling time for each strain. C. Mistranslation rates as determined by MS analysis of the whole proteome. Percent mistranslation was calculated from the number of unique peptides containing the substitution out of all unique peptides containing the target amino acid, displayed on the y-axis. The amino acid substitution searched for is indicated on the x-axis. Points indicate mistranslation rate for one replicate, bars represent the median rate for the strain, and colors represent strains as labeled in legend.

When calculating the background levels of each amino acid substitution in the No Sub strain, I expect a low rate, typically less than 1%. The identifications within this background level are likely a combination of true basal mistranslation and false identifications. Mistranslation rates detected in the No Sub strain for both the proline to alanine and the proline to serine substitutions were very low at less than 0.5%, as expected. Background levels of alanine to serine mistranslation detected in the No Sub strain were significantly higher than expected at 3.2%. This high background rate could be due to the detection of another modification with the same mass shift, such as oxidation. However, importantly, I observed that the mistranslation rate in the Ala to Ser strain was over three times higher than the background rates detected in the No Sub strain, indicating that true mistranslation was occurring.

4.3.2 *Phosphoproteome data quality and reproducibility*

After establishing mistranslation rates in each strain, I prepared samples to examine the impact on the phosphoproteome. Using six replicates of each strain, I prepared phosphoproteome samples and analyzed them by MS. To assess the quality of the phosphoproteome data captured for each sample, I calculated the number of peptides, reproducibility of identification, and reproducibility of quantification (Figure 4.2). These statistics fell within a consistent range across all samples, aside from one outlier sample from the No Sub strain, for which quantification was poorer than all other samples. This sample was excluded from all the analyses, leaving five replicate samples of the No Sub strain, and six replicates of the other three strains.

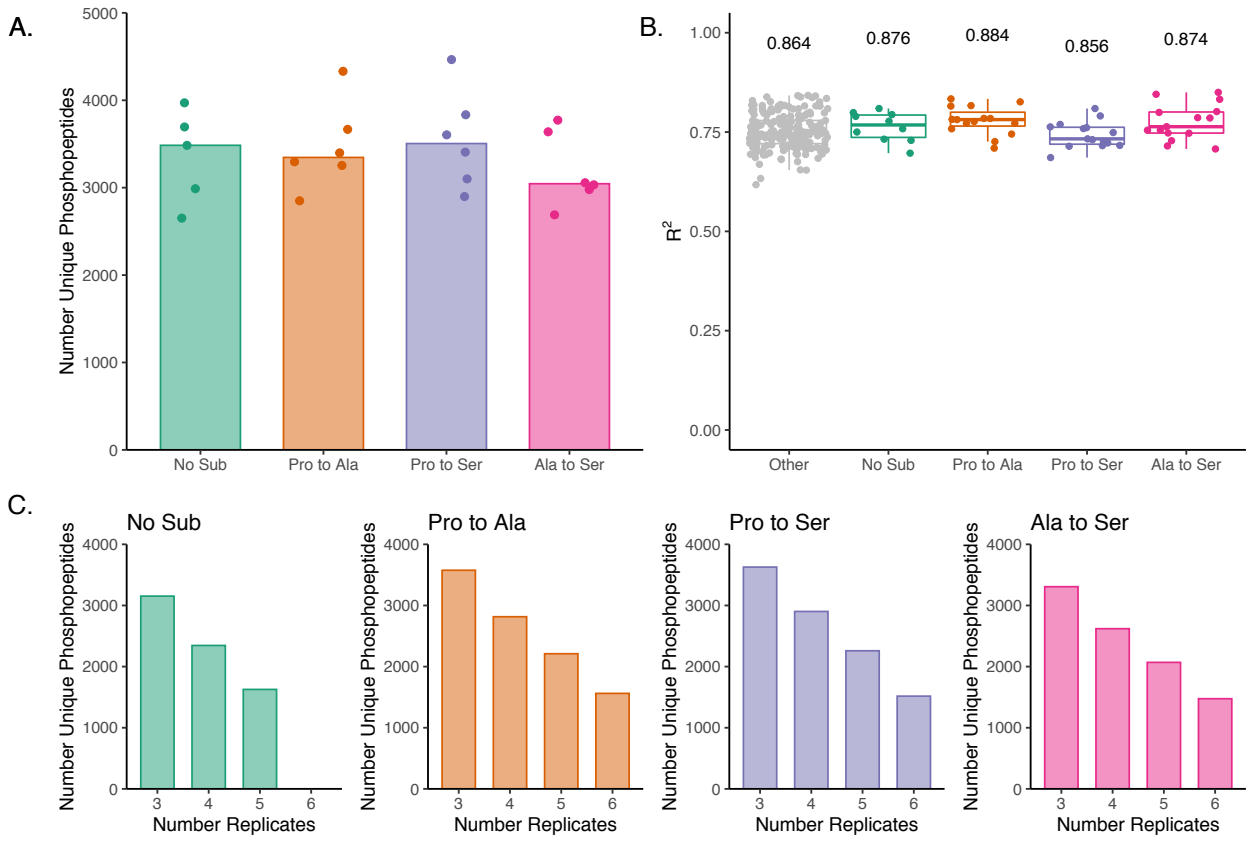


Figure 4.2 Reproducible identification and quantification of phosphopeptides. A. Number of unique phosphopeptides identified in each sample. Yeast strains are indicated on the x-axis, by the type of amino acid substitution introduced. Bars represent median number of identifications per

strain, points represent number identifications in each replicate. B. Reproducibility of phosphopeptide quantification. Boxplot shows distribution of R^2 values, representative of the correlation of quantification between two replicates within the same strain, as indicated on the x-axis. Points represent individual R^2 values for each pairwise combination of samples. "Other" shows R^2 values representative of the correlation of quantifications between replicates of different strains. C. Reproducibility of phosphopeptide identification. Cumulative number of unique phosphopeptides identified in three or more replicates, four or more replicates, etc. as indicated by x-axis.

After removal of the outlier sample, approximately three to four thousand unique phosphopeptides were identified in each sample. To evaluate quantitative reproducibility, I calculated an R^2 value representative of the correlation of phosphopeptide quantification across all pairwise combinations of samples within strains. I found that quantification was highly reproducible, with median R^2 values between 0.85 and 0.88 for comparisons within each strain. I also calculated an R^2 value across all pairwise combinations of samples between the different strains (annotated as "Other"), which I expected could be lower due to true differences in abundance. This value came out to a median of 0.86, a similar median value to those calculated within each strain. To assess reproducibility of phosphopeptide identifications, I determined the number of unique phosphopeptides that were observed in three or more replicates, as these peptides would be applicable to downstream statistical analyses. I found that the majority of identifications were captured within three or more replicates, with over three thousand unique phosphopeptides per strain meeting this requirement (Figure 4.2).

4.3.3 *Mistranslation-driven changes in abundance of phosphopeptides*

After establishing the quality of my dataset, I began to explore the hypothesis that phosphorylation is regulated in response to mistranslation. To identify regulation in abundance of phosphopeptides, I calculated fold change values between each mistranslating strain and the control strain. Phosphopeptides containing amino acid substitutions were removed for this analysis, so as to focus on changes in wildtype phosphopeptides captured in the two strains being compared. I also filtered out phosphopeptides that were identified less than three times in each strain, for each comparison. Next, I used the remaining phosphopeptides to calculate an average fold change between the mistranslating strain and the control strain. Regulated phosphopeptides were prioritized using a Student's t-test, as well as requiring a fold change value of at least 50%, either increasing or decreasing (Figure 4.3).

When comparing the Pro to Ala strain against the No Sub strain, 2818 unique phosphopeptides were assessed. I found that 3.6% of these phosphopeptides (n = 102) were down-regulated, and 3.2% (n = 89) were up-regulated relative to the No Sub strain. In my assessment of the Pro to Ser strain against the No Sub strain, I analyzed a total of 2838 unique phosphopeptides. My analysis identified that 4.5% (n = 128) of these phosphopeptides were down-regulated, and 4.2% (n = 120) were up-regulated. Lastly, I analyzed a total of 2715 unique phosphopeptides between the Ala to Ser strain and the No Sub strain. In the Ala to Ser strain, I found that 4.3% (n = 116) of these phosphopeptides were down-regulated, while 3% (n = 82) were up-regulated (Supplementary Table 4.1).

Interestingly, while the number and percentage of regulated phosphopeptides were similar between the Pro to Ala and Ala to Ser strain, both of these numbers were higher in the Pro to Ser strain. This higher level of phospho-regulation was particularly notable given that the Pro to Ser

strain displayed a more severe growth defect than the Pro to Ala and Ala to Ser strains, which displayed very similar and more modest growth defects. Another trend I observed in this dataset was that in all mistranslating strains, there were a higher number and percentage of down-regulated phosphopeptides than up-regulated phosphopeptides. Collectively, these results prompted further investigation of what types of phosphopeptides were being regulated, and whether there were commonalities or differences in regulation between mistranslating strains.

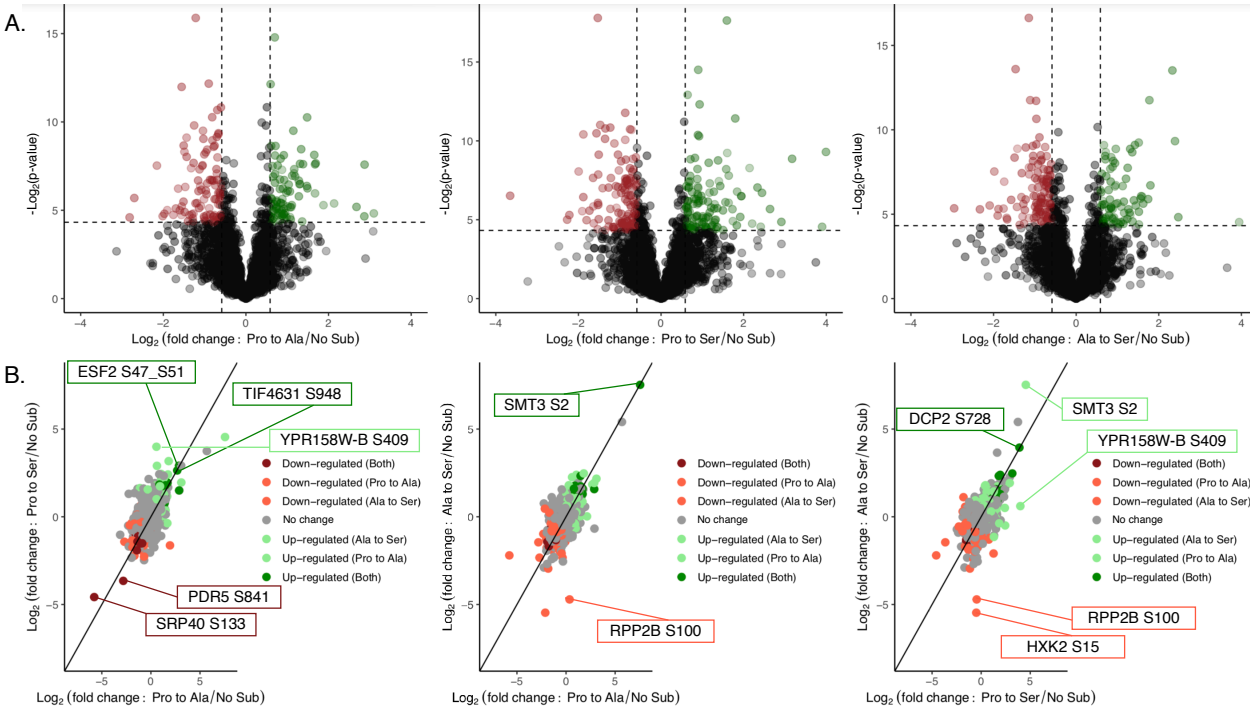


Figure 4.3 Abundance of phosphopeptides is regulated in response to mistranslation. A. Volcano plots showing differential abundance of phosphopeptides, prioritized by Student's t-test and 50% change in abundance. Fold change was calculated between each mistranslating strain and the No Sub strain. The first plot is Pro to Ala, the second plot is Pro to Ser, and the third plot is Ala to Ser. Change in abundance is indicated by Log_2(fold change : Mistranslating Strain/No Sub) on the x-axis, and Log_2(p-value), on the y-axis. X-intercepts annotated by dashed lines show a minimum threshold of 50% increase or decrease in abundance. Y-intercept annotates p-values < 0.05. Red points represent phosphopeptides down-regulated, green points are phosphopeptides up-

regulated. B. Scatterplots comparing fold change values from the volcano plots between mistranslating strains. Color of points indicates whether the phosphopeptide was up- or down-regulated in only one strain, up- or down-regulated in both strains, or displaying no change, as described in legend. Diagonal line represents x = y. The first plot compares phosphopeptides from the Pro to Ala strain on the x-axis and Pro to Ser strain on the y-axis. The second plot compares phosphopeptides from the Pro to Ala strain on the x-axis and Ala to Ser strain on the y-axis. The third plot compares phosphopeptides from the Pro to Ser strain on the x-axis and Ala to Ser strain on the y-axis. Phosphopeptides of interest are annotated by the boxes indicating the protein and phosphorylated residue.

After the volcano plot analysis, I visualized the fold change values (relative to the No Sub strain) from each combination of mistranslating strains on a scatter plot to identify phosphopeptides that were strongly up- or down-regulated in both strains, as well as phosphopeptides that were regulated in response to one amino acid substitution but not the other (Figure 4.3). One example of common regulation between strains was a peptide with phosphorylation at serine 841 from the multi-drug transporter protein PDR5, which was found to be strongly down-regulated in both the Pro to Ala and Pro to Ser strains. Interestingly, removal of this phosphosite by substitution to an alanine residue has been demonstrated to significantly increase resistance to several drugs in yeast (Rahman *et al.*, 2020). In my comparison of phosphopeptides from the Pro to Ala strain and the Ala to Ser strain, I found a peptide containing phosphorylation of serine 2 in ubiquitin-like protein SMT3, a protein with known roles in cell cycle regulation (Sheng *et al.*, 2002).

This analysis also identified phosphopeptides that were strongly regulated in one mistranslating strain, but not another, such as a peptide containing phosphorylation of serine 15 on yeast protein HXK2. This phosphopeptide was strongly down-regulated in the Ala to Ser strain, but did not change in abundance in the Pro to Ser strain. This protein, also known as hexokinase 2, is an enzyme that carries out multiple functions depending on its localization in either the cytoplasm or the nucleus. Phosphorylation of serine 15 has been shown to shuttle this kinase between these subcellular localizations. Overall, this type of analysis allowed us to identify phosphopeptides that are regulated similarly and differently between strains as interesting targets for downstream validation experiments.

4.3.4 *GO analysis of regulated phosphoproteins*

After identifying phosphopeptides that were regulated in response to the different types of mistranslation, I wondered what types of proteins were being phosphorylated and whether there were any trends in Gene Ontology present. I first generated lists of proteins that contained phosphopeptides identified as significantly up-regulated or down-regulated by the volcano plot analysis (Figure 4.3). These protein lists from each mistranslating strain were entered into Venn diagrams to identify which phosphoproteins were commonly regulated across all types of mistranslation, as well as those that were regulated in specific strains. Venn diagrams were generated for all up-regulated proteins and all down-regulated proteins (Figure 4.4). I then performed a GO analysis of the proteins identified as commonly regulated across all strains, as well as those regulated specifically in one strain.

For the lists of proteins with up-regulated phosphopeptides, the Venn Diagram analysis identified 22 phosphoproteins that were up-regulated in all the mistranslating strains. I additionally identified 29 phosphoproteins regulated only in the Pro to Ala strain, 25 phosphoproteins in the

Ala to Ser strain, and 59 phosphoproteins in the Pro to Ser strain. These lists were then analyzed for GO Enrichment using the GOrilla online tool (Eden *et al.*, 2009), in order to identify any significant trends among up-regulated proteins in function, subcellular compartment, or biological process. This analysis identified GO terms enriched both specifically to the mistranslation type, and in common between the mistranslation types (Figure 4.4).

Interestingly, up-regulated phosphorylated proteins unique to each substitution were significantly enriched for different cellular processes. In the strain substituting alanine at proline sites, many of the processes and functions associated with the identified proteins were related to kinases and regulation of phosphorylation. In the strain replacing proline residues with serine, my analysis highlighted metabolic processes as significantly enriched. In the strain with alanine to serine mistranslation, proteins associated with cell cycle and mitosis processes were significantly enriched. GO enrichment was also performed on the set of proteins that was identified as commonly up-regulated among all three mistranslating strains. This common set of proteins was enriched in subcellular components including ribonucleoprotein granules and polysomes, and functions including protein binding, RNA binding and heat shock protein binding.

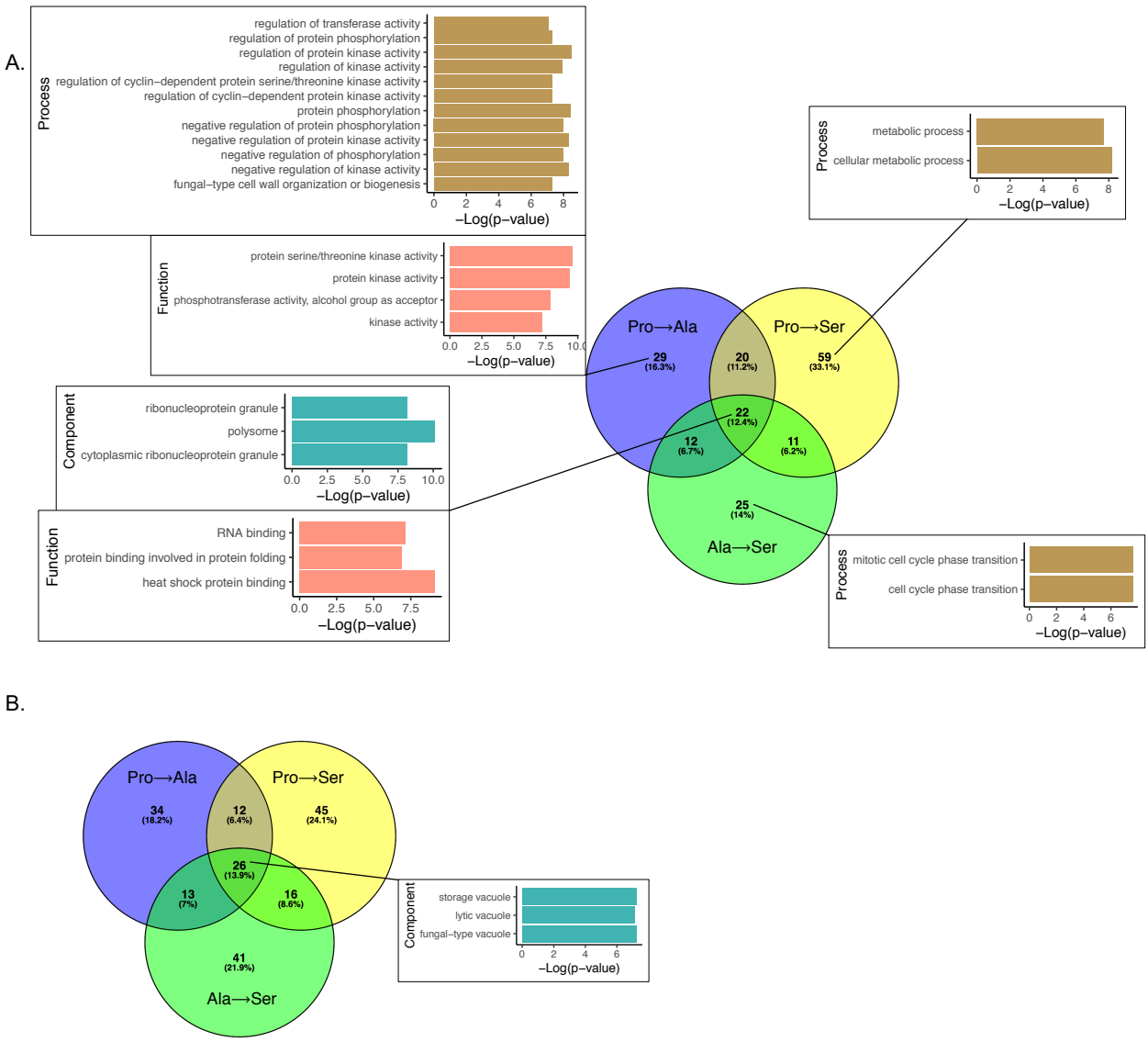


Figure 4.4 Significantly regulated phosphoproteins are enriched for different gene ontologies. A. Venn Diagram generated from lists of phosphoproteins up-regulated in each mistranslating strain with respect to control strain (*Oliveros*, *2007-2015*). GO terms significantly enriched within the specified Venn Diagram sets are shown with -Log(p-value). B. Venn diagram generated from lists of down-regulated phosphoproteins. GO terms were only significantly enriched in the set of proteins commonly regulated in all strains.

The same combined Venn Diagram and GO enrichment analysis was applied to the set of proteins with down-regulated phosphopeptides. In contrast to the up-regulated proteins, no GO terms were significantly enriched in the sets of proteins that were regulated uniquely to one substitution type. However, the set of proteins that was commonly down-regulated in all the strains was enriched for subcellular compartments related to the vacuole. The result that there were few GO terms enriched in the set of down-regulated phosphoproteins was surprising, given that there were more down-regulated than up-regulated phosphopeptides across all mistranslating strains. After observing the lack of significant GO term enrichment in the down-regulated set, I decided to analyze regulation at an amino acid sequence-specific level by investigation of common phosphorylation motifs.

4.3.5 *Changes in phosphorylation of known motifs*

Because the amino acid substitutions within my study involved both proline and serine, I hypothesized that phosphorylation of known phosphomotifs would be increased or decreased with respect to the No Sub strain, particularly those containing proline. To assess this, I compared the distributions of fold change values for phosphopeptides with or without the motif in question. I first focused this analysis on the known proline-directed motif of a phosphoserine or phosphothreonine followed by a proline residue (notated as [S/T]P, Figure 4.5). This motif is targeted by several kinases within the mitogen-activated protein kinase (MAPK) family, including KSS1, HOG1, and FUS3 (Mok *et al.*, 2010). These kinases are critical to signaling pathways controlling a wide array of cellular functions such as filamentous growth, pheromone response and mating, and osmoregulation. Kinases from the cyclin-dependent kinase (CDK) family also target this proline-directed motif, including the Pho80-Pho85 complex and CDC28. The Pho80-Pho85

complex plays an important role in the starvation response and calcium signaling, while CDC28 regulates cell cycle and metabolism.

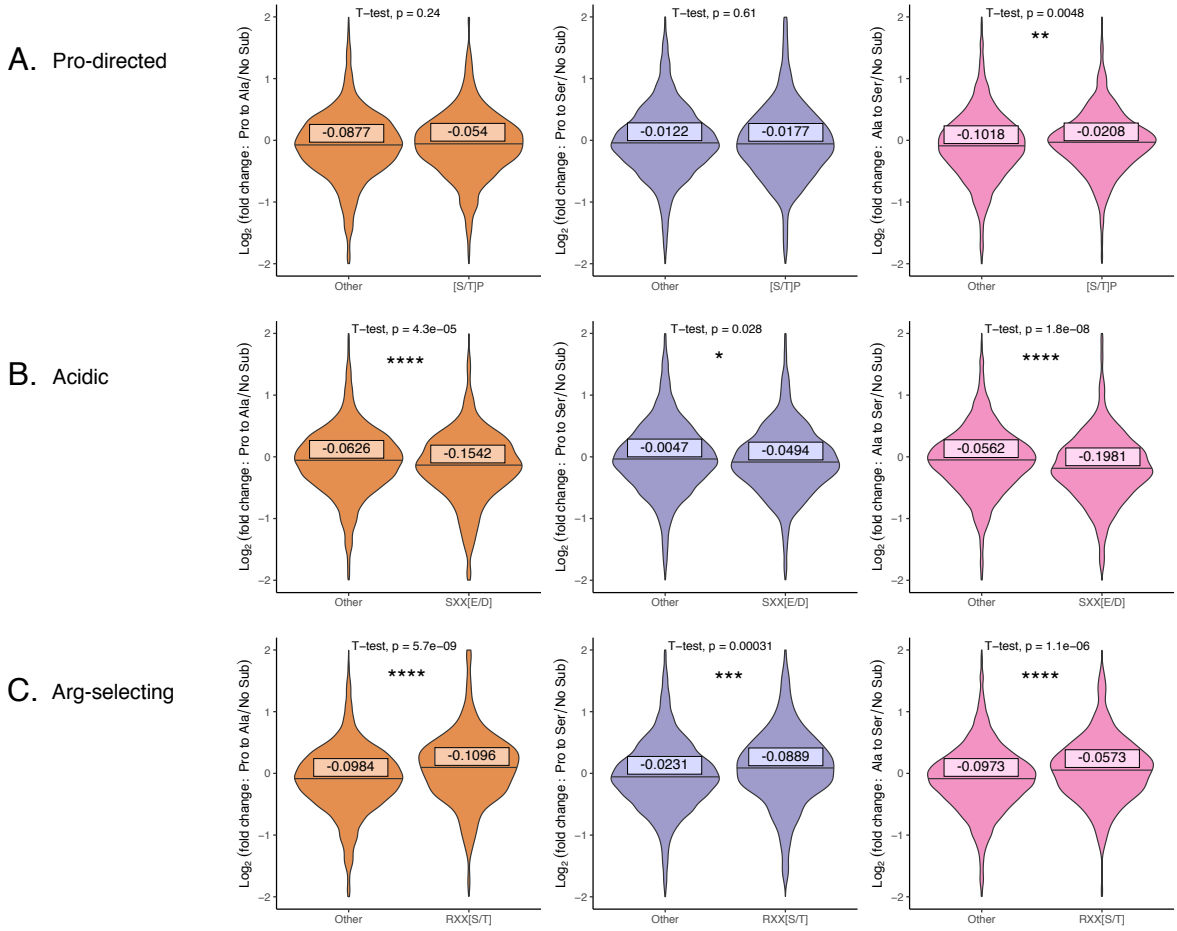


Figure 4.5 Differences in phosphorylation of specific motifs in response to mistranslation. Fold change values were calculated between phosphopeptides in each mistranslating strain with respect to the No Sub strain. Phosphopeptides were separated by whether or not they contained the motif in question, and distributions of fold change values were compared. Groups of phosphopeptides are annotated as "Other" for those that do not contain the phosphomotif, or as containing the phosphomotif specified on the x-axis. Distribution of Log_2(fold change : Mistranslating Strain/No Sub) is visualized by violin plot along the y-axis, the horizontal line indicates 50% quantile, the boxed number is the mean fold change for that distribution. Significant

differences were assigned based on Student's t-test, p-value is shown with asterisks indicating significance (* : $p \leq 0.05$, ** : $p \leq 0.01$, *** : $p \leq 0.001$, **** : $p \leq 0.0001$). The first column of plots in orange shows Pro to Ala (n = 3829 unique phosphopeptides), the second column in purple shows Pro to Ser (n = 3887 unique phosphopeptides), and the third column in pink shows Ala to Ser (n = 3677 unique phosphopeptides). A. Proline-directed phosphomotif [S/T]P, B. Acidic phosphomotif SXX[E/D], C. Arg-selecting phosphomotif RXX[S/T].

Given that proline residues are removed by the substitutions at play in both the Pro to Ala and Pro to Ser strains, I predicted that either of these strains could display decreased phosphorylation of this proline-directed motif. Surprisingly, in both of these strains my analysis showed that there was no significant difference between the distribution of fold changes calculated for phosphopeptides containing the [S/T]P motif and those without the motif (Figure 4.5). In the Ala to Ser strain, however, fold change values were significantly higher in [S/T]P-containing phosphopeptides than all other phosphopeptides, though the difference was modest. The result that the two strains with substitutions involving proline displayed no regulation of this motif could be a preliminary indication that removal of proline residues hinders regulation by proline-directed kinases.

I next examined a known acidic motif, a phosphoserine followed by two unspecified amino acids and then a glutamic acid or aspartic acid (notated as SXX[E/D]). Acidic motifs are known to be targeted by casein kinases such as CKA1, a protein associated with cell growth and proliferation processes (Songyang *et al.*, 1996). Another yeast kinase that targets acidic motifs is CDC5, a protein with important roles in the mitotic cell cycle (Fiol *et al.*, 1988). My analysis showed that

the distribution of fold change in phosphopeptides containing the SXX[E/D] motif was significantly lower than that of all other phosphopeptides, and this trend was observed across all mistranslating strains (Figure 4.5).

Lastly, I examined the arginine-selecting motif, comprised of an arginine residue followed by two unspecified amino acids and a phosphoserine or phosphothreonine. Many yeast kinases target this arginine-selecting motif, including kinases involved in translation such as the regulatory protein PSK2 (Mok *et al.*, 2010). Across all my motif analyses, the impact regarding arginine-selecting motifs was the most pronounced. In all the mistranslating strains, the distribution of fold change values was significantly higher in phosphopeptides containing this motif than those without the motif. This feature of the data warrants further exploration in future studies, and can be used to determine which phosphosites, corresponding phosphoproteins, and respective kinases are driving this trend. I also made these comparisons between each combination of mistranslating strains, rather than comparing to the No Sub strain (Supplementary Figure 4.2). While these comparisons are more difficult to interpret, an interesting finding was that no significant differences in phosphomotif phosphorylation were observed aside from the [S/T]P motif and SXX[E/D] motif, in the comparison between the Pro to Ser and Ala to Ser strains.

4.4 DISCUSSION

Taken together, my analyses revealed several insights into the relationship between phosphorylation and amino acid substitution, at the peptide-level, protein-level, and site-level. My volcano plot analysis revealed that across all mistranslating strains, there were always a higher number and proportion of down-regulated phosphopeptides than up-regulated phosphopeptides. This finding was somewhat surprising, as I expected a strong up-regulation of phosphorylation responding to stress. The greater response in down-regulated phosphopeptides indicates that decreasing phosphorylation of particular sites plays an equally or more important role in responding to amino acid substitutions. I also found that in the Pro to Ser strain, the number and proportion of regulated phosphopeptides were higher than those of the other strains. This result made sense when taken together with the fact that the Pro to Ser strain grew more slowly than the other two mistranslating strains, perhaps because a greater stress is caused by this type of amino acid substitution.

I also used results from the above analysis to identify phosphopeptides that were commonly or differentially regulated between the different mistranslating strains. Several sites were identified as highly up-regulated or down-regulated in the two mistranslating strains being compared, while others were identified as highly regulated in one strain but unchanged in the other. For next steps, the identification of these phosphosites can be used to design validation experiments to understand their functions in response to different types of mistranslation. In another downstream analysis I used the results from the volcano plot to apply GO enrichment to the proteins corresponding to significantly regulated phosphopeptides. I found that terms related to heat shock proteins were enriched in the set of commonly up-regulated phosphoproteins, which made sense given that a misfolding/aggregation stress would occur upon introduction of amino acid substitutions. One

surprising finding from my GO enrichment analysis was that terms involved in phosphorylation were highlighted only in the Pro to Ala strain. This result indicated that this amino acid substitution was potentially the most severe perturbation to phosphorylation, in comparison with the Ala to Ser and Pro to Ser strains. While I anticipated that all our substitution types would impact mistranslation, I would have predicted that the amino acid substitutions introducing random serine residues across the proteome would have had a stronger impact. Another unexpected finding from this analysis was that for the proteins with down-regulated phosphopeptides, no significant enrichment of GO terms was observed specific to each strain. The only significant GO enrichment within the down-regulated phosphoproteins was in the group commonly down-regulated between the three mistranslating strains, where vacuole-related terms were highlighted. This finding was notable given that the vacuole in yeast cells plays a central role in protein degradation, and could suggest that regulation of phosphorylation in vacuolar proteins is a response to amino acid substitutions and the facilitates the degradation necessary to clear out mistranslated proteins. The lack of significant GO enrichment within the phosphoproteins down-regulated uniquely to each amino acid substitution type indicates that down-regulation of phosphorylation is occurring on proteins with diverse functions, processes, or localizations.

From my analyses of specific phosphomotifs, I was also surprised to find no significant changes in proline-directed motifs in the Pro to Ala or Pro to Ser strain. Since both substitutions remove proline residues, I anticipated that this would change or decrease global phosphorylation of these residues, but there was only a very modest difference observed in the third strain, Ala to Ser. For the acidic phosphorylation motif, I observed a significant decrease in phosphorylation across all mistranslating strains. By contrast, I observed significant increases in phosphorylation

across all strains, with respect to the arginine-containing motif. The arginine-selecting motif is targeted by many kinases, and this trend warrants further investigation in future studies.

As a future direction for this study, I have also explored the possibility of generating new phosphosites through amino acid substitutions introducing serine residues across the proteome. Though the findings are preliminary, I do find evidence supporting the formation of new phosphosites in our strains, and future work could focus on improving the confidence in identification and localization of these sites. Given the preliminary evidence of new phosphosites, another downstream application of this system could be as a model for aberrant phosphorylation, which is an area of research with relevance to cancer and other diseases.

4.5 METHODS

4.5.1 *Yeast strains and growth*

Yeast strain BY4742 (*MATα his3Δ0 leu2Δ0 lys2Δ0 ura3Δ0*) is a derivative of S288C (Winzeler *et al.*, 1997). Yeast strains were grown at 30°C in synthetic media supplemented with nitrogenous bases and amino acids. Growth curves were generated by diluting saturated cultures to an OD_{660} of 0.1 and incubating at 30°C with agitation in a BioTek microplate spectrophotometer for 24 hrs. OD_{660} was measured every 15 minutes. Doubling time was calculated using the R package 'growthcurver' (Sprouffske *et al.*, 2016).

4.5.2 *DNA constructs*

URA3-containing centromeric plasmids expressing $tRNA^{Ser}$ (pCB3076), $tRNA^{Pro}_{UGG\ G3:U70}$ (pCB2948) and $tRNA^{Ser}_{UGG,\ G26A}$ (pCB4023) are described in (Berg *et al.*, 2017; Hoffman *et al.*, 2017; Berg *et al.*, 2019). $tRNA^{Ser}_{AGC,\ G26A}$ was engineered using two-step mutagenic PCR with primers XK8036/UG5954 and XK8037/UG5953 (Supplementary Table 4.2) during the first round with pCB3076 as the template. Products from the first reaction were amplified with outside primers UG5953/UG5954 and cloned into YCplac33 as an *Eco*RI fragment to give pCB4632.

4.5.3 *MS sample preparation*

Yeast strains were grown by diluting saturated cultures to an OD_{660} of 0.1 and incubating at 30°C with agitation, until they reached an OD_{660} of 1.0. Cultures were harvested by centrifuging and decanting spent media. Cell pellet was washed once by resuspending in ice-cold water, centrifuging and decanting water. Pellets were snap frozen in liquid nitrogen and stored at -80°C. Pellets were thawed on ice and resuspended in a lysis buffer (8 M Urea, 150 mM NaCl, 100 mL Tris pH 8.2). Cells were lysed by bead beating with 0.5 mm-diameter zirconia/silica beads for four 1-minute cycles, resting 1 minute on ice in between cycles. Lysate was cleared of debris by

centrifugation, and protein concentration in cleared lysate was quantified using the Pierce™ BCA Protein Assay Kit. Lysate was reduced by adding dithiothreitol (DTT) to a final concentration of 5mM and incubating at 55°C with agitation for 30 minutes. After reduction, sample was alkylated by adding iodoacetamide (IAA) to a final concentration of 15mM and incubating at room temperature for 30 minutes. Alkylation reaction was quenched by adding DTT to a final concentration of 5mM. Reduced and alkylated samples were stored at -20°C. Whole proteome and phosphoproteome samples were prepared by the R2-P1 and R2-P2 methods (Leutert *et al.*, 2019).

4.5.4 *MS methods*

Phosphoproteome samples were subjected to a 90-minute LC-MS/MS run on a Thermo Easy-nLC online with an Orbitrap Exploris 480 mass spectrometer. Samples were loaded onto a 100-µm internal diameter (ID) x 3-cm trap column filled with 3-µm Reprosil C18 beads (Dr. Maisch GmbH). Phosphopeptides were separated via a 65-minute gradient of increasing concentration of 80% acetonitrile in 0.125% formic acid on a 100-µm ID x 30-cm analytical column filled with 1.9-µm Reprosil C18 beads (Dr. Maisch). An MS1 survey scan was used to perform data-dependent acquisition for the top N most abundant precursors in a 3-second duty cycle. We used an MS1 scan over the range of 350 – 1500 m/z with the following parameters: 120,000 resolution, 300% normalized automated gain control (AGC), and auto-injection time (IT). Top N precursors were isolated within a 2.0 m/z window and subjected to 30 normalized collision energy (NCE). MS/MS scans were performed at 30,000 resolution, 100% normalized AGC, and auto-IT. Targeted precursors were placed on dynamic exclusion for 45 seconds, to prevent redundant sampling.

4.5.5 *MS data analysis*

MS data were searched with Comet (2015.02.rev.5; (Eng *et al.*, 2013) against the *Saccharomyces cerevisiae* protein sequence database (downloaded from Uniprot in July 2014) using LysC enzyme specificity and allowing for two missed cleavages. For the whole proteome data, the Comet configuration file allowed for the variable modifications of methionine oxidation, protein N-terminal acetylation and the amino acid substitution relevant to each strain (Pro to Ala, Pro to Ser, or Ala to Ser). For the phosphoproteome data, the Comet configuration file included the same variable modifications in addition to phosphorylation on serine, tyrosine, or threonine residues. Carbamidomethylation of cysteine was included as static a modification. Other parameters included 20 ppm tolerance for precursor mass and 0.02 Dalton tolerance for fragment ions. Percolator (Käll *et al.*, 2007) was used to filter results to a PSM-level FDR of 1% estimated via the target-decoy method (Elias *et al.*, 2007). PSM-level quantification values were aggregated to peptide-level for fold change calculations and motif analyses.

4.6 CONTRIBUTIONS

Bianca Ruiz and Matthew D. Berg prepared all whole proteome and phosphoproteome samples together. Matthew D. Berg engineered yeast strains, and provided doubling time and percent mistranslation data, as illustrated in Figure 4.1. Bianca Ruiz performed phosphoproteome data analyses, with assistance by Ian Smith.

Chapter 5. NEXT STEPS FOR HIGH-THROUGHPUT STUDIES OF PROTEINS

5.1 GENERATING PROTEIN VARIANTS WITH NONCANONICAL AMINO ACIDS

In chapter 2, I describe the work I have done to develop noncanonical amino acids for generation of protein variants across the proteome. I characterized over twenty noncanonical compounds for both toxicity and incorporation into proteins in *E. coli.* While several of the ncAAs in our panel were known to incorporate in the literature, others were under-studied. I foresee many possible applications. Within our lab, this study provides a resource for experiments that apply these compounds, particularly within Miró assays where a proteome-wide selection is applied. One potential functional selection uses the thermal proteome profiling (TPP) method (Savitski *et al.*, 2014). This method was originally designed to identify drug targets, by applying a gradient of temperature conditions to the proteomes of cells treated with and without a drug of interest for characterization to identify any proteins that were stabilized upon treatment with the drug.

In our lab, TPP has been applied to proteomes with amino acid substitutions facilitated by both ncAAs and genetic methods to identify destabilizing amino acid substitutions. While we have conducted TPP with several ncAAs in our lab, most of the compounds screened in Chapter 2 await experimental selections. The ncAA incorporation data captured in Chapter 2 can be useful to a TPP experiment using any of these compounds in *E. coli*, by providing data on the concentration of ncAA required to reach a desired incorporation level. Many other selections could be applied as well; for example, my proteome-wide pH selection would be particularly interesting to those ncAAs that change the charge state of the residue. Nevertheless, a TPP screen of all the ncAAs in our panel would be a great start for generating hypotheses, finding interesting substitutions for validation, and further developing the computational tools we use to analyze these data.

5.2 FUNCTIONAL SELECTIONS FOR HIGH-THROUGHPUT STUDIES OF PROTEINS

In Chapter 3 I developed a new functional selection applicable to the proteome. When designing a proteome-wide selection, the key is to come up with a selection that will be generalized enough to reveal properties across all types of proteins. For this purpose, a great place to start is by implementing selections that have been previously used to study individual purified proteins. The advantages of this approach are that tried-and-true selections on biochemical properties, such as thermal stability assays, are relatively straightforward and low-cost to apply, and that they are guaranteed to provide information about virtually any protein.

Inspired by the elegance and power of a thermal selection to study entire proteomes with TPP, I wondered if I could apply another established selection that had the potential to reveal information about all proteins – pH gradients. Protein chemists have used pH to study the properties of proteins for decades (Womack *et al.*, 1979; Kumar *et al.*, 2004). Many of these studies investigated the activities or other properties of purified proteins, while others focused on more chemical properties such as solubility. The behavior of proteins in different pH environments has far-reaching biological relevance, and treatment of proteins with a gradient of pH was relatively reasonable to design and implement in the lab at a proteome scale.

In order to pilot this method and determine if a proteome-scale selection was possible, I paralleled existing TPP protocols as they had been implemented in our lab and replaced the temperature selection with a pH selection. Keeping as many steps as possible constant between TPP and the new pH method ensured that I could compare the methods and determine what types of different information pH would reveal. Similar to TPP, after applying the selection, the soluble fraction of the proteome was TMT-labeled, the samples from the ten conditions were combined, and analyzed by mass spectrometry. One of the main differences between these methods

manifested at the data analysis step. TPP typically results in a sigmoidal curve, whereas most proteins show a bell-shaped solubility curve against pH. I found that spline fitting rather than fitting to a sigmoidal worked best for this type of data. In summary, this project resulted in experimental and computational tools for a proteome-scale pH selection, capable of producing solubility profiles for the entire proteome.

I found that the solubility of proteins in specific ranges of pH often reflected the pH of their microenvironments within the cell. An example of this finding was that proteins from one of the most alkaline cellular environments, the mitochondria, were more soluble at a higher pH range than the rest of the proteome. Similarly, I found that proteins from the vacuole, one of the most acidic organelles, were able to remain soluble at the lowest pH conditions in my experiment. These findings suggest that proteins have intrinsic properties that allow them to fold and function in such different conditions.

Next steps I envision for this dataset would be to compare known properties of proteins from different organelles, to assess what features are enabling proteins to remain folded, soluble, and functioning in diverse environments. Some interesting properties to explore would be amino acid composition, distribution of structural features such as disordered regions, beta sheets and alpha helices, and complexing to proteins or other biomolecules such as nucleic acids. This dataset would also be a good candidate for machine learning, in order to analyze all of these features together.

Moreover, this method has been demonstrated to generate rich data about protein solubility across the entire proteome and can be applied to high throughput studies of amino acid substitutions. The pH selection would be particularly interesting for substitutions that change the charge or hydrophobicity of a residue. Substitution data could also be re-integrated with the

original datasets, to learn more about how proteins remain soluble in diverse environments, and how amino acid substitutions can change this ability.

5.3 DEVELOPING BIOLOGICAL SELECTIONS TO UNDERSTAND AMINO ACID SUBSTITUTIONS IN PROTEINS

Proteome-wide screens of amino acid substitutions can be performed in the context of more sophisticated and fine-tuned selections. Interesting data can be taken from these experiments with relevance to disease or understanding entire pathways, or networks within a cell or an organism. In Chapter 4, I describe a project that produced such data, by determining whether or not global signaling via phosphorylation is disrupted by proteome-wide amino acid substitutions. Phosphorylation across the proteome is inarguably a very important post-translational modification. Phosphorylation of a single amino acid can re-direct signaling pathways or change the conformation of a protein involved in a cellular structure or function. In a single experiment, I measured how the phosphoproteome responds to random substitutions that were likely disrupting structure and function, and how substitutions adding potential artificial phosphosubstrates or disturbing phosphomotifs impact normal phosphorylation. While interesting biological insights can be extracted from this data, there are some challenges associated with data analysis and interpretation, making this part still a work in progress. Overall, the project served to continue building upon methods for proteome-scale selections, provided rich data that revealed interesting insights about phosphorylation, and helped generate hypotheses for downstream validation experiments. I am refining the analysis pipelines for this selection data and foresee that ideas and methods from this project can inform new proteome-wide studies of amino acid substitutions.

Examination of the phosphoproteome interactions with amino acid substitutions requires a nuanced and critical interpretation, because I essentially let natural biological processes carry out the selection. Instead of applying heat or chemical conditions, the endogenous signaling system applies the selection. While this type of readout provides more direct indications of the biology,

the open-endedness of such a selection also limits the conclusions I can make from the proteome-wide screen alone. In next steps for this project, I can design experiments to validate the impact of amino acid substitutions on phosphorylation, for example by hard coding a substitution observed from the proteomic data and expressing the protein variant for a more precise readout. More specifically, a validation could involve engineering a protein to have a specific alanine to serine substitution where I had observed phosphorylation and verify if that phosphorylation occurs in the absence of global mistranslation stress. Kinase-substrate interactions can also be examined more closely by *in vitro* phosphorylation assays, to determine if specific amino acid substitutions truly disrupt phosphorylation of a particular motif in a reproducible manner.

Another selection is probing the impact of amino acid substitutions on protein-protein interactions, for which the Villén lab has conducted some studies using affinity purification. To expand these earlier studies, a complementary approach, chemical crosslinking proteomics, could be applied to proteome samples with or without random sites of amino acid substitution, and comparisons of these conditions could identify sites that are required for proper interactions with complexes or other binding partners. A similar application of crosslinking could be to explore protein interactions with DNA or RNA, and how those interactions are changed by proteome-wide amino acid substitutions. Taken together, these types of projects begin to paint a more comprehensive picture of the intrinsic properties of proteins, and how their many structures and functions come to be from the same twenty building blocks.

5.4 CLOSING REMARKS

When I think about proteins, I consider just how important they are to the normal and healthy functioning of a molecular interaction, subcellular localization, cellular phenotype, and ultimately an entire organism. I find it remarkable that a single amino acid substitution can cause a life-long disease for a person or dictate the major decision to undergo a prophylactic surgery. Amino acid substitutions in proteins can determine whether a couple decides to have a child given the presence of a hereditary disease in a family, or completely alter the course of a global pandemic. Our ability as scientists to predict, screen, and understand the effects of protein variants can result in saving lives, and this impact has been demonstrated time and time again.

As I have learned about proteins through my research, I have really marveled at the diversity of these molecules that make all of biology happen, all built by the same twenty compounds. How is it possible that combinations of the same twenty amino acids can make up a transmissible prion that causes a fatal disease in cattle, as well as an enzyme that oxidizes trace metals in the environment? Proteins can catalyze chemical reactions more efficiently than expert chemists trying to synthesize a particular compound in the lab, and so we turn to them to do this job orders of magnitude faster than we had even come close to. Every day, we use proteins to synthesize drugs, treat our food, facilitate bioremediation of environmental pollutants, and light up molecular biology in a fluorescent palette of colors. How is it that evolution has resulted in the existence of these incredibly sophisticated and diverse molecules, and how do proteins do what they do? Indeed, it is our duty as scientists to answer these questions, as our understanding can positively impact the world through the development of new medicines, cleaning up the environment, and in other ways we have not yet imagined.

FUNDING

This work was supported by funding from the National Science Foundation Graduate Research Fellowship (DGE-1256082), the W.M. Keck Foundation, and NIH/NIGMS, grant R35 GM119536.

BIBLIOGRAPHY

Audain, E., Ramos, Y., Hermjakob, H., Flower, D. R. and Perez-Riverol, Y. (2016) 'Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences', *Bioinformatics,* 32(6), pp. 821-7.

Balch, W. E., Morimoto, R. I., Dillin, A. and Kelly, J. W. (2008) 'Adapting proteostasis for disease intervention', *Science,* 319(5865), pp. 916-9.

Bearne, S. L. (2014) 'Illustrating the Effect of pH on Enzyme Activity Using Gibbs Energy Profiles', *J. Chem. Educ.,* 91(1), pp. 84-90.

Berg, M. D., Hoffman, K. S., Genereaux, J., Mian, S., Trussler, R. S., Haniford, D. B., O'Donoghue, P. and Brandl, C. J. (2017) 'Evolving Mistranslating tRNAs Through a Phenotypically Ambivalent Intermediate in', *Genetics,* 206(4), pp. 1865-1879.

Berg, M. D., Zhu, Y., Genereaux, J., Ruiz, B. Y., Rodriguez-Mias, R. A., Allan, T., Bahcheli, A., Villén, J. and Brandl, C. J. (2019) 'Modulating Mistranslation Potential of tRNA', *Genetics,* 213(3), pp. 849-863.

Blokhuis, A. M., Groen, E. J., Koppers, M., van den Berg, L. H. and Pasterkamp, R. J. (2013) 'Protein aggregation in amyotrophic lateral sclerosis', *Acta Neuropathol,* 125(6), pp. 777-94.

Bouchecareilh, M. and Balch, W. E. (2011) 'Proteostasis: a new therapeutic paradigm for pulmonary disease', *Proc Am Thorac Soc,* 8(2), pp. 189-95.

Brett, C. L., Donowitz, M. and Rao, R. (2006) 'Does the proteome encode organellar pH?', *FEBS Lett,* 580(3), pp. 717-9.

Browne, D. T., Kenyon, G. L. and Hegeman, G. D. (1970) 'Incorporation of monofluorotryptophans into protein during the growth of Escherichia coli', *Biochemical and Biophysical Research Communications,* 39, pp. 13 - 19.

Chait, B. T. (2006) 'Chemistry. Mass spectrometry: bottom-up or top-down?', *Science,* 314(5796), pp. 65-6.

Chavez, J. D., Keller, A., Wippel, H. H., Mohr, J. P. and *, J. E. B. (2021) 'Multiplexed Cross-Linking with Isobaric Quantitative Protein Interaction Reporter Technology', *Analytical Chemistry*.

Cowei, D. B., Cohen, G. N., Bolton, E. T. and De Robichon-Szulmajster, H. (1959) 'Amino acid analog incorporation into bacterial proteins', *Biochim Biophys Acta,* 34, pp. 39-46.

Dahlman, D. L. and Rosenthal, G. A. (1975) 'Non-protein aminoacid-insect interactions--I. Growth effects and symptomology of L-canavanine consumption by tobacco hornworm, Manduca sexta (L.)', *Comp Biochem Physiol A Comp Physiol,* 51(1A), pp. 33-6.

Dai, L. and Gao, G. F. (2021) 'Viral targets for vaccines against COVID-19', *Nat Rev Immunol,* 21(2), pp. 73-82.

Edbauer, D., Cheng, D., Batterton, M. N., Wang, C. F., Duong, D. M., Yaffe, M. B., Peng, J. and Sheng, M. (2009) 'Identification and characterization of neuronal mitogen-activated protein kinase substrates using a specific phosphomotif antibody', *Mol Cell Proteomics,* 8(4), pp. 681-95.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) 'GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists', *BMC Bioinformatics,* 10, pp. 48.

Elias, J. E. and Gygi, S. P. (2007) 'Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry', *Nat Methods,* 4(3), pp. 207-14.

Eng, J. K., Jahan, T. A. and Hoopmann, M. R. (2013) 'Comet: an open-source MS/MS sequence database search tool', *Proteomics,* 13(1), pp. 22-4.

Eraso, P. and Gancedo, C. (1987) 'Activation of yeast plasma membrane ATPase by acid pH during growth', *FEBS Lett,* 224(1), pp. 187-92.

Feng, L., Chan, W. W., Roderick, S. L. and Cohen, D. E. (2000) 'High-level expression and mutagenesis of recombinant human phosphatidylcholine transfer protein using a synthetic gene: evidence for a C-terminal membrane binding domain', *Biochemistry,* 39(50), pp. 15399-409.

Ferguson, J. (1964) 'Multivariable Curve Interpolation', *Journal of the ACM,* 11(2), pp. 221-228.

Fernandes, V. C., Golubeva, V. A., Di Pietro, G., Shields, C., Amankwah, K., Nepomuceno, T. C., de Gregoriis, G., Abreu, R. B. V., Harro, C., Gomes, T. T., Silva, R. F., Suarez-Kurtz, G., Couch, F. J., Iversen, E. S., Monteiro, A. N. A. and Carvalho, M. A. (2019) 'Impact of amino acid substitutions at secondary structures in the BRCT domains of the tumor suppressor BRCA1: Implications for clinical annotation', *J Biol Chem,* 294(15), pp. 5980-5992.

Fields, S. and Song, O. (1989) 'A novel genetic system to detect protein-protein interactions', *Nature,* 340(6230), pp. 245-6.

Fiol, C. J., Haseman, J. H., Wang, Y. H., Roach, P. J., Roeske, R. W., Kowalczuk, M. and DePaoli-Roach, A. A. (1988) 'Phosphoserine as a recognition determinant for glycogen synthase kinase-3: phosphorylation of a synthetic peptide based on the G-component of protein phosphatase-1', *Arch Biochem Biophys,* 267(2), pp. 797-802.

Fowden, L. (1956) 'Azetidine-2-carboxylic acid: a new cyclic imino acid occurring in plants', *Biochem J,* 64(2), pp. 323-32.

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D. and Fields, S. (2010) 'High-resolution mapping of protein sequence-function relationships', *Nat Methods,* 7(9), pp. 741-6.

Fowler, D. M. and Fields, S. (2014) 'Deep mutational scanning: a new style of protein science', *Nat Methods,* 11(8), pp. 801-7.

Fowler, D. M., Stephany, J. J. and Fields, S. (2014) 'Measuring the activity of protein variants on a large scale using deep mutational scanning', *Nat Protoc,* 9(9), pp. 2267-84.

Franken, H., Mathieson, T., Childs, D., Sweetman, G. M., Werner, T., Tögel, I., Doce, C., Gade, S., Bantscheff, M., Drewes, G., Reinhard, F. B., Huber, W. and Savitski, M. M. (2015) 'Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry', *Nat Protoc,* 10(10), pp. 1567-93.

Gilles, A. M., Marlière, P., Rose, T., Sarfati, R., Longin, R., Meier, A., Fermandjian, S., Monnot, M., Cohen, G. N. and Bârzu, O. (1988) 'Conservative replacement of methionine by norleucine in Escherichia coli adenylate kinase', *J Biol Chem,* 263(17), pp. 8204-9.

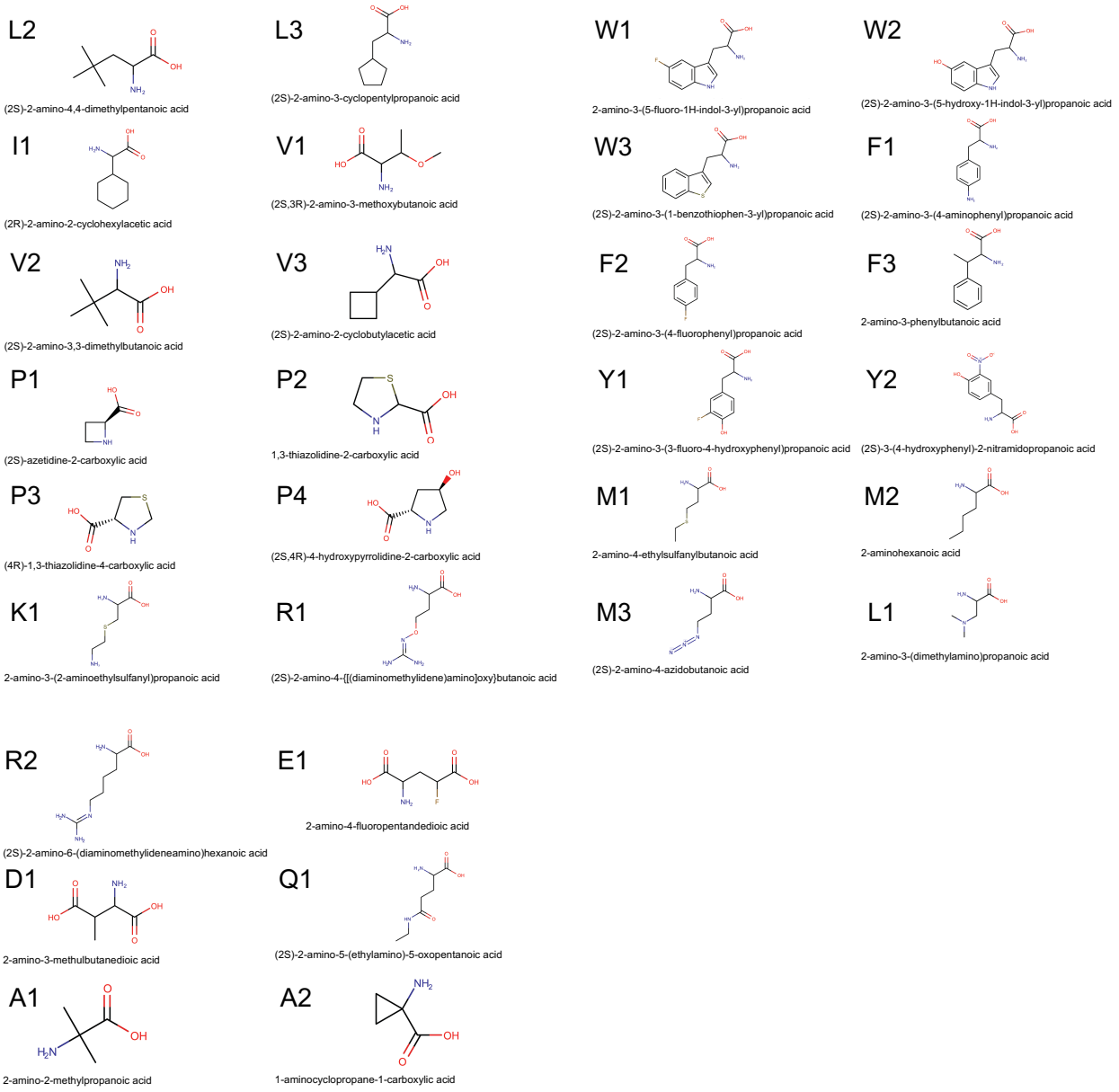
Gohar, A. V., Cao, R., Jenkins, P., Li, W., Houston, J. P. and Houston, K. D. (2013) 'Subcellular localization-dependent changes in EGFP fluorescence lifetime measured by time-resolved flow cytometry', *Biomed Opt Express,* 4(8), pp. 1390-400.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. and King, M. C. (1990) 'Linkage of early-onset familial breast cancer to chromosome 17q21', *Science,* 250(4988), pp. 1684-9.

Handford, J. I., Ize, B., Buchanan, G., Butland, G. P., Greenblatt, J., Emili, A. and Palmer, T. (2009) 'Conserved network of proteins essential for bacterial viability', *J Bacteriol,* 191(15), pp. 4732-49.
Hashikawa, N. and Sakurai, H. (2004) 'Phosphorylation of the yeast heat shock transcription factor is implicated in gene-specific activation dependent on the architecture of the heat shock element', *Mol Cell Biol,* 24(9), pp. 3648-59.
Herrick, J. B. (1910) 'Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. 1910', *Yale J Biol Med,* 74(3), pp. 179-84.
Hoffman, K. S., Berg, M. D., Shilton, B. H., Brandl, C. J. and O'Donoghue, P. (2017) 'Genetic selection for mistranslation rescues a defective co-chaperone in yeast', *Nucleic Acids Res,* 45(6), pp. 3407-3421.
INGRAM, V. M. (1958) 'Abnormal human haemoglobins. I. The comparison of normal human and sickle-cell haemoglobins by fingerprinting', *Biochim Biophys Acta,* 28(3), pp. 539-45.
Irvine, G. B., El-Agnaf, O. M., Shankar, G. M. and Walsh, D. M. (2008) 'Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases', *Mol Med,* 14(7-8), pp. 451-64.
Kaushik, S. and Cuervo, A. M. (2015) 'Proteostasis and aging', *Nat Med,* 21(12), pp. 1406-15.
Kim, H., Kim, S., Kim, D. and Yoon, S. H. (2020) 'A single amino acid substitution in aromatic hydroxylase (HpaB) of Escherichia coli alters substrate specificity of the structural isomers of hydroxyphenylacetate', *BMC Microbiol,* 20(1), pp. 109.
Krakauer, J., Long, Y., Kolbert, A., Thanedar, S. and Southard, J. (2015) 'Presence of L-canavanine in Hedysarum alpinum seeds and its potential role in the death of Chris McCandless', *Wilderness Environ Med,* 26(1), pp. 36-42.
Kumar, D. P., Tiwari, A. and Bhat, R. (2004) 'Effect of pH on the stability and structure of yeast hexokinase A. Acidic amino acid residues in the cleft region are critical for the opening and the closing of the structure', *J Biol Chem,* 279(31), pp. 32093-9.
Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. and MacCoss, M. J. (2007) 'Semi-supervised learning for peptide identification from shotgun proteomics datasets', *Nat Methods,* 4(11), pp. 923-5.
Leutert, M., Rodríguez-Mias, R. A., Fukuda, N. K. and Villén, J. (2019) 'R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies', *Mol Syst Biol,* 15(12), pp. e9021.
Li, X., You, R., Wang, X., Liu, C., Xu, Z., Zhou, J., Yu, B., Xu, T., Cai, H. and Zou, Q. (2016) 'Effectiveness of Prophylactic Surgeries in BRCA1 or BRCA2 Mutation Carriers: A Meta-analysis and Systematic Review', *Clin Cancer Res,* 22(15), pp. 3971-81.
Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J. and Fowler, D. M. (2018) 'Multiplex assessment of protein variant abundance by massively parallel sequencing', *Nat Genet,* 50(6), pp. 874-882.
Mintseris, J. and Gygi, S. P. (2020) 'High-density chemical cross-linking for modeling protein interactions', *Proc Natl Acad Sci U S A,* 117(1), pp. 93-102.
Mok, J., Kim, P. M., Lam, H. Y., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J. L., Sheu, Y. J., Sassi, H. E., Sopko, R., Chan, C. S., De Virgilio, C., Hollingsworth, N. M., Lim, W. A., Stern, D. F., Stillman, B., Andrews, B. J., Gerstein, M. B., Snyder, M. and Turk, B. E. (2010) 'Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs', *Sci Signal,* 3(109), pp. ra12.
Newman, J. (2004) 'Novel buffer systems for macromolecular crystallization', *Acta Crystallogr D Biol Crystallogr,* 60(Pt 3), pp. 610-2.
Nicaud, J. M., Mackman, N. and Holland, I. B. (1986) 'Current status of secretion of foreign proteins by microorganisms', *Journal of Biotechnology,* 3, pp. 255 - 270.
Old, J. M. and Jones, D. S. (1977) 'The aminoacylation of transfer ribonucleic acid', *Biochem. J.,* 165, pp. 367 - 373.
Oliveros, J. C. (*2007-2015*) *Venny. An interactive tool for comparing lists with Venn's diagrams.* Available at: https://bioinfogp.cnb.csic.es/tools/venny/index.html.
Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. and Mann, M. (2002) 'Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics', *Mol Cell Proteomics,* 1(5), pp. 376-86.
Orij, R., Brul, S. and Smits, G. J. (2011) 'Intracellular pH is a tightly controlled signal in yeast', *Biochim Biophys Acta,* 1810(10), pp. 933-44.
Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W. and Balch, W. E. (2009) 'Biological and chemical approaches to diseases of proteostasis deficiency', *Annu Rev Biochem,* 78, pp. 959-91.
Pratt, E. A. and Ho, C. (1975) 'Incorporation of fluorotryptophans into proteins of escherichia coli', *Biochemistry,* 14(13), pp. 3035-40.
Rahman, H., Rudrow, A., Carneglia, J., Joly, S. S. P., Nicotera, D., Naldrett, M., Choy, J., Ambudkar, S. V. and Golin, J. (2020) 'Nonsynonymous Mutations in Linker-2 of the Pdr5 Multidrug Transporter Identify a New RNA Stability Element', *G3 (Bethesda),* 10(1), pp. 357-369.
Ramos, Y., Gutierrez, E., Machado, Y., Sánchez, A., Castellanos-Serra, L., González, L. J., Fernández-de-Cossio, J., Pérez-Riverol, Y., Betancourt, L., Gil, J., Padrón, G. and Besada, V. (2008) 'Proteomics based on peptide fractionation by SDS-free PAGE', *J Proteome Res,* 7(6), pp. 2427-34.
Rappsilber, J., Mann, M. and Ishihama, Y. (2007) 'Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips', *Nat Protoc,* 2(8), pp. 1896-906.
Reitz, C., Fan, Q. and Neubauer, P. (2018) 'Synthesis of non-canonical branched-chain amino acids in Escherichia coli and approaches to avoid their incorporation into recombinant proteins', *Curr Opin Biotechnol,* 53, pp. 248-253.
Rosano, G. L. and Ceccarelli, E. A. (2014) 'Recombinant protein expression in Escherichia coli: advances and challenges', *Front Microbiol,* 5, pp. 172.
Rubenstein, E., Zhou, H., Krasinska, K. M., Chien, A. and Becker, C. H. (2006) 'Azetidine-2-carboxylic acid in garden beets (Beta vulgaris)', *Phytochemistry,* 67(9), pp. 898-903.
Saleh, A. M., Wilding, K. M., Calve, S., Bundy, B. C. and Kinzer-Ursem, T. L. (2019) 'Non-canonical amino acid labeling in proteomics and biotechnology', *J Biol Eng,* 13, pp. 43.
Samardzic, K. and Rodgers, K. J. (2019) 'Cell death and mitochondrial dysfunction induced by the dietary non-proteinogenic amino acid L-azetidine-2-carboxylic acid (Aze)', *Amino Acids,* 51(8), pp. 1221-1232.
Saunders, M., Wishnia , A. and Kirkwood, J. (1957) 'THE NUCLEAR MAGNETIC RESONANCE SPECTRUM OF RIBONUCLEASE', 79, pp. 3289–3290.
Savitski, M. M., Reinhard, F. B., Franken, H., Werner, T., Savitski, M. F., Eberhard, D., Martinez Molina, D., Jafari, R., Dovega, R. B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M. and Drewes, G. (2014) 'Tracking cancer drugs in living cells by thermal profiling of the proteome', *Science,* 346(6205), pp. 1255784.

Sharma, K., D'Souza, R. C., Tyanova, S., Schaab, C., Wiśniewski, J. R., Cox, J. and Mann, M. (2014) 'Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling', *Cell Rep,* 8(5), pp. 1583-94.
Sheng, W. and Liao, X. (2002) 'Solution structure of a yeast ubiquitin-like protein Smt3: the role of structurally less defined sequences in protein-protein recognitions', *Protein Sci,* 11(6), pp. 1482-91.
Shi, P. Y., Plante, J., Liu, Y., Liu, J., Xia, H., Johnson, B., Lokugamage, K., Zhang, X., Muruato, A., Zou, J., Fontes-Garfias, C., Mirchandani, D., Scharton, D., Kalveram, B., Bilello, J., Ku, Z., An, Z., Freiberg, A., Menachery, V., Xie, X., Plante, K. and Weaver, S. (2020) 'Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility', *Res Sq*.
Smith, I. R., Hess, K. N., Bakhtina, A. A., Valente, A. S., Rodríguez-Mias, R. A. and Villén, J. (2021) 'Identification of phosphosites that alter protein thermal stability', *Nat Methods,* 18(7), pp. 760-762.
Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J., DeMaggio, A. J., Hoekstra, M. F., Blenis, J., Hunter, T. and Cantley, L. C. (1996) 'A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1', *Mol Cell Biol,* 16(11), pp. 6486-93.
Sprouffske, K. and Wagner, A. (2016) 'Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves', *BMC Bioinformatics,* 17, pp. 172.
Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., Fowler, D. M., Parvin, J. D., Shendure, J. and Fields, S. (2015) 'Massively Parallel Functional Analysis of BRCA1 RING Domain Variants', *Genetics,* 200(2), pp. 413-22.
Sze, S. K., Ge, Y., Oh, H. and McLafferty, F. W. (2002) 'Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue', *Proc Natl Acad Sci U S A,* 99(4), pp. 1774-9.
Tischkowitz, M., Hamel, N., Carvalho, M. A., Birrane, G., Soni, A., van Beers, E. H., Joosse, S. A., Wong, N., Novak, D., Quenneville, L. A., Grist, S. A., Nederlof, P. M., Goldgar, D. E., Tavtigian, S. V., Monteiro, A. N., Ladias, J. A., Foulkes, W. D. and kConFab (2008) 'Pathogenicity of the BRCA1 missense variant M1775K is determined by the disruption of the BRCT phosphopeptide-binding pocket: a multi-modal approach', *Eur J Hum Genet,* 16(7), pp. 820-32.
Villén, J., Rodriguez-Mias, R. and Fields, S. (2017) *A mistranslation method for assessing the effects of amino acid substitutions on protein stability and function.* [Online].
Werner, T., Sweetman, G., Savitski, M. F., Mathieson, T., Bantscheff, M. and Savitski, M. M. (2014) 'Ion coalescence of neutron encoded TMT 10-plex reporter ions', *Anal Chem,* 86(7), pp. 3594-601.
Winzeler, E. A. and Davis, R. W. (1997) 'Functional analysis of the yeast genome', *Curr Opin Genet Dev,* 7(6), pp. 771-6.
Womack, F. C. and Colowick, S. P. (1979) 'Proton-dependent inhibition of yeast and brain hexokinases by aluminum in ATP preparations', *Proc Natl Acad Sci U S A,* 76(10), pp. 5080-4.
Yamashita, M. F., John B. (1984) 'Electrospray ion source. Another variation on the free-jet theme.', *J. Phys. Chem.,* 88(20), pp. 4451–4459.
Yu, J., Li, T., Liu, Y., Wang, X., Zhang, J., Shi, G., Lou, J., Wang, L. and Wang, C. C. (2020) 'Phosphorylation switches protein disulfide isomerase activity to maintain proteostasis and attenuate ER stress', *EMBO J,* 39(10), pp. e103841.

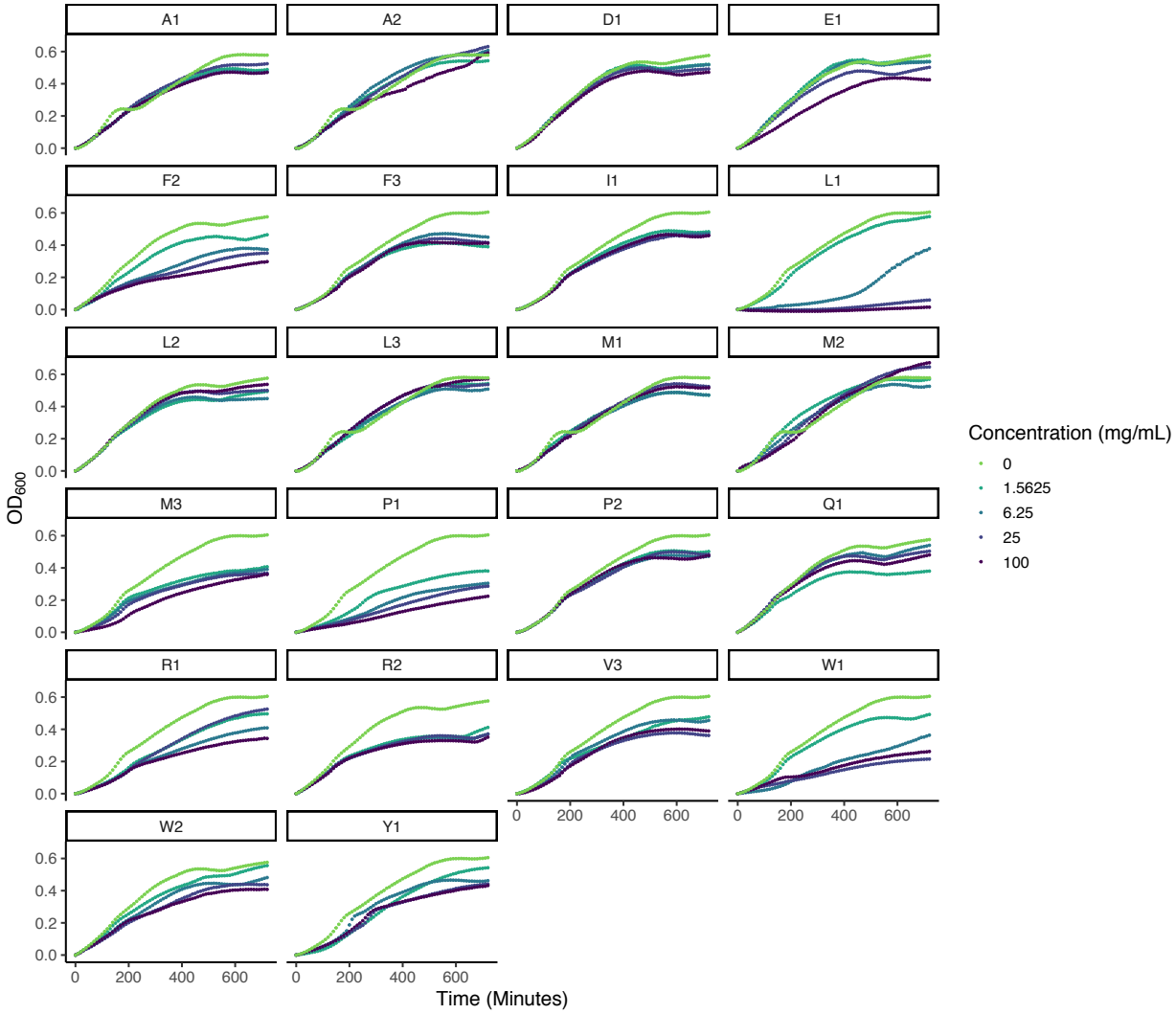
APPENDIX A



Supplementary Figure 2.1 Set of noncanonical amino acids screened in *E. coli*. Shorthand naming system used for Miró consists of the single-letter amino acid code of the residue being replaced by the ncAA, with a number distinguishing multiple ncAAs that target one canonical amino acid. For example, A1 and A2 are two ncAAs that replace alanine. Chemical structure and systematic IUPAC name are specified.

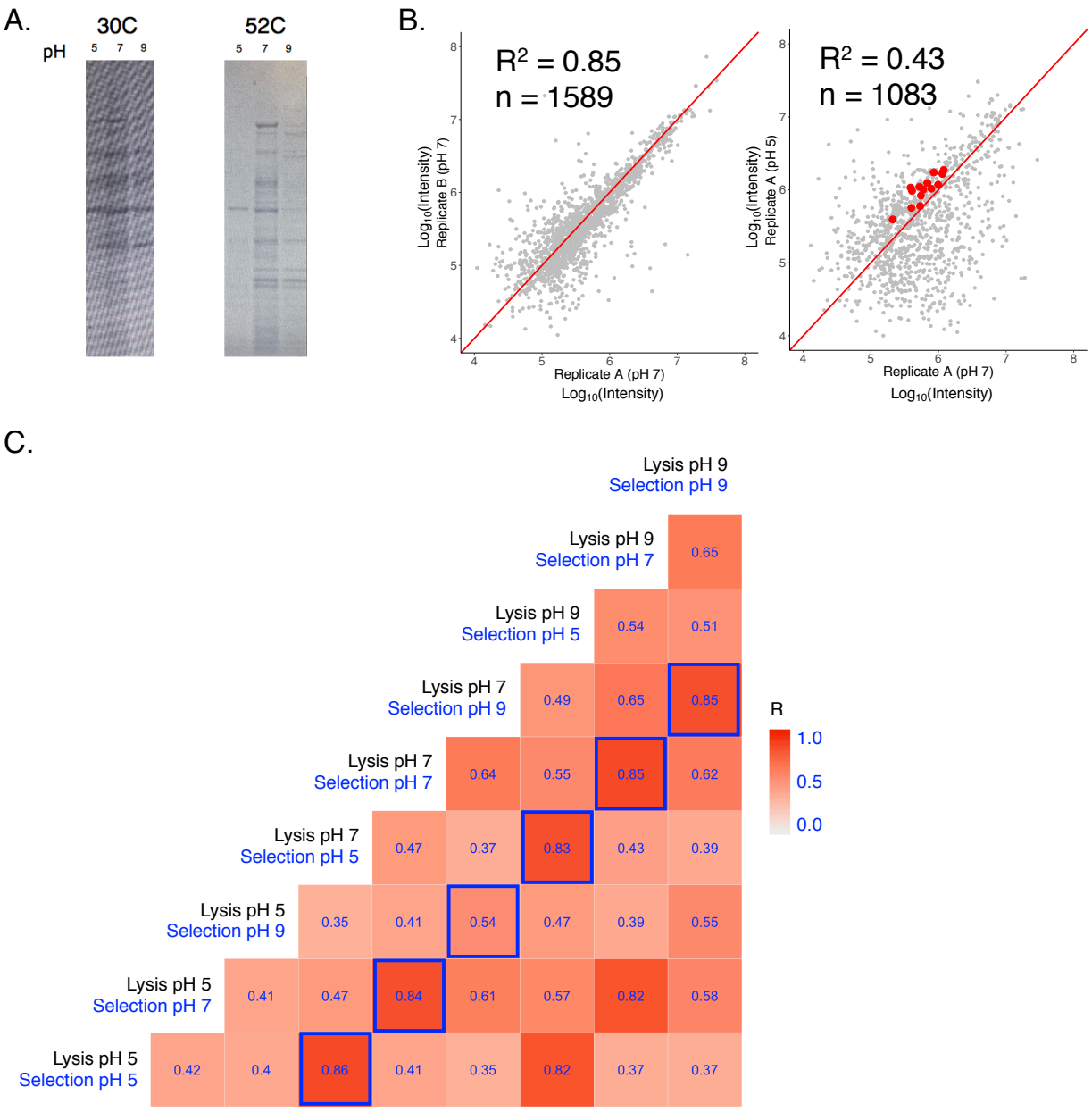
Supplementary Table 2.1 Concentrations of ncAA tested for toxicity and proteome incorporation in *E. coli*.

ncAA	Concentrations tested for toxicity (mg/mL)	Concentrations tested for incorporation (mg/mL)
A1	1.6, 6.3, 25, 100	100
A2	1.6, 6.3, 25, 100	100
D1	1.6, 6.3, 25, 100	200
E1	1.6, 6.3, 25, 100	100
F2	1.6, 6.3, 25, 100	100
F3	1.6, 6.3, 25, 100	100
I1	1.6, 6.3, 25, 100	200
L1	1.6, 6.3, 25, 100	50
L2	1.6, 6.3, 25, 100	500
L3	1.6, 6.3, 25, 100	500
M1	1.6, 6.3, 25, 100	500
M2	1.6, 6.3, 25, 100	100
M3	1.6, 6.3, 25, 100	100
P1	1.6, 6.3, 25, 100	5
P2	1.6, 6.3, 25, 100	200
Q1	1.6, 6.3, 25, 100	250
R1	1.6, 6.3, 25, 100	125
R2	1.6, 6.3, 25, 100	200
V3	1.6, 6.3, 25, 100	100
W1	1.6, 6.3, 25, 100	12.5
W2	1.6, 6.3, 25, 100	100
Y1	1.6, 6.3, 25, 100	125



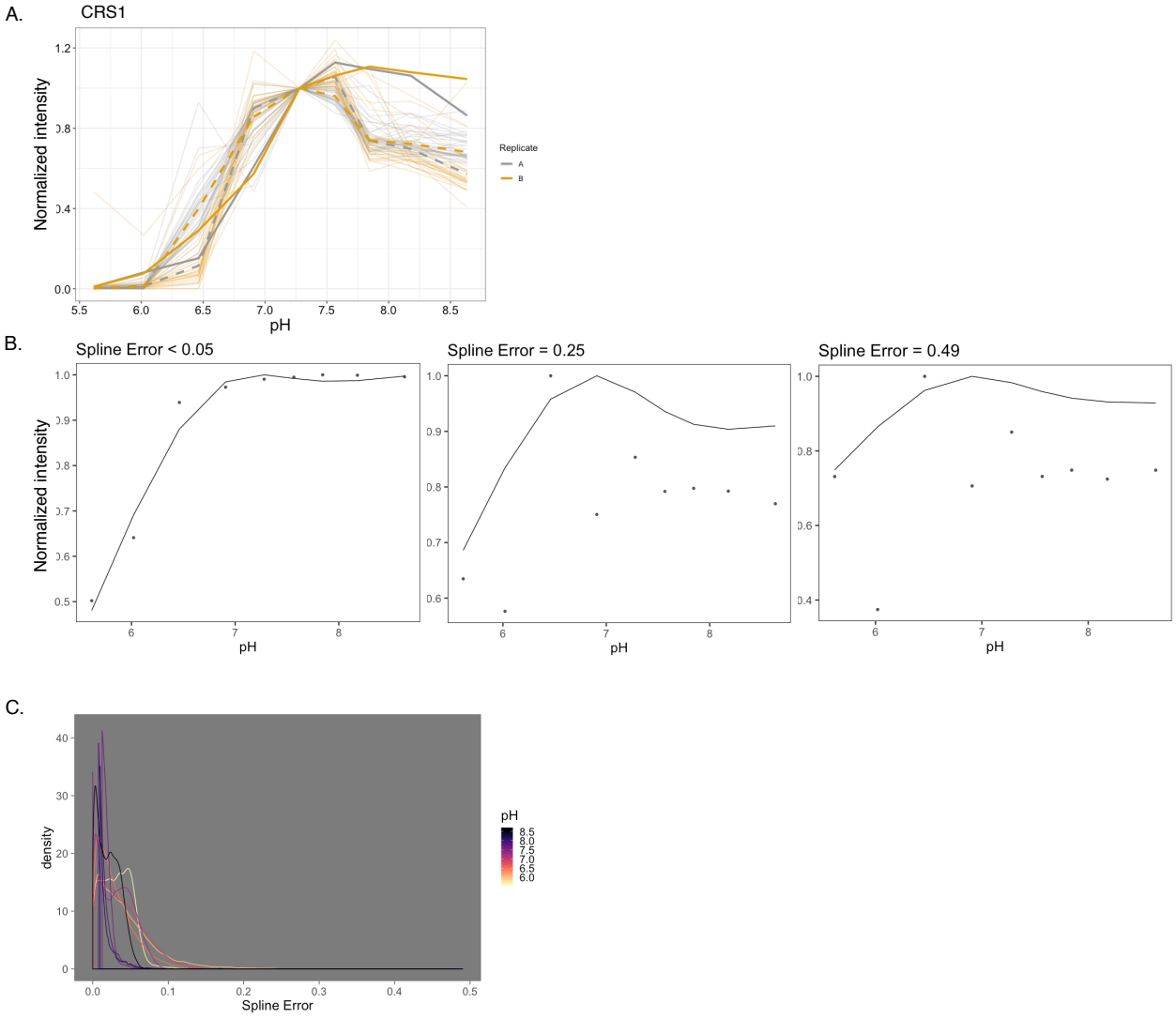
Supplementary Figure 2.2 *E.coli* growth with noncanonical amino acid treatment. Overnight culture of *E. coli* was diluted into a 96-well plate containing minimal media with various concentrations of ncAAs, and two replicate wells per condition. Cultures were incubated at 37ºC with agitation on a BioTek plate reader, and OD_{600} was measured every 15 minutes. OD_{600} measurements of two replicates per ncAA concentration were averaged and plotted against time. Plot title indicates ncAA, color of curves indicate concentration.

APPENDIX B



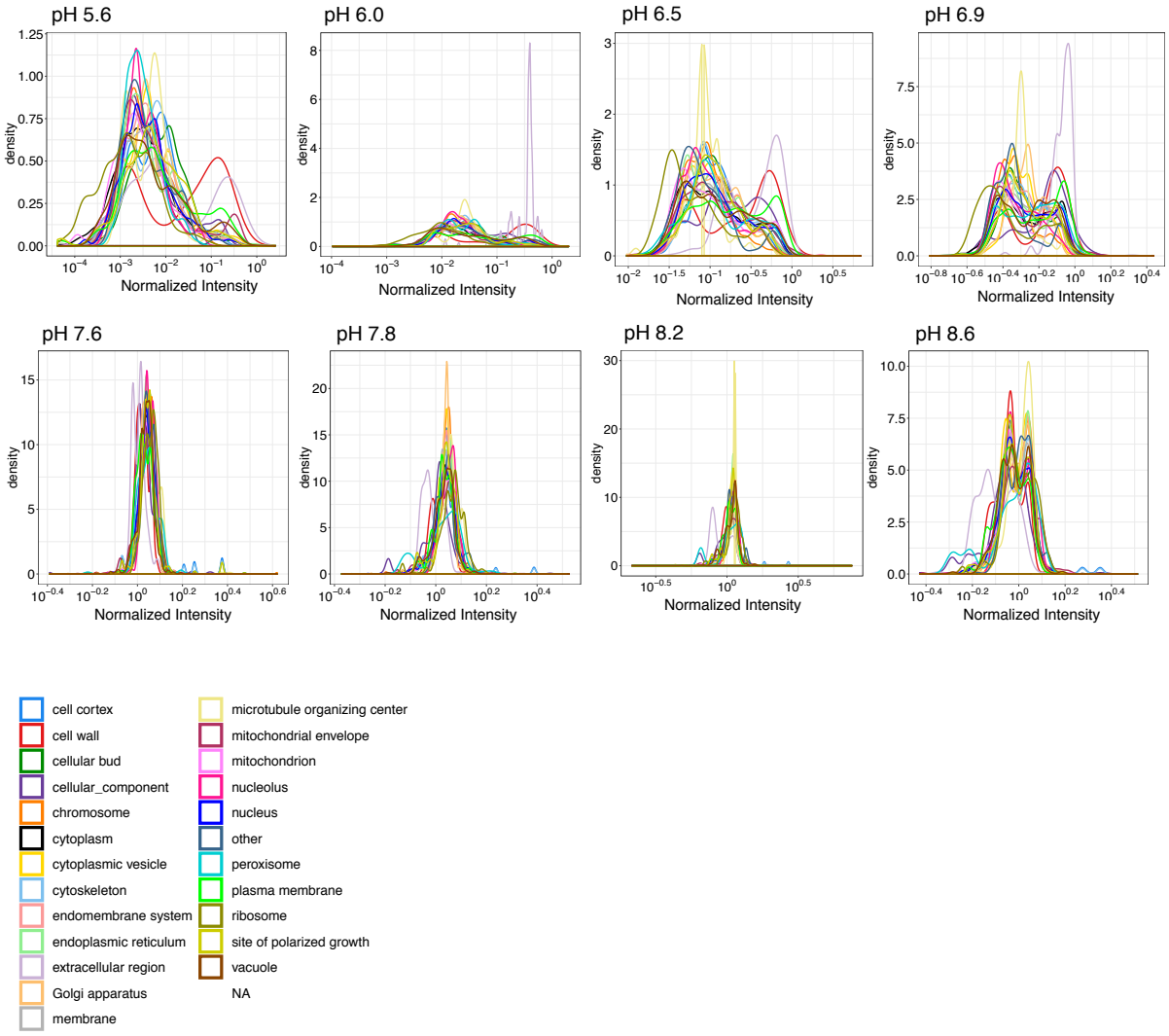
Supplementary Figure 3.1 Pilot experiments demonstrate that a pH selection can be applied to the whole proteome and help to determine experimental protocol. A. SDS-PAGE gels were generated for the soluble fractions of whole proteome samples subjected to pH 5, pH 7 or pH 9. Samples were incubated at either 30°C or 52°C. Differences in solubility across pH conditions is more pronounced in 52°C treated samples. B. Reproducibility of MS-measured peptide

quantification in proteome samples subjected to pH selections. First panel shows two replicates of samples treated at pH 7, second panel shows a pH 7 sample on the x-axis and pH 5 sample on the y-axis, diagonal red line represents y = x. Peptide quantification shows stronger correlation between two replicates treated with the same pH, indicating experimental reproducibility. Lower correlation in the comparison of pH 5 to pH 7 indicates that selection-driven differences in solubility are detectable by MS. Points representing peptides from the proteasome are highlighted in red to show reproducibility. C. Heatmap depicting correlation of peptide intensity values from proteome samples lysed in pH 5, 7 or 9, and then adjusted to a selection pH of 5, 7, or 9. Higher intensity of color indicates higher correlation, and Pearson's R value is indicated in each box. High correlation values on the diagonal of heat map boxed in blue indicate that selection occurs at the adjusted pH, not the lysis pH.



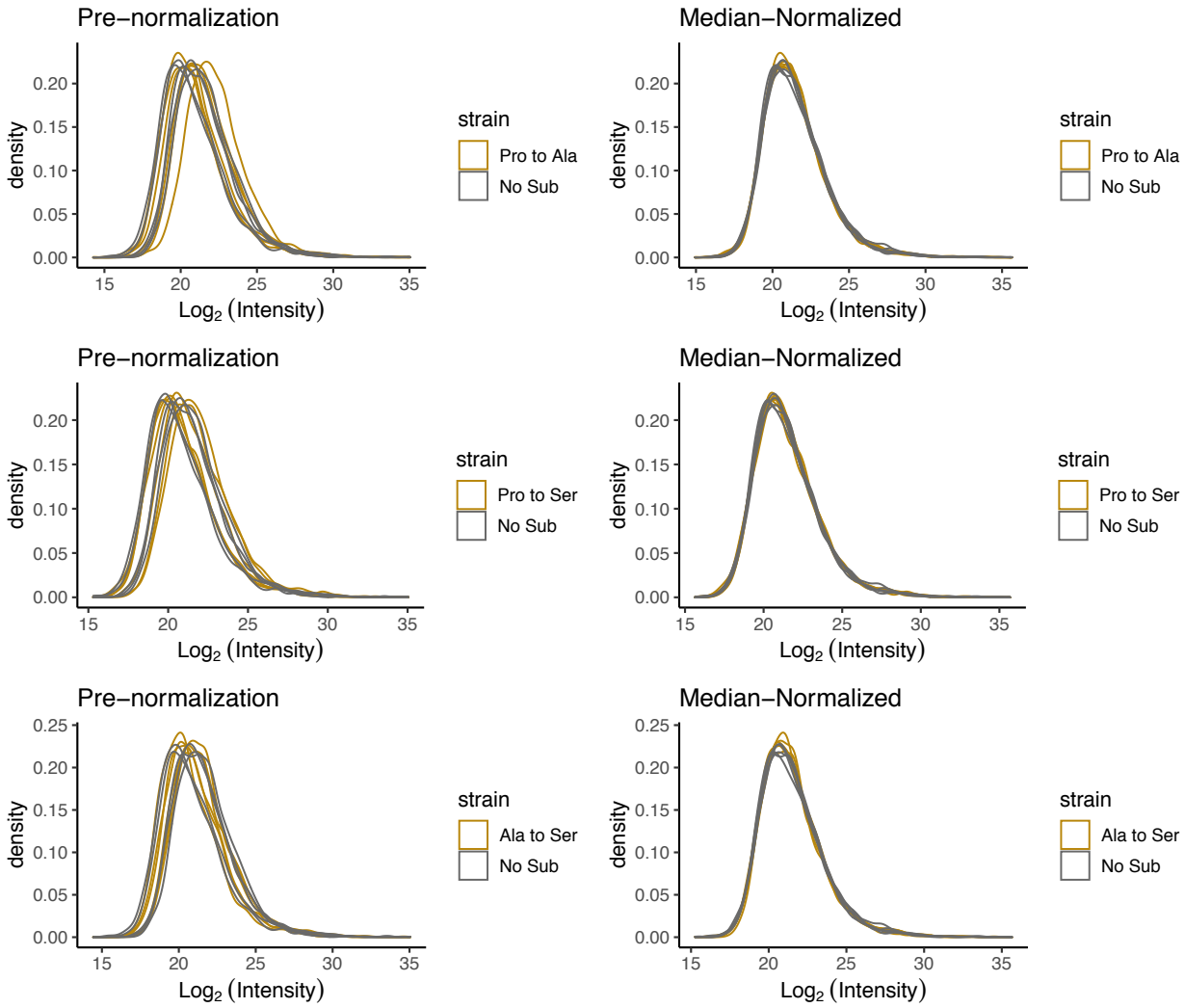
Supplementary Figure 3.2 Spline curve fits raw data more closely than normalizing to pH 7.2. A. Solubility profile of the protein CRS1 (YNL247W, a cysteinyl-tRNA synthetase) generated by normalizing all intensities to pH 7.2, to calculate the relative abundance values along the y-axis, with pH conditions indicated along the x-axis. Grey lines show data from biological replicate A, and yellow lines show data from replicate B. Fine lines represent PSM-level measurements, bold dashed lines represent these data aggregated to the protein-level, and bold solid lines represent the solubility profile of the entire proteome. B. Example plots depicting error between raw peptide-level abundance measurements and spline-fit curves. Points represent raw abundance, lines

represent spline curve. C. Density plot showing the error calculated between raw abundance measurements and a spline-fit curve. Each color represents a different pH. Spline error is higher in more acidic conditions, and lower in the basic conditions.



Supplementary Figure 3.3 Acidic pH is more selective for protein solubility and reveals the most differences. Distribution of abundance plotted for different subcellular localizations, indicated by different colors. Each plot represents data from one pH. Wider and more varied distributions show that lower pH is most selective and captures stark differences in solubility between proteins localized to unique cellular compartments.

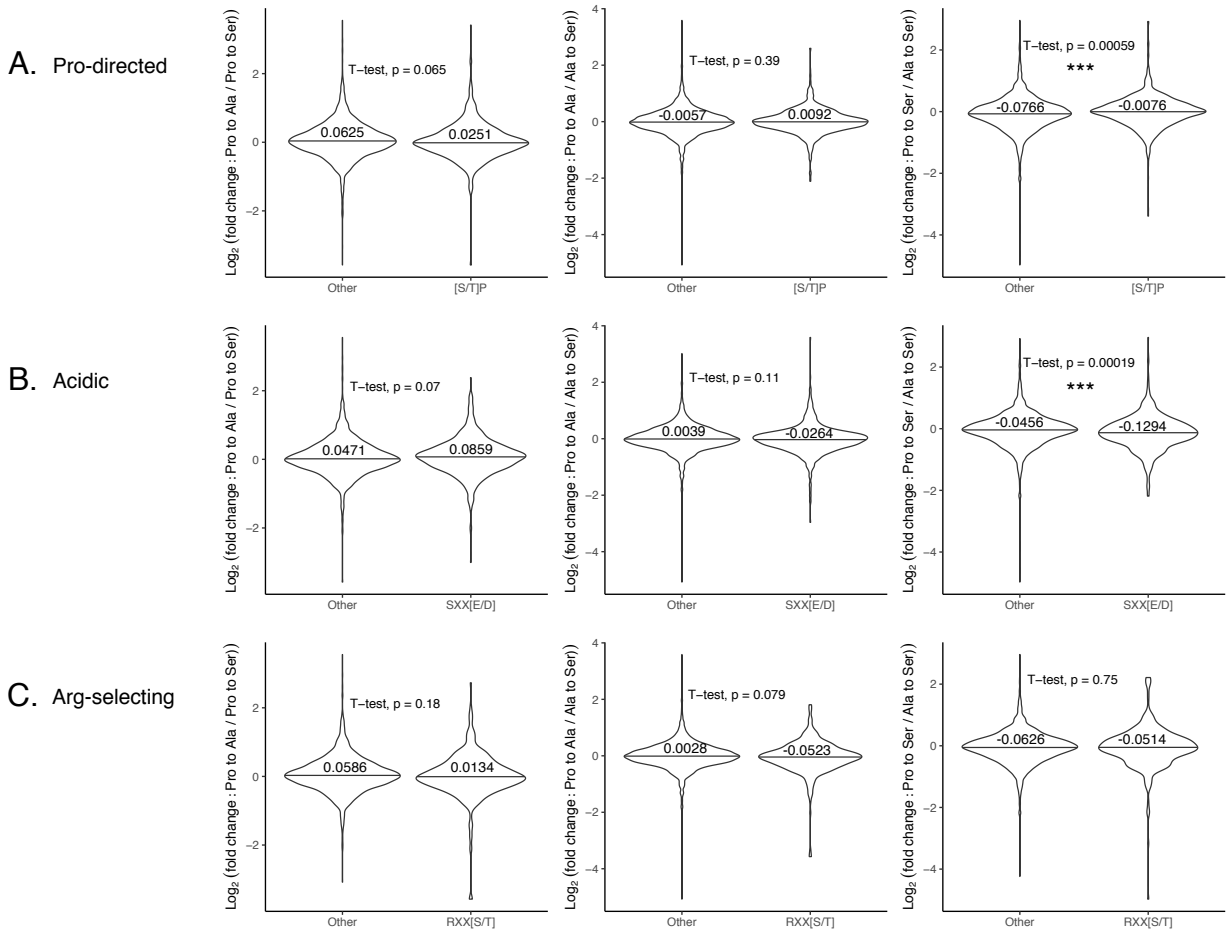
APPENDIX C



Supplementary Figure 4.1 Median normalization of phosphopeptide quantification. Distributions of phosphopeptide quantification are shown before and after median normalization, for each mistranslating strain with the control strain. These normalizations were applied before running the t-test to identify differentially abundant phosphopeptides.

Supplementary Table 4.1 Volcano Plot Statistics. Summary of volcano plot statistics in Figure 4.3.

Statistic	Pro to Ala	Pro to Ser	Ala to Ser
Number down-regulated phosphopeptides	102	128	116
Percent down-regulated phosphopeptides	3.6%	4.5%	4.3%
Number non-significant phosphopeptides	2627	2590	2517
Percent non-significant phosphopeptides	93.2%	91.3%	92.7%
Number up-regulated phosphopeptides	89	120	82
Percent up-regulated phosphopeptides	3.2%	4.2%	3%
Total phosphopeptides	2818	2838	2715



Supplementary Figure 4.2 Differences in phosphorylation of specific motifs compared between two mistranslating strains. Fold change values were calculated between phosphopeptides in each combination of mistranslating strains. Phosphopeptides were separated according to presence of the motif in question, and distribution of fold change values were compared. Groups of phosphopeptides are annotated as "Other" for those that do not contain phosphomotif, or as containing the phosphomotif specified on the x-axis. Distribution of Log_2(fold change) is visualized by violin plot along y-axis, number label is the mean fold change for that distribution. Significant differences were assigned based on Student's t-test, p-value is shown with asterisks indicating significance (*** : $p \leq 0.001$). First column of plots shows Log_2(fold change: Pro to Ala / Pro to Ser), second column of plots shows Log_2(fold change: Pro to Ala / Ala to Ser), third column of plots shows Log_2(fold change: Pro to Ser / Ala to Ser). A. Proline-directed

phosphomotif [S/T]P, B. Acidic phosphomotif SXX[E/D], C. Arg-selecting phosphomotif RXX[S/T].

Supplementary Table 4.2 Primers used in this study.

Primer Number	Sequence	Description
UG5953	TCTAAGCTTCGGACGATTGCCAACCGCC GAA	*SUP17* ($tRNA^{Ser}$)
UG5954	CTGCAGAATTCCGCGGAAATTAGCACGG CC	*SUP17* ($tRNA^{Ser}$)
XK8036	ACAGACTAGCAATCTGTTGGGCTCTGCC C	$tRNA^{Ser}_{AGC,G26A}$
XK8037	CCAACAGATTGCTAGTCTGTTGCCTTAA CCAC	$tRNA^{Ser}_{AGC,G26A}$

VITA

Bianca Ynez Ruiz was born in Phoenix, Arizona, on August 25, 1992. Growing up in the military, she and her family lived in Texas, Colorado, and Southern California. She graduated from Lancaster High School in the Antelope Valley of Los Angeles in the spring of 2010, and began pursuing her undergraduate degree at Cal State Fullerton later that year. During this time, she began her academic research career in the environmental microbiology laboratory of Dr. Hope Johnson, studying manganese-oxidizing bacteria. In the spring of 2016, Bianca graduated from Cal State Fullerton with a B.S. in Molecular Biology and Biotechnology, and a Minor in Chemistry. Later that year, she began pursuing her graduate education in the Department of Genome Sciences at the University of Washington. She conducted research in the laboratories of Dr. Stanley Fields and Dr. Judit Villén, and defended her thesis on December 1, 2021.