Large-Scale Analysis of Protein Folding and Stability Changes

Associated with Breast Cancer

by

Fang Liu

Department of Chemistry
Duke University

Date:_______________________

Approved:

___________________________
Michael C. Fitzgerald, Supervisor

___________________________
Katherine J. Franz

___________________________
Terrence G. Oas

___________________________
Warren S. Warren

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University

2018
ABSTRACT

Large-Scale Analysis of Protein Folding and Stability Changes

Associated with Breast Cancer

by

Fang Liu

Department of Chemistry
Duke University

Date:_______________________

Approved:

__________________________________________
Michael C. Fitzgerald, Supervisor

__________________________________________
Katherine J. Franz

__________________________________________
Terrence G. Oas

__________________________________________
Warren S. Warren

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University

2018
Abstract

Proteomic methods for disease state characterization and biomarker discovery have traditionally utilized quantitative mass spectrometry methods to identify proteins with altered expression levels in disease states. Unfortunately, these studies have not been as useful as expected at identifying disease-related proteins that can be exploited for diagnostic and therapeutic purposes, presumably due to the indirect link between a protein’s expression level and its function. Investigated here is the use of thermodynamic stability measurements to probe a more biologically relevant dimension of the proteome. It has the potential to become a new strategy for disease state characterization and to help elucidate the molecular basis of the disease. This thesis outlines the use of two discovery based techniques and one validation based technique to study protein folding and stability changes associated with breast cancer.

The first part of this dissertation describes the application of a mass spectrometry-based technique, stable isotope labeling with amino acids in cell culture and stability of proteins from rates of oxidation (SILAC-SPROX), in a comparison of the MCF-7 versus BT-474 breast cancer cell lines and a comparison of the MCF-7 versus MDA-MB-468 breast cancer cell lines. This work enabled ~1000 proteins to be assayed for breast cancer-related thermodynamic stability differences. The 242 and 445 protein
hits identified with altered stabilities in these comparative analyses created distinct molecular markers to differentiate the three cell lines.

The second part of this dissertation describes the development of a SILAC-based limited proteolysis (SILAC-LiP) strategy. The applicability of the protocol was demonstrated in a proof-of-principle study using proteins from a yeast cell lysate and a ubiquitous ligand. The SILAC-LiP protocol was further applied in a comparison of the MCF-7 versus MCF-10A cell lines. This work identified ~200 proteins with cell line dependent conformational changes, as determined by their differential susceptibility to proteolytic digestion using the nonspecific protease, proteinase K. The overlap between the SILAC-LiP hits reported here and the SILAC-SPROX hits previously identified in these same cell lines was relatively small (~20%). Thus, this work indicates that the SILAC-SPROX and SILAC-LiP techniques can be used together to provide complementary information on the disease states.

Furthermore, the protein hits identified in both the SILAC-SPROX and SILAC-LiP experiments included a large fraction (~70%) with no significant expression level changes. This suggests protein folding and stability measurements can provide information about disease states that is orthogonal to that obtained in protein expression level analyses.

The last part of this dissertation focuses on the establishment of targeted mass spectrometry-based validation assays for the protein biomarker candidates with altered
thermodynamic stabilities identified in the SILAC-SPROX experiments. Application of the PAB-SPROX protocol on the MCF-7 cell lysate enabled reproducible identification and quantitation of a subset of prioritized target peptides.
Dedication

I would like to dedicate this dissertation to my parents who always love and support me. I could not have become who I am without them.
# Contents

Abstract........................................................................................................................................... iv

List of Tables ..................................................................................................................................... xiii

List of Figures .................................................................................................................................... xiv

List of Abbreviations ........................................................................................................................ xvi

Acknowledgements ............................................................................................................................ xviii

1. Introduction ................................................................................................................................. 1

   1.1 Protein Folding and Stability Measurements ............................................................................ 1

   1.2 Approaches for Protein Stability Measurements ...................................................................... 2

      1.2.1 Overview ........................................................................................................................... 2

      1.2.2 Stability of Proteins from Rates of Oxidation (SPROX) ...................................................... 7

         1.2.2.1 General Workflow ........................................................................................................ 7

         1.2.2.2 Advantages and Limitations ........................................................................................... 9

         1.2.2.3 Targeted SPROX .......................................................................................................... 10

         1.2.2.4 Using SPROX for Disease State Characterization ......................................................... 11

      1.2.3 Limited Proteolysis (LiP) .................................................................................................... 12

         1.2.3.1 General Workflow ........................................................................................................ 12

         1.2.3.2 Advantages and Limitations ........................................................................................... 13

         1.2.3.3 Targeted LiP .................................................................................................................. 14

         1.2.3.4 Using LiP for Disease State Characterization ................................................................. 15

   1.3 Overview of Breast Cancer ...................................................................................................... 15
# Table of Contents

1.3.1 Early Diagnosis and Protein Biomarker Discovery ........................................... 15  
1.3.2 Molecular Classification of Breast Cancer ......................................................... 18  
1.4 Dissertation Focus ................................................................................................. 20  

2. Large-Scale Analysis of Protein Folding and Stability Changes Associated with Breast Cancer Using SILAC-SPROX ................................................................. 23  

2.1 Introduction ........................................................................................................... 23  
2.2 Experimental ......................................................................................................... 25  

2.2.1 Cell Culture and Cell Lysate Preparation ......................................................... 25  
2.2.2 SILAC-SPROX Analyses .................................................................................... 27  
2.2.3 Proteomic Sample Preparation .......................................................................... 28  
2.2.4 LC-MS/MS Analyses ......................................................................................... 28  
2.2.5 Proteomic Data Analysis ..................................................................................... 29  

2.3 Results .................................................................................................................... 34  

2.3.1 Experimental Design ......................................................................................... 34  
2.3.2 Proteomic Coverage ......................................................................................... 37  
2.3.3 Hit Identification ................................................................................................ 38  
2.3.4 Classification of Hit Behavior ............................................................................. 43  
2.3.5 Protein Expression Level Analysis........................................................................ 45  

2.4 Discussion .............................................................................................................. 46  

2.4.1 Comparison between Luminal-A and -B Subtypes ......................................... 46  
2.4.2 Comparison between Luminal-A and Basal Subtypes ...................................... 52  

2.5 Conclusions .......................................................................................................... 56
3. Development of a SILAC-Based Limited Proteolysis Approach ........................................ 58

3.1 Introduction ...................................................................................................................... 58

3.2 Experimental .................................................................................................................. 60

3.2.1 Yeast Cell Culture .......................................................................................................... 60

3.2.2 Yeast Cell Lysate Preparation ....................................................................................... 61

3.2.3 SILAC-LiP Analyses ....................................................................................................... 62

3.2.4 LC-MS/MS Analyses ...................................................................................................... 63

3.2.5 Proteomic Data Analysis ............................................................................................... 63

3.3 Results and Discussion .................................................................................................... 64

3.3.1 Experimental Design ..................................................................................................... 64

3.3.2 Proteomic Coverage ....................................................................................................... 68

3.3.3 Discovery of ATP-Interacting Proteins ......................................................................... 70

3.3.4 Overlap with SPROX .................................................................................................... 72

3.4 Conclusions ...................................................................................................................... 75

4. Large-Scale Analysis of Protein Folding and Stability Changes Associated with Breast Cancer Using SILAC-LiP .................................................................................................................. 76

4.1 Introduction ...................................................................................................................... 76

4.2 Experimental .................................................................................................................. 77

4.2.1 Cell Culture and Cell Lysate Preparation ..................................................................... 77

4.2.2 SILAC-LiP Analyses ....................................................................................................... 78

4.2.3 LC-MS/MS Analyses ...................................................................................................... 79

4.2.4 Proteomic Data Analysis ............................................................................................... 80
List of Tables

Table 1: List of FDA-approved protein markers for breast cancer currently used in clinical practice. ............................................................................................................................................. 17

Table 2: Molecular classification of breast cancer........................................................................................................................................... 20

Table 3: Summary of the proteomic data obtained in the SILAC-SPROX experiments for differential thermodynamic stability analysis in the MCF-7 vs. BT-474 and MCF-7 vs. MDA-MB-468 cell line comparisons........................................................................................................................................ 37

Table 4: Cytoskeletal proteins identified with altered thermodynamic stability in the SILAC-SPROX study of BT-474 versus MCF-7 cell line comparison......................................................... 49

Table 5: Protein hits with chaperone activity identified in the SILAC-SPROX study of BT-474 versus MCF-7 cell line comparison................................................................. 50

Table 6: Protein hits involved in the ubiquitin proteasome pathway that were identified in the SILAC-SPROX study of MDA-MB-468 versus MCF-7 cell line comparison.......... 55

Table 7: Proteomic coverage obtained in the ATP-binding study performed with SILAC-LiP................................................................................................................................................. 70

Table 8: Summary of the 27 newly discovered ATP-interacting proteins that were identified in at least two of the SIALC-LiP, SILAC-SPROX and iTRAQ-SPROX experiments........................................................................................................................................ 74

Table 9: Proteomic coverage obtained in the cell line comparisons performed with SILAC-LiP. ............................................................................................................................................. 85

Table 10: Summary of the PRM precursors that were targeted in the PAB-SPROX experiments. Bolded and underlined methionine residues were modified with PAB... 111

Table 11: Summary of the proteomic data reported in the previous SILAC-SPROX studies(2, 35) for differential thermodynamic stability analysis of breast cancer cell lines. ........................................................................................................................................ 113

Table 12: Summary of H/L ratios generated for the eight target peptides using the PAB-SPROX protocol........................................................................................................................................ 119
List of Figures

Figure 1: Schematic representation of the proteomic approaches for protein folding measurements. ................................................................. 6

Figure 2: Global distributions of the log2(normalized H/L) values determined for all the non-methionine-containing peptides identified at all the denaturant concentrations in the (A) MCF-7 versus BT-474 and (B) MCF-7 versus MDA-MB-468 cell line comparisons. .......................................................................................................................... 33

Figure 3: Schematic representation of the experimental workflow used in this work. Shown at the bottom are the protein unfolding curves extracted from the corresponding SILAC-SPROX behaviors shown for a wild type methionine-containing peptide. ........... 36

Figure 4: Representative SILAC-SPROX data and associated protein unfolding curves obtained in this work ................................................................................................................................. 40

Figure 5: Evaluation of the expression level data generated in this work.............................. 45

Figure 6: Protein classes observed in the proteins assayed and identified as hits with altered thermodynamic stabilities in the (A) BT-474 versus MCF-7 and (B) MDA-MB-468 versus MCF-7 cell line comparisons ........................................................................................................ 51

Figure 7: Schematic representation of the SILAC-based limited proteolysis workflow used in this work................................................................................................................................. 65

Figure 8: Chemical structure of AMP-PNP, the non-hydrolyzable ATP mimic used in the ATP-binding study. ................................................................. 67

Figure 9: Volcano plot of statistical significance against fold change between single and double digestion .................................................................................................................. 68

Figure 10: Pie chart showing the distribution of known ligands for the hit proteins identified in the ATP-binding study. .................................................................................................................. 71

Figure 11: Comparison of SILAC-LiP, SILAC-SPROX and iTRAQ-SPROX ATP-binding results .......................................................................................................................... 73

Figure 12: Schematic representation of the SILAC-based limited proteolysis workflow used in this work .......................................................................................................................... 82
Figure 13: Global analysis of the H/L ratios generated for the assayed peptides in the MCF-7 vs. MCF-10A comparison. ................................................................. 85

Figure 14: Global analysis of the H/L ratios generated for the assayed peptides in the control experiment. ................................................................. 87

Figure 15: Bar graph showing the number of protein hits that were identified with multiple peptides in the MCF-7 vs. MCF-10A comparison. ........................................ 90

Figure 16: Expression profile generated for a subset of protein hits identified in the MCF-10A vs. MCF-7 comparison. ................................................................. 93

Figure 17: Bar graph showing the distribution of the relative expression levels of the protein hits identified in the MCF-7 vs. MCF-10A comparison. ........................ 94

Figure 18: Pie chart showing the distribution of different protein structural elements to which the peptide hits identified in the MCF-7 vs. MCF-10A comparison mapped. .... 95

Figure 19: Crystal structure showing the nine antiparallel helices that comprise the 14-3-3σ homodimer (PDB 1YZ5). ................................................................. 96

Figure 20: Schematic representations of the folded three-dimensional structures of selected protein hits identified in the MCF-7 vs. MCF-10A comparison. ................ 100

Figure 21: Schematic representation of PAB labeling reaction and gas phase fragmentation behavior of resulting PAB-labeled methionine residues. ...................... 108

Figure 22: Venn diagram showing the overlap among the 325 methionine-containing peptide hits commonly assayed in all of the four cell line comparisons. .............. 114

Figure 23: SILAC-SPROX data for the MKETAEAYLGK peptide from 78 kDa glucose-regulated protein in the four cell line comparisons. ........................................ 117

Figure 24: General workflow used in the PAB-SPROX experiments for the validation of SILAC-SPROX hits. ................................................................. 118
List of Abbreviations

MS  mass spectrometry
LC-MS/MS  liquid chromatography tandem mass spectrometry
QqQ-MS  triple quadrupole mass spectrometer
iTRAQ  isobaric tags for relative and absolute quantitation
TMT  tandem mass tags
SILAC  stable isotope labeling with amino acids in cell culture
SRM  selective reaction monitoring
PRM  parallel reaction monitoring
SPROX  stability of proteins from rates of oxidation
LiP  limited proteolysis
ER  estrogen receptor
PR  progesterone receptor
HER2  human epidermal growth factor receptor-2
ATP  adenosine triphosphate
AMP-PNP  adenylyl imidodiphosphate
TEAB  triethylammonium bicarbonate
SDS  sodium dodecyl sulfate
TCEP-HCl  tris(2-carboxyethyl)phosphine hydrochloride
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMTS</td>
<td>methylmethane thiosulfonate</td>
</tr>
<tr>
<td>TFA</td>
<td>trifluoroacetic acid</td>
</tr>
<tr>
<td>GdmCl</td>
<td>guanidine hydrochloride</td>
</tr>
<tr>
<td>TCA</td>
<td>trichloroacetic acid</td>
</tr>
<tr>
<td>CID</td>
<td>collision induced dissociation</td>
</tr>
<tr>
<td>HCD</td>
<td>higher energy collision-induced dissociation</td>
</tr>
<tr>
<td>NCE</td>
<td>normalized collision energy</td>
</tr>
</tbody>
</table>
Acknowledgements

First, I would like to give my sincere thanks to my advisor, Professor Michael C. Fitzgerald, for his guidance and motivation throughout my PhD research. There are frustrating moments during the PhD research, it was his encouragement and optimism that helped me gain back confidence. This thesis would not have been possible without his support.

I also would like to thank my committee members, Professor Terrence G. Oas, Professor Katherine J. Franz and Professor Warren S. Warren, for their valuable advice and guidance of my research towards the right direction.

I would like to acknowledge both current and past Fitzgerald lab members. In particular, I am grateful to Dr. Yingrong Xu for her mentoring when I joined the lab, Dr. Jagat Adhikari for the many enlightening discussions for the successful development of the SILAC-LiP strategy, and Dr. Lorrain Jin for her great help in the use of targeted mass spectrometry. I am also thankful to the present lab members He Meng, Arthur Ma, Aurora Cabrera, and the former lab members Dr. M. Ariel Geer Wallace, Dr. Ryenne N. Ogburn and Dr. Julia H. Roberts.

I would like to give special thanks to the Duke Proteomics Facility, especially Sarah Rains and Dr. Will Thompson, for acquiring the LC-MS/MS data.
Finally, I would like to express my deepest gratitude to my dearest parents for their love and support throughout my life. Thank you both for giving me strength to chase my dreams. Also to my boyfriend, Siyu, thank you for keeping me happy everyday.
1. Introduction

Portions of this chapter can also be found in the papers “Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using Limited Proteolysis”(1) and “Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using SILAC-SPOX”(2).

1.1 Protein Folding and Stability Measurements

Proteins are important molecules in living organisms. Under physiological conditions, proteins exist in a dynamic equilibrium between fully folded native state and various partially or globally unfolded non-native states. The stability of a non-native conformation in this ensemble is determined by its free energy relative to the folded native state of the protein (i.e., the folding free energy ΔGf). Similarly, the free energy difference between the folded and globally unfolded states defines the thermodynamic stability of a protein. Protein folding and thermodynamic stability is fundamental to protein function. Changes in the protein folding stability can arise from a number of biologically significant phenomena such as mutations, modifications and altered interactions with cellular targets. Therefore, measurements of protein folding stability are essential for understanding the critical role of protein folding in the regulation of protein functions and various biologically significant phenomena.

This dissertation is largely focused on understanding the changes in protein folding stability associated with breast cancer. The goal of this dissertation is to make
proteome-wide measurements of protein folding stability in breast cancer to identify novel molecular signatures of the disease. For decades, protein folding stability measurements have been widely used in studies of protein structure and function. For example, they have been an essential part of fundamental studies to better understand the molecular basis of protein folding and function, and they have been a critical part of more applied studies aimed at developing protein-targeted drugs. The work in this dissertation extends the use of protein folding stability measurements to characterize disease states. It has the potential to create a new avenue for molecular characterization of disease states and biomarker discovery for diagnostic and therapeutic purposes.

1.2 Approaches for Protein Stability Measurements

1.2.1 Overview

The traditional methods for making protein stability measurements have relied on various calorimetric(3-5) or spectroscopic(4, 6-10) techniques. The calorimetric approach involves thermal denaturation of proteins with increasing temperature, during which the melting temperature ($T_m$) and the heat capacity changes ($\Delta C_p$) of proteins from the folded to unfolded states can be measured by isothermal calorimetry (ITC) or differential scanning calorimetry (DSC). Fluorescence spectroscopy and circular dichroism (CD) spectroscopy are among the most frequently reported spectroscopic approaches for evaluations of the thermodynamic parameters associated with protein folding and protein-ligand binding reactions. The intrinsic fluorescence properties of
aromatic residues such as tryptophan or tyrosine are sensitive to changes in their local geometry. By measuring the fluorescence intensity of tryptophan or tyrosine probes as a function of pH, temperature or chemical denaturation unfolding conditions, the equilibrium unfolding/refolding reactions of proteins can be monitored. CD spectroscopy measures the difference in absorption of the left and right circularly polarized light by the chromophores. It measures how secondary structure changes at different denaturing conditions (e.g., different pH, temperature, or chemical denaturant concentrations). The intrinsic backbone CD can be monitored with the far UV (180-260 nm wavelengths) and the CD bands from aromatic chromophores can be followed using near UV (250-350 nm wavelengths) measurements.

These traditional methods usually require high concentration and large amounts of purified protein material (e.g., a typical DSC experiment requires 1-2 mL purified protein solution with concentration of at least 1 mg/ml)(4). For the fluorescence technique, it may also require mutagenesis to incorporate tryptophan or tyrosine in the absence of a native buried fluorophore in protein sequence, which can perturb the native structure of the protein and induce changes in the observed stability. The experimental limitations associated with these traditional methods have precluded their applications in the proteome-wide measurements of protein folding and thermodynamic stability, which is the focus of this dissertation.
Over the past decade, a series of mass spectrometry (MS)-based proteomic techniques have been introduced to study protein folding in unpurified biological samples including cell lysates and even intact cells (11-19). By eliminating the need for relatively large amounts of purified protein material, these new MS-based proteomic techniques have enabled the large-scale and high-throughput measurements of protein folding stability in their biological context. Proteins can undergo a wide range of conformational changes from more global unfolding events where all higher-order structure is lost to more local unfolding events where only some specific elements of secondary structure are partially unfolded. Some of the recently established MS-based techniques are analogous to the traditional spectroscopy-based methods in that they also exploit the chemical or thermal denaturation properties of proteins to generate quantitative measures of the more global unfolding/refolding properties of proteins, while the other MS-based proteomic techniques are more qualitative and have utilized limited proteolysis strategies under native conditions (or at very low concentrations of denaturant) to probe the more local unfolding/refolding properties of proteins. The basic workflow used for these techniques is shown in Figure 1.

The quantitative approaches that focus on the more global unfolding/refolding reactions of proteins typically involve the use of a series of chemical or heat denaturation to promote the global unfolding of proteins in solution. To differentiate unfolded forms of proteins from their native forms, a chemical or enzymatic reaction is
performed to selectively modify the unfolded forms of proteins where the reaction sites are only exposed during protein unfolding. The extent of the reaction as a function of the chemical denaturant concentration or the temperature is quantified with MS-based proteomic platforms. Examples include the Stability of Proteins from Rates of Oxidation (SPROX)(11, 20), Pulse Proteolysis (PP)(12-13, 21-22) and Thermal Proteome Profiling (TPP)(17) techniques. The different techniques have utilized the chemical denaturant dependence of a covalent modification reaction (SPROX) or a proteolytic digestion (PP), or utilized the temperature dependence of a protein precipitation reaction (TPP) to probe the equilibrium unfolding/refolding reactions of proteins.

A main class of the qualitative techniques that focus on the more local unfolding/refolding reactions of proteins has utilized limited proteolysis strategies to probe the differential stabilities of proteins under different conditions. Examples include the Drug Affinity Responsive Target Stability (DARTS)(15) and Limited Proteolysis (LiP)(14) techniques. Instead of using a series of chemical or heat denaturation that shifts the protein unfolding/refolding equilibrium to the unfolded state, these qualitative methods involve just one chemical denaturant or temperature point at which difference can be observed (typically under native condition). Therefore, the methods are only useful to qualitatively detect changes in the protein stabilities.
Figure 1: Schematic representation of the proteomic approaches for protein folding measurements.

Prior to the work described in this dissertation, most of the above techniques have been utilized to understand the protein folding properties induced by specific ligands. They have proven useful for identifying the protein targets of drugs (15, 17, 23-26) and other biologically important molecules (11-12, 16, 20, 22, 27-31). There were only two studies focused on the global analysis of protein folding stability associated with different biological states. One study utilized a LiP approach to study the structural transitions of proteins from yeast cultures upon a metabolic transition, while the other study utilized the SPROX technique to investigate breast cancer-related protein folding properties using cell culture models of the disease. Such applications are relatively new,
albeit with great potential to provide a novel and complementary approach for disease state characterization. Researchers are still investigating the benefits and drawbacks of using a specific technique in a given context. The work in this dissertation is aimed at extending the use of proteome-wide thermodynamic profiling strategies for disease state characterization and providing a new paradigm for the characterization of disease states. The techniques that will be used in this work include the SPROX and LiP approaches. The basic principles of the SPROX and LiP techniques are reviewed below.

1.2.2 Stability of Proteins from Rates of Oxidation (SPROX)

1.2.2.1 General Workflow

The initial step in SPROX involves distributing a complex protein mixture of interest (e.g., a cell lysate sample) into a series of buffers containing increasing concentrations of a chemical denaturant (e.g., urea or guanidine hydrochloride) and allowing the proteins to reach their unfolding/refolding equilibrium. The globally protected methionine (Met) residues in proteins are only exposed and will be modified by a chemical reaction as the denaturant concentration increases. The reaction used in SPROX is the oxidation of the thioether group in the side chain of the exposed methionine residues in proteins by hydrogen peroxide. The hydrogen peroxide concentration and the reaction time are chosen such that the pseudo-first-order oxidation reaction of an unprotected methionine residue will proceed for 3-5 half-lives. The hydrogen peroxide reaction conditions also ensure that the system exists in the so-
called EX2 regime (i.e., the modification reaction rate is much slower than the protein unfolding/refolding rate). After the oxidation reaction is quenched with 6-fold excess amount of reducing agent (e.g., methionine), the proteins in each denaturant-containing buffer are subjected to a quantitative bottom-up proteomics analysis.

Ultimately, the extent of oxidation as a function of the denaturant concentration can be evaluated for the methionine-containing peptides identified and quantified in the LC-MS/MS readout. An increase in the oxidized methionine-containing peptide signal or a decrease in the wild-type methionine-containing peptide signal across the denaturant concentrations can be observed for the globally protected methionine residues in proteins. Biophysical properties of the chemical denaturant-induced equilibrium unfolding/refolding reactions (e.g., transition midpoints $C_{1/2}$) can be assessed for the proteins (or protein domains) from which the peptides are derived. The quantitation in the SPROX protocol can be achieved through an isobaric mass tagging strategy (TMT-10plex or iTRAQ-8plex)(11) or SILAC labeling of cells(20).

In ideal cases where the chemical denaturant-induced equilibrium unfolding/refolding properties of a protein (or protein domain) can be well-modeled by a reversible, two-state transition (i.e., no partially folded intermediate states of the protein are populated) and the intrinsic modification rate of an unprotected methionine residue is slower than the protein refolding rate, the $C_{1/2}$ values can report back to the protein folding free energies. In the case of proteins that unfold in a highly concerted yet
not in a two-state manner, the $C_{1/2}$ values cannot provide meaningful determination but can be used to assess relative stability and free energy changes under different conditions (e.g., with and without a ligand). The free energy changes in the case of ligand binding studies (i.e., binding free energy) can be further used to determine dissociation constant under the assumption that the ligand exclusively binds to the native state of the protein. In the case of large multidomain proteins that do not unfold in a highly cooperative manner, the $C_{1/2}$ values provide unique biophysical information about the specific protein domains to which the methionine-containing peptide probes map.

### 1.2.2.2 Advantages and Limitations

One unique advantage of SPROX is that methionine residues from different domains of a protein can have different behaviors in response to the denaturant dependent oxidation. This property of SPROX provides the ability to measure protein domain specific stability characteristics. Another unique advantage of SPROX is that the non-methionine-containing peptides can act as internal standards to account for experimental variations and protein expression level differences. The methionine-containing peptides from different denaturant concentrations identified in a bottom-up proteomics experiment thus can be normalized by the corresponding non-methionine-containing peptides.
SPROX requires the target protein to contain at least one globally protected methionine residue. The frequency of methionine residue in protein sequences is about 2.5%. Thus, a typical protein generally contains multiple methionine residues. However, one limitation of SPROX arises from the requirement of successful detection and quantitation of methionine-containing peptides in the bottom-up proteomic experiments. Because only ~20% of the peptides identified in a typical bottom-up proteomic experiment contain methionine residue. This limitation can be remedied by the use of a methionine-containing peptide enrichment strategy that typically increases the fraction of methionine-containing peptides to >70%(11, 32). Another approach for expanding the proteomic coverage in SPROX has incorporated dimethyl(2-hydroxy-5-nitrobenzyl)sulfonyl bromide (HNSB) to additionally label tryptophan residues in proteins and utilized a tryptophan-containing peptide enrichment strategy(19). A ~50% increase in proteomic coverage was successfully achieved using the hybrid protocol that combined tryptophan modification and methionine oxidation in one experiment(19).

1.2.2.3 Targeted SPROX

Recently, the SPROX technique has been coupled with targeted proteomics to enable targeted detection and quantitation of methionine-containing peptides using selective reaction monitoring (SRM) on a triple quadrupole mass spectrometer (QqQ-MS)(33). The strategy involves reaction of the tryptic peptides generated in the proteome-wide SPROX analysis with light and heavy phenacyl bromide (PAB) to
specifically label unoxidized methionine residues. The strategy has shown great utility in the validation of a specific methionine-containing peptide hit and protein folding stability measured in the shotgun proteomic analysis(34).

1.2.2.4 Using SPROX for Disease State Characterization

One important application of the proteome-wide SPROX technique is to study protein-ligand interactions. The technique was applied to characterize a number of well-established protein-ligand systems(11, 20, 23, 28). In these proof-of-principle studies, the technique was not only successful in identifying both known and novel binding partners, but also capable of quantifying affinities of the binding interactions. SPROX was further used in a number of less well-understood systems for protein target discovery efforts(25-26). Because the technique has enabled study of protein-ligand interactions directly in their biologically relevant context, both direct and indirect protein interactions can be detected.

Of particular significance to the current work is the other application of SPROX—as a proteome level screen for disease state characterization. Prior to the work in this thesis, SILAC-SPROX was used for the differential thermodynamic analysis of proteins in three well-studied cell culture models of breast cancer that have different molecular features and display different levels of invasiveness (MCF-10A, MCF-7 and MDA-MB-231)(35). The study successfully differentiated the three cell lines and identified a number of protein hits that have been previously associated with cancer through other
biochemical studies. Interestingly, ~45% of the protein hits identified in this study did not have altered expression levels, which highlights the complementary ability of protein stability measurements to traditional expression level studies.

Later during the current work, researchers start to investigate more disease states using such global thermodynamic stability profiling strategies. A study utilized iTRAQ-SPROX to identify age-related differences in mouse brain proteins. The study found that the large majority of the protein hits were destabilized in the old mice and the hits were enriched in proteins with slow turnover rates(36). Another study utilized the hybrid SPROX protocol to characterize the house dust mite proteome and found that allergenic proteins are both more stable and more abundant than other proteins in the proteome(37). Further extension of using proteome-wide thermodynamic profiling strategies for disease state characterization is the main focus of this dissertation.

1.2.3 Limited Proteolysis (LiP)

1.2.3.1 General Workflow

The LiP technique relies on the differential susceptibility of proteins to protease cleavage under native conditions(14). The limited proteolysis conditions with a non-specific protease (e.g., proteinase K) are controlled such that the extent of primary cleavages is largely determined by the protein structural features. Such cleavages have been shown to preferably occur in the locally unfolded regions of proteins, thus they can be exploited to probe the more local unfolding/refolding properties of proteins. After the
limited proteolytic digestion reaction is quenched with chemical denaturant, the protein samples are subjected to a trypsin digestion under denaturing conditions for bottom-up proteomics analysis. A control group is also generated from the same proteome with trypsin digestion only to correct for protein abundance differences, incomplete trypsin specificity and endogenous protease cleavages. Ultimately, proteins with altered conformations will exhibit a significant shift in the abundance of the corresponding semi- or fully-tryptic peptides. The quantitation in the LiP technique is accomplished using a label-free MS approach. As part of the work described in this thesis, SILAC-based quantitation strategy is incorporated into the LiP technique to enable more accurate quantitation.

1.2.3.2 Advantages and Limitations

The main advantage of native limited proteolysis is the ability to detect local folding/unfolding properties of proteins. These subtle conformation changes do not have a strong denaturant dependence on their energetics, thus typically go undetected using other approaches (e.g., SPROX). The protocol is also much easier in that it eliminates the need for a series of chemical or heat denaturation. However, it sacrifices the deeper information about protein thermodynamic property (e.g., folding free energy, dissociation constant). Another limitation about LiP is that not all proteins have protease susceptible regions under native conditions. This is improved in a later study that has performed limited proteolysis at a very low concentration of denaturant (1 M urea) to
assist unfolding of proteins and increase their proteolytic susceptibility. The study also added a precipitation step with trichloroacetic acid (TCA) after limited proteolysis to remove partially digested and undigested proteins, which can largely reduce sample complexity (16). In addition, hit selection on the peptide level with only one data point is prone to error, thus multiple replicates should be performed. Lastly, semi-tryptic peptides tend to be singly charged during electrospray ionization, which may make their sequencing difficult in shotgun proteomics experiments. Researchers are investigating strategies for the enrichment of semi-tryptic peptides.

1.2.3.3 Targeted LiP

The proteome-wide LiP technique was originally developed to be amenable to both targeted and shotgun proteomic analysis with the use of a label-free quantitation strategy. The sensitivity and background filtering capabilities of SRM MS is especially useful to reproducibly probe the proteolytic patterns generated by limited proteolysis. Peptides with protease susceptibility changes in different conditions are first identified through a discovery phase using shotgun proteomics and then validated by SRM measurements to assess their precise abundance differences. Using such a strategy the study has successfully proved the ability of LiP to capture both subtle and pronounced structural changes of proteins on a large scale.
1.2.3.4 Using LiP for Disease State Characterization

The proteome-wide LiP technique was established with a proof-of-principle study that investigated the conformational changes of proteins from yeast cultures subjected to a metabolic transition. For the first time, the study assessed the structural features of more than 1000 yeast proteins on a large scale and detected altered conformations for ~300 proteins upon a metabolic switch from glucose- to ethanol-based growth. This study again demonstrated the feasibility and benefits to use proteome-wide protein stability measurements to characterize different biological states. Furthermore, the LiP technique offers the possibility to probe the more local unfolding/refolding properties of proteins, thus it can be used together with approaches that are sensitive to changes in the more global unfolding/refolding properties of proteins to provide complementary information about disease states.

1.3 Overview of Breast Cancer

1.3.1 Early Diagnosis and Protein Biomarker Discovery

Breast cancer is the leading cause of cancer death among women worldwide(38). The 5-year survival rate for breast cancer is close to 97% when tumors are confined to breast tissue, but decreases dramatically to 23% when tumors have metastasized to other organs at the time of diagnosis(39). Thus early detection has the potential of not only saving lives but also ameliorating the sufferings of patients for better quality of life. This has prompted research efforts directed towards better understanding of molecular
mechanisms involved in the cancer development to identify effective biomarkers and therapeutic targets.

Tremendous biomarker discovery efforts have been made over the past few decades in the field of proteomics. These efforts mainly involve protein expression profiling studies using microarrays and mass spectrometry to understand the differences between a normal and a disease state(40-51). The results of these studies have helped provide a better understanding of the biology of breast cancer, but they have not been as useful as expected at identifying disease-related biomarkers that can be exploited for diagnostic and therapeutic purposes. A striking discrepancy exists between the large number of reported candidate markers that have come out of these studies and the number of markers that make it into clinical practice(52). Table 1 lists the handful of US Food and Drug Administration (FDA) approved protein biomarkers in current clinical use by 2013(53).

One possible reason for this discrepancy may be the indirect link between a protein’s expression level and its function; that is, disease-related changes in protein function may not always be reflected in the changes in their abundance. Furthermore, functionally relevant proteins with the same expression levels in different biological states can go undetected using the expression level profiling strategies. To obtain more robust biomarkers that can perform better in their translation into diagnostic or
therapeutic practice, the need arises for approaches capable of probing more biologically relevant dimension of the proteome.

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Clinical use</th>
<th>Specimen</th>
<th>Methodology</th>
<th>Year first approved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulating Tumor Cells (EpCAM, CD45, cytokeratins 8, 18+, 19+)</td>
<td>Prediction of cancer progression and survival</td>
<td>Whole blood</td>
<td>Immunomagnetic capture/ immune-fluorescence</td>
<td>2005</td>
</tr>
<tr>
<td>Estrogen receptor (ER)</td>
<td>Prognosis, response to therapy</td>
<td>FFPE tissue</td>
<td>Immunohistochemistry</td>
<td>1999</td>
</tr>
<tr>
<td>Progesterone receptor (PR)</td>
<td>Prognosis, response to therapy</td>
<td>FFPE tissue</td>
<td>Immunohistochemistry</td>
<td>1999</td>
</tr>
<tr>
<td>HER-2/neu</td>
<td>Assessment for therapy</td>
<td>FFPE tissue</td>
<td>Immunohistochemistry</td>
<td>1998</td>
</tr>
<tr>
<td>CA15-3</td>
<td>Monitoring disease response to therapy</td>
<td>Serum, plasma</td>
<td>Immunoassay</td>
<td>1997</td>
</tr>
<tr>
<td>CA27-29</td>
<td>Monitoring disease response to therapy</td>
<td>Serum</td>
<td>Immunoassay</td>
<td>1997</td>
</tr>
</tbody>
</table>
Later, activity-based protein profiling (ABPP) was introduced for such a purpose (54). ABPP provides a more direct measure of protein function. It relies on the design of active-site-directed covalent probes and mass spectrometry platforms to interrogate specific enzyme families in complex biological mixtures. The technique has been used to detect catalytic activity differences of a number of enzyme classes in different biological systems that include disease states (54). These studies are attractive because they can probe the functional significance of proteins with a range of expression levels. However, ABPP requires development of specialized probes for different enzyme families and only one specific enzyme class can be probed at a time.

Another functionally relevant avenue to characterizing disease states beyond expression level profiling is to exploit the protein folding and stability property. Strategies for proteome-wide protein stability measurements described earlier in this chapter are expected to provide novel information for detecting changes in disease states that is not limited to specific protein class.

1.3.2 Molecular Classification of Breast Cancer

Breast cancer is highly heterogeneous in that it can exhibit a number of different biological characteristics. Thus, there are a number of different subtypes of the disease. For example, breast cancer has been classified into five subtypes based on molecular profiling using DNA microarrays: luminal-A, luminal-B, basal-like, HER2-enriched, and normal breast-like groups (55-56). The luminal subtypes, which are the most common
subtypes among breast cancers\(^57\), display a high expression of hormone receptors (estrogen receptor/progesterone receptor, ER/PR) and associated genes, and they generally carry a good prognosis. Compared to the luminal-A subtype, the luminal-B subtype exhibits relatively high proliferation rate and less favorable prognosis\(^58\).

Luminal tumors usually response well to hormone therapy yet poorly to conventional chemotherapy. The basal-like subtype has high expression of basal markers (such as cytokeratins 5, 6, 14, 17) and proliferation related genes and typically no expression of hormone receptors or HER2, a member of the epidermal growth factor family of receptors. Although the basal subtype is only found in \(~15\%\) of breast cancer diagnoses, it has been associated with aggressive behavior and poor prognosis\(^57\). Given the triple negative receptor status, basal tumors are not amenable to conventional targeted breast cancer therapies, leaving chemotherapy the only option. The HER-2 enriched subtype also carries a poor prognosis. But unlike the basal subtype, molecularly targeted agents such as the anti-HER2 monoclonal antibody, trastuzumab, are available for HER2 over-expression cancers. The normal-like subtype is similar to the luminal-A subtype: hormone receptors positive and HER2 negative with low levels of the protein Ki-67, which helps control how fast cancer cells grow. The normal-like subtype resembles the normal breast profiling with slightly worse prognosis than the luminal-A subtype\(^59\).

Table 2 summarizes the five breast tumor molecular subtypes\(^59\). The heterogeneity underscores the importance of developing disease biomarkers and
therapeutic targets that are subtype specific. Thus, it is necessary to study the different subtypes using the global thermodynamic profiling strategies to obtain a better understanding of the disease and to uncover unique and shared molecular features in these subtypes.

Table 2: Molecular classification of breast cancer.

<table>
<thead>
<tr>
<th>Subtypes</th>
<th>Immunoprofile</th>
<th>Grade</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal-A</td>
<td>ER+, PR+, HER2-, Ki67-</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Luminal-B</td>
<td>ER+, PR+, HER2+/-, Ki67+</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Basal-like</td>
<td>ER-, PR-, HER2-</td>
<td>3</td>
<td>Poor</td>
</tr>
<tr>
<td>HER2-enriched</td>
<td>ER-, PR-, HER2+</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Normal breast-like</td>
<td>ER+, PR+, HER2-, Ki67-</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

1.4 Dissertation Focus

The main focus of this dissertation is the development and application of proteome-wide stability profiling strategies to characterize breast cancer. One goal of the work is to add to a growing body of evidence that global thermodynamic profiling strategies can be used to classify disease states. It may identify protein hits that are known to play a role in tumorigenesis and cancer progression. In such a case, the additional stability information can help elucidate the molecular basis for the
disregulation and/or dysfunction of these proteins in cancer. It may also identify novel protein hits that will otherwise go undetected in an expression level study. In such a case, the study opens the door for a new region of the proteome that has not been associated with the disease and provides additional insights into the disease.

Of particular significance to the work in this dissertation is one previous SILAC-SPROX analysis of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell lines to study disease-related changes in protein folding and stability (35). The work in Chapter 2 of this dissertation extends this initial SILAC-SPROX study to several additional cancer subtypes in order to identify subtype specific thermodynamic stability profiles. Chapter 3 describes the adaptation of the recently established LiP protocol that incorporates the SILAC-based quantitation strategy to facilitate more accurate and reproducible quantitation in the bottom-up shotgun proteomics analysis. The SILAC-LiP protocol developed as part of this work was demonstrated in a proof-of-principle study using a ubiquitous ligand, adenosine triphosphate (ATP), to detect its binding interactions in a yeast cell lysate. With the feasibility of the SILAC-LiP approach established, Chapter 4 describes the use of SILAC-LiP to study the conformational changes of proteins involved in breast cancer pathogenesis using the same two cell lines that were used in the previous SILAC-SPROX study by Adhikari and coworkers (35). One goal for this part of the work was to compare the SILAC-LiP protein hits with the SILAC-SPROX protein hits. The types of protein conformational changes probed by the SPROX and LiP
techniques are different. Thus, the two techniques are expected to provide complementary biologically relevant information on the disease states. Chapter 5 is focused on the use of the targeted PAB-SPROX strategy to develop validation assays for the differentially stabilized protein hits identified in the SILAC-SPROX analysis. The goal for this part of work was to prioritize a subset of protein hits and establish PRM assays using cell culture models. Work in the appendix is the application of the SPROX and TPP protocols in the detection of novel protein targets of small molecules, including inositol pyrophosphate in the context of human colon cancer cell lines and phosphatidylinositol 3-phosphate in the context of malaria parasite.
2. Large-Scale Analysis of Protein Folding and Stability Changes Associated with Breast Cancer Using SILAC-SPROX

The work described in this chapter comes largely from the research paper titled “Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using SILAC-SPROX” that was published in 2017 on the Journal of Proteome Research (vol. 16(9), p. 3277-3286)(2).

2.1 Introduction

Protein biomarker discovery efforts in breast cancer and other disease states have typically relied on the large-scale protein expression level profiling strategies using various mass spectrometry-based proteomic approaches. Such protein biomarker studies of breast cancer have included analyses of cell culture models(39, 48-49, 60-64) and clinical specimens of the disease (e.g., breast tissues(62, 65-69) and various types of biofluids(70-75)). While these studies can provide useful information, expression levels are not directly tied to functional differences. In contrast, the link between protein folding stability and function is more direct. Because changes in protein folding stability can result from a number of different biologically significant phenomena such as point mutations, post-translational modifications, or altered binding interactions with cellular ligands, they have the potential to be a general probe of protein function and to be sensitive to a number of different types of functionally relevant and disease-related changes.
The work described in this chapter was designed to extend the initial SILAC-SPROX study of three breast cancer cell lines (MCF-7, MCF-10A, and MDA-MB-231)(35) to an additional two breast cancer cell lines (BT-474 and MDA-MB-468). The original study successfully used protein folding and stability measurements to differentiate the MCF-7, MCF-10A, and MDA-MB-231 breast cancer cell lines and identified novel molecular signatures of breast cancer(35). However, the high heterogeneity of breast cancer makes it necessary to study more subtypes to obtain a better understanding of the disease. To this end, this work seeks to characterize several additional cancer subtypes using the Stable Isotope Labeling with Amino Acids in Cell Culture and Stability of Proteins from Rates of Oxidation (SILAC-SPROX) technique(20).

Described here are the results of SILAC-SPROX analyses performed in two different comparisons. Protein folding stability differences were studied in a comparison of two luminal breast cancer subtypes, luminal-A and -B (i.e., MCF-7 and BT-474 cells, respectively), and in a comparison of a luminal-A and basal subtype of the disease (i.e., MCF-7 and MDA-MB-468 cells, respectively). The 242 and 445 protein hits identified with altered stabilities in these comparative analyses included a large fraction with no significant expression level changes. This suggests thermodynamic stability measurements create a new avenue for protein biomarker discovery and understanding the molecular basis of disease. A number of the identified protein hits have been previously found to be biologically relevant to the disease. This not only substantiates
the biological significance of the protein hits identified using the SILAC-SPROX approach, but it also establishes a potential biophysical link between the differentially stabilized proteins and the specific disease states. Also included in this work is the use of a new automated data analysis pipeline for hit identification in SILAC-SPROX experiments.

2.2 Experimental

2.2.1 Cell Culture and Cell Lysate Preparation

The breast cancer cell lines BT-474 and MDA-MB-468 were cultured following the American Type Culture Collection (ATCC) guidelines. The heavy SILAC-labeled breast cancer cell line, MCF-7, was cultured using heavy-labeled lysine and arginine according to the established protocols(76). Starting cells for each of the cell lines were acquired from the ATCC (provided by the Cell Culture Facility at Duke University). The BT-474 cells were maintained in RPMI 1640 medium (Sigma) containing 2.5 g/L glucose (Sigma), 10 mM HEPES (Gibco), 1 mM sodium pyruvate (Gibco), 10% fetal bovine serum (FBS) (Hyclone) supplemented with 10 µg/mL of insulin (Gibco). The MDA-MB-468 cells were maintained in Leibovitz’s L-15 medium (Gibco) containing 10% FBS.

The MCF-7 cells were initially propagated in DMEM medium (Gibco) containing 1 mM sodium pyruvate, 0.1 mM non-essential amino acids (NEAA) (Gibco) and 10% FBS supplemented with 10 µg/mL of insulin. The MCF-7 cells were passaged twice in this medium and then transferred to DMEM medium containing all of the above except
10% dialyzed FBS was used instead of the regular serum. The cells were adapted well in this medium before passaging on the heavy SILAC medium. The heavy SILAC medium comprised of the SILAC DMEM (Thermo Scientific), 10% dialyzed FBS (Sigma), 10 µg/mL of insulin with 500 µL of $^{13}$C₆ Arginine (86.2 mg/L) and 500 µL of $^{13}$C₆$^{15}$N₂ Lysine (181.2 mg/L) (Cambridge Isotope Laboratories, MA) and 500 µL of Proline (200 mg/L) (Sigma). The MCF-7 cells were passaged at least four times in the heavy SILAC medium before they were harvested.

All cells were maintained in a humidified 37 °C incubator with 5% CO₂ except the MDA-MB-468 cells were cultured in the absence of CO₂. All cells were washed twice with PBS before they were harvested. The BT-474 and MDA-MB-468 cells were harvested with 0.25% (w/v) Trypsin/0.53 mM EDTA solution (Gibco). The MCF-7 cells were harvested with HyQtase solution (Hyclone). The harvested cells were pelleted at 1000 rpm for 5 min. The resulting cell pellets were washed with PBS and stored at -80 °C.

The cells were lysed by mechanical disruption in the presence of 1 mm diameter zirconia/silica beads (Biospec) using a Disruptor Genie (Scientific Industries). This involved 10 cycles consisting of 20 s of disruption and 1 min of cooling on ice. The lysis buffer was 20 mM phosphate buffer (pH 7.4) containing a cocktail of protease inhibitors that included the following: 1 mM 4-(2-aminoethyl)benzenesulfonyl uoride hydrochloride (AEBSF), 500 µM Bestatin, 15 µM E-64, 20 µM Leupeptin, and 10 µM
Pepstatin A (Thermo Pierce). Lysed cells were centrifuged at 15,000 g for 15 min at 4 °C, and the supernatant was used for subsequent analyses.

2.2.2 SILAC-SPROX Analyses

The SILAC-SPROX technique was used in two different comparative studies (including one involving heavy-labeled MCF-7 and light-labeled BT-474 cell lines and one involving heavy-labeled MCF-7 and light-labeled MDA-MB-468 cell lines). Each comparative analysis was performed in triplicate as previously described(20). Briefly, in each comparison, the total protein concentrations in the light- and heavy-labeled cell lysate samples were determined using a Bradford assay and normalized to the same concentration (4–6 mg/mL total protein depending on the experiment). Aliquots (20 µL) of the light- and heavy-labeled cell lysate samples were each mixed with 75 µL of a series of 20 mM phosphate buffers (pH 7.4) containing increasing concentrations of urea (0–9 M). The final urea concentrations ranged from 0 to 7 M. Each mixture was equilibrated for 16–18 h, and 5 µL of 30% (w/w) hydrogen peroxide (Sigma) was added to each of the protein samples in the denaturant-containing buffers. After 6 min, the reaction in each protein sample was quenched with 734 µL of 375 mM L-methionine (Sigma). The light- and heavy-labeled cell lysate samples that corresponded to the same denaturant concentration were combined, and the proteins were precipitated after addition of 318 µL of TCA (1 g/mL) (Sigma) and an overnight equilibration at 0 °C. The
resulting protein pellets were subjected to a quantitative, bottom-up proteomics analysis using SILAC quantitation.

### 2.2.3 Proteomic Sample Preparation

The protein pellets from each denaturant-containing buffer were dissolved in 60 µL of 0.5 M triethylammonium bicarbonate (TEAB) buffer (Sigma) containing 0.1% sodium dodecyl sulfate (SDS). The disulfide bonds were reduced with 5 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP) (Thermo Fisher, Inc.) for 1 hour at 60 °C. The samples were then treated with 10 mM methylmethane thiosulfonate (MMTS) (Sigma) for 15 min at room temperature. Ultimately, the protein samples were digested with trypsin using an enzyme/substrate ratio of 1/50 (w/w) at 37 °C with overnight incubation. The proteolytic digestion reaction was quenched upon acidification (pH ~ 2–3) with trifluoroacetic acid (TFA) (Sigma).

### 2.2.4 LC-MS/MS Analyses

The peptide mixtures were desalted using C18 resin (The Nest Group) according to the manufacturer’s protocol. LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive Plus high-resolution mass spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. Samples were trapped on a Symmetry C18 300 mm × 180 µm trapping column for 3 min at 5 µL/min (99.9/0.1 v/v water/acetonitrile 0.1% formic acid) and separated on a 75 µm × 250 mm column packed with a 1.7 µm Acquity
HSST3 C18 stationary phase (Waters Corp.). Peptides were separated using a gradient of 3 to 30% acetonitrile with 0.1% formic acid over 90 min at a flow rate of 0.4 µL/min with a column temperature of 55 °C. Data collection was performed in a data-dependent acquisition (DDA) mode with a resolution of 70,000 (at m/z 200) for full MS scan from m/z 375–1600 with a target AGC value of $1 \times 10^6$ ions, followed by 20 product ion scans at a resolution of 17,500 (at m/z 200), using an AGC target value of $1 \times 10^5$ ions, a maximum fill time of 60 ms, and a normalized collision energy of 30 V. A 30 s dynamic exclusion was employed to decrease MS/MS oversampling.

### 2.2.5 Proteomic Data Analysis

The raw LC-MS/MS data files were searched using MaxQuant 1.5.2.8(77) against the 20,265 human proteins in the 2014-04 release of the UniProt Knowledgebase downloaded on 5/16/2014. Searches were performed with fixed MMTS modification on cysteine and SILAC labeling of lysine and arginine, variable oxidation of methionine, deamidation of asparagine and glutamine, and acetylation of the protein N-terminus. Trypsin was set as enzyme, and up to two missed cleavages were allowed. The mass tolerance for precursor ions was set to 20 ppm for the first search where initial mass recalibration was performed, and a 4.5 ppm precursor mass tolerance was used for the main search. The mass tolerance for fragment ions was set to 0.02 Da. Also included were a match between runs and re-quantification of the searched peptides. The rest of the parameters were set at the default settings. In cases where a given peptide was
matched to multiple protein isoforms or multiple members of a protein family, the peptide was assigned to the leading razor protein listed by the MaxQuant algorithm. The false discovery rate for peptides and proteins identification was set at <1%.

Data analysis was accomplished using a Mathematica 11.0 script that was developed in-house and designed with an improved data analysis workflow compared that described in our earlier SILAC-SPROX work(20, 35). Briefly, in each biological replicate, only peptides identified with heavy to light (H/L) ratios > 0 were used for the following analysis. To normalize protein expression level differences in the cell lines, the H/L ratios of non-methionine-containing peptides from a given protein were divided by the median H/L ratio of all non-methionine-containing peptides from that protein. The H/L ratios of the methionine-containing peptides were also normalized by the corresponding protein medians. The normalized H/L ratios measured for a given methionine-containing peptide and charge state at a given denaturant concentration were averaged and log 2 transformed. Only methionine-containing peptides identified in seven or more denaturant concentrations were considered for hit selection, which initially involved identification of methionine-containing peptides with two or more consecutive log2(normalized H/L) values that were significantly different from 0. Log2(normalized H/L) values that were significantly different from 0 were considered to be those that fell below or above the 5th and 95th percentiles, respectively, of log2(normalized H/L) values of all the non-methionine-containing peptides detected.
across all the denaturant concentrations (Figure 2). The 5th and 95th percentiles for the log$_2$(normalized H/L) values varied slightly depending on the biological replicate (Figure 2), but they were generally close to −1 and 1, respectively. Subsequently, the log$_2$(normalized H/L) values versus denaturant concentration data obtained for oxidized and wild type methionine-containing peptides with 2 or more consecutive log$_2$(normalized H/L) values that were significantly different from 0 were fitted to Equations 1 and 2, respectively.

Equation 1

$$\text{Log}_2(H/L) = \text{Log}_2\left(\frac{1 - A \cdot e^{1 + e^{-\frac{-k_{ox}t}{RT}}}}{1 - A \cdot e^{1 + e^{-\frac{-k_{ox}t}{RT}}}}\right)$$

Equation 2

$$\text{Log}_2(H/L) = \frac{1.44k_{ox}t}{1 + e^{-\frac{\Delta G + m_2[D]}{RT}}} - \frac{1.44k_{ox}t}{1 + e^{-\frac{\Delta G + \Delta \Delta G + m_1[D]}{RT}}}$$

Each equation represents the ratio of the two SPROX curves expected for the heavy- and light-labeled peptide. In each case, $A$ is the extent of the oxidization (i.e., the amplitude of the SPROX curve), $k_{ox}$ is the pseudo-first-order rate constant associated with the oxidation reaction of an unprotected methionine residue, which has been previously determined to be $0.0065 \text{ s}^{-1}$ when the concentration of H$_2$O$_2$ is 0.5 M(78), $t$ is the oxidation reaction time (360 s), $\Delta G$ is the protein folding free energy, $\Delta \Delta G$ is the
change in the folding free energy, \( m_1 \) and \( m_2 \) are \( \delta \Delta G / \delta [\text{Urea}] \), representing the steepness of the transition regions of the two SPROX curves, \( R \) is the ideal gas constant, and \( T \) is temperature. Constraints were applied in order to avoid overfitting and ensure the physical significance of the parameters.

The goodness of fitting was evaluated by the \( R^2 \) extracted from the fitting and a customized deviation using Equation 3.

\[
CDev = \sqrt{\frac{\sum_{i=1}^{n} (A_{\text{Fit},i} - A_{\text{Exp},i})^2}{n - 1}} / \frac{\text{Max}(A_{\text{Exp},i}) - \text{Min}(A_{\text{Exp},i})}{n - 1}
\]

In Equation 3, \( n \) is the total number of denaturant concentrations identified, \( A_{\text{Exp},i} \) is the H/L ratio for the \( i^{th} \) data point from experiment, and \( A_{\text{Fit},i} \) is the H/L ratio for the \( i^{th} \) data point calculated from the fitted model. Smaller \( CDev \) indicates better fitting. Data sets with \( R^2 > 0.8 \) and \( CDev < 0.2 \) were classified as good fitting.

In order to achieve optimal fitting results, the data was fitted an additional \( n \) times with the systematic removal of one point, where \( n \) is the total number of denaturant concentrations identified. The best regression was selected by the highest \( R^2 \) value.
Figure 2: Global distributions of the log\(_2\) (normalized H/L) values determined for all the non-methionine-containing peptides identified at all the denaturant concentrations in the (A) MCF-7 versus BT-474 and (B) MCF-7 versus MDA-MB-468 cell line comparisons. The arrows indicate the 5th and 95th percentile cut-off values in each biological replicate. The median log\(_2\) (normalized H/L) values were 0 in all distributions and the standard deviations ranged from 0.62-0.80.
After fitting the data for each methionine-containing peptide with 2 or more consecutive normalized H/L ratios different from 1 in each biological replicate, a single data set was generated for each of the assayed peptides by averaging the well-fit data ($R^2 > 0.8$ and $CDev < 0.2$) from all the biological replicates for that peptide. This averaged data set was again fitted to Equation 1 or 2, as described above. Ultimately, hit peptides were selected as those with averaged data sets that were well-fit to Equation 1 or 2. Transition midpoints (i.e., $C_{1/2}$ values) were calculated from the fitted parameters using Equations 4 and 5 for heavy- and light-labeled peptides, respectively.

$$C_{1/2,\text{Heavy}} = \frac{-0.28 - \Delta G - \Delta\Delta G}{m_1}$$  \hspace{1cm} \text{Equation 4}

$$C_{1/2,\text{Light}} = \frac{-0.28 - \Delta G}{m_2}$$  \hspace{1cm} \text{Equation 5}

2.3 Results

2.3.1 Experimental Design

The experimental workflow used in this study is shown in Figure 3. The cell culture models of breast cancer studied here included the MCF-7, BT-474, and MDA-MB-468 breast cancer cell lines. Each cell line represents a different tumor subtype and has different molecular features. The MCF-7 cell line is ER+/PR+/HER2− and reflective of the luminal-A subtype of breast cancer. The BT-474 cell line is ER+/PR+/HER2+ and is
representative of the luminal-B subtype of breast cancer. The MDA-MB-468 cell line is a basal-like subtype and triple negative (ER−/PR−/HER2−).

Two different comparative analyses were conducted using the SILAC-SPROX protocol (20) (Figure 3). In one analysis, the thermodynamic stabilities of proteins in the BT-474 cell lysate were compared to those in the MCF-7 cell lysate. In a second analysis, the thermodynamic stabilities of proteins in the MCF-7 cell lysate were compared to those in the MDA-MB-468 cell lysate. The MCF-7 cell line was heavy-labeled in both comparative analyses. Hit proteins with different stabilities in the two cell lines were identified based on the H/L ratios of the methionine-containing peptides identified and quantified in the SILAC-SRPOX experiment (Figure 3).
Figure 3: Schematic representation of the experimental workflow used in this work. Shown at the bottom are the protein unfolding curves extracted from the corresponding SILAC-SPROX behaviors shown for a wild type methionine-containing peptide.
2.3.2 Proteomic Coverage

Summarized in Table 3 is the proteomic coverage obtained in the SILAC-SPROX experiments performed in this work. In total, approximately 3500 methionine-containing peptides from close to 1000 proteins were assayed in the MCF-7 versus BT-474 cell line comparison, and about 3000 methionine-containing peptides from close to 800 proteins were assayed in the MCF-7 versus MDA-MB-468 cell line comparison. To be assayed for protein folding stability changes in the SILAC-SPROX experiment, a protein must be consistently identified with the same methionine-containing peptide from at least seven denaturant concentrations in the bottom-up shotgun proteomics analysis. Due to this limitation, the number of peptides and proteins assayed in the SILAC-SPROX experiment is significantly smaller (i.e., ~8-fold lower) than the total number of peptides and proteins successfully identified and quantified in the LC-MS/MS readout (Table 3).

Table 3: Summary of the proteomic data obtained in the SILAC-SPROX experiments for differential thermodynamic stability analysis in the MCF-7 vs. BT-474 and MCF-7 vs. MDA-MB-468 cell line comparisons.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Biological Replicate</th>
<th>Identified Peptides (Proteins)</th>
<th>Assayed Peptides (Proteins)</th>
<th>Hit Peptides (Proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF-7 versus BT-474</td>
<td>1</td>
<td>12388 (1954)</td>
<td>1172 (417)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16903 (2511)</td>
<td>1572 (554)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>22096 (2581)</td>
<td>2914 (765)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>28098 (3705)</td>
<td>3464 (957)</td>
<td>413 (257)</td>
</tr>
<tr>
<td>MCF-7 versus MDA-MB-468</td>
<td>1</td>
<td>13089 (1981)</td>
<td>629 (249)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10507 (1704)</td>
<td>796 (302)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>23999 (2642)</td>
<td>2833 (710)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>26571 (3460)</td>
<td>2980 (790)</td>
<td>1012 (450)</td>
</tr>
</tbody>
</table>
The frequency of methionine residues in protein sequences is about 2.5%. Therefore, a typical protein generally contains multiple methionine residues. In the case of small, single domain proteins with highly cooperative protein unfolding/folding reactions, the different methionine-containing peptide probes detected in the SILAC-SPROX experiment provide redundant biophysical information (i.e., report on the same chemical denaturation behavior). However, in the case of large, multi-domain proteins, the different methionine-containing peptide probes detected in SILAC-SPROX provide unique biophysical information about the specific protein domains to which they map (i.e., report on domain specific chemical denaturation behavior).

2.3.3 Hit Identification

In total, 413 methionine-containing peptides from 257 proteins were identified as hits in the MCF-7 versus BT-474 cell line comparison, and 1012 methionine-containing peptides from 450 proteins were identified as hits in the MCF-7 versus MDA-MB-468 cell line comparison. Representative peptide hits are shown in Figure 4. The peptide hits accounted for 12% and 34% of the assayed methionine-containing peptides in each comparison, respectively. The false positive rate of peptide hit discovery in the current study can be estimated by using the hit selection criteria described here to select hits among the non-methionine-containing peptides. Such an analysis of the non-methionine-containing peptide data from the two cell line comparisons revealed peptide hit rates of ~5%. This estimate of the false positive rate of peptide hit discovery is similar
to the 3.5% false positive rate of peptide hit discovery that was previously determined in a control experiment involving light- and heavy-labeled MCF-7 cells(20).

The 12 and 34% peptide hit rates observed in the two cell line comparisons described here are significantly above these estimated false positive rates of peptide hit discovery using the SILAC-SPROX protocol. The hit rate in the MCF-7 versus BT-474 cell line comparison is also similar to that observed in previous SILAC-SPROX analyses of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell lines (i.e., 10–12%)(35). The hit rate in the MCF-7 versus MDA-MB-468 cell line comparison is relatively high.
Figure 4: Representative SILAC-SPROX data and associated protein unfolding curves obtained in this work. (A) Data obtained on the peptide AMEVDERPTEQYSDIGGLDK from 26S protease regulatory subunit 6A in the MCF-7 versus MDA-MB-468 cell line comparison. (B) Data obtained on the peptide AM(ox)-EVDERPTEQYSDIGGLDK from 26S protease regulatory subunit 6A in the MCF-7 versus MDA-MB-468 cell line comparison. (C) Data obtained on the peptide DHASIQM(ox)NVAEVDKVTGR from 40S ribosomal protein S21 in the MCF-7 versus BT-474 cell line comparison. (D) Data obtained on the peptide NPPEEALEDTLNQVMVVF from Cullin-1 in the MCF-7 versus BT-474 cell line comparison. (E) Data obtained on a subset of the non-methionine-containing peptides from 26S protease regulatory subunit 6A in the MCF-7 versus MDA-MB-468 cell line comparison including: DSYLILETLPTYEYDSR (●), EKASIIFIDELDAIGTK (■), LKPGDLVGVNK (●), QTLYLPVIGLVDAAK (▲), and VDILDPALLR (▼). In (A)–(D), the solid lines represent the best fit of the data to Equation 1 or 2; the dashed curves represent the extracted SPROX curves; the vertical lines indicate the transition
midpoints of the SPROX curves; A red × indicates a data point that was not included in the regression analysis.

As part of this work, the hit identification strategy was improved by incorporating a non-linear regression analysis of the data to select hits. Compared to the visual inspection strategy employed in ref (35), the hit identification strategy described here is more statistically rigorous and allows one to quantitatively evaluate the quality of data. When applied to the SILAC-SPROX data reported in ref (35), we found a large overlap with the peptide hits reported therein (i.e., ~50%, see Appendix C for additional details). The peptide hits in ref (35) that were not identified as hits with the new hit identification strategy primarily included (1) those with too few data points (i.e., the new regression analysis was only applied to peptides with H/L data from at least seven (instead of six) denaturant concentrations) and (2) those with poor quality data sets (i.e., the data was not well-fit to Equation 1 or 2).

The confidence level associated with a peptide hit identified in SILAC-SPROX is raised when both wild type and oxidized forms of the same peptide are consistently identified as hits (i.e., both forms of the peptide show the same stabilizing (or destabilizing) SILAC-SPROX behavior) (Figures 4A and 4B). The MCF-7 versus BT-474 and MCF-7 versus MDA-MB-468 cell line comparisons performed here identified 39 and 121 such hit peptide pairs (respectively), among which 4 and 103 pairs (respectively) showed consistent behaviors. These consistent peptide pairs act as “internal controls”
for the quantitative changes in their thermodynamic stabilities and help substantiate the results reported here. The hit peptide pairs with inconsistent data are likely false positives, and the peptides in these pairs were excluded from the final hit lists used in the following discussion. The final hit lists included 341 peptides from 242 proteins in the MCF-7 versus BT-474 cell line comparison, and 972 peptides from 445 proteins in the MCF-7 versus MDA-MB-468 cell line comparison.

The confidence level associated with a peptide hit identified in SILAC-SPROX is also raised when it is identified in multiple biological replicates or with multiple peptide probes that include the same methionine residue. In both of the cell line comparisons performed here, approximately 15% of the methionine-containing peptide hits were identified as hits in two or more biological replicates. Unfortunately, the remaining 85% of the methionine-containing peptide hits were only assayed in one biological replicate. However, 22% and 30% of the protein hits from the MCF-7 versus BT-474 and MCF-7 versus MDA-MB-468 cell line comparisons (respectively) were identified with multiple methionine-containing peptide probes, even though the methionine-containing peptides were only identified as a hit in one biological replicate. In some cases, multiple unique peptide sequences sharing the same methionine residue (in the same wild type (or oxidized) form) were assayed. A total of 5 of the 21 and 23 of the 43 sets of overlapping methionine-containing peptides in the MCF-7 versus BT-474 and MCF-7 versus MDA-MB-468 cell line comparisons (respectively) yielded consistent results. Approximately,
50% of the protein hits in each comparison were identified with a single peptide hit in one replicate.

### 2.3.4 Classification of Hit Behavior

The peptide hits identified in this work displayed the same three types of SILAC-SPROX behaviors as those observed in the previous SILAC-SPROX analyses of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell lines(35). The first type of behavior results when a methionine-containing peptide is globally protected in both cell lines, but the transition midpoints are shifted (see Figures 4A and 4B). The second type of behavior results when a methionine-containing peptide is globally protected in one cell line and it is only locally (or not at all) protected in the other cell line (Figure 4C). The third type of behavior results when a methionine-containing peptide undergoes different degrees of local protection between the two cell lines or undergoes a transition from local protection in one cell line to no protection in the other cell line (Figure 4D). These different types of peptide hits all indicate differential stabilizations of the proteins from which the peptides are derived.

The differential stabilizations in these proteins can result in shifts in the transition midpoints ($\Delta C_{1/2}$) and/or changes in the cooperativity of their folding/unfolding reactions ($m$-values). We note that $\Delta C_{1/2}$ and corresponding $\Delta \Delta G$ values can be determined only for those peptides that come from globally protected regions in proteins (Figures 4A and 4B). Also of note is that the accuracies of the $m$-value...
determinations in this work were limited due to the number of data points in each chemical denaturation set.

A total of 106 of the 341 peptide hits identified in the MCF-7 versus BT-474 cell line comparison and 467 of the 972 peptide hits identified in the MCF-7 versus MDA-MB-468 cell line comparison showed the first type of SILAC-S PROX behavior (Figures 4A and 4B) and had sufficient data points for protein stability changes quantification. Two-thirds of the 106 peptide hits with quantifiable protein stability changes in the MCF-7 versus BT-474 cell line comparison were the result of a protein stabilization in the BT-474 cell line. This is similar to that observed in the previous SILAC-S PROX analyses of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell lines in which ~50% of the protein hits were the result of stabilizations (35). Approximately 90% of the protein stabilizations and destabilizations detected in the MCF-7 versus BT-474 cell line comparison were the result of $C_{1/2}$ value shifts of 1.2 to 5.6 M. The median $\Delta C_{1/2}$ and $\Delta \Delta G$ values were 2.7 M and 1.7 kcal/mol, respectively. In contrast, 445 (95%) of the 467 peptide hits with quantifiable protein stability changes in the MCF-7 versus MDA-MB-468 cell line comparison resulted from protein stabilizations in the MDA-MB-468 cells. In general, the magnitudes of the detected stability changes for the MCF-7 versus MDA-MB-468 cell line comparison were similar to those above.
2.3.5 Protein Expression Level Analysis

The median H/L ratio of all non-methionine-containing peptides from a given protein can be used to evaluate the relative protein expression levels between the light- and heavy-labeled cell lines. We compared the expression level data generated in this work with that previously reported in the literature for the cell lines in this study (60) and found that they were in general agreement (Figure 5A). We further examined proteins with expression level changes >2-fold in this work and in ref (60) and also found a large degree of overlap (∼35%) between our data and the data in ref (60) (Figures 5B and 5C).

![Figure 5: Evaluation of the expression level data generated in this work. (A) Global distribution of the differences between the expression level data generated in this work and those reported in ref (60) for the MCF-7 versus BT-474 (solid line) and MCF-7 versus MDA-MB-468 (dashed line) cell line comparisons. Expression level](image-url)
data reported in ref (60) were processed to generate ratios of protein abundance in the MCF-7 cell line versus the other cell line (BT-474 or MDA-MB-468). (B) and (C) Venn diagrams showing the overlaps between proteins with expression level changes >2-fold identified in this work and those in ref (60) for the MCF-7 versus BT-474 (B) and MCF-7 versus MDA-MB-468 (C) cell line comparisons.

In both of the comparative analyses performed here, ~70% of the protein hits identified with thermodynamic stability differences between the two cell lines did not have significant (i.e., less than 2-fold) changes in their expression levels. This is similar to that observed in the previous SILAC-SPROX analyses of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell lines, where it was also noted that a large fraction (~45%) of the differentially stabilized proteins identified therein did not have significant changes in their expression levels(35). These results suggest that protein folding stability can provide information about disease states that is orthogonal to that obtained in protein expression level analyses, and thus may be a useful protein property to exploit in protein biomarker discovery studies.

2.4 Discussion
2.4.1 Comparison between Luminal-A and -B Subtypes

The differentially stabilized proteins identified in the BT-474 versus MCF-7 comparison were subjected to bioinformatics analyses using PANTHER to investigate their potential biological context(79-80). The glycolysis pathway was significantly overrepresented (i.e., 2.9-fold) in the protein hits (p-value <0.05). Increased glycolysis is
observed in cancer cells where they use this metabolic pathway for ATP generation (i.e., the Warburg effect). Notably, seven glycolytic enzymes that play important roles in the glycolytic metabolism (i.e., ALDOA, GAPDH, PGK1, PGAM1, ENO1, PKM, and LDHB) were identified with altered thermodynamic stabilities in the BT-474 versus MCF-7 comparison. Considering the close link between protein folding and function, this may indicate an altered glycolytic activity between the two cell lines, which can result from the higher proliferation rate observed in the luminal-B subtype of breast cancer. Indeed, it has been shown that phosphatidylinositol 3-kinase (PI3K) is often activated in ErbB2-overexpressing breast cancer cells and results in an overactive PI3K/AKT/mTOR pathway, which plays a critical role in controlling cell cycle, growth, and survival. Such an overactive PI3K/AKT/mTOR pathway has been shown to increase glucose uptake and lead to a switch from mitochondrial respiration to lactate production(81). A previous study has also shown that the PI3K/mTOR/p70S6K signaling pathway plays an enhanced role in the anchorage-independent growth of ErbB2 overexpressing breast cancer cells(82). Inhibition of glycolysis has been suggested to be a promising therapeutic strategy for cancer treatment, particularly in the case of cancers showing a high dependence on this bioenergetic metabolism(83).

A gene ontology term analysis of the protein hits identified in the BT-474 versus MCF-7 comparison reveals about a 2-fold increase in the fraction of cytoskeletal proteins as compared to those assayed (Figure 6A). These included 22 of the protein hits
described here, such as tubulins (e.g., TUBA1B, TUBB4B, TUBB2A, and TUBB6) as well as many actin related proteins (e.g., ACTG1, MYL3, MYH9, etc.; see Table 4). These cytoskeleton proteins are highly versatile in the regulation of cell cycle, cellular morphogenesis, and cell migration. They are responsible for driving chromosomal separation and cell division in the cell cycle, promoting stable cell–cell adhesions through interactions with cadherins during cellular morphogenesis, and forming protrusions to disseminate to the surrounding tissue during cell migration(84). The observation of cytoskeleton proteins being differentially stabilized in the two luminal subtypes may reflect a greater involvement of this protein class to meet the higher proliferation requirement in the luminal-B subtype.

Proteins with chaperone activity were also enriched in the hits identified in the BT-474 versus MCF-7 comparison (Figure 6A). These included a total of 18 protein hits (Table 5). Increased chaperone activity has been shown to be characteristic of tumor growth due to their primary function in facilitating protein folding. HSPs have been shown to have critical functions associated with the cell cycle and proliferative response(85-86). It is also interesting to note that six of the eight subunits of T-complex protein 1, which mediates the folding of the major cytoskeletal proteins tubulins and actins and has been implicated in the cancer cell proliferation, were identified as hits in this work(87). Our work suggests that the functions of these proteins may be disease state specific and result from differential stabilizations.
Table 4: Cytoskeletal proteins identified with altered thermodynamic stability in the SLAC-SPROX study of BT-474 versus MCF-7 cell line comparison.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Protein Name</th>
<th>Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTG1</td>
<td>Actin, cytoplasmic 2</td>
<td>Actin and actin related protein</td>
</tr>
<tr>
<td>MYL3</td>
<td>Myosin light chain 3</td>
<td>Actin family cytoskeletal protein</td>
</tr>
<tr>
<td>MYH9</td>
<td>Myosin-9</td>
<td>Actin binding motor protein</td>
</tr>
<tr>
<td>CLIC1</td>
<td>Chloride intracellular channel protein 1</td>
<td>Cytoskeletal protein</td>
</tr>
<tr>
<td>KRT1</td>
<td>Keratin, type II cytoskeletal 1</td>
<td>Intermediate filament</td>
</tr>
<tr>
<td>KRT18</td>
<td>Keratin, type I cytoskeletal 18</td>
<td>Intermediate filament</td>
</tr>
<tr>
<td>DYN1C1H1</td>
<td>Cytoplasmic dynein 1 heavy chain 1</td>
<td>Microtubule binding motor protein</td>
</tr>
<tr>
<td>KIF5B</td>
<td>Kinesin-1 heavy chain</td>
<td>Microtubule binding motor protein</td>
</tr>
<tr>
<td>EPPK1</td>
<td>Epiplakin</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>CORO1B</td>
<td>Coronin-1B</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>TMOD3</td>
<td>Tropomodulin-3</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>DBNL</td>
<td>Drebrin-like protein</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>DSP</td>
<td>Desmoplakin</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>FLII</td>
<td>Protein flightless-1 homolog</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>CTNNA1</td>
<td>Catenin alpha-1</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>TAGLN2</td>
<td>Transgelin-2</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>CTTN</td>
<td>Src substrate cortactin</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>DSTN</td>
<td>Destrin</td>
<td>Non-motor actin binding protein</td>
</tr>
<tr>
<td>TUBA1B</td>
<td>Tubulin alpha-1B chain</td>
<td>Tubulin</td>
</tr>
<tr>
<td>TUBB4B</td>
<td>Tubulin beta-4B chain</td>
<td>Tubulin</td>
</tr>
<tr>
<td>TUBB2A</td>
<td>Tubulin beta-2A chain</td>
<td>Tubulin</td>
</tr>
<tr>
<td>TUBB6</td>
<td>Tubulin beta-6 chain</td>
<td>Tubulin</td>
</tr>
</tbody>
</table>
**Table 5: Protein hits with chaperone activity identified in the SILAC-SPROX study of BT-474 versus MCF-7 cell line comparison.**

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Protein Name</th>
<th>Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>YWHAE</td>
<td>14-3-3 protein epsilon</td>
<td>Chaperone</td>
</tr>
<tr>
<td>YWHAZ</td>
<td>14-3-3 protein zeta/delta</td>
<td>Chaperone</td>
</tr>
<tr>
<td>HSPD1</td>
<td>60 kDa heat shock protein, mitochondrial</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>HSP90B1</td>
<td>Endoplasmin</td>
<td>Hsp90 family chaperone</td>
</tr>
<tr>
<td>HSPB1</td>
<td>Heat shock protein beta-1</td>
<td>Chaperone</td>
</tr>
<tr>
<td>HSP90AB1</td>
<td>Heat shock protein HSP 90-beta</td>
<td>Hsp90 family chaperone</td>
</tr>
<tr>
<td>NPM1</td>
<td>Nucleophosmin</td>
<td>Chaperone</td>
</tr>
<tr>
<td>FKBPIA</td>
<td>Peptidyl-prolyl cis-trans isomerase FKBPIA</td>
<td>Chaperone</td>
</tr>
<tr>
<td>FKBP3</td>
<td>Peptidyl-prolyl cis-trans isomerase FKBP3</td>
<td>Chaperone</td>
</tr>
<tr>
<td>FKBP4</td>
<td>Peptidyl-prolyl cis-trans isomerase FKBP4</td>
<td>Chaperone</td>
</tr>
<tr>
<td>PFDN2</td>
<td>Prefoldin subunit 2</td>
<td>Chaperone</td>
</tr>
<tr>
<td>STIP1</td>
<td>Stress-induced-phosphoprotein 1</td>
<td>Chaperone</td>
</tr>
<tr>
<td>TCP1</td>
<td>T-complex protein 1 subunit alpha</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>CCT2</td>
<td>T-complex protein 1 subunit beta</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>CCT5</td>
<td>T-complex protein 1 subunit epsilon</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>CCT3</td>
<td>T-complex protein 1 subunit gamma</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>CCT8</td>
<td>T-complex protein 1 subunit theta</td>
<td>Chaperonin</td>
</tr>
<tr>
<td>CCT6A</td>
<td>T-complex protein 1 subunit zeta</td>
<td>Chaperonin</td>
</tr>
</tbody>
</table>
Figure 6: Protein classes observed in the proteins assayed and identified as hits with altered thermodynamic stabilities in the (A) BT-474 versus MCF-7 and (B) MDA-MB-468 versus MCF-7 cell line comparisons. Each protein class contained at least four proteins. Protein classes enriched in the hits are indicated with a “*”.

Considering the different expression levels of HER2 in the MCF-7 (HER2-) and BT-474 (HER2+) breast cancer cell lines, the protein hits identified in the comparison were interrogated for potential direct and/or indirect interactions with HER2 using the
STRING database analysis(88). Two protein hits (plakoglobin and α-catenin), which are both in the catenin family, are known to have interactions with HER2. The catenin family is responsible for connecting cadherin receptor cytoplasmic domains to actin filaments in adherens junctions of epithelial cells. It includes β-catenin, which binds to the cytoplasmic domain of cadherin, α-catenin, which binds to β-catenin and actin, γ-catenin (plakoglobin), which is homologous to β-catenin, and δ-catenin. The cadherin–catenin complex (known as adherens junctions) plays an important role in stabilizing adhesion between cells and regulating cell growth. Direct association of HER2 with β-catenin and plakoglobin has been demonstrated in human cancer cells where the authors suggested that the HER2-mediated signaling might regulate the cell adhesion and invasive growth of cancers(89). Additionally, the actin binding protein vinculin was also identified with altered stability in the MCF-7 versus BT-474 cell line comparison. Thus, these protein hits together suggest changes in cell–cell adhesion between the two cell lines, which may be the result of the distinct proliferation signature in the MCF-7 and BT-474 cell lines.

2.4.2 Comparison between Luminal-A and Basal Subtypes

The MCF-7 versus MDA-MB-468 cell line comparison yielded a high hit rate (34%) and a large fraction of protein stabilizations (95%) in the MDA-MB-468 cell line compared to the MCF-7 cell line. This is in contrast to the other cell line comparison performed here as well as to the previous SILAC-SPROX analyses of the MCF-10A,
MCF-7 and MDA-MB-231 breast cancer cell lines in ref (35), where approximately equal
distribution between protein stabilizations and destabilizations was observed. One
explanation for the large number of stabilizations observed in the MDA-MB-468 cell line
is the more extreme phenotypic differences between this cell line and the MCF-7 cell
line. The basal subtype is proliferative, high-grade, and associated with a poor
prognosis, while the luminal-A subtype is usually low-grade and displays a relatively
favorable survival. Our work suggests that this thermodynamic stability signature of
stabilizations may be a major biophysical distinction between the luminal-A and basal
subtypes that the two cell lines represent. However, additional experiments on more cell
lines and tumor tissue samples are needed to further investigate this hypothesis. The
protein hits with altered thermodynamic stabilities identified in the MDA-MB-468
versus MCF-7 comparison covered a wide range of protein classes, including nucleic
acid binding proteins (21%), hydrolase (10%), transferase (9%), enzyme modulator
proteins (8%), cytoskeletal proteins (8%), oxidoreductase (7%), and chaperone proteins
(7%) (Figure 6B). However, no one class appeared to be overrepresented in the hits.

Notably, 14 of the protein hits identified in the MDA-MB-468 versus MCF-7
comparison are indicated in the ubiquitin proteasome pathway (UPP) as revealed by the
PANTHER pathway analysis (Table 6). The ubiquitin proteasome system is responsible
for the degradation of most intracellular proteins and therefore is an important regulator
of many critical cellular processes, including cell cycle progression, proliferation,
differentiation, and apoptosis. Since cancer is characterized by uncontrolled cellular proliferation or by a failure of cells to undergo apoptosis, the alteration of proteasome activity has been implicated in several cancers. Increasing evidence has shown that cancer cells exhibit high proteasome activity to support their uncontrolled cell division by getting rid of tumor suppressors such as p53, pro-apoptosis proteins such as Bax, and cell cycle inhibitors such as p27(90-91). Up-regulation of proteins involved in UPP has also been observed in cell transformation and tumorigenesis(92). It is interesting to note that only two proteins (UBA1, PSMC3) were annotated with association to the UPP among the differentially stabilized proteins identified in the BT-474 versus MCF-7 comparison. This suggests a similar degree of involvement for UPP in the two luminal subtypes and a very different degree of involvement for UPP between the luminal and basal subtypes of breast cancer.
Table 6: Protein hits involved in the ubiquitin proteasome pathway that were identified in the SILAC-SPROX study of MDA-MB-468 versus MCF-7 cell line comparison.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Protein Name</th>
<th>Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSMC6</td>
<td>26S protease regulatory subunit 10B</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMC1</td>
<td>26S protease regulatory subunit 4</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMC3</td>
<td>26S protease regulatory subunit 6A</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMC4</td>
<td>26S protease regulatory subunit 6B</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMC2</td>
<td>26S protease regulatory subunit 7</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMC5</td>
<td>26S protease regulatory subunit 8</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>PSMD1</td>
<td>26S proteasome non-ATPase regulatory subunit 1</td>
<td>Enzyme modulator</td>
</tr>
<tr>
<td>PSMD12</td>
<td>26S proteasome non-ATPase regulatory subunit 12</td>
<td>Enzyme modulator</td>
</tr>
<tr>
<td>PSMD14</td>
<td>26S proteasome non-ATPase regulatory subunit 14</td>
<td>Metalloprotease</td>
</tr>
<tr>
<td>PSMD6</td>
<td>26S proteasome non-ATPase regulatory subunit 6</td>
<td>-</td>
</tr>
<tr>
<td>SAE1</td>
<td>SUMO-activating enzyme subunit 1</td>
<td>Ligase</td>
</tr>
<tr>
<td>UBA2</td>
<td>SUMO-activating enzyme subunit 2</td>
<td>Ligase</td>
</tr>
<tr>
<td>UBE2N</td>
<td>Ubiquitin-conjugating enzyme E2 N</td>
<td>-</td>
</tr>
<tr>
<td>UBA1</td>
<td>Ubiquitin-like modifier-activating enzyme 1</td>
<td>Ligase</td>
</tr>
</tbody>
</table>

A major difference between the MCF-7 and MDA-MB-468 cells is that the MCF-7 cells have high expression of hormone receptors while the triple negative MDA-MB-468 cells lack specific cell surface receptors. This difference is likely to contribute to the protein hits observed in our work. Indeed, we found six protein hits that have previously known direct interactions with ER, including chaperone proteins (HSPA4, HSPA8, HSP90AA1), 26S proteasome (PSMC5, PSMD1), and MAPK1. These proteins are involved in many different pathways, such as those that regulate cell proliferation (the MAPK signal pathway). Changes in the thermodynamic stabilities of these proteins may induce a series of alterations in their downstream targets and the associated pathways, resulting in the high hit rate observed here.
2.5 Conclusions

This study extends the use of the thermodynamic stability profiling strategy to more breast cancer subtypes (i.e., luminal-A, luminal-B, and basal-like). The use of well-established cell lines in this study provides not only a reliable source of highly homogeneous cells but also an opportunity to validate the findings in this proof-of-principle work with previous studies involving these cell lines. Indeed, the thermodynamic stability measurements described here successfully differentiated the different cancer subtypes studied here and identified a subset of protein hits, a number of which have been previously found to be biologically relevant to the disease. These protein hits included many that are associated with cell proliferation processes, such as cell cycle, growth factor signaling, and metabolism, which is one of the major differences among the breast cancer subtypes in this study. Therefore, this study demonstrates the potential of using differential thermodynamic stability profiling as an unbiased approach for the discovery of disease-related biomarkers and the elucidation of biologically important aspects of diseases.

Notably, a significant fraction (~70%) of the protein hits did not have altered protein expression levels. This makes the differential thermodynamic profiling strategy especially attractive because such a strategy can not only establish a potential biophysical link between the differentially stabilized proteins and the specific disease states but also provide information about disease states that is orthogonal to that
obtained in protein expression level analyses. Therefore, it can be used to complement protein expression level studies especially when protein expression profiles fail to discriminate some disease states.
3. Development of a SILAC-Based Limited Proteolysis Approach

3.1 Introduction

The SILAC-SPROX technique described in Chapter 2 probes the equilibrium unfolding transitions in proteins that result from large amplitude motions (e.g., the unfolding/refolding of a protein domain or the entire protein). However, more local conformational changes can also occur as a result of ligand binding and/or disease states to serve specific functions (e.g., molecular recognition). Such local conformational changes have been documented in a number of well-studied model proteins. For example, the F helix region of holomyoglobin undergoes a conformational transition from a folded helical conformation to one or more locally disordered states upon the dissociation of heme to form apomyoglobin\(^{(93-94)}\). The hydrophobic surface area exposed in these local conformational changes is relatively small thus they do not have a strong denaturant dependence to their energetics (i.e., they have a small \(m\)-value\(^{(95)}\) and will go undetected using the SILAC-SPROX technique. A method that is sensitive to these local conformational changes could give insight into their critical roles in protein regulation and function.

The work described in this chapter is an adaptation of a recently established strategy for large-scale screening of both local and global conformational changes of proteins\(^{(14)}\). The strategy utilizes conformation specific limited proteolysis (LiP) of proteins in their native states by a non-specific protease to probe stability differences.
Proteins with altered conformations are expected to exhibit differential proteolytic patterns (i.e., differential abundance in the peptides that map to the protein regions undergoing conformational changes). In the original demonstration of the LiP technique, such abundance changes were quantified using a spectral counting approach that relies on counting and comparing the number of matched tandem mass spectra of a given peptide(14). Such a label-free quantitation strategy suffers from the commonly employed dynamic exclusion of ions that have already been selected for fragmentation, which results in a relatively poor accuracy. Developed in this chapter is a SILAC-based quantitation strategy for performing LiP experiments.

As part of the work, the so-called SILAC-LiP strategy was demonstrated in a proof-of-principle study to identify adenosine triphosphate (ATP) interacting proteins in yeast. The goal of this study was to demonstrate the ability of the method to discover protein targets of the selected ligand. A total of 79 proteins were found to have ATP-induced stability changes. A number of the identified protein hits (15 out of 79) were previously annotated as ATP-binding proteins. The other 64 non-annotated ATP-binding proteins included 20 proteins that were found to be ATP-sensitive in the earlier SPROX studies(20, 28).
3.2 Experimental

3.2.1 Yeast Cell Culture

The SILAC labeled yeast cells were cultured by Dr. Jagat Adhikari from the Fitzgerald lab. Briefly, the yeast deletion strain BY4739 (Open Biosystems, Lafayette, CO), an auxotroph for lysine, was streaked on a petri dish containing synthetic complete (SC) media with Difco™-agar and L-lysine. Per liter media for the plate was comprised of 1.7 g of yeast nitrogen base (Amresco), 5 g of ammonium sulfate (Sigma Aldrich), 20 g of Difco™-agar (BD, USA), 2 g of glucose (Sigma), 1.92 g of synthetic drop out mix without lysine (Sunrise Science Product, San Diego, CA), and 30 mg of L-lysine (Sigma Aldrich). Per liter SC media for the subsequent culture consisted of 1.7 g of yeast nitrogen base, 5 g of ammonium sulfate, 20 g of glucose, 1.92 g of synthetic drop out mix without lysine. After 3 days of growth at 30 °C, an isolated colony was selected and inoculated into 10 mL of SC media containing 30 mg/L of light L-lysine (light SC media). The 10 mL culture was incubated overnight at 30 °C. A 100 µL aliquot of the cell culture was transferred into 100 mL of the light SC media, while another 100 µL aliquot was transferred into 100 mL of SC media containing 30 mg/L of heavy lysine (13C6, 15N2) (Cambridge Isotopes Laboratories, MA) (heavy SC media). The two 100 mL cell cultures were incubated overnight at 30 °C before 100 µL of the light and heavy cell cultures were transferred into 100 mL of the corresponding light and heavy SC media. This step was repeated 3 times. Finally, when the measured OD600 was 6-8, 100 µL of the light and
heavy cell cultures were transferred to 1 L of the corresponding light and heavy SC media and grown overnight at 30 °C. Portions of the 1 L overnight culture (250 mL each) were centrifuged at 1000 g to generate light and heavy lysine labeled cell pellets. The cell pellets were stored at -20 °C.

### 3.2.2 Yeast Cell Lysate Preparation

For each experiment, one light and one heavy lysine labeled yeast cell pellet (stored at -20 °C) was thawed on ice. A total of 250 μL of pre-chilled lysis buffer was added to each pellet to generate light and heavy cell lysates. The lysis buffer in the ATP-binding experiment was 30 mM Tris-HCL buffer (pH 7.5) containing a cocktail of protease inhibitors (same composition as described in Chapter 2). The lysis buffer in the control experiment was 20 mM HEPES buffer (pH 7.5) containing 150 mM KCl, 10 mM MgCl₂, and a cocktail of protease inhibitors (same composition as described in Chapter 2). Cell lysis was accomplished through a total of 20 cycles consisting of mechanical disruption for 20 s and cooling on ice for 1 min using a Disruptor Genie (Scientific Industries) and glass beads (0.5 mm). Lysate was centrifuged at 14,000 g for 15 min at 4°C and the supernatant was saved for subsequent SILAC-LiP analyses. The total protein concentration in each supernatant was determined by a Bradford assay and normalized to 2.5 mg/mL (for the ATP-binding experiment) or 2 mg/mL (for the control experiment).
3.2.3 SILAC-LiP Analyses

In the ATP-binding experiment, the heavy lysate was mixed with 250 mM MgCl$_2$ and 206 mM adenylyl imidodiphosphate (AMP-PNP, a non-hydrolyzable ATP analog) ligand stock solution at a volume ratio of 8:1:1. The light lysate was mixed with 250 mM MgCl$_2$ and 30 mM Tris-HCl buffer at the same ratio. The resulting (+) and (-) ligand samples were equilibrated on ice for 2 hours. 50 µL aliquot of each sample was treated with proteinase K (pK) (Sigma) at an enzyme/substrate ratio of 1/100 (w/w) for 5 min at room temperature. This proteolysis conditions (i.e., enzyme concentrations, reaction times, and temperatures) were the same as those previously described (14). The reaction was quenched upon addition of guanidine hydrochloride (GdmCl) (EMD Millipore) crystals (final concentration 7.6 M) and heating at 99 °C for 3 min. The (+) and (-) ligand samples (i.e., heavy and light cell lysates, respectively) were then combined as a double digestion group. Similarly, another 50 µL aliquot of each sample was denatured with GdmCl without pK treatment. The samples were then combined as a single digestion group.

The combined light and heavy lysates were reacted with 5 mM TCEP (Thermo Fisher, Inc.) for 1 hour at 60 °C to reduce the disulfide bonds. The samples were treated with 10 mM MMTS (Sigma) for 10 min at room temperature. The samples were then diluted with 0.5 M TEAB buffer (pH 8.5) (Sigma) such that the final concentration of GdmCl was less than 2 M. Ultimately, the protein samples were digested with trypsin.
using an enzyme/substrate ratio of 1/50 (w/w) at 37 °C with overnight incubation. The proteolytic digestion reaction was quenched upon acidification (pH ~ 2–3) with TFA (Sigma).

The SiLAC-LiP protocol used in the control experiment was identical to that described above for the ATP-binding experiment, except that no ligand was used.

3.2.4 LC-MS/MS Analyses

The peptide mixtures were desalted using C18 MacroSpin columns (The Nest Group) according to the manufacturer’s protocol. LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive Plus high-resolution mass spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. The instrument parameters were the same as those described in Chapter 2. Each sample was analyzed in triplicate.

3.2.5 Proteomic Data Analysis

The raw LC-MS/MS data files were searched using MaxQuant 1.5.2.8(77) against the 6619 Saccharomyces cerevisiae proteins in the 2012-01 release of the UniProtKB/SwissProt database downloaded on 2/28/2012. Searches were performed with fixed MMTS modification on cysteine and SILAC labeling of lysine, variable oxidation of methionine and deamidation of asparagine and glutamine. Trypsin was set as the enzyme with semi-specificity and up to two missed cleavages allowed. The mass tolerance for precursor ions was set to 20 ppm for the first search, where initial mass
recalibration was performed, and a 10 ppm precursor mass tolerance was used for the main search. The mass tolerance for fragment ions was set to 0.02 Da. Also included were a match between runs and requantification of the searched peptides. The rest of the parameters were set at the default settings. In cases where a given peptide was matched to multiple protein isoforms or multiple members of a protein family, the peptide was assigned to the leading razor protein listed by the MaxQuant algorithm. Peptides and proteins identified with false discovery rates <1%, and positive heavy to light (H/L) ratios were used for subsequent data analysis.

The H/L ratios were transformed to log2 values, and the resulting values obtained for a given peptide at all charge states and from three technical replicates were used in the subsequent analyses to select peptide hits. Peptide hits were only selected from those that were identified at least twice in both the double and single digestion groups. For peptide hit selection, a Student’s two-tailed t test was used with R (https://www.r-project.org/) to identify hit peptides with significantly different H/L ratios in the doubly and singly digested samples. Hit peptides were selected as those with a p-value less than 0.05.

3.3 Results and Discussion
3.3.1 Experimental Design
Figure 7: Schematic representation of the SILAC-based limited proteolysis workflow used in this work.
The experimental workflow used in this study is shown in Figure 7. The non-hydrolyzable ATP mimic, AMP-PNP (Figure 8), was used as the test ligand to identify ATP-interacting proteins in yeast cell lysates. As part of the experimental workflow (Figure 7), the heavy- and light-labeled protein samples (e.g., cell lysate) were incubated with and without ligand, respectively. The resulting (+) and (-) ligand samples were each subjected to a double and a single digestion. For the double digestion, a limited proteolytic digestion with proteinase K was performed under native conditions on both the light- and heavy-labeled cell lysate samples. The experimental conditions of the limited proteolytic digestion with proteinase K, which were similar to those previously described(14), were controlled such that the extent of primary cleavages is largely determined by the protein conformation. After the proteinase K digestion was quenched, the light- and heavy-labeled samples were combined, and the resulting protein mixture was subjected to a trypsin digestion under denaturing conditions to generate peptides suitable for bottom-up proteomic analysis. For the single digestion, proteins were only subjected to the trypsin digestion, serving as a control for endogenous protease cleavages, incomplete labeling in heavy-labeled cells, and protein abundance differences across samples.
Proteins with ligand-induced conformational changes were identified based on the proteinase K-induced H/L ratio changes (p < 0.05) of the peptides identified and quantified in the LC-MS/MS readout. A semi-tryptic peptide with a decreased H/L ratio, or a fully-tryptic peptide with an increased H/L ratio, suggested protease protection in the heavy-labeled (+) ligand sample. This is expected as direct binding of the ligand to the native state of a protein typically increases the stability of the protein (or protein folding domain to which the ligand binds). Conversely, a semi-tryptic peptide with an increased H/L ratio, or a fully-tryptic peptide with a decreased H/L ratio, indicated greater protease susceptibility in the heavy-labeled (+) ligand sample. This can happen when ligand binds to one region of a protein inducing conformational changes that are destabilizing another region of the protein (allosteric effects)(96). Alternatively, a direct ligand binding interaction with one protein can cause another protein to be destabilized due to the disruption of a protein-protein interaction. Furthermore, cases in which ATP...
interacts with non-native states of the protein could also result in an apparent destabilization(12, 97).

### 3.3.2 Proteomic Coverage

The proteomic coverage obtained in this study is summarized in Table 7. In total, approximately 2000 peptides from ~300 proteins were assayed for their proteinase K sensitivity. These assayed peptides included those that were successfully detected and quantified at least twice in the bottom-up shotgun proteomics analyses of both the double and single digestion groups. A total of 207 peptides from 96 proteins were identified as hits with statistically significant (p-value < 0.05) H/L ratio changes in the presence of AMP-PNP (Figure 9). The peptide hits accounted for 10% of those assayed.

![Volcano plot](image)

**Figure 9**: Volcano plot of statistical significance against fold change between single and double digestion. The horizontal dashed line represents a p-value = 0.05.
In order to assess the false positive rate of peptide hit discovery using the SILAC-LiP protocol described here, a control experiment was performed using the light- and heavy-labeled yeast cell lysates in the absence of any ligand. Similar to the ATP-binding experiment, approximately 2000 peptides from ~300 proteins were assayed in the control experiment, and 183 peptides from 95 proteins were identified as hits in the control experiment (Table 7). This resulted in a false positive rate of 8.9%, which is close to the hit rate in the ATP-binding experiment. This control experiment only involved the analysis of one biological replicate, albeit with three technical replicates. Different LC-MS/MS analyses on the same sample (technical replicates) tend to yield data with less variance compared to different LC-MS/MS analyses on different samples (biological replicates). Since hit selection was based on a Student’s t test that took both sample mean and variance into consideration, using data solely from technical replicates can potentially result in more false positives. Therefore, multiple biological replicates should be performed for the SILAC-LiP protocol, which will be discussed in Chapter 4.

Nevertheless, a total of 69 peptide hits in the control experiment also displayed significantly altered (p-value < 0.05) H/L ratios in the ATP-binding experiment. The proteinase K-induced H/L ratio changes in these peptides are not likely to be ATP-related. Thus, they were excluded from the final hit list used in the following discussions. The final hit list contained 138 peptide hits from 79 unique proteins.
Table 7: Proteomic coverage obtained in the ATP-binding study performed with SILAC-LiP.

<table>
<thead>
<tr>
<th></th>
<th>ATP-binding</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Double digestion</td>
<td>Single digestion</td>
</tr>
<tr>
<td>Total peptides (proteins) identified</td>
<td>3043 (384)</td>
<td>3075 (412)</td>
</tr>
<tr>
<td>Assayed peptides (proteins)</td>
<td>2076 (307)</td>
<td>2059 (308)</td>
</tr>
<tr>
<td>Peptide (protein) hits</td>
<td>207 (96)</td>
<td>183 (95)</td>
</tr>
</tbody>
</table>

3.3.3 Discovery of ATP-Interacting Proteins

A total of 58 of the 307 unique proteins assayed in this study were previously annotated as ATP-binding proteins in the Saccharomyces Genome Database (SGD). Only 15 of these 58 proteins were identified as hits in the work described here. This result is similar to those reported in a number of recent studies of ATP-binding proteins using SPROX(20, 28) or other chemical proteomics approaches(98-99) where approximately 15-30% of the assayed annotated ATP-binding proteins were identified as hits. We hypothesize that the other 43 known ATP binding proteins assayed here did not undergo conformational changes that can induce a protease susceptibility shift measurable by the SILAC-LiP protocol.

The newly discovered ATP-interacting proteins account for 81% of the identified protein hits in this work. This is also similar to the above-mentioned studies in which a
large fraction (60-72%) of the identified protein hits were also not previously known to bind ATP. Interestingly, approximately 70% of the newly discovered ATP-interacting proteins in this work have known interactions with other nucleotides and co-factors (e.g., NAD), DNA and RNA based on GO-term analyses (Figure 10). The mode of protein-adenine recognition for ATP is similar to that of other adenine containing molecules such as NAD. Thus, it is not surprising to find that some of the protein hits in our ATP-binding study were known to bind co-factors containing the adenosine moiety. It is also not surprising that many RNA-binding proteins shown as hits because there is increasing evidence that RNA-binding proteins participate in many other cellular processes and require ATP for function(100-101).

Figure 10: Pie chart showing the distribution of known ligands for the hit proteins identified in the ATP-binding study.
3.3.4 Overlap with SPROX

The newly discovered ATP-interacting proteins in this work can be cross-validated by other energetics-based approaches. Summarized in Figure 11A is the overlap between the proteomic coverage obtained here and those previously reported in the SILAC-SPROX(20) and iTRAQ-SPROX(28) experiments. A total of 135 proteins were assayed in all three studies. In these 135 proteins is a subset of 90 proteins that were identified as hit proteins in at least one of the three studies (Figure 11B). These 90 protein hits included 36 proteins that appeared as hits in at least two of the three studies. A total of 27 of these 36 protein hits were not previously known to bind ATP. The SILAC-LiP, SILAC-SPROX and iTRAQ-SPROX techniques employ different probes and mass spectrometry-based readouts. Thus, it is relatively unlikely that these 27 newly discovered ATP-interacting proteins are false positives. Table 8 lists these 27 proteins with their ATP binding properties corroborated by the indicated experiments.
Figure 11: Comparison of SILAC-LiP, SILAC-SPROX and iTraq-SPROX ATP-binding results. (A) Venn diagram showing the overlap among the assayed proteins in the three studies. A total of 135 proteins were assayed in all three studies. (B) Venn diagram showing the overlap among the 90 protein hits coming from the 135 commonly assayed proteins.
Table 8: Summary of the 27 newly discovered ATP-interacting proteins that were identified in at least two of the SIALC-LiP, SILAC-SPROX and iTRAQ-SPROX experiments.

<table>
<thead>
<tr>
<th>Protein Accession</th>
<th>Protein Name</th>
<th>SILAC-LiP</th>
<th>SILAC-SPROX</th>
<th>iTRAQ-SPROX</th>
</tr>
</thead>
<tbody>
<tr>
<td>P23248</td>
<td>40S ribosomal protein S1-B</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Q01855</td>
<td>40S ribosomal protein S15</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P07281</td>
<td>40S ribosomal protein S19-A</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P05750</td>
<td>40S ribosomal protein S3</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P05694</td>
<td>5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P41805</td>
<td>60S ribosomal protein L10</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P14126</td>
<td>60S ribosomal protein L3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P05738</td>
<td>60S ribosomal protein L9-A</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P00924</td>
<td>Alcohol dehydrogenase 1</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P10081</td>
<td>ATP-dependent RNA helicase eIF4A</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P39522</td>
<td>Dihydroxy-acid dehydratase, mitochondrial</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P02994</td>
<td>Elongation factor 1-alpha</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P32324</td>
<td>Elongation factor 2</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P00924</td>
<td>Enolase 1</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P00925</td>
<td>Enolase 2</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P19097</td>
<td>Fatty acid synthase subunit alpha</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P00359</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase 2</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P10591</td>
<td>Heat shock protein SSA1</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P10592</td>
<td>Heat shock protein SSA2</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P04807</td>
<td>Hexokinase-2</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P15180</td>
<td>Lysine--tRNA ligase, cytoplasmic</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P14832</td>
<td>Peptidyl-prolyl cis-trans isomerase</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P38013</td>
<td>Peroxiredoxin type-2</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P00950</td>
<td>Phosphoglycerate mutase 1</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P06169</td>
<td>Pyruvate decarboxylase isozyme 1</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>P41920</td>
<td>Ran-specific GTPase-activating protein 1</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q12118</td>
<td>Small glutamine-rich tetratricopeptide repeat-containing protein 2</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.4 Conclusions

The recently established LiP protocol was used in combination with the SILAC-based quantitation strategy to identify ATP-interacting proteins in the *Saccharomyces cerevisiae* proteome. The study enabled 307 proteins to be assayed for their interactions (both direct and indirect) with ATP. A total of 79 proteins were found to have ATP-induced stability changes. These 79 protein hits included 15 previously annotated ATP-binding proteins and 20 newly discovered ATP-interacting proteins that also showed ATP binding properties in the earlier SPROX experiments. The results of this study not only added additional proteomic coverage into the *Saccharomyces cerevisiae* ATP-interactome, but they also helped demonstrate the capability of the SILAC-LiP protocol for protein target discovery.
4. Large-Scale Analysis of Protein Folding and Stability Changes Associated with Breast Cancer Using SILAC-LiP

The work described in this chapter comes largely from the research paper titled “Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using SILAC-LiP” that was published in 2016 on the Journal of Proteome Research (vol. 15(12), p. 4666-4674(1).

4.1 Introduction

The work in this chapter describes an application of the SILAC-LiP protocol developed in Chapter 3 to study the conformational changes of proteins involved in breast cancer pathogenesis. The cell culture models of the disease used in the current work, including the non-tumorigenic MCF-10A breast cell line and the non-invasive, estrogen receptor positive MCF-7 breast cancer cell line, were the same as those characterized in an earlier SILAC-SPROX study that profiled the chemical denaturant-induced equilibrium unfolding properties of the proteins in these cell lines (35). One goal of the work in this chapter was to compare the protein hits with altered thermodynamic stabilities identified in the SILAC-SPROX study to those protein hits identified in the SILAC-LiP experiment described here. The types of protein conformational changes probed by the SPROX and LiP techniques are different. Thus, the identified protein hits in the LiP experiment are expected to create novel molecular signatures of breast cancer.
In total, the SILAC-LiP approach described here identified ~200 proteins with cell-line-dependent conformational changes, as determined by their differential susceptibility to proteolytic digestion using the nonspecific protease, proteinase K. The protease susceptibility profiles of the proteins in these cell lines were compared to thermodynamic stability and expression level profiles previously generated for proteins in these same breast cancer cell lines (35, 49). The comparisons revealed that there was little overlap between the proteins with protease susceptibility changes and the proteins with thermodynamic stability and/or expression level changes. Thus, the large-scale conformational analysis described here provides unique insight into the molecular basis of the breast cancer phenotypes in this study.

4.2 Experimental

4.2.1 Cell Culture and Cell Lysate Preparation

Heavy-SILAC-labeled breast cancer cell line MCF-7 was cultured using heavy-labeled lysine and arginine according to the established protocols (76) as described in Chapter 2. The breast epithelial cell line MCF-10A was cultured following the American Type Culture Collection (ATCC) guidelines. Starting cells were acquired from ATCC (provided by the Cell Culture Facility at Duke University). The MCF-10A cells were maintained in MEGM medium containing MEBM (CC-3151) and all the components of the kit except gentamycin (CC-4136) from Lonza, along with 100 ng/mL of cholera toxin (Sigma). All cells were maintained in a humidified 37 °C incubator with 5% CO₂. All
cells were washed twice with PBS before they were harvested with HyQtase solution (Hyclone). The harvested cells were pelleted at 1000 rpm for 5 min. The resulting cell pellets were washed with PBS and stored at -80 °C.

Cell lysis was performed in the same manner as that described in Chapter 2. The lysis buffer was 20 mM HEPES buffer (pH 7.5) containing 150 mM KCl, 10 mM MgCl₂, and a cocktail of protease inhibitors (same composition as described in Chapter 2). Lysed cells were centrifuged at 14,000 g for 15 min at 4 °C, and the supernatant was used for subsequent analyses. The total protein concentration in each supernatant was determined using a Bradford assay and normalized to 2 mg/mL.

4.2.2 SILAC-LiP Analyses

The SILAC-LiP technique was applied in a comparative study involving heavy-labeled MCF-7 and light-labeled MCF-10A cell lines. The study was performed with five biological replicates. The heavy-labeled MCF-7 and light-labeled MCF-10A cell lysates generated for each biological replicate were each divided into two 50 µL aliquots. One aliquot of each cell lysate was subjected to limited proteolysis under native conditions followed by tryptic digestion under denaturing conditions (double digestion), while the other aliquot of each cell lysate was only subjected to tryptic digestion (single digestion). The proteolysis conditions (i.e., enzyme concentrations, reaction times, and temperatures) were the same as those described in Chapter 3.
Specifically, in the double digestion group, each of the two cell lysates was treated with proteinase K (pK) (Sigma) at an enzyme/substrate ratio of 1/100 (w/w) for 5 min at room temperature. The reaction was quenched upon addition of guanidine hydrochloride (GdmCl) (EMD Millipore) crystals (final concentration 7.6 M) and heating at 99 °C for 3 min. In the single digestion group, both the light- and heavy-labeled cell lysates were denatured by GdmCl without pK treatment. The light- and heavy-labeled lysates that were singly digested were combined, as were the light- and heavy-labeled lysates that were doubly digested. The combined light and heavy lysates were subjected to a standard bottom-up proteomics sample preparation procedure as described in Chapter 3, which involved reduction of disulfide bonds by TCEP (Thermo Fisher, Inc.), modification of free cysteine residues by MMTS (Sigma), addition of TEAB buffer (pH 8.5) (Sigma) to dilute GdmCl, and protein digestion by trypsin (Sigma Aldrich).

The SILAC-LiP technique was also applied in a control study involving heavy- and light-labeled MCF-7 cell lines. The study was performed in the same manner as that described above for the MCF-7 vs. MCF-10A cell line comparison. Three biological replicates of the control experiment were performed.

4.2.3 LC-MS/MS Analyses

The peptide mixtures were desalted using C18 MacroSpin columns (The Nest Group) according to the manufacturer’s protocol. LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive Plus high-resolution mass
spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. The instrument parameters were the same as those described in Chapter 2. Each sample was analyzed in triplicate.

4.2.4 Proteomic Data Analysis

The resulting raw data files included 30 raw LC-MS/MS data files for the MCF-7 vs. MCF-10A cell line comparison (double and single digestion, five biological replicates, three technical replicates for each sample) and 18 raw LC-MS/MS data files for the control experiment (double and single digestion, three biological replicates, three technical replicates for each sample). The data in each experiment were searched using MaxQuant 1.5.2.8(77) against the 20265 human proteins in the 2014-04 release of the UniProt Knowledgebase downloaded on 5/16/2014. Searches were performed with fixed MMTS modification on cysteine and SILAC labeling of lysine and arginine, variable oxidation of methionine, deamidation of asparagine and glutamine, and acetylation of the protein N-terminus. Trypsin was set as the enzyme with semi-specificity and up to two missed cleavages allowed. The mass tolerance for precursor ions was set to 20 ppm for the first search, where initial mass recalibration was performed, and a 10 ppm precursor mass tolerance was used for the main search. The mass tolerance for fragment ions was set to 0.02 Da. Also included were a match between runs and requantification of the searched peptides. The rest of the parameters were set at the default settings. In cases where a given peptide was matched to multiple protein isoforms or multiple
members of a protein family, the peptide was assigned to the leading razor protein listed by the MaxQuant algorithm. Peptides and proteins identified with false discovery rates <1%, and positive heavy to light (H/L) ratios were used for subsequent data analysis.

For each biological replicate, the data in the technical replicates were averaged to generate a single H/L ratio for each charge state of each identified peptide. A median H/L ratio was determined for each protein based on all the identified peptides from a given protein in the single digestion group. The H/L ratios of the peptides from a given protein were subsequently divided by the median H/L ratio for that protein, which effectively normalized for protein abundance differences in the two different cell lines (i.e., all the H/L ratios were normalized to 1). These normalized H/L ratios were transformed to log₂ values, and the resulting values obtained for a given peptide at all charge states and from five biological replicates were used in subsequent analyses to select peptide hits. Peptide hits were only selected from those that were identified at least twice in both the double and single digestion groups. For peptide hit selection, a Student’s two-tailed t test was used with R (https://www.r-project.org/) to identify hit peptides with significantly different H/L ratios in the doubly and singly digested samples. Hit peptides were selected as those with a p-value less than 0.05.

4.3 Results and Discussion

4.3.1 Experimental Design
Figure 12: Schematic representation of the SILAC-based limited proteolysis workflow used in this work.
The experimental workflow used in this study is shown in Figure 12. The cell culture models of breast cancer analyzed here included a model of normal mammary epithelial cells (MCF-10A) and a breast cancer cell line (MCF-7). The MCF-7 cell line is estrogen receptor positive and reflective of the luminal subtype of breast cancer. The conformational properties of proteins in the MCF-10A cell lysate were compared to those in the MCF-7 cell lysate where the MCF-7 cell line was heavy-labeled.

As part of the experimental workflow (Figure 12), each cell line in the comparison was subjected to a double and a single digestion. Proteinase K and trypsin were used in the double digestion, and only trypsin was used in the single digestion. The initial step in the workflow involved performing a limited proteolytic digestion with proteinase K under native conditions on both the light- and heavy-labeled cell lines. The experimental conditions of the limited proteolytic digestion with proteinase K, which were the same as those described in Chapter 3, were controlled such that the primary cleavages are dictated by the structural features of proteins. After the proteinase K digestion was quenched, the light- and heavy-labeled samples were combined, and the resulting protein mixture was subjected to a trypsin digestion under denaturing conditions to generate peptides suitable for bottom-up proteomic analysis. The single digestion group, which included proteins subjected to the trypsin digestion, served as a control for endogenous protease cleavages, incomplete labeling in heavy-labeled cells, and protein abundance differences across samples.
Proteins with different conformations in the two cell lines were identified based on the proteinase K-induced H/L ratio changes (p < 0.05) of the peptides identified and quantified in the LC-MS/MS readout. A semi-tryptic peptide with an increased H/L ratio, or a fully tryptic peptide with a decreased H/L ratio, indicated greater protease susceptibility in the heavy-labeled MCF-7 cell line. Conversely, a semi-tryptic peptide with a decreased H/L ratio, or a fully tryptic peptide with an increased H/L ratio, suggested protease protection in the heavy-labeled MCF-7 cell line.

4.3.2 Proteomic Coverage

Summarized in Table 9 is the proteomic coverage obtained in the comparative analysis performed in this work. In total, approximately 4300 peptides from ∼550 proteins were assayed for their proteinase K sensitivity. These assayed peptides included those that were successfully detected and quantified at least twice in the bottom-up shotgun proteomics analyses of both the double and single digestion groups. A total of 489 peptides from 203 proteins were identified as hits with statistically significant (p-value < 0.05) proteinase K-induced H/L ratio changes in the MCF-7 vs. MCF-10A comparison (Figure 13). The peptide hits accounted for ∼11% of those assayed.
Table 9: Proteomic coverage obtained in the cell line comparisons performed with SILAC-LiP.

<table>
<thead>
<tr>
<th></th>
<th>MCF-7 vs. MCF-10A</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Double digestion</td>
<td>Single digestion</td>
</tr>
<tr>
<td>Total peptides (proteins) identified</td>
<td>7408 (812)</td>
<td>10994 (979)</td>
</tr>
<tr>
<td>Assayed peptides (proteins)</td>
<td>4342 (556)</td>
<td>3296 (478)</td>
</tr>
<tr>
<td>Peptide (protein) hits</td>
<td>489 (203)</td>
<td>201 (113)</td>
</tr>
</tbody>
</table>

Figure 13: Global analysis of the H/L ratios generated for the assayed peptides in the MCF-7 vs. MCF-10A comparison. Volcano plot of statistical significance against fold change between single and double digestion. The horizontal dashed line represents a p-value = 0.05. The gray points (n = 3853) represent non-hit peptides (p-value > 0.05). The green points (n = 304) represent peptide hits (p-value < 0.5) observed
in four or five of the biological replicates. The blue points (n = 114) represent peptide hits (p-value < 0.05) observed in two or three of the biological replicates. The red points (n = 39) represent peptide hits (p-value < 0.05) observed in one biological replicate.

In order to assess the false positive rate of peptide hit discovery using the SILAC-LiP protocol described here, a control experiment was performed using light- and heavy-labeled MCF-7 cells. Approximately 3300 peptides from ~480 proteins were assayed in the control experiment, and 201 peptides from 113 proteins were identified as hits in the control experiment (Table 9 and Figure 14). This resulted in a false positive rate for peptide hit discovery of ~6%. These peptide hits likely arose from the random errors associated with SILAC quantitation. A total of 32 peptide hits in the control experiment also displayed significantly altered (p-value < 0.05) H/L ratios in the MCF-7 vs. MCF-10A comparison. The proteinase K-induced H/L ratio changes in these peptides are not likely to be disease-related. Thus, they were excluded from the final hit list used in the bioinformatics analyses described below.
Figure 14: Global analysis of the H/L ratios generated for the assayed peptides in the control experiment. Volcano plot of statistical significance against fold change between single and double digestion in the control experiment using heavy- and light-labeled MCF-7 cell lines. The horizontal dashed line represents a p-value of 0.05. The grey points (n = 3095) represent non-hit peptides (p-value > 0.05). The green points (n = 141) represent peptides hits (p-value < 0.5) observed in three biological replicates. The blue points (n = 46) represent peptide hits (p-value < 0.05) observed in two biological replicates. The red points (n = 14) represent peptide hits (p-value < 0.05) observed in one biological replicate.

The fold-change values between the double and single digestions for the final hit peptides spanned a relatively wide range from 1.1- to 8.5-fold (Figure 13). However, the median fold-change value of the hit peptides identified in this work was relatively small, at 1.3-fold. This relatively small value suggests that the protease susceptibility differences of the proteins to which these hit peptides mapped were relatively small.
One limitation of the current control experiment is that it does not account for potential error introduced by protein abundance changes. The false discovery of peptide hits in the SILAC-LiP experiment could result from errors in the determination of protein median H/L ratios, which are ultimately used for the normalization of protein abundance differences. To assess the impact of this potential error on our hit discovery, the fraction of hits with no significant (i.e., less than 2-fold) protein expression level changes (69%) was compared to the fraction of assayed proteins with no significant protein expression level changes (67%). This analysis indicated the protein hits were not enriched in proteins that had significant expression level changes (i.e., those would be more prone to errors in the determination of a median H/L ratio). Also, the relative standard deviations (RSD) associated with the protein median H/L ratios were similar in the MCF-7 vs. MCF-10A comparison (median RSD = 0.18) and the control experiment (median RSD = 0.15).

Hit selection at the peptide level is prone to error. While all the peptide hits identified in this work had p-values < 0.05, not all were identified in all five biological replicates. Approximately 9% of the peptide hits were identified based on data from only one biological replicate, and another 25% of the peptide hits were identified based on data from two or three biological replicates. The strongest peptide hits were a subset of approximately 66% of the peptide hits that were determined to be hits based on data from four or five independent biological replicates (see Figure 13).
Also strong were those protein hits that result from multiple peptide hits. Since the protease-induced cleavage in LiP only occurs at sites in a protein where conformational changes occur, not all peptides assayed from a hit protein are expected to be a hit. However, in cases where different peptides from the same protein show altered protease susceptibility, the confidence in the given protein hit increases. In this study, approximately half of the protein hits were determined based on two or more peptide hits (Figure 15). For example, the ribosomal protein RPS12 was a protein hit as determined from three different peptide hits. Among these three peptides, a semi-tryptic peptide SNCDEPMYVK (residues 54–63) showed an increased H/L ratio, while the fully tryptic peptide QAHLCVLASNCDEPMYVK (residues 46–63) encompassing the semi-tryptic peptide showed a decreased H/L ratio. Together, these peptides suggest a conformational change at this site of RPS12 between the MCF-10A and MCF-7 cell lines.
Figure 15: Bar graph showing the number of protein hits that were identified with multiple peptides in the MCF-7 vs. MCF-10A comparison.

The observation of a semi-tryptic peptide and its corresponding fully tryptic peptide showing opposite changes in H/L ratios is especially useful, because such peptide pairs can act as “internal controls” for the quantitative changes in their proteolytic susceptibility. Unfortunately, the detection and quantitation of such peptide pairs is a challenge using the SILAC-LiP approach described here. This is largely because semi-tryptic peptides tend to be singly charged during electrospray ionization, which makes their sequencing difficult. In fact, such singly charged ions are not generally selected for sequencing in shotgun proteomics experiments. Moreover, many semi-tryptic peptides do not contain lysine or arginine, which not only further reduces their ionization efficiency but also makes them difficult to quantify using SILAC.
Of the 4342 peptides assayed in our experiments, only 13% of them were semi-tryptic peptides. The percentage of hit peptides identified here that were semi-tryptic was also similar (i.e., 12%). Not surprisingly, only a small number of the tryptic and semi-tryptic peptide hits identified in this work (39 and 38, respectively) had corresponding semi-tryptic and tryptic peptides that were successfully assayed. In total, only seven pairs were identified as hits based on both tryptic and semi-tryptic peptides (i.e., the tryptic and semi-tryptic peptides showed opposite and significant (p-value < 0.05) changes in their H/L ratios).

It is noteworthy that the rest of the tryptic/semi-tryptic peptide pairs are not necessarily false positives. Due to the non-specificity of proteinase K, a given tryptic peptide can be cleaved into several different semi-tryptic peptides, whose relative proportions may be dictated by the protein conformational differences. Thus, it is possible that semi-tryptic peptides that map to tryptic peptide hits did not contain the proteinase K site that was differentially cleaved. Similarly, it is also possible for semi-tryptic peptide hits to map to tryptic peptide non-hits. For this reason, tryptic and semi-tryptic peptide hits mapping to semi-tryptic and tryptic peptide non-hits (respectively) were not removed from the hit list.

### 4.3.3 Expression Level Analysis

The H/L ratios obtained in the single digestion can be used to assess the relative abundance of proteins in each cell line. The expression level data generated in this work
was in general agreement with that previously reported in the literature for the cell lines in this study(35, 49) (Figure 16). For example, the $\sigma$ isoform of the 14-3-3 proteins has been directly implicated in breast cancer as a negative cell cycle regulator(102). Protein expression level studies have found that 14-3-3$\sigma$ is down-regulated in transformed mammary epithelial cells(103). The expression level information obtained in this work indicates a similar down-regulation of 14-3-3$\sigma$ in the MCF-7 cells compared with MCF-10A cells. Interestingly, $\sim69\%$ of the protein hits identified with conformational differences between the two cell lines did not have significant (i.e., less than 2-fold) changes in their expression levels (Figure 17). This is similar to that observed in the SILAC-SPROX analyses of the cell lines used in these studies (i.e., $\sim50\%$ of the differentially stabilized proteins in the SILAC-SPROX experiments did not have significant changes in their expression levels)(35). These results suggest that conformational changes can provide information about disease states that is orthogonal to that obtained in protein expression level analyses.
Figure 16: Expression profile generated for a subset of protein hits identified in the MCF-10A vs. MCF-7 comparison. The protein hits highlighted here are those with protein expression level data (i.e., Log2(H/L) values) available both in this work and in the literature. The numbers in the boxes represent ratios of protein expression levels in the MCF-7 cell line versus in the MCF-10A cell line (log2 transformed).

<table>
<thead>
<tr>
<th>Gene Names</th>
<th>SILAC-Lip</th>
<th>Expression Level Study</th>
<th>Gene Names</th>
<th>SILAC-Lip</th>
<th>Expression Level Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPS16</td>
<td>-1.44</td>
<td>-0.26</td>
<td>HSPA4</td>
<td>-0.96</td>
<td>-0.34</td>
</tr>
<tr>
<td>RPS18</td>
<td>-1.21</td>
<td>-0.19</td>
<td>HSPB1</td>
<td>2.08</td>
<td>2.54</td>
</tr>
<tr>
<td>RPS2</td>
<td>-1.52</td>
<td>-0.28</td>
<td>MIF</td>
<td>-0.11</td>
<td>0.51</td>
</tr>
<tr>
<td>RPS21</td>
<td>-0.37</td>
<td>-0.31</td>
<td>MDH2</td>
<td>0.62</td>
<td>1.22</td>
</tr>
<tr>
<td>RPS28</td>
<td>-0.87</td>
<td>0.15</td>
<td>KRT18</td>
<td>3.35</td>
<td>3.22</td>
</tr>
<tr>
<td>RPS6</td>
<td>-2.45</td>
<td>-0.53</td>
<td>NCL</td>
<td>-1.25</td>
<td>0.10</td>
</tr>
<tr>
<td>RPS7</td>
<td>-1.64</td>
<td>0.37</td>
<td>HSPD1</td>
<td>0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>RPL3</td>
<td>2.17</td>
<td>-0.65</td>
<td>PPLA</td>
<td>-0.85</td>
<td>0.21</td>
</tr>
<tr>
<td>RPL4</td>
<td>1.17</td>
<td>-0.64</td>
<td>FKBP4</td>
<td>1.17</td>
<td>1.22</td>
</tr>
<tr>
<td>RPL5</td>
<td>-1.00</td>
<td>-0.64</td>
<td>PRDX1</td>
<td>0.22</td>
<td>-0.13</td>
</tr>
<tr>
<td>HSPA5</td>
<td>-0.25</td>
<td>0.41</td>
<td>PGK1</td>
<td>0.09</td>
<td>0.24</td>
</tr>
<tr>
<td>CTSD</td>
<td>1.35</td>
<td>2.26</td>
<td>PGAM1</td>
<td>0.05</td>
<td>-0.28</td>
</tr>
<tr>
<td>EEF1A1</td>
<td>-0.70</td>
<td>0.14</td>
<td>PFN1</td>
<td>-0.09</td>
<td>-0.54</td>
</tr>
<tr>
<td>EEF1A2</td>
<td>2.78</td>
<td>4.08</td>
<td>PSMA2</td>
<td>0.21</td>
<td>0.33</td>
</tr>
<tr>
<td>EEF1B2</td>
<td>-0.37</td>
<td>0.55</td>
<td>PDIA3</td>
<td>0.40</td>
<td>-0.48</td>
</tr>
<tr>
<td>EEF2</td>
<td>-0.58</td>
<td>0.23</td>
<td>PDIA4</td>
<td>1.08</td>
<td>0.78</td>
</tr>
<tr>
<td>HSP90B1</td>
<td>-0.10</td>
<td>-0.40</td>
<td>S100A11</td>
<td>-0.12</td>
<td>0.34</td>
</tr>
<tr>
<td>ECHS1</td>
<td>0.97</td>
<td>1.04</td>
<td>MAT2A</td>
<td>1.17</td>
<td>-1.05</td>
</tr>
<tr>
<td>EIF4A1</td>
<td>-0.96</td>
<td>-0.67</td>
<td>HSPA9</td>
<td>-0.33</td>
<td>-0.17</td>
</tr>
<tr>
<td>FASN</td>
<td>0.02</td>
<td>1.44</td>
<td>STIP1</td>
<td>0.68</td>
<td>0.89</td>
</tr>
<tr>
<td>GAPDH</td>
<td>0.14</td>
<td>0.90</td>
<td>CCT2</td>
<td>-0.85</td>
<td>0.00</td>
</tr>
<tr>
<td>RTCB</td>
<td>-0.24</td>
<td>0.16</td>
<td>TALDO1</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>0.40</td>
<td>1.31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 17: Bar graph showing the distribution of the relative expression levels of the protein hits identified in the MCF-7 vs. MCF-10A comparison.

4.3.4 Hit Characterization

Approximately 58% of the hit peptides identified here were more protected from proteolytic cleavage, with the rest becoming less protected from proteolytic cleavage in the MCF-7 cells as compared to the MCF-10A cells. The positions of 190 hit peptides in 80 proteins with available structures in the Protein Data Bank were examined. Most of the hit peptides (i.e., 149 out of 190) mapped to loops, turns, or regions with no secondary structure (Figure 18). This is consistent with these regions of protein structure being more susceptible to proteolytic cleavages(104).
Figure 18: Pie chart showing the distribution of different protein structural elements to which the peptide hits identified in the MCF-7 vs. MCF-10A comparison mapped.

For example, the loop that connects the third and fourth helixes in the 14-3-3σ (i.e., SNEEGSEEKGPVEYR, residues 69–85) was found to be a site where an altered proteolysis pattern was observed in this work (i.e., decreased protease susceptibility in the MCF-7 cells) (Figure 19). This loop region appears as an unstructured region in the X-ray crystallographic structure of the dimeric 14-3-3σ(105-107). Yang and co-workers suggest that the conformational flexibility at the dimerization interface enhanced by this loop region could facilitate binding of 14-3-3σ to proteins of various sizes and sequences(105). Our finding that this loop region is more protected from proteolytic cleavage in the MCF-7 cell line than in the MCF-10A cell line suggests that these
previously reported protein–protein interactions involving this loop region in 14-3-3σ(105) may be disease-related.

Figure 19: Crystal structure showing the nine antiparallel helices that comprise the 14-3-3σ homodimer (PDB 1YZ5). The regions to which the peptide hits mapped are colored in red.

The limited proteolysis-based approaches generally assume that the differential protease susceptibility is due to alterations in the protein conformation. However, there is the possibility that intrinsic modifications in proteins such as post-translational modifications can prevent the access of proteinase K to its protein substrate by stearic hindrance. In such cases, the different cleavage patterns may result solely from different post-translational modifications on the protein in the different disease states rather than differences in protein conformation. In order to assess this possibility, we matched the peptide hits identified in this work to the known post-translational modification sites in
proteins in the MCF-10A and MCF-7 cell lines using the PhosphoSitePlus knowledgebase (108) (http://www.phosphosite.org/). A total of 26 peptide hits from 19 proteins have known phosphorylation sites in the MCF-10A and/or MCF-7 cell lines. These peptides, which account for ~6% of the peptide hits, may be a result of differential post-translational modifications in the two cell lines inducing different proteinase K cleavages. Additional experiments are needed to evaluate the effect of phosphorylation on the proteolysis.

An analysis of the final hit proteins using PANTHER (79) revealed that the protein complex biogenesis process was significantly overrepresented by the protein hits (p-value = 0.01). Molecular chaperones are one of the major classes of proteins related to this process. This finding is consistent with the results of previous work that found increased chaperone activity to be characteristic of breast tumorigenesis (85, 109-110).

Many heat shock proteins, including HSPA8, HSPA9, HSPD1, HSPB1, HSP90B1, HSP90AB1, HSP90AA1, HSPH1, and HSPA1A, displayed different protease susceptibility in the two different cell lines. In agreement with this observation, it has been suggested that the HSP family members are crucial for tumor growth by promoting cell proliferation as well as by inhibiting death pathways (86, 111-112). Targeting HSPs has thus been considered for therapeutic intervention in cancers (113).

Gene ontology (GO) classification of the protein hits also revealed that a large fraction of the hits are nucleic acid binding proteins (~30%). Given their pivotal role in
the translation of genetic expression into a functional phenotype, it is not surprising that many ribosomal proteins identified as hits in our study have been previously associated with tumorigenesis. It has been suggested that one of the effects of oncogenic signaling appears to be altered recruitment of mRNAs to ribosomes, which will in turn affect proteins involved in growth regulation and cell-cell interaction(114). The MDM2/MDMX-p53 negative feedback loop can be regulated by a number of ribosomal proteins, such as RPS3, RPS7, and RPS20(115-117), which were identified as hits in our study. Another hit protein, RPS6, has been found to attenuate KRAS-induced DNA damage and p53-mediated tumor suppression during development of pancreatic cancer(118). The 60S RPL32 and 40S RPS16, which were also hits in our study, were reported to have altered expression in a model that recapitulates the progression to an androgen-independent state of prostate tumor cells(119). Studies have also shown that knockdown of RPL13, another hit in our study, resulted in drastic attenuation of human gastrointestinal cancer cell growth with significant G1 and G2/M arrest of the cell cycle(120). The differential proteolysis patterns of the ribosomal proteins observed here are consistent with the current understanding of ribosomal proteins in the development of cancer.

4.3.4 Correlating Conformational Changes with Protein Function

Some of the hit proteins were enzymes with known catalytic activities. For example, lactate dehydrogenase A (LDHA) catalyzes the transformation of pyruvate
into L-lactate with concomitant inter-conversion of NADH and NAD⁺. One of the principal biochemical characteristics of cancer cells compared to normal cells is a metabolic switch from oxidative phosphorylation to increased glycolysis (i.e., the Warburg effect). LDHA is the final enzyme in the glycolysis pathway and has been shown to play an important role in the development, invasion, and metastasis of malignancies(121).

In our study, two peptides, NRVICGCNLDSAR (residues 156−169) and RVHPVSTMIK (residues 269−278), from the substrate binding domain of LDHA were identified with decreased protease susceptibility in the MCF-7 cell line compared to the MCF-10A cell line (Figure 20A). Structural studies have revealed that His 193, Asp 166, Arg 169, and Thr 244 all make major contributions to the catalytic geometry in the active site of LDHA(122). The active site loop residues 99−110 enclose the active site and are responsible for the NADH-cofactor binding. A peptide adjacent to this active site loop (IVSGKDYNVTANSKLVIITAGAR, residues 77−99) also displayed decreased sensitivity to protease attack in the MCF-7 cells compared with MCF-10A cells. Considering the close link between protein folding and function, we speculate that this greater protection from protease cleavage in the substrate-binding domain of LDHA may reflect an increased stability that results in increased enzymatic activity in the MCF-7 cells. In agreement with this hypothesis, it has been shown that LDH activity (in pyruvate
reducing direction) is higher in breast tumors than normal tissues (123), which is consistent with cancer cells utilizing LDH for anaerobic glycolysis.

Figure 20: Schematic representations of the folded three-dimensional structures of selected protein hits identified in the MCF-7 vs. MCF-10A comparison. The structure of LDHA in complex with NADH and the inhibitor oxamate (PDB 1I10) is shown in part A. The structure of GAPDH in complex with NAD (PDB 1ZNQ) is shown in part B. The structure of PKM2 in complex with Mg²⁺, K⁺, the inhibitor oxalate, and the allosteric activator fructose 1,6-bisphosphate (FBP) (PDB 1T5A) is shown part C. The structure of PGK1 in complex with 3-phosphoglycerate (3PG) and ADP (PDB 2XE7) is shown in part D. In each case, the regions to which the peptide
hits mapped are colored in red. Images were generated in KiNG (Kinemage, Next Generation)(124).

Several glycolytic enzymes were also identified as hits based on peptides derived from their substrate and/or cofactor binding sites. For example, two GAPDH peptide hits, RVIISAPSADAPMFVM and LISWYDNEFGYSNRVVDLMHMA, are located near the NAD binding site (Figure 20B). Also, two peptide hits in pyruvate kinase, VNFAMNGK and GDLGIEIPA, are located near binding sites of the substrate and an allosteric activator, respectively (Figure 20C). Similarly, three phosphoglycerate kinase 1 peptides (VVMRVDFNVPVNQITQNQR, ACANPAAGSVILLENLR, and AHSSMVGNLQKAGGFLMKK) are located close to the substrate binding site (Figure 20D). The altered proteolytic patterns observed in the above enzymes may suggest that the protein–ligand interactions close to these sites are disease-related.

4.3.5 Comparison with SILAC-SPROX Results

One goal of this work was to compare the protein hits identified using the SILAC-LiP approach described here to those previously identified using the SILAC-SPROX approach(35). A total of 243 proteins were assayed in both the SILAC-LiP and SILAC-SPROX studies. In these 243 proteins is a subset of 122 proteins that were identified as hit proteins in either one or both of the studies. These 122 protein hits included 84 proteins that appeared as hits only in the SILAC-LiP experiment, 14 proteins that appeared as hits only in the SILAC-SPROX experiment, and 24 proteins that
appeared as hits in both studies. The overlap between the SILAC-SPROX and SILAC-LiP protein hits was relatively small (~20%). This is not surprising, as the two techniques probe different properties of proteins. SPROX is sensitive to changes in the more global unfolding/refolding properties of proteins. More local conformational changes, which do not have a strong denaturant dependence on their energetics(95), typically go undetected in SPROX. However, local conformational changes may be more or less sensitive to protease digestion. Changes in the global stability of a protein (or protein domain) may not always be differentially susceptible to protease. Thus, it is not surprising that the two techniques provide complementary biologically relevant information on the disease states.

The fundamental differences between SILAC-SPROX and SILAC-LiP also make it difficult to correlate the SILAC-SPROX behavior of proteins with their SILAC-LiP behavior. Peptides derived from different domains of proteins may display very different behavior (i.e., stabilization or destabilization in SPROX, less or more proteolytic susceptibility in LiP). Even when the same peptide probe is detected as a hit in the two experiments, there may not be a direct correlation between thermodynamic stability and protease susceptibility. Indeed, there are examples in the literature where proteins with increased global stability have been found to be less susceptible to proteolytic digestion under native conditions(125-126) and where proteins with increased global stability have been found to be more susceptible to proteolytic digestion(127-129).
Among the 24 overlapping hit proteins in the SILAC-LiP and SILAC-SPROX experiments, only 5 were identified as hits based on the same peptide probes. However, a clear trend was not observed between the SILAC-SPROX and SILAC-LiP behavior of these peptides. For example, the peptides SFYPEEVSSMLTK and YDDMATCMK from HSPA8 and YWHAQ (respectively) were stabilized in SILAC-SPROX and more protected from proteolytic digestion in SILAC-LiP in the MCF-7 cell line as compared to the MCF-10A cell line, and the peptide AAHSEGNTTAGLDLMR from CCT2 was destabilized in SILAC-SPROX and less protected from proteolytic digestion in SILAC-LiP in the MCF-7 cell line than in the MCF-10A cell line. However, the peptides NALESYAFNMK and VVDLMAMHASKE from HSPA1A and GAPDH (respectively) were destabilized in SILAC-SPROX and also more protected from proteolytic digestion in SILAC-LiP in the MCF-7 cell line as compared to the MCF-10A cell line.

Altered protease susceptibility in LiP can be detected only at sites in a protein where the conformational changes occur, while altered thermodynamic stability in SPROX can be probed by any global protected methionine-containing peptide from a given protein (or protein domain). On the one hand, this creates a challenge in SILAC-LiP experiments of having to detect a tryptic (or semi-tryptic) peptide from the specific site of the conformational change. On the other hand, if such peptides are successfully detected in the SILAC-LiP experiment, they can provide specific information about the site of the conformational change. Interestingly, a number of the SILAC-LiP peptide hits
identified here mapped to regions of protein structure that are at or near known ligand binding sites (see e.g., Figure 20). This suggests that the conformational changes detected here may result in or be the result of altered ligand binding interactions.

4.4 Conclusions

A comparative, global proteomic analysis of the protein conformational differences between the MCF-10A and MCF-7 breast cancer cells resulted in the identification of a subset of proteins with differential protease susceptibility. A bioinformatics analysis revealed that nucleic acid binding proteins represented the largest fraction of the hits showing altered conformations in the different disease states. Also of note is the overrepresentation of the protein complex biogenesis process.

Changes in protein conformation can result from a variety of disease-related phenomena. However, the biological phenomena responsible for the conformational changes identified in this work cannot be directly determined from the data in this study. Additional biophysical and biochemical studies are needed on the protein hits in this study to better understand the molecular basis of the increased/decreased proteolytic stabilities detected here. The utility of the detected changes in proteolytic susceptibility as biomarkers of disease also requires the analysis of clinical samples. One attractive feature of the global conformational analysis performed here is that a relatively small amount of material (less than 200 µg of total protein) is required for
analysis. This greatly facilitates the extension of this methodology to the analysis of clinical samples.

5.1 Introduction

The SILAC-SPROX technique has been utilized to identify breast cancer-related changes in protein folding and stability in the previous work by Adhikari and coworkers (35) as well as in the work described in Chapter 2 of this thesis (2). These studies used several well-established cell culture models of the disease including the MCF-10A, MCF-7, BT-474, MDA-MB-468 and MDA-MB-231 cell lines that represent different subtypes of the disease. Four comparative studies were performed each involving a heavy-labeled MCF-7 cell line and one of the other four cell lines that was light-labeled. In these comparative studies, it was possible to differentiate the different breast cancer cell lines based on their thermodynamic stabilities measured by SILAC-SPROX. A number of protein hits were identified with altered thermodynamic stabilities that were specific to a given breast cancer cell line, indicating their potential to serve as biomarkers for subtype specific disease-related changes.

However, real malignant tumors are more heterogeneous in their cellular composition and include various cell populations compared to cancer cell lines. Therefore, the identified protein hits need further validation in clinical samples before their translation into diagnostic or therapeutic practice. The limited availability of
clinical samples makes it desirable to use validation approaches that are capable of acquiring reproducible proteomic results with limited amount of materials.

The work described in this chapter utilizes a recently reported targeted mass spectrometry-based SPROX approach (PAB-SPROX)(33) to develop validation assays for a set of prioritized hit proteins identified in the previous SILAC-SPROX studies. The PAB-SPROX technique is fundamentally identical to SILAC-SPROX except that it uses a phenacyl bromide (PAB) labeling strategy to specifically label unoxidized methionine residues (Figure 21). The PAB-labeled methionine side chain undergoes gas-phase fragmentation through collision induced dissociation (CID), and produces a fixed mass neutral loss and a specific product ion (Figure 21). This makes it possible to detect the transition between the sulfonium precursor ion and the post-fragmentation product ion using targeted mass spectrometry (e.g., PRM, SRM).

Described here is an initial effort to select protein targets for PAB-SPROX validation and to develop PRM assays using a MCF-7 cell lysate. Once established the assays can be applied (1) to the five breast cancer cell lines previously used in the SILAC-SPROX studies for further hit validation and (2) to clinical samples derived from the corresponding breast cancer subtypes for additional validation. One attractive feature of the PAB-SPROX technique is that a relatively small amount of material (~10 µg in each denaturant-containing sample) is required for analysis(33). This greatly facilitates the extension of this methodology to the analysis of clinical samples.
Figure 21: Schematic representation of PAB labeling reaction and gas phase fragmentation behavior of resulting PAB-labeled methionine residues.
5.2 Experimental

5.2.1 Cell Culture and Cell Lysate Preparation

The MCF-7 breast cancer cell line was cultured in the same manner as that described in Chapter 2. Cell lysis was also performed in the same manner as that described in Chapter 2. The lysis buffer was 20 mM phosphate buffer (pH 7.4) containing a cocktail of protease inhibitors (same composition as described in Chapter 2). Lysed cells were centrifuged at 14,000 g for 15 min at 4 °C, and the supernatant was used for subsequent analyses. The total protein concentration in the supernatant was 7.3 mg/mL as determined by a Bradford assay.

5.2.2 PAB-SPROX Analyses

A mock experiment involving only MCF-7 cell lysate was performed using the PAB-SPROX protocol as previously described(33). Briefly, the MCF-7 cell lysate was divided into two portions. Aliquots (20 µL) of each portion were each mixed with 75 µL of two 20 mM phosphate buffers (pH 7.4) containing different concentrations of urea. The final urea concentrations in the two urea-containing buffers were 0 and 2.4 M. Each mixture was equilibrated for 1 h, and 5 µL of 30% (w/w) hydrogen peroxide (Sigma) was added to each of the protein samples in the denaturant-containing buffers. After 6 min, the reaction in each protein sample was quenched with 1 mL of 300 mM L-methionine (Sigma). Proteins in the denaturant-containing buffer were precipitated upon addition of 250 µL of TCA (1 g/mL) (Sigma) and overnight equilibration at 0 °C.
The resulting protein pellets were each dissolved in 30 µL of 0.5 M TEAB buffer (Sigma) containing 0.1% SDS. The disulfide bonds were reduced with 5 mM TCEP (Thermo Fisher) for 1 h at 60 °C. The samples were then treated with 10 mM MMTS (Sigma) for 15 min at room temperature. Ultimately, the protein samples were digested overnight with trypsin using an enzyme/substrate ratio of 1/50 (w/w) at 37 °C. After addition of 20 µL of acetic acid and 30 µL of acetonitrile, the tryptic peptides were derivatized with either light or heavy PAB by adding 10 µL of 1 M solutions of 12C6-PAB or 13C6-PAB. The reactions were allowed to proceed for 24 h. The light- and heavy-labeled samples at the same denaturant concentration were combined. The combined samples were desalted using C18 resin (The Nest Group) according to the manufacturer’s protocol and submitted to LC–MS/MS analysis.

5.2.3 LC-MS/MS Analyses

LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive HF high-resolution mass spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. The instrument was set using parallel reaction monitoring (PRM) method. The method was based on a 90 min 3–30% acetonitrile gradient. Targeted precursors were summarized in Table S10. MS2 resolution was 30,000 (at m/z 200) with an AGC target value of 5 × 10⁴ ions, a maximum ion trap time of 100 ms, an isolation window of 1.0 m/z and a stepped normalized collision energy (NCE) of 10 V and 30 V.
Table 10: Summary of the PRM precursors that were targeted in the PAB-SPROX experiments. Bolded and underlined methionine residues were modified with PAB.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Peptide sequence</th>
<th>Protein name</th>
<th>Gene name</th>
<th>z</th>
<th>Precursor m/z</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDA-MB-468 vs. MCF-7</td>
<td>NVLLDPQLVPGGASEM AVAHALTEK</td>
<td>T-complex protein 1 subunit gamma</td>
<td>CCT3</td>
<td>3</td>
<td>912.4739 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>914.4807 (heavy)</td>
</tr>
<tr>
<td></td>
<td>GVIVDKDFSHPQMPK</td>
<td>T-complex protein 1 subunit epsilon</td>
<td>CCT5</td>
<td>3</td>
<td>605.9748 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>607.9815 (heavy)</td>
</tr>
<tr>
<td></td>
<td>MEGPLSVFGDR</td>
<td>T-complex protein 1 subunit alpha</td>
<td>TCP1</td>
<td>2</td>
<td>663.3134 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>666.3234 (heavy)</td>
</tr>
<tr>
<td></td>
<td>MSVQPTVSLGGFEITPPV VLR</td>
<td>Nucleophosmin</td>
<td>NPM1</td>
<td>2</td>
<td>1173.1323 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1176.1424 (heavy)</td>
</tr>
<tr>
<td>MDA-MB-231 vs. MCF-7</td>
<td>QSFTMVADTPENLR</td>
<td>LIM and SH3 domain protein 1</td>
<td>LASP1</td>
<td>2</td>
<td>864.4009 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>867.4110 (heavy)</td>
</tr>
<tr>
<td>MCF-10A vs. MCF-7</td>
<td>MSASDPNSSIILTDTAK</td>
<td>Tryptophan--tRNA ligase, cytoplasmic</td>
<td>WARS</td>
<td>2</td>
<td>951.9431 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>954.9532 (heavy)</td>
</tr>
<tr>
<td>BT-474 vs. MCF-7</td>
<td>AQAGEGVRSPMQLELR</td>
<td>Pyridoxal kinase</td>
<td>PDXK</td>
<td>3</td>
<td>653.0035 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>655.0102 (heavy)</td>
</tr>
<tr>
<td>All four comparisons</td>
<td>MKETAEAYLGK</td>
<td>78 kDa glucose-regulated protein</td>
<td>HSPA5</td>
<td>3</td>
<td>453.5602 (light)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>455.5669 (heavy)</td>
</tr>
</tbody>
</table>
The raw LC−MS/MS data files were imported into Skyline 4.1.0.11796 (MacCoss Lab, Department of Genome Sciences, UW). Both the post neutral loss production ion and fragmented production ions were used for peptide identification. Peak areas of the post neutral loss production ion and other qualified fragment ions were extracted and integrated across the elution profile to generate H/L ratios for the targeted peptides from each denaturant concentration.

5.3 Results and Discussion

5.3.1 Selection of SILAC-SPROX Hits for Validation

The selection of protein targets for PAB-SPROX validation was based on the LC-MS/MS data reported in ref (35) and (2). The two studies applied the same SILAC-SPROX protocol(20) to a series of different breast cancer cell lines to identify disease-related changes in protein folding and stability. The first study focused on the MCF-7, MCF-10A, and MDA-MB-231 cell lines(35), while the second study focused on the MCF-7, BT-474, and MDA-MB-468 cell lines (Chapter 2)(2). A total of four comparative analyses were performed in the two studies, each involving a heavy-labeled MCF-7 cell line and one of the other four cell lines that was light-labeled. The hit identification strategies used in the two studies were slightly different. Ref (35) utilized a visual inspection of the data to select hits, while ref (2) utilized a regression analysis to automate data analysis.
To make the protein hits comparable in the two studies, the LC-MS/MS data from the first study (35) were re-analyzed using the hit identification strategy described in the second study (2). A large overlap (~50%) was observed between the hits identified using the new hit identification strategy and those previously reported in ref (35). The peptide hits in ref (35) that were not identified as hits with the new hit identification strategy primarily included (1) those with too few data points (i.e., the new regression analysis was only applied to peptides with H/L data from at least seven (instead of six) denaturant concentrations) and (2) those with poor quality data sets (i.e., the data was not well-fit to Equation 1 or 2). The number of peptide hits identified with the new hit identification strategy is summarized in Table 11. The assayed peptides in Table 11 included those methionine-containing peptides that were identified from seven or more denaturant concentrations in the SILAC-SPROX experiment.

Table 11: Summary of the proteomic data reported in the previous SILAC-SPROX studies (2, 35) for differential thermodynamic stability analysis of breast cancer cell lines.

<table>
<thead>
<tr>
<th>Cell line comparison</th>
<th>Total peptides (proteins) identified</th>
<th>Total peptides (proteins) assayed</th>
<th>Total peptide (protein) hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT-474 vs. MCF-7</td>
<td>28098 (3705)</td>
<td>3464 (957)</td>
<td>341 (242)</td>
</tr>
<tr>
<td>MDA-MB-468 vs. MCF-7</td>
<td>26571 (3460)</td>
<td>2980 (790)</td>
<td>972 (445)</td>
</tr>
<tr>
<td>MCF-10A vs. MCF-7</td>
<td>16582 (2431)</td>
<td>2053 (620)</td>
<td>208 (129)</td>
</tr>
<tr>
<td>MDA-MB-231 vs. MCF-7</td>
<td>18147 (2872)</td>
<td>2049 (650)</td>
<td>177 (127)</td>
</tr>
</tbody>
</table>
A total of 161 protein hits with altered thermodynamic stabilities were assayed with the same 325 methionine-containing peptide probes in all of the four cell line comparisons; that is, they were identified as hits in at least one of the four cell line comparisons. Thus, only these 161 protein hits could have their thermodynamic stability profiles unambiguously determined in all of the five breast cancer cell lines. The overlap among the 325 methionine-containing peptide probes is shown in Figure 22.

Figure 22: Venn diagram showing the overlap among the 325 methionine-containing peptide hits commonly assayed in all of the four cell line comparisons. The red circles represent cell line specific hits.

One goal in this PAB-SPROX validation study was to identify protein hits that showed cell line specific stability changes (i.e., those that were assayed with the same
methionine-containing peptide probes in all four comparative studies and only identified as hits in one comparison). Such protein hits are especially attractive because they may be useful biomarkers for the diagnosis and treatment of specific breast cancer subtypes. For example, the 26 peptide hits that were specific to the MCF-10A vs. MCF-7 comparison changed their stabilities upon transitions from normal cells to cancer cells, and showed the same stabilities in all four cancer cell lines. Thus these 26 peptide hits have the potential to serve as biomarkers for carcinogenesis. Similarly, the 23 peptide hits that were specific to the BT-474 vs. MCF-7 comparison could be used as biomarkers for the luminal-B subtype of breast cancer.

The cell line specific protein hits received higher priority for PAB-SPROX validations if (1) they were identified as hits in multiple biological replicates, (2) they showed the first type of SILAC-SPROX hit behaviors as described in Chapter 2 (see Figures 4A and 4B), or (3) they have been previously found in other studies to play a role in breast cancer. The final protein target list in this initial work is summarized in Table 10. Three peptides from different subunits of T-complex protein 1 and one peptide from nucleophosmin were targeted for MDA-MB-468 specific stability changes. One peptide from LIM and SH3 domain protein 1 was targeted for MDA-MB-231 specific stability changes. One peptide from tryptophan--tRNA ligase (cytoplasmic) was targeted for MCF-10A specific stability changes. One peptide from pyridoxal kinase was targeted
for BT-474 specific stability changes. These proteins have been previously linked to breast cancer (87, 109, 130-133).

The MKETAEAYLGK peptide from 78 kDa glucose-regulated protein was identified as a hit in all four comparisons (Figure 22). Since each of the four cell line comparisons involved a heavy-labeled MCF-7 cell line and one of the other four cell lines that was light-labeled, this enabled determination of the relative protein stability among the five cell lines. The SILAC-SPROX data for the MKETAEAYLGK peptide in the four cell line comparisons were shown in Figure 23 (data from ref (35) and (2)). The relative thermodynamic stability for the MKETAEAYLGK peptide could be determined as: BT-474 > MDA-MB-468 > MDA-MB-231 > MCF-7 > MCF-10A. Such a stability ranking was also included in the target list for PAB-SPROX validation (Table 10).
Figure 23: SILAC-SPROX data for the MKETAEAYLGK peptide from 78 kDa glucose-regulated protein in the four cell line comparisons. The red solid lines represent the best fit of the data to Equations 1 or 2; the green curves represent the extracted SPROX curves; the vertical dashed lines indicate the transition midpoints of the SPROX curves; the horizontal dashed lines indicate the cutoffs for significant changes in log₂(H/L) values. A red × indicates a data point that was not included in the regression analysis.

5.3.2 Experimental Design

The PAB-SPROX technique was employed here to develop the validation assays for the selected protein targets as shown in Table 10. The general workflow used in the PAB-SPROX experiments is shown in Figure 24. The work described in this chapter was
designed as a pilot study that only involved the use of two urea concentrations and the MCF-7 breast cancer cell line.

Figure 24: General workflow used in the PAB-SPROX experiments for the validation of SILAC-SPROX hits.
5.3.3 PAB-SPROX Results

All the peptide targets listed in Table 10 were identified unambiguously in the pilot study described here. That is, the elution profiles of the post neutral loss production ions were consistent between the light and heavy PAB labeled methionine-containing peptides, and there were other fragment ions and precursor ions eluting around the same time to support the identification. Most of the H/L ratios for the eight target peptides were close to 1 at both urea concentrations (Table 12). This is expected because the light and heavy PAB labeled methionine-containing peptides were derived from the same MCF-7 cell lysate. The deviations from 1 observed for some H/L ratios should result from the variability introduced during the processing steps (e.g., during TCA precipitation). This can be controlled for in future studies by normalizing the data to a control (non-hit) protein that is either in the test sample or spiked into the sample after the SPROX reaction is complete.

Table 12: Summary of H/L ratios generated for the eight target peptides using the PAB-SPROX protocol.

<table>
<thead>
<tr>
<th>Peptide sequence</th>
<th>Gene name</th>
<th>H/L (0 M)</th>
<th>H/L (2.4 M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVLLDPQLVPGGGAEMAVAHALTEK</td>
<td>CCT3</td>
<td>0.92</td>
<td>N/A</td>
</tr>
<tr>
<td>GVIVDSDKFHPOQPK</td>
<td>CCT5</td>
<td>0.91</td>
<td>0.45</td>
</tr>
<tr>
<td>MEGPLSVFGDR</td>
<td>TCP1</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>MSVQPSTLVSGFEITPPVVL</td>
<td>NPM1</td>
<td>0.87</td>
<td>1.11</td>
</tr>
<tr>
<td>QSFTMVADTPENLR</td>
<td>LSP1</td>
<td>0.97</td>
<td>1.65</td>
</tr>
<tr>
<td>MSASDPNSSIFLTDTAK</td>
<td>WARS</td>
<td>1.06</td>
<td>1.47</td>
</tr>
<tr>
<td>AQAGEGVRPSMPQLELR</td>
<td>PDXK</td>
<td>0.69</td>
<td>1.25</td>
</tr>
<tr>
<td>MKETAEAYLGK</td>
<td>HSPA5</td>
<td>1.03</td>
<td>1.33</td>
</tr>
</tbody>
</table>
5.4 Conclusions

Previous SILAC-SPROX studies of the MCF-10A, MCF-7, BT-474, MDA-MB-468 and MDA-MB-231 breast cancer cell lines identified a subset of protein hits with cell line specific stability changes\(^2\), \(^{35}\). These proteins have the potential to be useful biomarkers for the diagnosis and treatment of specific breast cancer subtypes. As a first step toward the validation of these protein biomarker candidates in clinical samples, the work described in this chapter was focused on the use of the PAB-SPROX protocol\(^{33}\) in the cell culture models of breast cancer to establish parallel reaction monitoring (PRM) assays for the target proteins of interest. Application of the PAB-SPROX protocol on the MCF-7 cell lines enabled reproducible identification and quantitation of the eight target peptides prioritized in this work. The validation assays established in this study can be applied in the future work to validate the altered stabilities observed for these protein targets in the corresponding cell line comparisons.
6. Conclusions and Future Directions

This thesis was largely focused on understanding the changes in protein folding stability associated with breast cancer. Such an understanding is very important for the characterization of breast cancer, because changes in protein folding stability can occur as a result of a wide range of biologically significant phenomena such as point mutations, post-translational modifications, or altered binding interactions with cellular ligands. However, the applications of large-scale protein folding stability measurements to disease states are very recent and some important preliminary questions still remain to be answered. These questions include, but are not limited to, the applicability of using thermodynamic stability profiles to differentiate the disease states, the benefits and drawbacks of using a specific technique in a given context, the ability to discover novel and more biologically relevant information about disease states, etc.

One goal of this work was to provide data that could help address the above questions. The major contribution of this work included (1) the extension of using proteome-wide thermodynamic profiling strategies for breast cancer characterization, where the extension was mainly focused on the study of more disease subtypes and the use of additional approaches, (2) the development of a SILAC-based limited proteolysis strategy capable of probing the more local unfolding/refolding properties of proteins on a large scale, and (3) the establishment of targeted mass spectrometry-based validation assays for protein biomarker candidates with altered thermodynamic stabilities.
The work presented in Chapter 2, in combination with the results obtained in previous SILAC-SPROX on several additional breast cancer subtypes, provides some answers to the above questions about the ability of thermodynamic stability measurements to differentiate the disease states and to discover novel biologically relevant information about disease states. The high heterogeneity of breast cancer makes it necessary to study different subtypes to obtain a better understanding of the disease. As part of this work protein folding stability differences were studied in a comparison of two luminal breast cancer subtypes, luminal-A and -B (i.e., MCF-7 and BT-474 cells, respectively), and in a comparison of a luminal-A and basal subtype of the disease (i.e., MCF-7 and MDA-MB-468 cells, respectively). The 242 and 445 protein hits identified with altered stabilities in these comparative analyses successfully created distinct molecular markers to differentiate the three cell lines. A large fraction (~70%) of the differentially stabilized proteins did not have significant expression level changes. This suggests thermodynamic stability measurements can provide information about disease states that is orthogonal to that obtained in protein expression level analyses.

To answer the question about technique selection for disease state characterization, the work described in Chapter 3 was focused on the development of a SILAC-based limited proteolysis strategy. The applicability of the protocol was demonstrated in a proof-of-principle study using proteins from a yeast cell lysate and a ubiquitous ligand (ATP). The types of protein conformational changes probed by the
SPROX and LiP techniques are different. SPROX is sensitive to changes in the more global unfolding/refolding properties of proteins, while LiP is sensitive to changes in the more local unfolding/refolding properties of proteins. The two techniques are expected to provide complementary information on the disease states. Because the more local conformational changes typically do not have a strong denaturant dependence to be detected by SPROX, but may be more or less sensitive to protease digestion to be detected by LiP. Similarly, changes in the global stability of a protein (or protein domain) may not always be differentially susceptible to protease. Such changes can detected by SPROX but not LiP.

To test this hypothesis, the work included in Chapter 4 applied the SILAC-LiP protocol to study the conformational changes of proteins involved in breast cancer pathogenesis using the same two cell lines that were used in the previous SILAC-SPROX study. Indeed, the overlap between the SILAC-SPROX and SILAC-LiP protein hits was found to be relatively small (∼20%). The study has also identified molecular signatures that were different from those discovered in SILAC-SPROX study, thus providing additional insight into the molecular basis of the disease. Interestingly, similar to SILAC-SPROX, a significant fraction (∼69%) of the protein hits identified with conformational differences did not have altered expression levels. Therefore, this work demonstrates the benefits of using both SILAC-SPROX and SILAC-LiP techniques for disease state characterization to provide complementary biologically relevant information.
An important next step is to validate the stability differences observed in these discovery-based experiments with clinical samples. Summarizing the results of SILAC-SPROX analyses performed to date on the MCF-10A, MCF-7, BT-474, MDA-MB-468 and MDA-MB-231 breast cancer cell lines, a subset of protein hits were identified with cell line specific stability changes. These proteins have the potential to be useful biomarkers for the diagnosis and treatment of specific breast cancer subtypes. Chapter 5 described initial efforts to develop and validate assays for the protein biomarker candidates with cell line specific stability changes. Application of the PAB-SPROX protocol on the MCF-7 cell line enabled reproducible identification and quantitation of the eight target peptides prioritized in the work described in Chapter 5. This is encouraging as it brings us one step closer to the ultimate goal of translating the differential stability profiling strategy into clinical practice.

In conclusion, this thesis demonstrates the potential of using proteome-wide thermodynamic profiling strategies to provide a new paradigm for the characterization of disease states. There remains a need for application of the established validation assays on the same cell line comparisons performed in the previous SILAC-SPROX study. This can help differentiate the true positives (i.e., proteins with altered stabilities in the cell line comparisons) from the false positives and further narrow down the candidate list for future clinical studies. In theory, the described approaches can be directly applied to the analysis of clinical samples provided sufficient amounts of
protein can be obtained. A typical SILAC-LiP experiment requires ~200 µg of total protein, while a typical SILAC-SPROX experiment requires ~2-3 mg of total protein. With the aid of targeted mass spectrometry, the total amount of required material can be significantly reduced (~20-fold less) in a PAB-SPROX experiment. Therefore, we anticipate the ultimate application of the described protein stability profiling strategies to the analysis of clinical samples should happen in the near future.
Appendix A: Identification of Protein Targets for Inositol Pyrophosphate \( \text{InsP}_8 \) in Human Colon Cancer Cell Lines Using iTRAQ-SPROX

Introduction

Described here is a completed collaboration between the Fitzgerald lab and the inositol signaling group at NIEHS (Research Triangle Park, NC). In this work, the iTRAQ-SPROX technique was utilized to identify potential protein targets of 1,5-bisdiphosphoinositol 2,3,4,6-tetrakisphosphate (\( \text{InsP}_8 \)) (Figure S1). The inositol pyrophosphates (including \( \text{InsP}_7 \) and \( \text{InsP}_8 \)) comprise a unique class of cell signaling molecules and are characteristic of containing high-energy diphosphate groups. They have been shown to play a dynamic role in metabolism regulation at the molecular (phosphate homoeostasis), cellular (energetic) and organismal (insulin signaling) levels(134). They are largely synthesized from \( \text{InsP}_6 \) by two classes of enzymes, the IP6Ks (\( \text{InsP}_6 \) kinases) and the PPIP5Ks (\( \text{InsP}_7 \) kinases) (Figure S1).

In a study of inositol pyrophosphates and their roles in Pi homeostasis, our collaborators showed that levels of both 5-\( \text{InsP}_7 \) and ATP decrease upon Pi starvation and subsequently recover during Pi replenishment using a human intestinal epithelial cell line (HCT116)(135). Interestingly, in this study Gu and coworkers(135) also found that \( \text{InsP}_8 \) reacts more dramatically to fluctuations in extracellular [Pi] than does 5-\( \text{InsP}_7 \). In order to understand this novel \( \text{InsP}_8 \) response, here the iTRAQ-SPROX technique was used to identify \( \text{InsP}_8 \)-interacting proteins (both directly and indirectly).
Supplemental Figure 1: Inositol pyrophosphate biosynthetic pathway.

**Experimental**

**Cell Lysate Preparation**

Both the wild type and PPIP5Ks knockout HCT116 cell pellets were cultured by Dr. Chunfang Gu. Cell lysates were prepared as described in Chapter 2. The lysis buffer was 20 mM phosphate buffer (pH 7.4) containing 10 mM NaF and a cocktail of protease inhibitors (same composition as described in Chapter 2). Lysed cells were centrifuged at 14,000 g for 15 min at 4 °C, and the supernatant was used for subsequent analyses. The total protein concentration was determined by a Bradford assay and normalized to 10 mg/mL.
iTRAQ-SPROX Analyses

In Experiment 1, a total of 5 sets of iTRAQ-SRPOX analyses were conducted including 1 on the wild type cell lysate (WT), 2 on the two different PPIP5Ks knockout cell lysates (KO1-InsP$_8$ and KO2-InsP$_8$), and 2 on the two PPIP5Ks knockout cell lysates each incubated with InsP$_8$ (KO1+InsP$_8$ and KO2+InsP$_8$). Specifically, the two PPIP5Ks knockout cell lysates were each divided into two 180 µL portions. A 20 µL aliquot of a 5 mM InsP$_8$ solution was added to one portion of the cell lysate to generate the “KO+InsP$_8$” sample, and 20 µL of water was added to the other portion of the cell lysate to generate the “KO-InsP$_8$” sample. A 180 µL aliquot of the wild type cell lysate was also mixed with 20 µL of 5 mM InsP$_8$ to generate the “WT+InsP$_8$” sample. This was because the endogenous concentration of InsP$_8$ in mammalian cells is much lower (0.1-0.2 µM) than that used in this experiment.

In Experiment 2, a total of 2 sets of iTRAQ-SRPOX analyses were performed including one on the wild type cell lysate (WT-InsP$_8$) and one on the wild type cell lysate incubated with InsP$_8$ (WT+InsP$_8$). Similar to Experiment 1, the wild type cell lysate was divided into two 180 µL portions. A 20 µL aliquot of a 5 mM InsP$_8$ solution was added to one portion of the cell lysate to generate the “WT+InsP$_8$” sample, and 20 µL of water was added to the other portion of the cell lysate to generate the “WT-InsP$_8$” sample.

Both experiments were performed using identical iTRAQ-SPROX protocol as previously described(11, 136) except that the two experiments employed slightly
different final denaturant concentrations. Briefly, the samples were equilibrated for 1 h at room temperature before aliquots (20 µL) of each sample were combined with 75 µL of a series of 20 mM phosphate buffers (pH 7.4) containing different concentrations of urea. The final urea concentrations in the two experiments performed here are shown in Table S1. The final concentration of InsPs was 100 µM in both experiments. Each mixture was equilibrated for 1 h at room temperature, and 5 µL of 30% (w/w) hydrogen peroxide (Sigma) was added to each of the protein samples in the denaturant-containing buffers. After 6 min, the reaction in each protein sample was quenched with 1 mL of 300 mM L-methionine (Sigma). The proteins in each urea-containing buffer were precipitated upon addition of 210 µL of TCA (1 g/mL) (Sigma) and an overnight incubation on ice.

<table>
<thead>
<tr>
<th>Final urea concentrations (M)</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>iTRAQ reporter ion m/z</td>
<td>113</td>
<td>114</td>
</tr>
</tbody>
</table>

The resulting protein pellets from each denaturant-containing buffer were dissolved in 30 µL of 0.5 M TEAB buffer (Sigma) containing 0.1% SDS. The disulfide bonds were reduced with a final concentration of 5 mM TCEP (Thermo Fisher, Inc.) for 1 hour at 60 °C. The samples were then treated with 10 mM MMTS (Sigma) for 15 min at room temperature. Ultimately, the protein samples were digested with trypsin using an
enzyme/substrate ratio of 1/50 (w/w) at 37 °C with overnight incubation. The proteolytic digestion reaction was quenched upon acidification (pH ~ 2–3) with TFA (Sigma).

Each of the resulting peptide mixtures was labeled with 0.5 unit of an iTRAQ-8plex tag (AB Sciex) dissolved in 50 µL of isopropanol (VWR International) according to the manufacturer’s instruction. The labeling scheme can be found in Table S1. The labeling reactions were allowed to proceed for 2 h at room temperature. Within each set, 10 µL or 30 µL aliquots from each of the 8 different iTRAQ labeled samples were combined to generate the non-enriched and Met-enriched samples, respectively. The combined non-enriched samples (80 µL) were subjected to a desalting protocol using the C18 MacroSpin columns (The Nest Group), while the combined Met-enriched samples (240 µL) were subjected to a methionine-containing peptide enrichment protocol using the Pi3 Methionine Selective Resins (The Nest Group). Both protocols were performed according to the manufacturer’s instructions. In the end, the non-enriched samples were each subjected to a single LC-MS/MS analysis, while the Met-enriched samples were each subjected to two LC-MS/MS analyses.

**LC-MS/MS Analyses**

LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive Plus high-resolution mass spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. The instrument parameters were the same as those described in Chapter 2. Peak lists were
extracted from the LC-MS/MS data and searched against the SwissProt Homo Sapiens database (Version 2016-04-13) using Proteome Discoverer (Version 2.1.1.21). All data was searched with the fixed MMTS modification on cysteine and iTRAQ 8-plex modifications on lysine and N-termini. The oxidation of methionine and deamidation of asparagine and glutamine were set as variable modifications in the search. Up to three missed tryptic cleavages were allowed. The parameters included a 10 ppm mass tolerance window for precursor masses and 0.02 Da for fragment mass tolerance. Only peptide spectra with FDR < 5%, isolation interference ≤30%, and eight iTRAQ reporter ion intensities that summed to >1000 were used in subsequent analyses of the data.

In addition to the above LC-MS/MS analyses performed in the data-dependent acquisition (DDA) mode where MS/MS fragmentation spectra are generated for all detected precursors, the Met-enriched samples in Experiment 1 were also analyzed on the same instrument using parallel reaction monitoring (PRM) method. Four methionine-containing peptides from PKM were targeted and their precursors were summarized in Table S2. The method was based on a 90 min 3–30% acetonitrile gradient, and a normalized collision energy of 30 V was used. The raw LC-MS/MS data files were imported into Skyline 3.7.1.11271 (MacCoss Lab, Department of Genome Sciences, UW). The fragmented product ions were used for peptide identification. An extracted ion chromatogram (EIC) was generated from the ion signals of each iTRAQ reporter ion.
The iTRAQ reporter ion intensities were measured by integrating their EIC peaks across the elution profiles.

Supplemental Table 2: Summary of the PRM precursors from pyruvate kinase PKM that were targeted in Experiment 1. Bolded and underlined amino acids were modified with iTRAQ-8plex.

<table>
<thead>
<tr>
<th>Peptide sequence</th>
<th>Charge</th>
<th>Precursor m/z</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDVIVLTGWRPGFTNTMR</td>
<td>2</td>
<td>1284.1889</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>856.4617</td>
</tr>
<tr>
<td>VNFAMNVGK</td>
<td>2</td>
<td>794.4605</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>529.9761</td>
</tr>
<tr>
<td>MQHLLAR</td>
<td>2</td>
<td>586.8474</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>391.5674</td>
</tr>
<tr>
<td>ITLDNAYMEK</td>
<td>2</td>
<td>903.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>602.6691</td>
</tr>
</tbody>
</table>

Data Analysis

The iTRAQ-SPROX data analysis was performed as previously described (34, 36). Briefly, in each analysis, the eight iTRAQ reporter ion intensities in each product ion mass spectrum were averaged and the raw intensity of each reporter ion in the product ion mass spectrum was divided by the average value to generate so-called N1-normalized values. All of the N1-normalized values for non-methionine containing peptides in the non-enriched sample were averaged for each reporter ion. Summarized in Table S3 are the set of eight average values (so-called N2 normalization factors) that were generated in each SPROX analysis. The N1-normalized values for each methionine-
containing peptide were then divided by the corresponding N2 normalization factors to generate the N2-normalized reporter ion intensities.

**Supplemental Table 3: Summary of the N2 normalization factors and their standard deviations for each iTRAQ-SRPOX analysis.**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Sample</th>
<th>113</th>
<th>114</th>
<th>115</th>
<th>116</th>
<th>117</th>
<th>118</th>
<th>119</th>
<th>121</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WT+InsPs</td>
<td>0.85</td>
<td>0.92</td>
<td>0.97</td>
<td>1.06</td>
<td>1.13</td>
<td>1.10</td>
<td>0.93</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.12</td>
<td>0.12</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>KO1-InsPs</td>
<td>0.86</td>
<td>0.78</td>
<td>1.10</td>
<td>1.05</td>
<td>1.12</td>
<td>1.14</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.12</td>
<td>0.14</td>
<td>0.10</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>KO2-InsPs</td>
<td>0.74</td>
<td>0.93</td>
<td>1.01</td>
<td>1.14</td>
<td>1.15</td>
<td>1.04</td>
<td>0.90</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.12</td>
<td>0.25</td>
<td>0.11</td>
<td>0.14</td>
<td>0.13</td>
<td>0.11</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>KO1+InsPs</td>
<td>0.98</td>
<td>0.92</td>
<td>1.00</td>
<td>1.03</td>
<td>1.19</td>
<td>1.03</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.14</td>
<td>0.11</td>
<td>0.12</td>
<td>0.11</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>KO2+InsPs</td>
<td>0.76</td>
<td>0.90</td>
<td>1.04</td>
<td>1.15</td>
<td>1.16</td>
<td>1.08</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.13</td>
<td>0.14</td>
<td>0.11</td>
<td>0.13</td>
<td>0.10</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>WT-InsPs</td>
<td>0.99</td>
<td>1.21</td>
<td>0.85</td>
<td>0.98</td>
<td>1.03</td>
<td>1.02</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.17</td>
<td>0.17</td>
<td>0.19</td>
<td>0.14</td>
<td>0.16</td>
<td>0.14</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>WT+InsPs</td>
<td>0.97</td>
<td>1.14</td>
<td>0.85</td>
<td>1.09</td>
<td>1.06</td>
<td>1.04</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.15</td>
<td>0.13</td>
<td>0.28</td>
<td>0.10</td>
<td>0.11</td>
<td>0.14</td>
<td>0.14</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Ultimately, the chemical denaturation data sets (i.e., N2-normalized iTRAQ reporter ion intensity as a function of the denaturant concentration) were fitted to a four-parameter sigmoidal equation (Equation S1) using a JAVA-based program (developed in house) that utilized the Nelder and Mead Simplex method for regression analysis(137).

\[ y = A + \frac{B - A}{1 + e^{-\left(C_{1/2} - x\right)/b}} \]  

Equation S1

In Equation S1, A and B are the pre- and post-transition baselines (respectively), \( C_{1/2} \) is the transition midpoint, and \( b \) represents the steepness of the transition. Each data set was fitted an additional 8 times with the systematic removal of one point. The best regression was selected by the highest \( R^2 \) value and output by the program. PSMs with poor data (\( R^2 < 0.8 \)) were removed from the subsequent analysis. If a methionine-containing peptide was identified with multiple PSMs (i.e., multiple chemical denaturation data sets) that were well fit to Equation S1, the N2-normalized reporter ion intensities were averaged to generate a single set of iTRAQ reporter ion intensities at the 8 denaturant concentrations and the averaged data was fit to Equation S1 as described above to extract a single \( C_{1/2} \) value.

**Hit Identification**

Six comparative analyses were performed based on the thermodynamic stability data generated in Experiment 1. One comparative analysis was performed based on the
thermodynamic stability data generated in Experiment 2 (Table S4). Peptide hit determination was based on (1) the magnitude of the $\Delta C_{1/2}$ value between the two samples that were being compared and (2) the magnitude of the N2-normalized value differences at or between the transition regions of the chemical denaturation curves generated for the two samples that were being compared. A pooled standard deviation ($S_{\text{pooled}}$) of $C_{1/2}$ measurements for all the well-fit PSMs can be determined for each comparison using Equation S2. Statistically significant $\Delta C_{1/2}$ values were determined to be those $> 2 * S_{\text{pooled}}$. In Equation S2, $x_i$, $x_j$ and $x_k$ refer to the $C_{1/2}$ values from different PSMs for a given peptide, $N_1$, $N_2$, $N_3$ are the number of PSMs for a given peptide, and $n_t$ is the total number of unique peptides.

$$S_{\text{pooled}} = \sqrt{\frac{\sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{N_2}(x_j - \bar{x}_2)^2 + \sum_{k=1}^{N_3}(x_k - \bar{x}_3)^2 + \cdots}{N_1 + N_2 + N_3 + \cdots - n_t}}$$  \hspace{1cm} \text{Equation S2}$$

Statistically significant N2-normalized value differences were taken as those with $\text{Diff}_{\text{prob}} < 0.1$, where $\text{Diff}_{\text{prob}}$ corresponds to the probability that the measured N2-normalized value difference would randomly occur in the data sets. The specific hit selection criteria used for each comparison was summarized in Table S4.
Supplemental Table 4: Summary of the hit criteria used in the iTRAQ-SPROX analysis.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Comparison</th>
<th>(2*S_{pooled})</th>
<th>-log_{10}(Diff_{prob})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WT+InsP_{8} vs. KO1-InsP_{8}</td>
<td>0.98</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>WT+InsP_{8} vs. KO2-InsP_{8}</td>
<td>1.04</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>WT+InsP_{8} vs. KO1+InsP_{8}</td>
<td>0.98</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>WT+InsP_{8} vs. KO2+InsP_{8}</td>
<td>1.02</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>KO1-InsP_{8} vs. KO1+InsP_{8}</td>
<td>1.00</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>KO2-InsP_{8} vs. KO2+InsP_{8}</td>
<td>1.12</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>WT-InsP_{8} vs. WT+InsP_{8}</td>
<td>1.02</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Results and Discussion

Experimental Design

This study was designed to identify InsP_{8}-interacting proteins using the wild type and PPIP5Ks knockout HCT116 cell lysates. The two PPIP5Ks knockout HCT116 cells were cultured using different protocols. A total of 5 and 2 sets of iTRAQ-SRPOX analyses were performed as previously described in Experiment 1 and 2, respectively. The general workflow employed in the iTRAQ-SPROX experiment for InsP_{8} binding analyses is shown in Figure S2. The thermodynamic stability data generated in Experiment 1 were used to conduct two sets of comparative analyses, each involving a wild type cell lysate (WT+InsP_{8}), a PPIP5Ks knockout cell lysate (KO1-InsP_{8} or KO2-InsP_{8}) and a PPIP5Ks knockout cell lysate that was incubated with InsP_{8} (KO1+InsP_{8} or KO2+InsP_{8}).
KO2+InsPs), to identify proteins interacting with PPIP5Ks and/or InsPs (Figure S3). The ligand concentration employed in the iTRAQ-SPROX experiment is typically in large excess over the protein concentration and the ligand endogenous concentration to detect weak binding interactions. Thus, the wild type cell lysate in Experiment 1 was spiked in with the same amount of InsPs as were used in the PPIP5Ks knockout cell lysates. Experiment 2 was designed to assess the effect of this excessive InsPs on protein stability.

Supplemental Figure 2: General workflow used in the iTRAQ-SPROX experiments for the detection of protein targets of InsPs.
Supplemental Figure 3: Summary of the comparative analyses performed in Experiment 1. Comparisons between WT+InsP₈ and KO-InsP₈ can help identify proteins interacting with PPIP5Ks and/or InsP₈. Comparisons between WT+InsP₈ and KO+InsP₈ can help identify proteins interacting with PPIP5Ks. Comparisons between KO+InsP₈ and KO-InsP₈ can help identify proteins interacting with InsP₈.

Proteomic Coverage

In the seven sets of iTRAQ-SPROX analyses performed here, the proteomic coverage ranged from 1000-1500 proteins based on 2000-3300 methionine (Met)-containing peptides (Table S5). The coverage only included Met peptides with high quality chemical denaturation data (FDR <5%, isolation interference ≤30%, R² >0.8, no missing iTRAQ reporter ion intensity and their sum intensity >1000).

Each set of the comparative analyses performed in Experiment 1 included three comparisons. The “WT+InsPs” versus “KO-InsPs” comparison is expected to identify proteins interacting with PPIP5Ks and/or InsPs. The “WT+InsPs” versus “KO+InsPs” comparison is expected to identify proteins interacting with PPIP5Ks. The “KO-InsPs” versus “KO+InsPs” comparison is expected to identify proteins interacting with InsPs.
each of these 6 comparisons, about 1100-1800 Met peptides from 600-900 unique proteins were matched between the two samples such that their interactions with PPIP5Ks and/or InsPs can be probed (Table S5). Each of these 6 comparisons identified a subset of peptide and protein hits, resulting in a hit rate ranging from 4.7-17.3% (Table S5, Figures S4 and S5).

**Supplemental Table 5: Summary of the proteomic data obtained in the iTRAQ-SPROX experiments for the detection of protein targets of InsPs.**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Comparison</th>
<th>Identified Met peptides (proteins)</th>
<th>Matched Met peptides (proteins)</th>
<th>Hit peptides (proteins)</th>
<th>Hit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WT+InsPs</td>
<td>3293 (1550)</td>
<td>1153 (611)</td>
<td>93 (86)</td>
<td>8.1%</td>
</tr>
<tr>
<td></td>
<td>KO1-InsPs</td>
<td>2030 (1073)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>WT+InsPs</td>
<td>3293 (1550)</td>
<td>1162 (592)</td>
<td>107 (101)</td>
<td>9.2%</td>
</tr>
<tr>
<td></td>
<td>KO1+InsPs</td>
<td>2042 (1038)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KO1+InsPs</td>
<td>2042 (1038)</td>
<td>1311 (691)</td>
<td>62 (56)</td>
<td>4.7%</td>
</tr>
<tr>
<td></td>
<td>KO1-InsPs</td>
<td>2030 (1073)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WT+InsPs</td>
<td>3293 (1550)</td>
<td>1721 (879)</td>
<td>237 (210)</td>
<td>13.8%</td>
</tr>
<tr>
<td></td>
<td>KO2-InsPs</td>
<td>3035 (1453)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>WT+InsPs</td>
<td>3293 (1550)</td>
<td>1797 (923)</td>
<td>124 (116)</td>
<td>6.9%</td>
</tr>
<tr>
<td></td>
<td>KO2+InsPs</td>
<td>2982 (1432)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>KO2+InsPs</td>
<td>2982 (1432)</td>
<td>1636 (864)</td>
<td>283 (144)</td>
<td>17.3%</td>
</tr>
<tr>
<td></td>
<td>KO2-InsPs</td>
<td>3035 (1453)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>WT+InsPs</td>
<td>2397 (1144)</td>
<td>1152 (622)</td>
<td>176 (163)</td>
<td>15.3%</td>
</tr>
<tr>
<td></td>
<td>WT-InsPs</td>
<td>1955 (976)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Figure 4: Volcano plots showing the thermodynamic stability measurements in the (A) “WT+InsP₈” versus “KO1-InsP₈” comparison, (B) “WT+InsP₈” versus “KO1+InsP₈” comparison, and (C) “KO1-InsP₈” versus “KO1+InsP₈” comparison. The vertical lines represent the S_{pooled} hit criteria, and the horizontal lines represent the Diff_{prob} hit criteria. The black and red points represent the non-hit and hit peptides, respectively.
Supplemental Figure 5: Volcano plots showing the thermodynamic stability measurements in the (A) “WT+InsP8” versus “KO2-InsP8” comparison, (B) “WT+InsP8” versus “KO2+InsP8” comparison, and (C) “KO2-InsP8” versus “KO2+InsP8” comparison. The vertical lines represent the $S_{\text{pooled}}$ hit criteria, and the horizontal lines represent the $\text{Diff}_{\text{prob}}$ hit criteria. The black and red points represent the non-hit and hit peptides, respectively.
The protein hits identified in the “WT+InsP₈” versus “KO-InsP₈” comparison are expected to include those identified in both the “WT+InsP₈” versus “KO+InsP₈” and “KO-InsP₈” versus “KO+InsP₈” comparisons. Figure S6 shows the overlaps of protein hits among the three comparisons for the two different PPIP5Ks knockout cell lysate samples. The overlaps were not as large as expected. Perhaps the simultaneous absence of PPIP5Ks and Ins₈ would have different downstream effects than the sole absence of PPIP5Ks or Ins₈. It could also result from the false discovery associated with the iTRAQ-SPROX technique.

Supplemental Figure 6: Venn diagrams showing the overlaps of protein hits among the three comparisons involving (A) PPIP5Ks knockout cell lysate 1 and (B) PPIP5Ks knockout cell lysate 2.
It is worth noting that the two different PPIP5Ks knockout cells were cultured using different protocols. Thus, the same comparisons involving different PPIP5Ks knockout cell lysates (e.g., “WT+InsPs” versus “KO1-InsPs” comparison and “WT+InsPs” versus “KO2-InsPs” comparison) cannot be used as a direct measure of the biological variability, as can be seen from the quite different hit rates.

Experiment 2 compared the thermodynamic stabilities of proteins in the wild type cell lysate with endogenous [InsPs] (WT-InsPs) to those in the same cell lysate, albeit with excessive InsPs (WT+InsPs), to evaluate the effect of excess InsPs on protein stability as were used in Experiment 1. In total, about 600 proteins were probed for potential interactions with the InsPs using ~1100 Met peptides. A total of 176 unique peptides from 163 proteins were identified as hits (Table S5, Figure S7). These protein hits could result from (1) weak binding interactions with InsPs that were only detectable with excessive InsPs, (2) nonspecific binding interactions with InsPs, and (3) false discovery by the iTRAQ-SPROX technique.
Supplemental Figure 7: Volcano plot showing the thermodynamic stability measurements in the “WT+InsP₈” versus “WT-InsP₈” comparison. The vertical lines represent the $S_{\text{pooled}}$ hit criteria, and the horizontal line represents the $\text{Diff}_{\text{prob}}$ hit criteria. The black and red points represent the non-hit and hit peptides, respectively.

**Protein Target of Interest: Pyruvate Kinase**

Of particular interest to our collaborators’ work is the protein hit pyruvate kinase PKM2. Four peptides from PKM2 were identified as hits in the different comparisons performed in Experiment 1 (Figure S8). Among these peptide hits, peptide GDVVIVLTGWRPGSGFTNTMR is especially interesting to us because it locates at the allosteric activator fructose 1,6-bisphosphate (FBP) binding site of pyruvate kinase (Figure S9). Pyruvate kinase catalyzes the last step of glycolysis, transferring the
phosphate from phosphoenolpyruvate (PEP) to ADP to yield ATP and pyruvate. It has been hypothesized that FBP binding induces conformational changes that promote the association of the protein into homotetramers that comprise the most active form of the enzyme(138-140). However, our collaborators found hyperactive metabolism and hyperactive pyruvate kinase in InsP₈-deficient cells. Thus they hypothesized that InsP₈ competes with FBP and turns the active tetramers into inactive monomers. If this were true, PKM2 would only be active in the “KO1-InsP₈” and “KO2-InsP₈” samples in Experiment 1. This also implies that InsP₈-bound monomer should be more thermodynamically stable than the FBP-bound tetramer.
Supplemental Figure 8: Peptide hits from pyruvate kinase identified in Experiment 1 including (A) MQHLIAR in the “KO1-InsPs” vs. “KO1+InsPs” comparison, (B) GDVVIVLTGWRPGSGFTNTMR in the “WT+InsPs” vs. “KO1+InsPs” comparison, (C) ITLDNAYMEK in the “WT+InsPs” vs. “KO2-InsPs” comparison, (D) GDVVIVLTGWRPGSGFTNTMR in the “WT+InsPs” vs.
“KO2+InsP₈” comparison, and (E) VNFAMNVGK in the “KO2-InsPs” vs. “KO2+InsPs” comparison. iTRAQ-SPROX data from the former one in each comparison are represented by black open circles, while data from the latter one in each comparison are represented by red filled circles. The black and red solid curves represent the best fit of the corresponding data to Equation S1. The black and red dashed lines indicate the respective C₁/₂ values. An ‘X’ represents a data point that was overlooked during fitting.
Supplemental Figure 9: Schematic representations of the folded three-dimensional structure of the protein hit PKM2 identified in Experiment 1. The PKM2 homotetramer is shown in part A (PDB 1T5A) with the allosteric activator fructose 1,6-bisphosphate (FBP) binding site highlighted. In part B, the regions to which the peptide hits mapped are colored in red in the PKM2 monomer, which is in complex with Mg$^{2+}$, K$^{+}$, FBP, and the inhibitor oxalate.
Peptide GDVVIVLTGWRPGSGFTNTMR was found to be more stable in both the “WT+InsPs” vs. “KO1+InsPs” comparison (ΔC_{1/2} = 1.4 M urea, Figure S8B) and the “WT+InsPs” vs. “KO2+InsPs” comparison (ΔC_{1/2} = 1.1 urea, Figure S8D), which is very consistent. However, its hit behaviors cannot be evaluated in other comparisons in Experiment 1 due to missing data or bad quality data. Therefore, additional LC-MS/MS analyses using PRM methods were performed to specifically acquire more data on these four peptides from PKM2. Summarized in Table S6 are their C_{1/2} values determined by fitting the averaged data generated from triplicate non-targeted analyses and triplicate targeted analyses to Equation S1. Unfortunately, the direction of shift in the C_{1/2} values of peptide GDVVIVLTGWRPGSGFTNTMR is not consistent between “WT+InsPs” vs. “KO1-InsPs” comparison and the “WT+InsPs” vs. “KO2-InsPs” comparison. Thus no clear conclusions can be drawn from this data set.

In order to evaluate the effect of excessive InsPs (as were used in Experiment 1) on the thermodynamic stabilities of proteins (especially on pyruvate kinase), Experiment 2 compared the thermodynamic stabilities of proteins in the wild type cell lysate with endogenous [InsP8] to those in the same cell lysate, albeit with excessive InsPs. The comparison identified a subset of peptide and protein hits. However, no peptide hits were identified from pyruvate kinase.
**Conclusions**

The iTRAQ-SPOX technique was used to identify InsP₈-interacting proteins in the wild type and PPIP5Ks knockout HCT116 cell lysates. A total of 5 sets of iTRAQ-SPOX analyses were conducted in Experiment 1, including 1 on the wild type cell lysate, 2 on the two different PPIP5Ks knockout cell lysates, and 2 on the two PPIP5Ks knockout cell lysates each incubated with InsP₈. This enabled two sets of comparative analyses to be performed. Each set of comparisons compared any two of the wild type cell lysate, the PPIP5Ks knockout cell lysate, and the ligand-incubated PPIP5Ks knockout cell lysate. Each of these 6 comparisons identified a subset of peptide and protein hits, resulting in a hit rate ranging from 4.7-17.3%.

Of particular interest to our collaborators was one of the protein hits, pyruvate kinase PKM2, because they found hyperactive metabolism and hyperactive pyruvate kinase in InsP₈-deficient cells. However, no clear conclusions can be drawn on the thermodynamic stability of this protein due to the inconsistent data obtained with the
two different PPIP5Ks knockout cell lysates. An additional comparative analysis was performed in Experiment 2 that compared the thermodynamic stabilities of proteins in the wild type cell lysate with endogenous [InsP8] to those in the same cell lysate, albeit with excessive InsPs. The aim of Experiment 2 was to evaluate the effect of excessive InsPs (as were used in Experiment 1) on the thermodynamic stabilities of proteins (especially on pyruvate kinase). No peptide hit was identified from pyruvate kinase.
Appendix B: Identification of Protein Targets for Phosphatidylinositol 3-Phosphate in Malaria Parasites Using TMT-SPROX and TPP

Introduction

Described here is part of an ongoing collaboration between the Fitzgerald lab and the Derbyshire lab at Duke University Chemistry Department. The aim of this work was to identify potential protein targets of phosphatidylinositol 3-phosphate (PI3P, Figure S10) in malaria parasites. It has been shown that PI3P lipid binding can target malaria proteins to the host cell(141). Two different techniques were used including the TMT-SPROX and Thermal Proteome Profiling (TPP)(17) techniques.

Supplemental Figure 10: Chemical structure of D-**myo**-phosphatidylinositol 3-phosphate (PI3P).

Experimental

Malaria Parasite Lysate Samples

The malaria parasite lysate samples were provided by Kuan-Yi Lu in the Derbyshire lab. The lysis buffer was 20 mM phosphate buffer (pH 7.4) containing a
cocktail of protease inhibitors (same composition as described in Chapter 2). The cell lysis did not use any detergent as usually applied in the preparation of malaria parasite samples. Only sonication was involved in the cell lysis procedure to preserve the native structure of proteins.

**TMT-SPROX Analyses**

Experiment 1 included only one set of TMT-SPROX analysis performed on the malaria parasite lysate sample (total protein concentration 3.2 mg/mL) in the absence of any ligand. Experiment 2 included two sets of TMT-SPROX analyses performed on the malaria parasite lysate samples (total protein concentration 8 mg/mL) that were incubated with or without PI3P. Specifically, two 198 µL portions of the lysate sample were used in Experiment 2. A 22 µL aliquot of a 5 mM PI3P solution prepared in water was added to one portion of the lysate sample to generate the (+) ligand sample, and 22 µL of 5 mM PI was added to the other portion of the lysate sample to generate the (−) ligand sample. The (+) and (−) ligand samples were equilibrated for 1 h at room temperature before proceeding to TMT-SPROX analysis.

Compared to the iTRAQ-SPROX protocol as described in Appendix A, the TMT-SPROX protocol used in this work has the following improvements: (1) TCEP is used instead of methionine for H₂O₂ quenching to avoid the potential reaction between iTRAQ reagent and free methionine, (2) a filter-aided sample preparation (FASP)(142) is used to reduce the manipulation processes in SPROX experiments, and (3) TMT-10plex
is used instead of iTRAQ-8plex for quantitation because of its increased multiplex capability.

Specifically, aliquots (20 µL) of each sample were combined with 25 µL of a series of 20 mM phosphate buffers (pH 7.4) containing different concentrations of guanidine hydrochloride (GdmCl). The final GdmCl concentrations in the two experiments performed here are shown in Table S7. The final concentration of PI3P was 180 µM in Experiment 2. Each mixture was equilibrated for 1 h at room temperature, and 5 µL of 30% (w/w) hydrogen peroxide (Sigma) was added to each of the protein samples in the denaturant-containing buffers. After 3 min, the reaction in each protein sample was quenched with 500 µL of a 500 mM TCEP solution.

### Supplemental Table 7: Summary the GdmCl concentrations and corresponding TMT labeling scheme used in the 2 TMT-SPROX experiments.

<table>
<thead>
<tr>
<th>Final GdmCl concentrations (M)</th>
<th>Experiment 1</th>
<th>1.0</th>
<th>1.3</th>
<th>1.5</th>
<th>1.8</th>
<th>2.1</th>
<th>2.4</th>
<th>2.6</th>
<th>2.9</th>
<th>3.2</th>
<th>3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>TMT reporter ion m/z</td>
<td></td>
<td>126</td>
<td>127</td>
<td>127</td>
<td>128</td>
<td>128</td>
<td>129</td>
<td>129</td>
<td>130</td>
<td>130</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>C</td>
</tr>
</tbody>
</table>

The samples were then prepared for bottom-up shotgun proteomic analysis using FASP protocol as previously described(142). Briefly, protein samples were transferred into 10 Amicon Ultra-0.5 mL centrifugal filters with 10 kDa molecular weight cut off and centrifuged at 14,000 g for 30 min at room temperature. The filters were
washed twice with 200 µL of 8 M urea buffer freshly prepared in 0.1 M Tris·HCl (UA buffer, pH 8.5). Each wash was followed by centrifugation at 14,000 g for 15 min at room temperature. Disulfide bonds in proteins were reduced with 100 µL of 5 mM TCEP solution freshly prepared in UA buffer for 1 h at room temperature. Cysteine residues were then reacted with 100 µL of 20 mM MMTS solution freshly prepared in UA buffer for 10 min at room temperature before the filter units were centrifuged at 14,000 g for 15 min. The filters were washed three times with 100 µL of 100 mM TEAB buffer, followed by centrifugation at 14,000 g for 15 min at room temperature. Protein digestion was performed with 120 µL of trypsin freshly prepared in 100 mM TEAB buffer at an enzyme/substrate ratio of 1/50 (w/w) with overnight incubation at 37 °C. Tryptic peptides were collected by centrifuging the filter units at 14,000 g for 15 min at room temperature. Another 50 µL of water was added to collect the residual peptides on the filters.

The tryptic peptides were labeled with 0.5 units of TMT 10-plex reagents (Thermo Fisher) according to the manufacturer’s instruction. The labeling scheme can be found in Table S7. The labeling reaction was allowed to proceed for 1 h at room temperature before the reaction was quenched with 4 µL of 5% hydroxylamine for 15 min at room temperature. Within each set, 24 µL or 116 µL aliquots from each of the 10 different TMT labeled samples were combined to generate the non-enriched and Met-enriched samples, respectively. The combined non-enriched and Met-enriched samples
were both desalted using C18 resin (The Nest Group) according to the manufacturer’s protocol. The non-enriched samples were then subjected to LC-MS/MS analysis, while the Met-enriched samples were subjected to a methionine-containing peptide enrichment procedure using the Pi3 Methionine Selective Resins (The Nest Group) according to the manufacturer’s protocol before proceeding to LC-MS/MS analysis.

**Thermal Proteome Profiling (TPP) Analyses**

Two sets of TPP analyses were performed on the same malaria parasite lysate sample (total protein concentration 8 mg/mL) as was used in TMT-SPROX Experiment 2 that were incubated with or without PI3P. The lysate sample was diluted to a total protein concentration of 3.6 mg/mL. Two 1030 µL portions of the lysate sample were used here. A 38 µL aliquot of a 5 mM PI3P solution prepared in water was added to one portion of the lysate sample to generate the (+) ligand sample, and 38 µL of 5 mM PI was added to the other portion of the lysate sample to generate the (−) ligand sample. The final concentration of PI3P was 180 µM. The (+) and (−) ligand samples were equilibrated for 1 h at room temperature before proceeding to TPP analysis.

The TPP analysis was performed as previously described(17). Briefly, each sample was divided into 10 aliquots of 100 µL, and transferred into 0.2 mL polymerase chain reaction (PCR) tubes (VWR International). Each time one aliquot from the (+) ligand sample and one aliquot from the (−) ligand sample were heated in parallel for 3 min to the respective temperature (40, 43, 46, 49, 52, 55, 58, 61, 64 and 67 °C), followed by
incubation at room temperature for 3 min. The samples were then transferred into 5 mL Ultra-Clear™ centrifuge tubes (Beckman Coulter), and centrifuged at 100,000 g for 20 min at 4 °C. Subsequently, 30 µL of the supernatant in each protein mixture was transferred into 1.7 mL Axygen™ microtubes (Thermo Fisher) for bottom-up shotgun proteomic analysis.

The initial step in the bottom-up proteomic sample preparation involved combining the 30 µL of the protein supernatant with 10 µL of 1 M TEAB buffer (pH 8.5), 5 µL of 2% SDS and 60 µL of water. Disulfide bonds in proteins were reduced with 2.5 µL of 200 mM TCEP for 1 h at 60 °C. Cysteine residues were then alkylated with 5 µL of 375 mM iodoacetamide prepared in 100 mM TEAB buffer for 30 min in the dark at room temperature. The proteins were precipitated after addition of 700 µL of pre-chilled acetone and an overnight equilibration at -20 °C.

The resulting protein pellets were dissolved in 100 µL of 100 mM TEAB buffer (pH 8.5). The protein samples were digested with trypsin freshly using an enzyme/substrate ratio of 1/50 (w/w) at 37 °C with overnight incubation. The tryptic peptides were labeled with 0.5 units of TMT 10-plex reagents (Thermo Fisher) according to the manufacturer’s instruction. The samples derived from 40, 43, 46, 49, 52, 55, 58, 61, 64 and 67 °C were labeled with the 126, 127N, 127C, 128N, 128C, 129N, 129C, 130N, 130C and 131 TMT tags, respectively. The labeling reaction was allowed to proceed for 1 h at room temperature before the reaction was quenched with 4 µL of 5%
hydroxylamine for 15 min at room temperature. 23 µL aliquots from each of the 10 different TMT labeled samples were combined. The combined samples were desalted using C18 resin (The Nest Group) according to the manufacturer’s protocol before proceeding to LC-MS/MS analysis.

**LC-MS/MS Analyses**

LC-MS/MS analyses were performed at Duke Proteomics Core Facility using a Q-Exactive Plus high-resolution mass spectrometer (Thermo Scientific, Inc.) with a nano-Acquity UPLC system (Waters Corp.) and a nanoelectrospray ionization source. The instrument parameters were the same as those described in Chapter 2. The non-enriched and Met-enriched samples generated in the TMT-SPROX Experiment 1 and the samples generated in the TPP analysis were each analyzed in triplicate. The non-enriched samples generated in the TMT-SPROX Experiment 2 were each subjected to a single LC-MS/MS analysis, while the Met-enriched samples generated in the TMT-SPROX Experiment 2 were each subjected to two LC-MS/MS analyses.

The samples generated in the TPP analysis were additionally analyzed at Duke Proteomics Core Facility using an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific, Inc.). Each sample was subjected to a single LC-MS/MS analysis using MS2 method with an isolation window of 1.2 Da, a normalized collision energy (NCE) of 30 V (HCD) and 50k resolution (at m/z 200) with MS2 scan in Orbitrap (OT). Each sample was also subjected to a single LC-MS/MS analysis using Synchronous
Precursor Selection (SPS, multinotch MS3) method, which included an iontrap (IT) CID scan at 35 V NCE as well as 10-ion SPS with 65V NCE HCD fragmentation and 50k MS3 OT scan from m/z 100-500 (for reporter ions only).

Peak lists were extracted from the LC−MS/MS data and searched using Proteome Discoverer (Version 2.1.1.21) against (1) the 5425 malaria parasite proteins from P. falciparum 3D7 strain in the release 33 (2017-06-26) of the PlasmoDB Plasmodium Genomics Resource and (2) the SwissProt Homo Sapiens database (Version 2016-04-13). Searches were performed with a fixed modification of cysteine residue by MMTS and iodoacetamide for the TMT-SPROX and TPP analysis, respectively. N-termini and lysine residues modified by TMT 10-plex, oxidation on methionine residues, and deamidation on asparagine and glutamine were set as variable modifications in the search. Up to three missed tryptic cleavages after R and K were allowed. For LC−MS/MS data obtained on the Q-Exactive Plus mass spectrometer, the parameters included a 10 ppm mass tolerance window for precursor masses and 0.02 Da for fragment mass tolerance. For LC−MS/MS data obtained on the Orbitrap Fusion Lumos Tribrid mass spectrometer, the parameters included a 10 ppm mass tolerance window for precursor masses and 0.6 Da for fragment mass tolerance. Only peptide spectra with FDR <5%, isolation interference ≤30%, and ten TMT reporter ion intensities that summed to >1000 were used in subsequent analyses of the data.
TMT-SPROX Data Analysis

The TMT-SPROX data analysis was performed in the same manner as that described in Appendix A for iTRAQ-SPROX data analysis. Summarized in Table S8 are the set of ten N2 normalization factors that were generated in each SPROX analysis. Experiment 1 only involved one set of TMT-SPROX analysis, thus no peptide and protein hits were determined for Experiment 1. In Experiment 2, peptide and protein hits were selected between the (−) and (+) ligand samples with a ΔC1/2 cutoff of 0.3 M (spooled = 0.15 M) and a Diffprob cutoff of 1.0.

Supplemental Table 8: Summary of the N2 normalization factors and their standard deviations for each TMT-SRPOX analysis.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>126</th>
<th>127N</th>
<th>127C</th>
<th>128N</th>
<th>128C</th>
<th>129N</th>
<th>129C</th>
<th>130N</th>
<th>130C</th>
<th>131</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N2</td>
<td>0.77</td>
<td>0.93</td>
<td>0.96</td>
<td>1.02</td>
<td>1.04</td>
<td>1.07</td>
<td>1.07</td>
<td>1.09</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.20</td>
<td>0.37</td>
<td>0.19</td>
<td>0.18</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
<td>0.21</td>
<td>0.19</td>
</tr>
<tr>
<td>(-) ligand</td>
<td>N2</td>
<td>0.90</td>
<td>0.78</td>
<td>0.96</td>
<td>0.99</td>
<td>1.05</td>
<td>1.13</td>
<td>1.06</td>
<td>1.07</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.19</td>
<td>0.22</td>
<td>0.22</td>
<td>0.19</td>
<td>0.14</td>
<td>0.15</td>
<td>0.15</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>(+) ligand</td>
<td>N2</td>
<td>0.93</td>
<td>0.77</td>
<td>0.89</td>
<td>0.97</td>
<td>0.91</td>
<td>1.10</td>
<td>1.17</td>
<td>1.12</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.22</td>
<td>0.22</td>
<td>0.21</td>
<td>0.19</td>
<td>0.15</td>
<td>0.15</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>

TPP Data Analysis

The TPP data analysis was performed with R (https://www.r-project.org/) and the TPP R package as provided in ref (143). Briefly, all of the product ion mass spectra generated for a given protein were summed for each TMT reporter ion. Protein fold changes were computed at different temperature points relative to the protein
abundance at the lowest temperature. These fold changes represent the relative amount of non-denatured protein at the corresponding temperature. A median fold change value was determined from all the proteins identified at a given temperature point. The median fold change versus temperature data was then fitted to Equation S3. In Equation S3, plateau represents the lower horizontal asymptote of the melting curve, a and b are constants, and T is the temperature. Shown in Figure S11 are the resulting “super melting curves” for the median fold change versus temperature data in the (−) and (+) ligand samples that were obtained on the different mass spectrometers with different methods.

\[
f(T) = \frac{1 - \text{plateau}}{1 + \exp\left(-\frac{a}{T} - b\right)} + \text{plateau}
\]  

Equation S3
Supplemental Figure 11: Super melting curves generated for the median fold change versus temperature data in the (−) and (+) ligand samples (red and black, respectively) that were obtained on the (A) Q-Exactive Plus mass spectrometer, (B) Orbitrap Fusion Lumos Tribrid mass spectrometer using MS2 method and (C) Orbitrap Fusion Lumos Tribrid mass spectrometer using SPS-MS3 method.

The correction factors were calculated as the ratios between the fitted fold change value and the actual fold change value at a given temperature point (Table S9). Ultimately, the fold change values for each protein were normalized by the corresponding correction factors. The normalized fold change versus temperature data sets were fitted to Equation S3 to generate melting curves, from which the melting
points of proteins (T_m) can be determined. Only proteins with thermal denaturation data that were well-fit (R^2 > 0.8) to Equation S3 in both the (-) and (+) ligand samples were considered for hit selection, which was based on a difference analysis as described in the hit identification section in Appendix A for iTRAQ-SPROX data analysis. Hit proteins were selected as those with Diffprob < 0.1.

**Supplemental Table 9: Summary of the correction factors for each TPP analysis.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>126</th>
<th>127N</th>
<th>127C</th>
<th>128N</th>
<th>128C</th>
<th>129N</th>
<th>129C</th>
<th>130N</th>
<th>130C</th>
<th>131</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-Exactive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-)</td>
<td>1.00</td>
<td>0.94</td>
<td>0.91</td>
<td>1.04</td>
<td>1.02</td>
<td>1.07</td>
<td>0.96</td>
<td>0.94</td>
<td>0.97</td>
<td>1.08</td>
</tr>
<tr>
<td>(+)</td>
<td>1.00</td>
<td>1.03</td>
<td>0.92</td>
<td>0.96</td>
<td>1.23</td>
<td>0.98</td>
<td>1.04</td>
<td>0.96</td>
<td>1.01</td>
<td>1.08</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-)</td>
<td>1.00</td>
<td>0.94</td>
<td>0.89</td>
<td>1.00</td>
<td>1.04</td>
<td>1.08</td>
<td>0.95</td>
<td>0.94</td>
<td>0.98</td>
<td>1.08</td>
</tr>
<tr>
<td>(+)</td>
<td>1.00</td>
<td>1.02</td>
<td>0.91</td>
<td>0.94</td>
<td>1.18</td>
<td>0.95</td>
<td>1.01</td>
<td>0.93</td>
<td>0.99</td>
<td>1.07</td>
</tr>
<tr>
<td>SPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-)</td>
<td>1.00</td>
<td>0.94</td>
<td>0.88</td>
<td>0.97</td>
<td>1.08</td>
<td>1.09</td>
<td>0.90</td>
<td>0.92</td>
<td>0.99</td>
<td>1.13</td>
</tr>
<tr>
<td>(+)</td>
<td>0.99</td>
<td>1.00</td>
<td>0.90</td>
<td>0.96</td>
<td>1.17</td>
<td>0.98</td>
<td>1.03</td>
<td>0.93</td>
<td>0.98</td>
<td>1.04</td>
</tr>
</tbody>
</table>
**Results**

**TMT-SPROX**

This study was designed to identify PI3P-interacting proteins in the malaria parasite. The malaria parasite lysate was prepared by sonication in the absence of any detergent. In Experiment 1, only one set of TMT-SPROX analysis was performed without using any ligand. The aim of Experiment 1 was to evaluate the efficiency of this lysate preparation protocol (i.e., proteomic coverage) as well as its ability to preserve the native structure of proteins (i.e., to produce SPROX denaturation curves). In Experiment 2, two sets of TMT-SPROX analyses were performed on the malaria parasite lysate samples that were incubated with or without the ligand. The aim of Experiment 2 was to identify proteins with ligand-induced stability changes. The general workflow employed in these TMT-SPROX experiments is shown in Figure S12.
Supplemental Figure 12: Schematic representation of the TMT-SPROX strategy applied in the detection of protein targets of PI3P in malaria parasite.
The proteomic coverage obtained in this work is summarized in Table S10. The coverage only included peptides with high quality chemical denaturation data (FDR <5%, isolation interference \( \leq 30\% \), no missing TMT reporter ion intensity and their sum intensity >1000). In total, approximately 1200-1600 malaria proteins were identified with 5800-8200 unique peptides, and 235-550 human proteins were identified with 843-1629 unique peptides. Among these identifications is a subset of about 1000-1600 methionine-containing peptides derived from 400-600 malaria proteins, and a subset of about 160-400 methionine-containing peptides derived from 60-200 human proteins. Fitting the chemical denaturation data from these methionine-containing peptides to Equation S1 resulted in \( \sim 70\% \) of high quality data (\( R^2 >0.8 \)). This is similar to those observed in other SPROX experiments, indicating that the lysate preparation protocol used here did not greatly perturb the native structures of proteins. However, as can be seen from Table S10, the proteomic coverage for the Met-enriched samples in all of the three TMT-SPROX analyses was unexpectedly low. Reasons causing this low proteomic coverage remain to be identified.
Supplemental Table 10: Summary of the proteomic data obtained in the TMT-SPROX experiments for the detection of protein targets of PI3P.

**Plasmodium Falciparum 3D7**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Identified peptides (proteins)</th>
<th>Non-enriched</th>
<th>Met-enriched</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identified peptides (proteins)</td>
<td>7603 (1554)</td>
<td>653 (378)</td>
<td>8250 (1595)</td>
</tr>
<tr>
<td></td>
<td>Met peptides (proteins)</td>
<td>1211 (483)</td>
<td>603 (363)</td>
<td>1654 (615)</td>
</tr>
<tr>
<td>2</td>
<td>Identified peptides (-) Ligand</td>
<td>5590 (1146)</td>
<td>292 (183)</td>
<td>5830 (1166)</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand (proteins)</td>
<td>5647 (1174)</td>
<td>595 (330)</td>
<td>6123 (1208)</td>
</tr>
<tr>
<td></td>
<td>Met peptides (-) Ligand</td>
<td>842 (369)</td>
<td>254 (179)</td>
<td>1060 (429)</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand (proteins)</td>
<td>884 (364)</td>
<td>522 (321)</td>
<td>1320 (495)</td>
</tr>
</tbody>
</table>

**Homo Sapiens**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Identified peptides (proteins)</th>
<th>Non-enriched</th>
<th>Met-enriched</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identified peptides (proteins)</td>
<td>1467 (487)</td>
<td>192 (110)</td>
<td>1629 (550)</td>
</tr>
<tr>
<td></td>
<td>Met peptides (proteins)</td>
<td>318 (131)</td>
<td>156 (96)</td>
<td>402 (201)</td>
</tr>
<tr>
<td>2</td>
<td>Identified peptides (-) Ligand</td>
<td>812 (228)</td>
<td>47 (27)</td>
<td>843 (235)</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand (proteins)</td>
<td>833 (230)</td>
<td>80 (47)</td>
<td>881 (245)</td>
</tr>
<tr>
<td></td>
<td>Met peptides (-) Ligand</td>
<td>139 (50)</td>
<td>39 (24)</td>
<td>168 (61)</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand (proteins)</td>
<td>155 (52)</td>
<td>69 (44)</td>
<td>201 (75)</td>
</tr>
</tbody>
</table>

In Experiment 2, 496 Met peptides from 219 malaria proteins and 75 Met peptides from 32 human proteins were matched between the (-) and (+) ligand samples such that their interactions with PI3P can be probed. A total of 7 peptides from 7 proteins (including 5 malaria proteins and 2 human proteins) were identified as hits (Figure S13). One of these 7 peptide hits showed inconsistent data (i.e., an oxidized form showing a decreasing trend as a function of chemical denaturant), thus it was excluded from the final hit list (Figure S14). One possible reason for the low hit rate in this work might be the use of GdmCl for protein denaturation. Urea should be used instead of
GdmCl when potential electrostatic interactions may affect protein-PI3P binding interactions.

Supplemental Figure 13: Volcano plot showing the thermodynamic stability measurements in the PI3P-binding study. The vertical lines represent the $S_{pooled}$ hit criteria, and the horizontal line represents the $\text{Diff}_{\text{prob}}$ hit criteria. The black and red points represent the non-hit and hit peptides, respectively.
Supplemental Figure 14: Peptide hits identified in the TMT-SPROX experiment for the detection of protein targets of PI3P. (A) Peptide hit DMNVIK from plasmepsin IV (PF3D7_1407800); (B) Peptide hit MNLWAVQK from adenosine deaminase (PF3D7_1029600); (C) Peptide hit ALDTSHTNVMAYSNCK from L-lactate
dehydrogenase (PF3D7_1324900); (D) Peptide hit ILLESNMVQR from importin alpha re-exporter (PF3D7_0932800); (E) Peptide hit NPDMANIQAYLK from glutaredoxin 1 (PF3D7_0306300); (F) Peptide hit VMALYDFQAR from spectrin alpha chain, erythrocytic 1 (P02549-1). In each case, data collected in the presence of PI3P are represented by black open circles, while data collected in the absence of PI3P are represented by red filled circles. The solid curves represent the best fit of the data to Equation S1. The black and red dashed lines indicate the respective C1/2 values. An ‘X’ represents a data point that was overlooked during fitting.
Supplemental Figure 15: Schematic representation of the TPP strategy applied in the detection of protein targets of PI3P in malaria parasite.
The experimental workflow used in the TPP analysis is shown in Figure S15. The proteomic coverage obtained in the TPP analysis is summarized in Table S11. Similar to TMT-SPROX, this coverage only considered peptides with high quality thermal denaturation data (FDR < 5%, isolation interference ≤ 30%, no missing TMT reporter ion intensity and their sum intensity > 1000).

**Supplemental Table 11: Summary of the proteomic data obtained in the TPP experiment for the detection of protein targets of PI3P.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Identified proteins (peptides)</th>
<th>Matched proteins</th>
<th>Proteins w/ high quality data (R² &gt; 0.8)</th>
<th>Protein hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-Exactive</td>
<td>(-) Ligand 1293 (6647)</td>
<td>1070</td>
<td>511</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand 1270 (6404)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion MS2</td>
<td>(-) Ligand 1060 (4822)</td>
<td>850</td>
<td>418</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand 1092 (4920)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion SPS</td>
<td>(-) Ligand 863 (3494)</td>
<td>673</td>
<td>307</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand 834 (3469)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>(-) Ligand 1543 (8294)</td>
<td>1205</td>
<td>590</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>(+) Ligand 1548 (8132)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The samples were initially analyzed on a Q-Exactive Plus mass spectrometer in triplicate. In total, over 1200 malaria proteins were identified with 6400-6600 unique peptides, which is comparable to that obtained on the non-enriched sample in the TMT-SPROX Experiment 1 (7603 peptides from 1554 malaria proteins, triplicate LC-MS/MS analyses). A total of 1070 malaria proteins were matched between the (-) and (+) ligand samples such that their interactions with PI3P can be probed. A total of 511 (48%) of the
1070 matched proteins had thermal denaturation data that were well-fit ($R^2 > 0.8$) to Equation S3 in both the (-) and (+) ligand samples. A total of 97 (19%) of the 511 proteins were determined as hits (Figure S16A). Representative hit protein is shown in Figure S17A.

Supplemental Figure 16: Volcano plots showing the thermal stability measurements based on the data collected on the (A) Q-Exactive Plus mass spectrometer, (B) Orbitrap Fusion Lumos Tribrid mass spectrometer using MS2 method, and (C) Orbitrap Fusion Lumos Tribrid mass spectrometer using MS3 method. The horizontal lines represent the Diff$_{prob}$ hit criteria. The black and red points represent the non-hit and hit proteins, respectively.
Of note is that the T_m value in the original work of TPP(17, 143) was taken as the temperature at which the value of a melting curve is 0.5, while in this work the T_m value was taken as the temperature at the inflection point of a melting curve (i.e., the temperature where the slope of a curve is the steepest). This is because we observed a number of melting curves with relatively high “plateau” (i.e., the lower horizontal asymptote of the melting curve). In such a case, the use of temperatures at which the value of a melting curve is 0.5 will be inaccurate (e.g., Figure S17B).

Supplemental Figure 17: Representative melting curves obtained in the TPP analysis are shown for (A) a hit vacuolar protein sorting–associated protein 29 (PF3D7_1406700) and (B) a non-hit translation initiation factor IF–2 (PF3D7_0607000). Data were collected on a Q-Exactive Plus mass spectrometer.

The issue of high “plateau” might result from the presence of near-isobaric ions that are isolated and fragmented together with the target precursor ions during MS2 (i.e.,
isolation interference). Reporter ions from the interference may skew the actual reporter ion intensities from the target precursor ions, thus affecting both accuracy and precision of quantification associated with isobaric labeling. Recent studies have shown that MS3 can help alleviate such interference and ratio distortion(144). To investigate this potential improvement by MS3, the samples were further analyzed on an Orbitrap Fusion Lumos Tribrid mass spectrometer using both MS2 and MS3 methods.

Using the MS2 method, over 1000 malaria proteins were identified with 4800-4900 unique peptides (Table S11). A total of 850 malaria proteins were matched between the (-) and (+) ligand samples. A total of 418 (49%) of the 850 matched proteins had thermal denaturation data that were well-fit to Equation S3 in both the (-) and (+) ligand samples. A total of 75 (18%) of the 418 proteins were determined as hits (Figure S16B). Using the MS3 method, over 800 malaria proteins were identified with ~3500 unique peptides. A total of 673 malaria proteins were matched between the (-) and (+) ligand samples. A total of 307 (46%) of the 673 matched proteins had thermal denaturation data that were well-fit to Equation S3 in both the (-) and (+) ligand samples. A total of 49 (16%) of the 307 proteins were determined as hits (Figure S16C). The overall proteomic coverage was reduced by ~20% with MS3 when compared to MS2 (Table S11).

A total of 265 proteins were assayed (i.e., $R^2 > 0.8$ in both the (-) and (+) ligand samples) in all three sets of data that were collected on the different mass spectrometers with different methods. In these 265 proteins is a subset of 77 proteins that were
identified as hit proteins in at least one data set. The overlap among these 77 protein hits is shown in Figure S18. The 6 proteins that were identified as hits in all three data sets were summarized in Table S12.

Supplemental Figure 18: Venn diagram showing the overlap among the 77 protein hits identified using data collected on the different mass spectrometers with different methods.

Supplemental Table 12: Summary of the ΔT<sub>m</sub> values measured for the 6 proteins that were identified as hits in all three data sets.

<table>
<thead>
<tr>
<th>Protein accession</th>
<th>Protein name</th>
<th>Q-Exactive</th>
<th>Fusion MS2</th>
<th>Fusion MS3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF3D7_0511800</td>
<td>inositol-3-phosphate synthase</td>
<td>3.2</td>
<td>3.2</td>
<td>2.9</td>
</tr>
<tr>
<td>PF3D7_0626800</td>
<td>pyruvate kinase</td>
<td>4.9</td>
<td>3.8</td>
<td>4.4</td>
</tr>
<tr>
<td>PF3D7_1015800</td>
<td>ribonucleoside-diphosphate reductase small chain, putative</td>
<td>5.7</td>
<td>4.1</td>
<td>4.6</td>
</tr>
<tr>
<td>PF3D7_1105000</td>
<td>histone H4</td>
<td>5.3</td>
<td>5.1</td>
<td>5.2</td>
</tr>
<tr>
<td>PF3D7_1117700</td>
<td>GTP-binding nuclear protein RAN/TC4</td>
<td>3.0</td>
<td>2.4</td>
<td>2.5</td>
</tr>
<tr>
<td>PF3D7_1408100</td>
<td>plasmepsin III</td>
<td>3.2</td>
<td>4.0</td>
<td>3.4</td>
</tr>
</tbody>
</table>
The “super melting curves” did not show a significant improvement in their lower horizontal asymptote (Figure S11) with MS3. One possible reason is that the “super melting curve” was generated using the median fold change value of all the proteins identified at a given temperature point. Thus, proteins with poor quality data ($R^2 < 0.8$) may compromise the resulting “super melting curve”. Whereas in the original work of TPP, the “super melting curve” was generated from a subset of proteins with fold changes at the seventh highest temperature point between 0.4 and 0.6, and at the 9th and 10th highest temperature points below 0.3 and 0.2, respectively (17). Indeed, when we focused on the proteins that were identified in both MS2 and MS3 with good quality data, we did observe a large fraction with improved baselines in MS3. An example is shown in Figures S19A and S19B. However, baselines for the rest of proteins did not show improvement using the MS3 method. An example is shown in Figures S19A and S19B, which is the same protein as that in Figure S17B. The issue of high “plateau” in these proteins could result from the incomplete precipitation rather than the isolation interference associated with isobaric labeling.

Figure S19 also revealed the variable influence of melting curve baselines on the determination of $T_m$. Thus, we compared the $\Delta T_m$ values for the same proteins determined by using data collected from different mass spectrometers using different methods (Figure S20). The relative poor correlation suggests the presence of large
variance among data collected from different mass spectrometers using different methods. More data are needed to further investigate this observation.

Supplemental Figure 19: Representative melting curves for prefoldin subunit 3, putative (PF3D7_0718500) and translation initiation factor IF-2 (PF3D7_0607000). Data were collected on an Orbitrap Fusion Lumos Tribrid mass spectrometer using both (A and C) MS2 and (B and D) MS3 methods.
Supplemental Figure 20: Correlation between the $\Delta T_m$ values determined by using data collected from different mass spectrometers using different methods. Pearson correlation coefficients are (A) 0.383, (B) 0.456 and (C) 0.456. The red points represent the proteins that were identified as hits in both data sets. The identity line is indicated in each plot.

**Conclusions**

The TMT-SPROX and TPP techniques were used to identify PI3P-interacting proteins in the malaria parasites. The lysate preparation protocol applied in this work...
was found to be compatible with SPROX (i.e., it was able to yield reasonable proteomic coverage and to preserve the native structure of proteins). The TMT-SPROX experiment enabled 219 malaria proteins to be probed for their interactions with PI3P using 496 Met peptides, while the TPP experiment enabled 300-500 malaria proteins to be probed for their interactions with PI3P. A total of 6 peptides from 6 proteins (including 5 malaria proteins and 1 human proteins) were identified as hits in the TMT-SPROX experiment. 16-19% of the matched malaria proteins that had good quality data were identified as hits in the TPP experiment. Further investigation of the biological significance of the identified hits in the TMT-SPROX and TPP experiments remain to be continued by other lab members.
Appendix C: Development of Data Analysis Pipelines for Hit Identification in SILAC-SPROX and SILAC-LiP Experiments

SILAC-SPROX or SILAC-LiP experiments generate large proteomic data sets that can contain over 100,000 peptide-to-spectrum matches (PSMs) using state-of-the-art mass spectrometers (e.g., Q-Exactive Plus high-resolution mass spectrometer). Thus, the data analysis demands of the SILAC-SPROX and SILAC-LiP approaches are substantial. The approaches also involve data analysis strategies that are different from conventional quantitative bottom-up proteomics analyses. Prior to the work in this thesis it typically took weeks to analyze a single data set, because the data analysis strategies initially employed in these approaches involved many manual manipulations and visual inspections of the data. There was a need for the development of automated data analysis pipelines in SILAC-SPROX and SILAC-LiP experiments. Such a pipeline would not only enable a better evaluation of the data quality, but it would also facilitate the adaptations of these approaches by other research labs. Therefore, a major part of this thesis was focused on the development of automated data analysis pipelines for hit identification in SILAC-SPROX and SILAC-LiP experiments to improve the data analysis efficiency. The established data analysis pipelines have greatly reduced the time required for analyzing a typical SILAC-SPROX or SILAC-LiP data set (from weeks to minutes). A large part of this time savings comes from an elimination of the visual
inspection steps that were previously required for data analysis. Elimination of the visual inspection steps also increased the statistical rigor of the data analysis.

**SILAC-SPROX**

The new hit identification strategy for SILAC-SPROX experiments was developed in collaboration with Mr. He Meng in the Fitzgerald lab. The initial steps both in the old (as described in ref (35)) and in the new hit selection criteria (as described in this thesis) were similar. They both involved (1) the normalization of a typical SILAC-SPROX data set (i.e., H/L ratio as a function of the denaturant concentration) by protein expression levels, (2) the selection of methionine-containing peptides that were identified from an adequate number of denaturant concentrations (only these peptides were deemed “assayed” and considered for hit selection), and (3) further selection among the assayed methionine-containing peptides for those with two or more consecutive log$_2$(normalized H/L) values that were significantly different from 0, which were considered to be those that fall below or above the 5th and 95th percentiles, respectively, of log$_2$(normalized H/L) values of all the non-methionine-containing peptides detected across all the denaturant concentrations. These methionine-containing peptides with two or more consecutive log$_2$(normalized H/L) values that were significantly different from 0 were considered to be “potential hit peptides”.

These “potential hit peptides” would become the final hit peptides in the old hit identification strategy (as described in ref (35)) if they did not show inconsistencies
among biological replicates (e.g., a peptide was identified as hit in one biological replicate but not in another, or the peptide retention times were not consistent in the LC-MS/MS analyses). Protein stability information can be subsequently determined from these final hit peptides. Compared to the old hit identification strategy, the new hit selection criteria described in this thesis included an additional regression analysis that fits the log₂(normalized H/L) value versus denaturant concentration data obtained from these “potential hit peptides” to the theoretical equations describing the chemical denaturant-induced equilibrium unfolding/refolding reactions of proteins (Equations 1 and 2, Chapter 2, page 31). The incorporation of regression analysis into SILAC-SPROX data analysis not only allows a quantitative and statistical evaluation of the data quality, but it also enables the extraction of meaningful thermodynamic information, including ΔC_{1/2} and ΔG_f, without the need for universally estimating m-values. Another difference is that the new hit identification strategy requires methionine-containing peptides to have H/L data from at least seven (instead of six in the old hit identification strategy) denaturant concentrations to be assayed for thermodynamic stability changes. The use of H/L ratio data from at least seven denaturant concentrations was deemed necessary in the new data analysis pipeline to provide an adequate number of data points for the regression analysis.

The non-linear regression analysis fits each data set to Equations 1 or 2 (Chapter 2, page 31) depending on whether a given peptide contains an oxidized or a wild type
methionine residue. It fits each data set n+1 times, once with all data points and then an additional n times, each time leaving out a different one of the n data points, where n is the total number of denaturant concentrations identified. The best regression was selected by the highest $R^2$ value. Compared to the protein level quantitation that integrates the results for all measured peptides of a given protein, the peptide level quantitation as is used in the SPROX technique is more sensitive to outliers. Thus, we applied such a leave-one-out strategy for the purpose of rescuing data sets that would otherwise be left out of the analysis and reduce the proteomic coverage. Unfortunately, such a leave-one-out strategy can potentially cause over fitting and result in large variation in the determination of $C_{1/2}$ values, especially when the data point being left out locates in the transition region. Thus, hits identified with such data points, should be carefully considered. A global analysis of the data distribution among replicates should be performed for outlier determination before its removal in the future data analysis. Another possible way is to always retain all data points during the regression analysis, albeit at the expense of proteomic coverage.

To evaluate the consistency of hit identification based on the regression analysis used in the new hit selection criteria and the visual inspection of data used in ref (35), the new hit identification strategy was applied to the MS data reported in the previous SILAC-SPROX analyses of the MCF-10A, MCF-7, and MDA-MB-231 breast cancer cell
lines (35). A large overlap (46-52%) was obtained between the peptide hits identified using the new hit identification strategy and those reported in the literature (Table S13).

The 66 and 69 peptide hits in the MCF-7 versus MCF-10A and MCF-7 versus MDA-MB-231 cell line comparisons (respectively), that were reported in ref (35) but were not identified as hits with the new hit identification strategy, included 20 and 28 peptides (respectively) that did not have sufficient data points for the regression analysis, and another 44 and 41 peptides (respectively) with sufficient data points for regression analysis yet with poor quality data (examples shown in Figure S21). Whether these previously reported peptide hits not being identified with the new hit identification strategy indicate false positives is an open question, especially for the 20 and 28 peptides that did not have sufficient data points for regression analysis. Shown in Figure S22 are representative peptide hits that were only identified with the new hit selection criteria.
Supplemental Table 13: Comparison of hit identification. Proteomic coverage for visual inspection was from ref (35). Proteomic coverage for regression analysis was obtained by applying the hit identification strategy described in this thesis to the data in ref (35). Assayed peptides are the methionine-containing peptides that were identified from at least six (for visual inspection) or seven (for regression analysis) denaturant concentrations in the SILAC-SPROX experiment.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>MCF-7 versus MCF-10A</th>
<th>MCF-7 versus MDA-MB-231</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visual Inspection</td>
<td>Regression Analysis</td>
</tr>
<tr>
<td>Assayed Peptides (Proteins)</td>
<td>2472 (772)</td>
<td>2053 (620)</td>
</tr>
<tr>
<td>Hit Peptides (Proteins)</td>
<td>138 (84)</td>
<td>208 (129)</td>
</tr>
<tr>
<td>Overlap Hit Peptides (Proteins)</td>
<td>72 (57)</td>
<td></td>
</tr>
<tr>
<td>% Overlap Hit Peptides (Overlap Hits/ Literature Hits)</td>
<td>52%</td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Figure 21: Representative peptide hits in the MCF-7 versus MCF-10A cell line comparison that were reported in ref (35) but were not identified as hits with the new hit selection criteria due to poor quality data. The red solid lines
represent the best fit of the data to Equations 1 or 2; the green curves represent the extracted SPROX curves; the vertical dashed lines indicate the transition midpoints of the SPROX curves; the horizontal dashed lines indicate the cutoffs for significant changes in \( \log_2(H/L) \) values. A red × indicates a data point that was not included in the regression analysis.

Supplemental Figure 22: Representative peptide hits in the MCF-7 versus MCF-10A cell line comparison that were only identified with the new hit selection criteria. The red solid lines represent the best fit of the data to Equations 1 or 2; the green curves represent the extracted SPROX curves; the vertical dashed lines indicate the transition midpoints of the SPROX curves; the horizontal dashed lines indicate the cutoffs for significant changes in \( \log_2(H/L) \) values. A red × indicates a data point that was not included in the regression analysis.
It is important to note that although the regression analysis offers a way to quantitatively score the data quality, the hit selection thresholds are crucial for controlling the assay stringency. For example, hits and non-hits that are close to the selected threshold should be carefully considered (see e.g., Figure S21E and Figure S22A). One attractive feature of the new data analysis pipeline is that the hit selection threshold can be adjusted to be more or less stringent.

**SILAC-LiP**

The work in Chapter 3 and 4 represents the first application of SILAC-based quantitation to the limited proteolysis technique. Hit peptides with significantly different H/L ratios in the doubly and singly digested samples were identified by performing a Student’s t-test on the H/L ratios obtained from multiple biological replicates. The use of a Student’s t-test for hit identification was similar to that in the original demonstration of the LiP technique (14). However, the original LiP study used a label-free quantitation strategy (i.e., spectral counting) that is more prone to quantitative errors. Thus, in the original LiP study, an additional spectral count fold change cutoff (> 2-fold) was used together with a p-value cutoff (< 0.05) for hit selection (14). As part of the work described in Chapters 3 and 4, we found the higher accuracy and precision associated with the SILAC quantitation rendered the fold change cut-off unnecessary as a selection criteria (see Chapter 3). Thus, the SILAC-LiP approach has the potential to reveal smaller protease susceptibility differences of proteins.
References


199


Biography

Fang Liu was born on October 7, 1990 in Tianjin, China. She attended University of Science and Technology of China (Hefei, China) from 2009-2013 and graduated in July 2013 with her Bachelor of Science degree in Chemistry. She attended graduate school at Duke University (Durham, NC) in August 2013 and worked as a graduate assistant in the lab of Dr. Michael C. Fitzgerald. She is expected to graduate with a Doctor of Philosophy degree in Chemistry and a Master of Science degree in Computer Science in the spring of 2018. While attending Duke University, she was awarded the Shaffer-Hunnicutt Fellowship in 2017. She has been a member of the American Society for Mass Spectrometry and the American Chemical Society from 2014-present.

Publications


