Systematic proteomic strategies

to map the human signaling landscape

Robert T. Lawrence

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Judit Villén, Chair

Elhanan Borenstein

Stanley Fields

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Systematic proteomic strategies to map the human signaling landscape

Robert T. Lawrence

Chair of the Supervisory Committee:

Assistant Professor Judit Villén, PhD

Department of Genome Sciences

Signal transduction is the process by which cells continuously sense and integrate environmental cues in order to make real-time decisions, e.g. how to direct metabolic flux or whether to enter mitosis. Constructing detailed maps of cellular signaling networks has proved a valuable way to summarize knowledge, formulate new hypotheses, and devise pharmacological control strategies for a myriad of human diseases. However, despite remarkable progress in this area, only recently have we begun to appreciate the true vastness and diversity of signaling landscapes. The capability to measure molecular systems at near genome scale represents a major paradigm shift for biology, and mass spectrometry (MS) approaches can now provide multiplexed quantitative measurements of thousands of cellular proteins and phosphorylation events in a single sample.

In this dissertation, I use MS-proteomics to study protein networks in cultured human cell lines. First, I describe an integrative proteomic analysis of twenty breast cancer cell lines. In this work I show that protein expression varies dramatically across cells derived from the same tissue type and across several canonical signaling pathways. I identify distinct patterns of protein expression that are found in triple negative breast cancer cells compared to luminal breast cancer. Further, I

suggest that genetic aberrations and protein expression are interconnected, and together they affect the responsiveness of cells to cancer therapeutics. Next, I used mass spectrometry to investigate phosphorylation-dependent signaling networks. I first performed a deep characterization of protein phosphorylation events in HeLa cells exposed to sixteen different stimuli (e.g. epidermal growth factor, tumor necrosis factor-alpha, osmotic stress), quantifying more than one hundred thousand phosphorylation sites. In this study I provide a detailed view of the vast signaling landscape present within an individual cell type and reveal the extent of regulatory cross-talk, whereby the same phosphorylation sites and proteins are regulated by multiple stimuli. To further dissect the topology, dynamics, and cross-talk of pathways frequently mutated in cancer I performed additional experiments using time-courses and systematic kinase inhibition against several key signaling nodes in the presence of growth factors and cellular stress in HeLa and MCF7 cells. Finally, I use the data collected during these experiments to evaluate and address several technical limitations of mass spectrometry-based phosphoproteomics analysis. I discuss a method to mine these large-scale data to rapidly generate targeted MS assays which now enable versatile, high-throughput, sensitive, and reproducible analysis of cellular signal transduction networks.

TABLE OF CONTENTS

LIST OF FIGURES

DEDICATION

In loving memory of my father, Dr. John C. Lawrence, Jr (1949-2006).

A scientist with the highest standards for rigor and integrity.

ACKNOWLEDGEMENTS

I am deeply indebted for the outstanding support and mentorship offered to me before and during my graduate studies. All of my success has depended upon individuals who have given me the opportunities and encouragement to continually challenge and improve my skills as an investigator.

I would not be here if it were not for my enlightening experiences as an undergraduate in Tom Skalak's lab at the University of Virginia and two years of postgraduate research. I particularly want to thank David James, who helped me obtain a fellowship to spend a year in his lab at the Garvan Institute in Australia. The passion and ambition of his group for pushing scientific boundaries by embracing new technology was contagious and a major reason I decided to pursue a graduate degree. Kyle Hoehn taught me how to work in the lab, how to identify testable scientific questions, and to design and execute complex experiments. These early experiences were invaluable throughout my graduate studies.

I would like to acknowledge the large network of individuals that have supported me during my time at the University of Washington. I had the privilege to work with excellent collaborators including Tony Blau, Thurl Harris, David Brautigan, Kyle Hoehn, Mike Czech and their respective lab members on many exciting biological questions. I want to thank the Molecular and Cellular Biology program, current and former program directors and administrative staff, and the Genome Sciences department for providing an incredibly supportive and well-organized training environment. I have made many great friends and colleagues during my time here. The 2011 MCB class, particularly Joe and Monica Sanchez who also joined the Genome Sciences department, and my former housemates Carissa Pilling, Jackie Lang, and Sergey Ovchinnikov; all have been great friends and have given me a valuable interdisciplinary perspective. I am also

grateful for mentorship from senior graduate students and postdocs; specifically Nate Peters, Roie Levy, Jarrett Egertson, and Jim Bollinger who have been role models for me.

I would also like to thank my committee. Mike MacCoss, John Scott, Elhanan Borenstein, and Stanley Fields, and Rich Gardner made critical contributions that have greatly enhanced the quality of my work. Thanks for the rigorous criticisms, which led my research in positive and unexpected directions.

The Villén lab has been an incredible interdisciplinary research environment. First I would like to thank Billy Edelman who has been a close friend since I joined the lab. It has been fun traveling through graduate school together. Danielle Swaney taught me the basic techniques I used most extensively in the lab as well and guided me through many technical aspects of mass spectrometry. Daniel Hernandez, Ariadna Llovet, and Brian Searle were collaborators on this work and I am indebted to them for their assistance. Sam Entwisle, Ricard Rodriguez, Miguel Martin have contributed with many helpful discussions. My advisor, Judit Villén, has been a major source of inspiration for me from the beginning. She is a fearless scientist, something I hope has rubbed off on me over the years. She gave me freedom to explore a medley of disparately related projects, to develop my own scientific style, and to make important decisions about how to focus my efforts. Training in Judit's lab has been an exhilarating journey to say the least, and I cannot imagine a better place to have started my career.

Lastly I would like to thank my family. My father, for whom this work is dedicated, instilled me with a curiosity for the world and its endless mysteries. I thank my brothers Matthew and Justin for many adventures and perseverance through difficult times, and my mother for her outright exuberance for life. Above all, I thank my fiancé, Devon McCurdy, for her unconditional love and support. Thanks for helping me remember the important things.

Chapter 1. Paradigms in signal transduction research

1.1 History of protein regulation and signal transduction

1.1.1 Discovery of reversible protein regulation

The first half of the 20th century was a time of rapid discovery in the field of metabolic biochemistry. Mechanisms for storing and releasing chemical energy via the hydrolysis of adenosine triphosphate (ATP), the means of generating reducing equivalents through glycolysis and the citric acid cycle, and the catalytic cycle of glycogen synthesis and degradation were meticulously elucidated. The latter process, now known as the "Cori cycle," was the basis of a Nobel prize awarded to Carl and Gerty Cori in 1947. Unknown at the time, the story of signal transduction began with their observation in the late 1920's that insulin and epinephrine have antagonistic effects on the utilization of glucose by peripheral tissues (Cori and Cori, 1928). Later, their extensive functional characterization of the enzyme glycogen phosphorylase would lead to the concepts of allostery and post-translational modification which are the basis of cellular signal transduction (Cohen, 2002a; Graves and Krebs, 1999). Despite nearly a century of work, the question of precisely how these hormones achieve metabolic control remains a highly active area of signal transduction research.

Glycogen phosphorylase catalyzes the reversible conversion of glucose-1-phosphate to $glycogen_{(n+1)}$ and inorganic phosphate. When purifying the enzyme from rabbit skeletal muscle, two forms could be isolated, *a* and *b* (Cori and Green, 1943). Phosphorylase *a* was highly active but could be de-activated if left at room temperature in a crude lysate prior to purification. It contained a phosphorous-enriched "prosthetic group" that was cleaved by a "prosthetic group removing enzyme" resulting in conversion to phosphorylase *b*. Phosphorylase *b* was only active in the presence of 5'-adenylic acid (AMP), and although it was not recognized as a general principle at the time, this was the first demonstration of allosteric protein regulation. In animals,

the reverse conversion of phosphorylase *b* to the active *a* was stimulated by epinephrine (Sutherland and Cori, 1951), but the mechanism was still fleeting until it was shown that the *in vitro* conversion of phosphorylase *b* to *a* required the presence of crude protein lysate, divalent metal cations, and ATP (Fischer and Krebs, 1955). It was then demonstrated using ^{32}P radioactive labeling that the reaction resulted in the direct transfer of phosphorous from ATP to the phosphorylase enzyme (Fischer et al., 1959). The prosthetic group turned out to be a covalently bound phosphate and the prosthetic group removing enzyme is now called a protein phosphatase. The enzyme responsible for activating phosphorylase *b* by phosphate incorporation was then called a protein kinase (Krebs and Fischer, 1956). While unappreciated for many years (perhaps still so), the demonstration of a reversible, enzymatic mechanism to control protein function represents a major paradigm shift in cell biology. This mode of regulation allows cells to simultaneously express proteins with antagonistic functions, and to respond rapidly to the environment without altering protein expression.

1.1.2 Early approaches to signaling pathway characterization

The development of the *in vitro* system for studying protein phosphorylation paved the way for the next wave of signaling research. By the latter half of the 20th century, many hormones and extracellular signaling molecules had been discovered, yet similar to epinephrine and insulin, the molecular mechanisms of signal transduction remained uncharacterized. By quantifying the incorporation or release of phosphorous from proteins in hormone-treated tissue extracts and identifying which small molecules were necessary for those reactions to occur, it has been feasible to map regulatory pathways from intracellular effector proteins to receptors on the cell surface. A ‘top-down’ discovery approach was often used to identify new effector molecules and phosphorylation mechanisms by looking at ^{32}P incorporation either proteome-wide or in subcellular fractions. For example, treatment of 3T3-L1 adipocytes with insulin resulted in rapid phosphate incorporation into many cellular proteins as resolved by two-dimensional gel

electrophoresis (Smith et al., 1979). The insulin receptor tyrosine kinase was subsequently identified by the same group by examining ^{32}P incorporation in the membrane fraction of the same experiments (Petruzzelli et al., 1982). A similar approach had been used to identify the epidermal growth factor (EGF) receptor tyrosine kinase activity in A-431 cell membranes (Carpenter et al., 1978).

A 'bottom-up' approach was often used to isolate the upstream regulators of a purified protein, and navigate back up the pathway towards the receptor. For example, microtubule associated protein kinase (MAPK) was initially characterized by incubating purified mictrotubule associated protein 2 (MAP-2) or ribosomal protein S6 with different fractions of insulin-treated 3T3-L1 cell extracts (Ray and Sturgill, 1987). Compared to S6, MAP-2 was phosphorylated by 3T3-L1 lysate collected in earlier time points after insulin stimulation and in different fractions of protein separated by phosphocellulose columns. One of the limitations of this strategy is that it is very often confounded by the presence of multiple signal transduction pathways (Lawrence, 1992). These general strategies of phosphorylation site discovery and characterization are still used routinely. The development of highly specific kinase inhibitors has provided an additional tool for characterizing signal transduction pathways (Cohen, 2002b). The catalog of commercially and clinically available kinase inhibitors has expanded dramatically in recent years.

1.1.3 Analytical techniques for signal transduction

A major drawback of this early work, particularly using a 'top-down' approach, was the time and material required to precisely identify proteins and determine the specific position phosphorylated within the proteins. It was very difficult to pinpoint the identity of the protein by any property other than its rough molecular weight and source. Protein sequencing was carried out by Edman degradation, which required approximately 100 grams of starting material and >1 year of work to characterize a single site of phosphorylation. These sequences enabled the development of

protein and phosphorylation site-specific antibodies. Once antibodies had been developed it became possible to detect and quantify phosphoproteins in many different treatments, tissue types, and organisms. However, the quality and availability of phospho-specific antibodies remains limited.

Mass spectrometry (MS) based approaches to protein characterization began to evolve rapidly around the turn of the 21st century (Aebersold and Mann, 2003). Essentially, a mass spectrometer measures the mass-to-charge ratio (m/z) and intensity of a range of ions as they are introduced through an ion source, usually emitted from a liquid chromatography column, enabling simultaneous identification and quantification. Proteins are first digested to peptides using high-fidelity proteases such as trypsin, which specifically cleaves proteins C-terminal to lysine and arginine. Mass spectra are typically acquired in tandem (MS/MS). First the mass spectrometer scans the m/z spectrum of the intact peptide ions (precursors). Next, precursor ions from this scan are fragmented against a collision gas and the m/z spectrum of the fragment ions (products) is measured. The m/z of the precursor ion and product ions are searched against a database of theoretical sequences for identification. The presence of phosphorylation results in a mass increase of 80 Da for all precursors and fragments containing the phosphate modification.

The recent success of MS-based approaches can be attributed to advances in sample preparation, instrumentation, genome sequencing, and statistical algorithms. In the last decade mass accuracy has improved approximately 1000-fold from +/- 0.5 Da to low ppm (+/- 0.001 Da), resulting in a greater percentage of identifiable spectra. Scanning speed has increased approximately 10-fold from 1-2 to 10-20 scans per second, resulting in more unique peptide spectra acquired. The sequencing of the human genome provided a detailed map of its protein-coding regions, and a comprehensive database from which to generate theoretical m/z spectra. Statistical algorithms for confident peptide identification and phosphorylation site localization were introduced (Eng et al., 1994; Elias and Gygi, 2007; Beausoleil et al., 2006). Methods to enrich

phosphorylated peptides with high yield and specificity were developed (Ficarro et al., 2002; Villén and Gygi, 2008). It is now feasible to accurately identify and measure more than 5,000 unique phosphorylated peptides per hour from less than 1 mg of tissue input, a number I expect to increase dramatically with the next iteration of technological improvements. Compared to techniques used in the latter half of the 20th century, this represents roughly a million fold reduction in sample input requirements and roughly a 50 million fold increase in throughput. These advances have revitalized the 'top-down' approach to signal transduction research.

1.2 Current challenges

Even in the early days, it was recognized in 2-dimensional gels that approximately 50% of proteins were phosphorylated (Pinna and Ruzzene, 1996) and approximately 100 phosphorylation sites had been characterized by the early 1990's (Roach, 1991). Using mass spectrometry, more than 100,000 phosphorylation sites have been characterized and more than half of human proteins are phosphorylated (Hornbeck et al., 2015; Lawrence et al., 2016). The initial observation of predicted proteins and discovery of new phosphorylation sites continues, but has begun to plateau. The "fact-finding" mission for protein phosphorylation (Graves and Krebs, 1999) is now coming to a close. One of the grand challenges that now lies ahead involves understanding how signal transduction systems orchestrate the function of thousands of cellular proteins to drive a multitude of complex cellular behaviors, and how the systems are broken in diseases such as cancer.

Phosphorylation sites, proteins, and processes are at the crossroads of multiple signal transduction pathways, and they are tuned to encode appropriate cell type-specific responses to stimuli (Cohen, 1992). Global quantification of proteins and phosphorylation events is an ideal way to observe and de-convolute this complexity. Many experimental frameworks have emerged to interrogate the various dimensions of the dynamic proteome, only a few of which I will highlight briefly here. One approach is to survey protein expression and phosphorylation across panels of

different cell lines, tissues, or species. Several of the most widely cited proteomics studies to date used this design (Huttlin et al., 2010; Kim et al., 2014; Lundby et al., 2012). It is a powerful way to understand what makes different tissues unique versus what they have in common. A limitation is that these studies are often just a snapshot of a dynamic living system, harvested in a single context (e.g. after fasting or at steady state growth).

Other studies have examined dynamic events in individual cell types, providing mechanistic insight into precisely how cells respond to certain stimuli. In this framework the proteome is typically measured after cells are stimulated with hormones or growth factors over a time course or with systematic perturbations to known regulatory hubs. Time course experiments in up to 9 time points have been effectively used in conjunction with proteomics to analyze the cellular response to many stimuli including epidermal growth factor, prostaglandin E, insulin, or nocodazole release (Olsen et al., 2006; de Graaf et al., 2014; Humphrey et al., 2013; Olsen et al., 2010). Temporal modeling is an ideal strategy to identify time-dependent properties and processes such as negative feedback mechanisms or the cell cycle. Systematic perturbation using kinase inhibitors has been used successfully for the analysis of cross-talk between different receptor tyrosine kinase networks and for inferring kinase-substrate relationships (Moritz et al., 2010; Terfve et al., 2015). Ligand dose-response curves have also been useful to determine the specific sensitivity and dynamic range of different nodes in cellular signal transduction networks (Hoehn et al., 2008), but have not yet been widely adopted in proteomics experiments.

1.3 Research Aims

In this body of work, I describe experiments performed in human cancer cell lines directed towards three general areas that systems biology approaches are beginning to answer. I also describe a method that helps bridge the gap between discovery and quantitative proteomics strategies, addressing some limitations of both.

1.3.1 The proteomic diversity of human breast cancer cells

Many maps of human signaling networks have been proposed, but how generalizable are these maps between different cell types? If cells express major differences in the abundance of signaling receptors, transducers, and effectors, then they are likely to elicit different responses to the same stimuli and would not have the same sensitivity to cancer treatments. I hypothesize that harnessing proteomic diversity will be key to personalized, precision cancer treatment. In Chapter 2 (Lawrence et al., 2015), I discuss the proteomic analysis of twenty breast cell lines, relating these measurements to genetic alterations, previously established subtypes, and drug sensitivity. I demonstrate the pronounced variability of signaling proteins and their effectors in cells of similar tissue of origin.

1.3.2 Towards a comprehensive map of the human cellular phosphoproteome

What is the extent of phosphorylation dependent signaling in an individual cell type? Hundreds of cellular stimuli and environmental perturbations have been described to date. While most of these have distinct molecular sensors and receptors, they are thought to connect to common signaling modules inside the cell for signal transmission. In the first part of Chapter 3, I discuss a deep characterization of protein phosphorylation in HeLa cells exposed to sixteen different stimuli selected to activate the full spectrum of signaling capabilities in human cells. These experiments reveal extensive signal integration at the level of individual phosphorylation sites. What are the mechanisms of signal integration, and how do the dynamics of signal transmission differ between stimuli? In the second part of Chapter 3, I discuss the results of experiments designed to systematically disrupt signal transduction through pathways commonly deregulated in cancer. In the third part of Chapter 3, I focus specifically on cross-talk between the EGF and IGF-1 pathways in MCF7 breast cancer cells using a panel of kinase inhibitors, time-course experiments, and time-staggered combinatorial stimulation with a phosphoproteomics readout. These experiments provide a wealth of data on cellular protein phosphorylation mechanisms.

1.3.3 Limitations of mass spectrometry based analysis

Challenges remain in the comprehensive analysis of signal transduction networks by mass spectrometry. In particular, data sparsity is a common problem in proteomics. Data acquisition strategies commonly used for phosphoproteomics are untargeted, so different sets of phosphorylated peptides are measured in each experiment, resulting in many missing values between experiments. Targeted data acquisition solves this problem, because the same peptides are deliberately measured in each experiment. However, deciding which phosphopeptide ions to target has been challenging. In Chapter 4, I describe a method to utilize aggregate previous knowledge to rapidly design customized signal transduction assays and demonstrate the utility of this approach.

Chapter 2. Diversity and specificity of cellular proteomes

This chapter is based on the following published article:

Robert T Lawrence, Elizabeth M Perez, Daniel Hernández, Chris P Miller, Kelsey M Haas, Hanna Y Irie, Su-In Lee, C. Anthony Blau, and Judit Villén. The proteomic landscape of triple-negative breast cancer. Cell Reports. 2015;11(4):630-44.

2.1 Summary

Triple-negative breast cancer is a heterogeneous disease characterized by poor clinical outcomes and a shortage of targeted treatment options. To discover molecular features of triple-negative breast cancer, we performed quantitative proteomics analysis of twenty human-derived breast cell lines and four primary breast tumors to a depth of more than 12,000 distinct proteins. We used this data to identify breast cancer subtypes at the protein level and demonstrate the precise quantification of biomarkers, signaling proteins, and biological pathways by mass spectrometry. We integrated proteomics data with exome sequence resources to identify genomic aberrations that affect protein expression. We performed a high-throughput drug screen to identify protein markers of drug sensitivity and understand the mechanisms of drug resistance. The genome and proteome provide complementary information that, when combined, provide a powerful engine for therapeutic discovery. This resource is available to the cancer research community to catalyze further analysis and investigation.

2.2 Introduction

A key challenge for medicine in the twenty-first century is to harness the predictive power of molecular data to eradicate cancer (Arteaga and Baselga, 2012; Vidal et al., 2012; Weinstein et al., 1997). Like other cancers, breast cancer is caused by a series of inherited and/or acquired

genetic aberrations that eventually lead to uncontrolled cell proliferation and metastasis. The diverse genetic “drivers” of breast cancer have been characterized in exquisite detail (Banerji et al., 2012; Curtis et al., 2012; Perou et al., 2000; Prat and Perou, 2011; The Cancer Genome Atlas Network, 2012; Vogelstein et al., 2013). However, characterization of the proteome has lagged behind.

At the functional level, relevant genomic aberrations affect cellular functions by altering the activity and abundance of proteins. These effects are context specific and very much depend on the unique catalog of proteins expressed by different cell types. For example, a mutation in the BRAF kinase might have different functional outcomes in skin cancer than in liver or breast cancer. In addition to driving cellular functions, proteins are the most actionable and druggable cellular components. Therefore, protein measurements are important to understand breast cancer and delineate breast cancer therapies.

In fact, protein measurements are being used today to classify breast cancer types according to their receptor status, in which the presence or absence of three cellular receptors (estrogen receptor ESR1, progesterone receptor PGR, and human epidermal growth factor receptor-2 ERBB2) is assessed via immunohistochemistry. Despite the reduced number of molecular features measured, this classification is the most useful today for chemotherapy selection. Irrespective of genomic aberrations, more than 80% of breast cancers express one or more of these receptors (Howlader et al., 2014) and are treatable by hormone deprivation and/or ERBB2 inhibition (Untch et al., 2014). Targeted therapies are not currently available for tumors that do not express these receptors, which are collectively referred to as triple-negative breast cancer (TNBC). TNBC is an important and unmet clinical problem. It tends to be more aggressive, is correlated with worse prognosis than receptor-positive subtypes (Hudis and Gianni, 2011), and is more common among young and African American women (Howlader et al., 2014). Identifying

subtypes within the TNBC type, and proteins within those subtypes that can serve as therapeutic targets will be extremely valuable.

Among protein measurements, reverse-phase protein arrays (RPPA) have been one the most widely adopted tools for integrated genomics and drug sensitivity analysis, but a key limitation of RPPA technology is its lack proteome coverage, generally less than two hundred analytes (Tibes et al., 2006). As such, mRNA expression has been used as a proxy for protein levels, despite mediocre quantitative concordance (Gygi et al., 1999; Maier et al., 2009). Both mRNA and protein expression using RPPA outperform genomic data as predictors of drug sensitivity and clinical outcomes (Costello et al., 2014; Yuan et al., 2014). These results highlight the potential of systematic protein expression analyses for breast cancer research in general and drug discovery in particular.

It is an excellent time to further investigate the triple-negative breast cancer proteome using more comprehensive techniques. Mass spectrometry in the form of “shotgun proteomics” is highly quantitative, and has reached the speed and sensitivity to measure proteomes at a depth comparable to gene expression studies (Kim et al., 2014; Wilhelm et al., 2014). In fact, proteomics is already making an impact in breast cancer research (Geiger et al., 2012a; Gholami et al., 2013; Kennedy et al., 2014), but yet, to show its full potential, proteomics needs to be integrated with other types of big data.

Here we present an integrative approach using quantitative mass spectrometry to characterize TNBC proteomes both as readouts of genetic abnormality and as predictors of drug sensitivity. The goal of this work is to refine our understanding of breast cancer biology as an integrated ‘proteogenomic’ landscape and to identify molecular diagnostic markers to improve drug selection in triple-negative breast cancer.

2.3 Results

2.3.1 The triple-negative breast cancer proteome

We assembled a panel of twenty human breast cell lines and four clinical tumors to analyze the proteomic landscape of TNBC (Figure 2.1a). These included 16 triple-negative cell lines covering mesenchymal, luminal, and basal-like subtypes, as well as 3 receptor-positive and 1 non-tumorigenic cell line to serve as a basis for comparison (Lehmann et al., 2011; Neve et al., 2006). Primary tumor tissues were derived from patients with metastatic triple-negative breast cancer (stage II-III). Cell lines were cultured and analyzed in duplicate to assess the precision of protein quantification. Proteins were digested in parallel with either lysyl-endopeptidase (LysC) or trypsin and separated at the peptide-level into five fractions to enhance proteome coverage (Figure 2.1b). We used liquid chromatography tandem mass spectrometry (LC-MS/MS) on a hybrid quadrupole-orbitrap mass spectrometer to acquire quantitative profiles of the peptides present in each fraction.

In total, more than 450 peptide fractions were analyzed, yielding approximately 20 million high-resolution mass spectra. Across the entire dataset, we identified 289,819 non-redundant peptide sequences mapping to at least 12,775 distinct proteins encoded by 11,466 genes (protein FDR <1%). To facilitate comparison of specific protein isoforms, we additionally retained in our data truncated protein isoforms having high sequence coverage, bringing the total proteins analyzed to 15,524. The median protein had 15 peptide matches, 4 isoform-specific peptide matches, and shared peptides with only one other protein in the dataset (Figure A.2.1). Median protein sequence coverage was 52%.

The number of proteins identified was consistent across cell lines, tissues, and replicates. On average, 80% of proteins were identified in both replicates. At least 9,000 proteins were found in

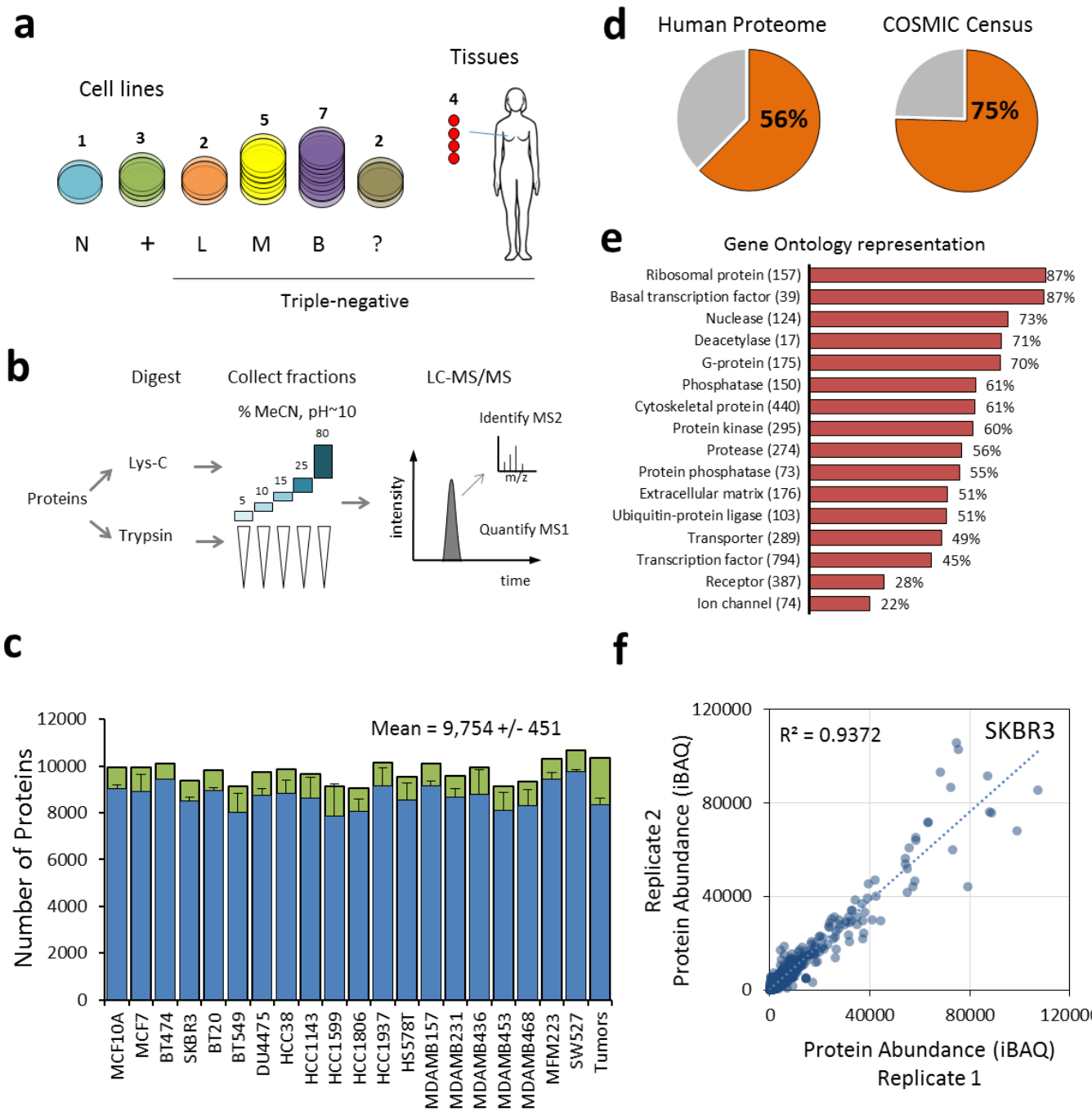


Figure 2.1 Mass spectrometry-based profiling of triple-negative breast cancer

(a) Overview of samples analyzed. N: normal epithelial, +: ER/PR/ERBB2+, L: luminal-like, M: mesenchymal-like, B: basal-like, ?: not matched. TNBC cell line classifications according to (Lehmann et al., 2011) **(b)** Workflow of proteomics sample preparation and data collection. **(c)** Average number of proteins identified in each replicate (blue bars), total number of proteins for each cell line (green bars). Error bars represent S.D. **(d)** Percent of identified proteins relative to the Uniprot/Swiss-Prot database (left) and the COSMIC census (right). **(e)** Number and percent representation of indicated gene ontology categories. **(f)** Representative scatter plot for cell line SKBR3 replicate protein measurements showing quantitative reproducibility of iBAQ protein abundance.

each cell line (Figure 2.1c), which agrees well with other recent deep proteome experiments (Beck et al., 2011; Geiger et al., 2012b; Gholami et al., 2013; Nagaraj et al., 2011). These proteins represent 56% of the 20,537 genes annotated in Uniprot/Swiss-Prot and at least 75% of genes included in the catalog of somatic mutations in cancer (COSMIC) (Figure 2.1d). As expected, we achieved near complete coverage of gene ontology categories involved in core cellular functions such as primary metabolism, protein synthesis, and general transcription, and lower coverage of tissue-specific categories such as transcription factors and receptors (Figure 2.1e).

To infer protein absolute abundances we used the intensity-based approach for absolute quantitation (iBAQ). Quantitative reproducibility between biological replicates was uniformly high across all cell lines, with an average R^2 equal to 0.92 (Figure 2.1f, Figure A.2.1). Proteins that were highly abundant and identified in all samples were the most reproducibly quantified (median CV = 16%, Figure A.2.1). By comparison, the average R^2 between different cell lines was 0.72, indicating significant differences in global protein expression.

The data presented here comprises more than 200,000 quantitative measurements of absolute protein abundance. Innovations in instrumentation and extensive peptide fractionation prior to analysis have greatly increased the sensitivity and reproducibility of “shotgun proteomics” analysis, and our quantitative results compared favorably with a recent targeted proteomics study on many of the same cell lines (Kennedy et al., 2014) completed by the CPTAC (Clinical Proteomic Tumor Analysis Consortium). To facilitate use and dissemination of the data, we have developed a web resource (https://zucchini.gs.washington.edu/BreastCancerProteome/) in which protein abundances can be queried, and correlated to genomic and drug sensitivity data, as presented below. To demonstrate the validity of our data set as a quantitative resource, we examined several clinical breast cancer biomarkers including ESR1, PGR, and ERBB2 (Figure 2.2). These measurements accurately reproduce the known classification of cell lines based on

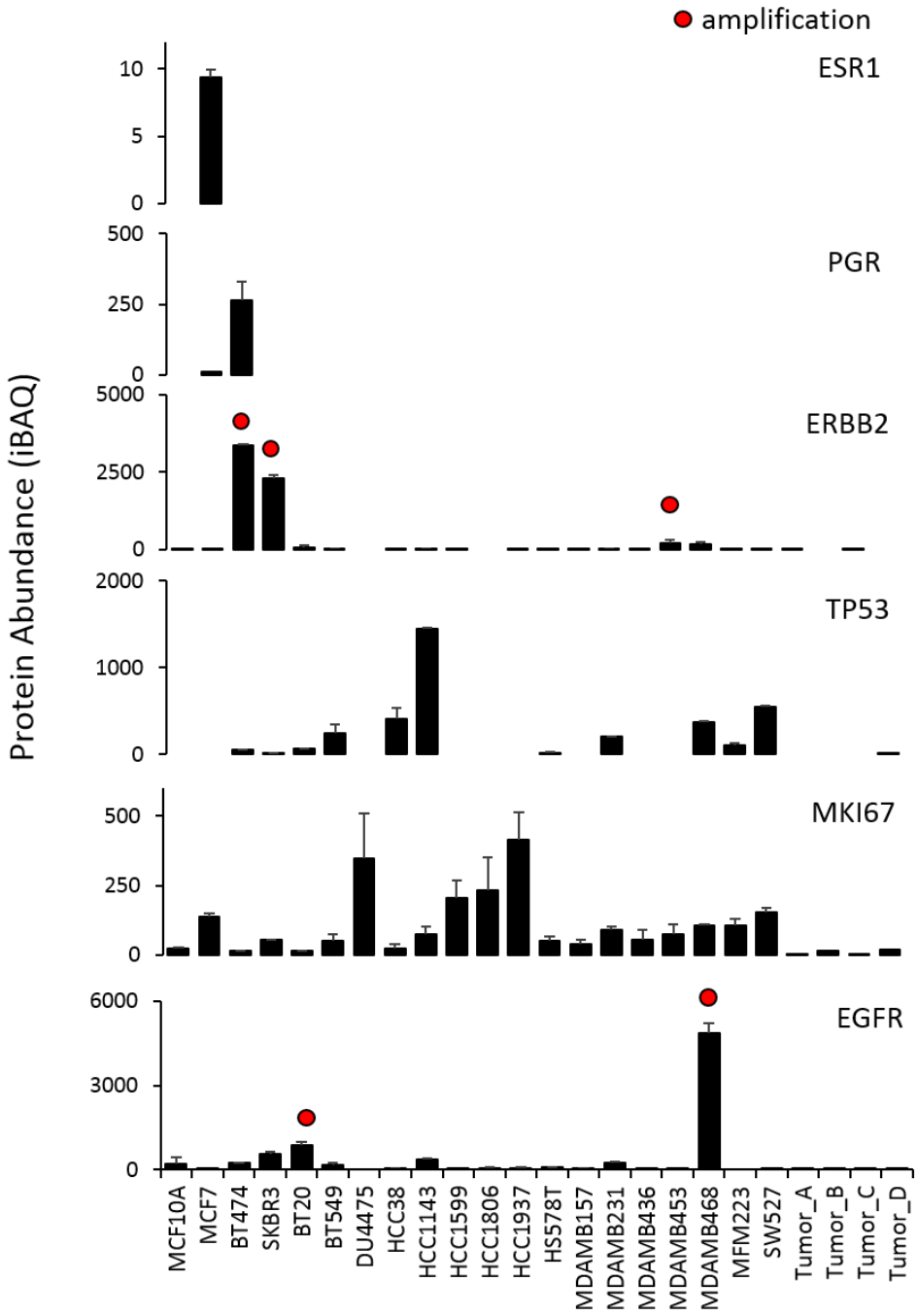


Figure 2.2 Quantification of clinical breast cancer biomarkers

ESR1: estrogen receptor, PGR: progesterone receptor, ERBB2: human epidermal growth factor receptor-2, TP53: tumor protein p53, MKI67: Ki-67 antigen, EGFR: human epidermal growth factor receptor. Sample labels are shown in the bottom panel. Absolute protein abundance was calculated using intensity-based absolute quantification (iBAQ). Error bars represent S.D. Red dots indicate gene copy number amplification (>7 copies).

immunocytochemistry (Subik et al., 2010) and correspond with known copy number amplifications. In contrast to antibody staining, which assesses the presence or absence of expression, mass spectrometry provides sensitive and precise quantitation over a broad range. This is an important consideration for markers such as Ki-67, which are dynamically expressed in all cells. As another example, the cell line MDA-MB-453 stains negative for ERBB2 (Vranic et al., 2011) and was classified as a TNBC cell line (Neve et al., 2006), despite bearing a copy number amplification. However, our results show that MDA-MB-453 expressed ERBB2 at levels 20-fold higher than the median, compared to several hundred-fold overexpression of ERBB2 by cell lines such as BT474 and SKBR3.

2.3.2 Quantitative analysis of TNBC proteomic subtypes

Molecular subtyping using gene expression or copy-number aberration has been used extensively to characterize clinical breast cancer specimens and cell lines (Banerji et al., 2012; Lehmann et al., 2011; Prat and Perou, 2011). We used hierarchical clustering to identify patterns based on correlation of protein expression profiles. This approach classified the panel of cell lines into two overarching groups containing four clusters (Figure 2.3a). To illustrate the relationship between driver gene alterations and proteome profiles, we show the most frequent census mutations and copy number aberrations for each cell line (Figure 2.3a, upper). Cell lines with similar genetic abnormalities tended to cluster together. As has been observed previously (The Cancer Genome Atlas Network, 2012), PIK3CA mutations were associated with luminal breast cancer subtypes (80% of the cell lines in cluster 1), whereas TP53 mutations were characteristic of triple-negative breast cancer (100% of the cell lines in clusters 3 and 4). Mutations in the tumor suppressor NF1 were exclusive to the mesenchymal-like subtype (cluster 4) and BCR mutations were exclusive to luminal cells (cluster 1).

Protein expression patterns within subtype clusters were still highly cell-type specific. To better illustrate this, we used principal component analysis (PCA) to project the distances between each proteome onto a two-dimensional coordinate system. Some of the sample proteomes formed tight clusters, while others were more distantly related to those in the same group (Figure 2.3b). Additional principal component dimensions are necessary to capture the proximity of cell lines such as MFM223, BT474, and HCC1599 to their respective subtypes. Intra-subtype correlation was also modest in earlier classification studies using mRNA expression (Lehmann et al., 2011), and the differences in mRNA may be further amplified at the protein level. The heterogeneity of protein expression underscores the importance of data-driven cell line selection in cancer research.

Accurate analysis of genes, transcripts, or proteins from heterogeneous clinical specimens represents a major challenge for precision medicine. The proteins expressed >10-fold in tumors versus the cell lines were enriched with proteins from blood cells and plasma ($p<0.001$). These proteins accounted for as much as 20% of the total proteome intensity from the tumors. Since TNBC cell lines should better represent the cellular component of the tumor we correlated tumor samples to the centroids from each cell line cluster to identify which proteomic subtype they belonged to, and found that they were all more similar to clusters 3 and 4, an observation which can also be made based on PCA (Figure 2.3b).

Nevertheless, many proteins significantly over- or under-expressed within each cluster could be identified. We were particularly interested in potential drug targets and proteins known to be involved in cancer biology. For example, the protein STAT5A, a pro-survival transcription factor, was expressed at high levels in the tumors and mesenchymal-like cell lines (Figure 2.3c). Using the first cluster as an example, we show how these proteins can be identified using our web-based resource (Figure A.2.2). The transcription factor FOXA1 was exclusively expressed by luminal-like cells, while TGFB1 was not found (Figure A.2.2). PPM1A, a protein involved in the

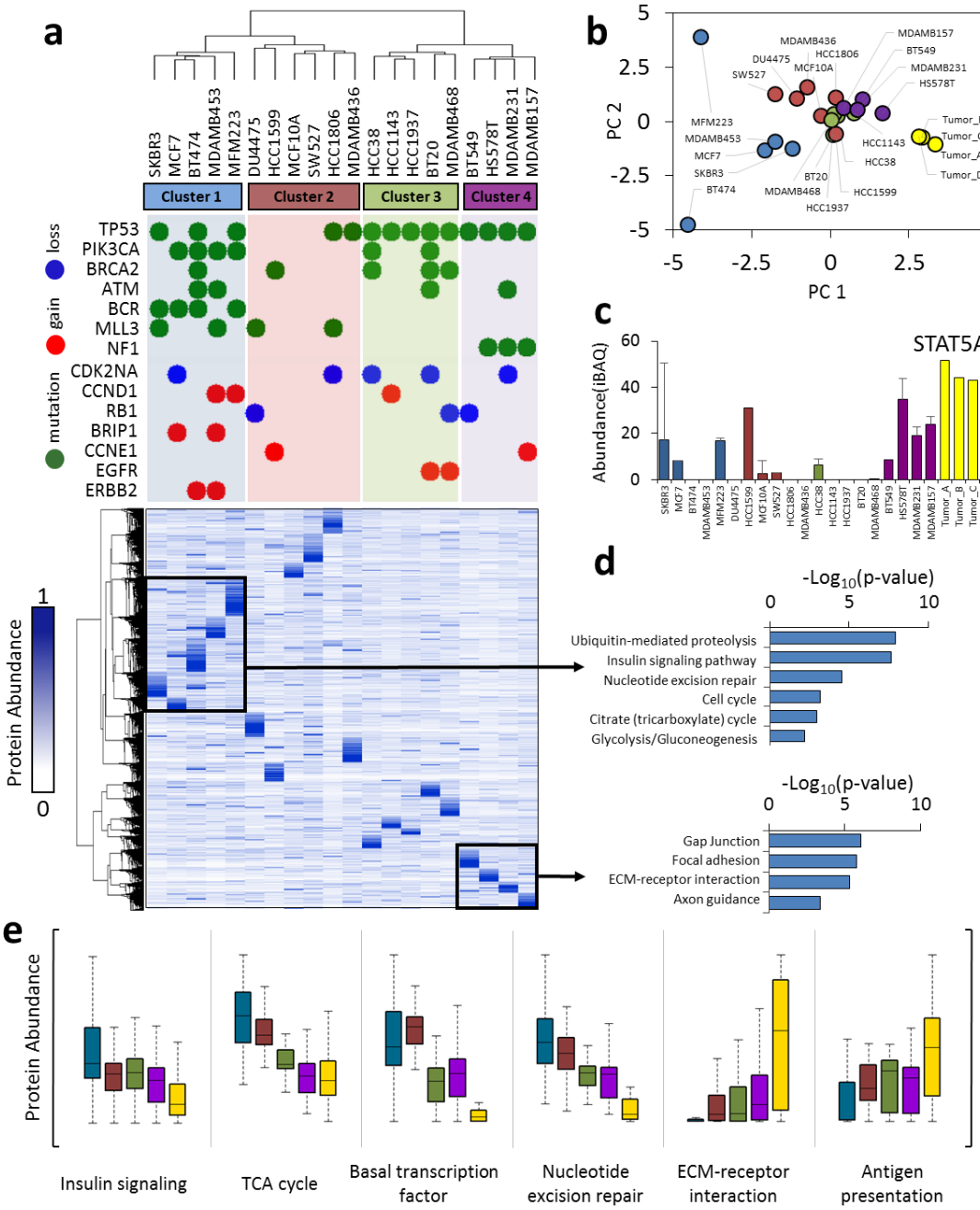


Figure 2.3 The triple-negative breast cancer proteome

(a) Hierarchical clustering of protein expression profiles computed using centered Pearson's correlation identified four proteome subtypes. Frequent genetic aberrations are overlaid onto the proteome clustering results. Green circles represent exonic mutations. Red and blue circles represent copy number gain (>7 copies) or loss (0 copies), respectively. Colored background shading corresponds to cluster membership. At the time of writing, exome sequence and copy number data were not available for MCF10A, SW527. **(b)** Scatter plot of principal component 1 and 2. Principal component analysis was performed using protein expression profiles. Each point represents a sample. Colors represent hierarchical cluster membership from (a). **(c)** Representative example of a protein upregulated in cluster 4 and tumors. STAT5A: signal transducer and activator of transcription 5A. Error bars represent S.D. **(d)** Biological pathways enriched from the indicated proteins clusters. Inverted $\log_{10}$ p-values are shown. **(e)** Distribution of protein abundances within each cluster (colors) for indicated biological processes. For all panels, cluster membership is indicated by the same colors used in (a), with tumor samples indicated in yellow.

suppression of TGF-β signaling pathways (Lin et al., 2006), was decreased in TNBC, while many proteins involved in immunity and metastasis such as POSTN, MYLK, and HLA-A were expressed at higher levels in TNBC (Figure A.2.3). Some of these proteins are thought to be provided by tumor-infiltrating immune cells and fibroblasts (Quail and Joyce, 2013), but here we show they are also abundant in the homogenous conditions of cell culture.

The composition of each cluster showed striking similarity to subtypes defined by mRNA expression arrays and morphological studies (Kenny et al., 2007; Lehmann et al., 2011; Neve et al., 2006). Cluster 1 contained the luminal breast cancer cell lines SKBR3, MCF7, and BT474 as well as “luminal-androgen-receptor” cell lines MFM-223 which expresses the androgen receptor protein, and MDA-MB-453 which overexpresses ERBB2 as described above. The set of proteins that were highly expressed by these cell lines was enriched for functions typically expected of cancer cells including insulin and ErbB signaling, glycolysis, and nucleotide excision repair (Figure 2.3d). Cluster 2, most similar to the “basal-like 2” gene expression subtype, contained, DU4475, SW527, HCC1806, MDA-MB-436, and the normal breast epithelial cell line MCF10A. Cluster 3 included all “basal-like 1” cell lines: HCC38, HCC1143, HCC1937, BT20, and MDA-MB-468. Cluster 4, containing BT549, HS578T, MDA-MB-231, and MDA-MB-157, was identical to “mesenchymal-like/claudin-low” subtype (Lehmann et al., 2011), all showing stellate morphology in 3D culture (Kenny et al., 2007), and high invasiveness in chamber assays (Neve et al., 2006) (Figure 2.3d). To better understand the biology of each subtype, we compared the distribution of protein abundance within gene ontology categories. Interestingly, luminal-like cells expressed higher levels of pathways associated with proliferation such as cell cycle, growth factor signaling, metabolism, and DNA damage repair mechanisms (Figure 2.3e, Figure A.2.3). TNBC cell types, particularly the tumors and more invasive cells, expressed higher levels of pathways associated with metastasis such as ECM-receptor interaction, cell adhesion, and angiogenesis (Figure 2.3e, Figure A.2.3). The expression of proliferation and metastasis pathways were mutually exclusive,

an observation also made in an analysis of mRNA expression profiles from claudin-low tumors (Prat et al., 2010). Thus, therapies targeting immune and metastatic signaling are an exciting avenue for TNBC treatment.

2.3.3 Differential expression of cancer signaling proteins

The cancer genome has been studied extensively (Futreal et al., 2004; Vogelstein et al., 2013). We sought to characterize the abundance of proteins derived from cancer census genes and signaling pathways (Figure 2.4, Figure A.2.4). The abundance of most signaling proteins spanned two to three orders of magnitude, but others were expressed similarly across all cell lines (Figure 2.4a-g). These proteins included several members of the RAS-MAPK pathway such as GRB2, HRAS/KRAS/NRAS, MEK1/2, and ERK1/2. In certain cases expression of these proteins was associated with proteomic-based breast cancer subtypes. For example, CHEK2, HMGA2, POT1, and IL6STwere highly expressed by members of clusters 1 through 4, respectively (Figure 2.4h-i). However, protein expression was generally variable and cell-type specific. MLL3 was specifically expressed by BT474, BT20, and tumor A, which were each from different clusters (Figure 2.4h). HCC1806 and MDA-MB-436 specifically lacked expression of the protein kinase AKT1/2 (Figure 2.4b). PKCα was expressed at high levels in each of the cell lines from cluster 4, but also was highly expressed in DU4475 (Figure A.2.4). These results show that despite overall concordance of whole proteome profiles with various cellular phenotypes, in most cases the expression of particular cancer proteins did not uniformly belong to one subtype or another. The identification of proteins with very specific outliers or large dynamic range provides a valuable resource for TNBC drug development efforts. EGFR, ERBB2, ESR1, and PGR exemplify these properties (Figure 2.4a, Figure A.2.4) and are already routine clinical targets in breast cancer, but there are many others.

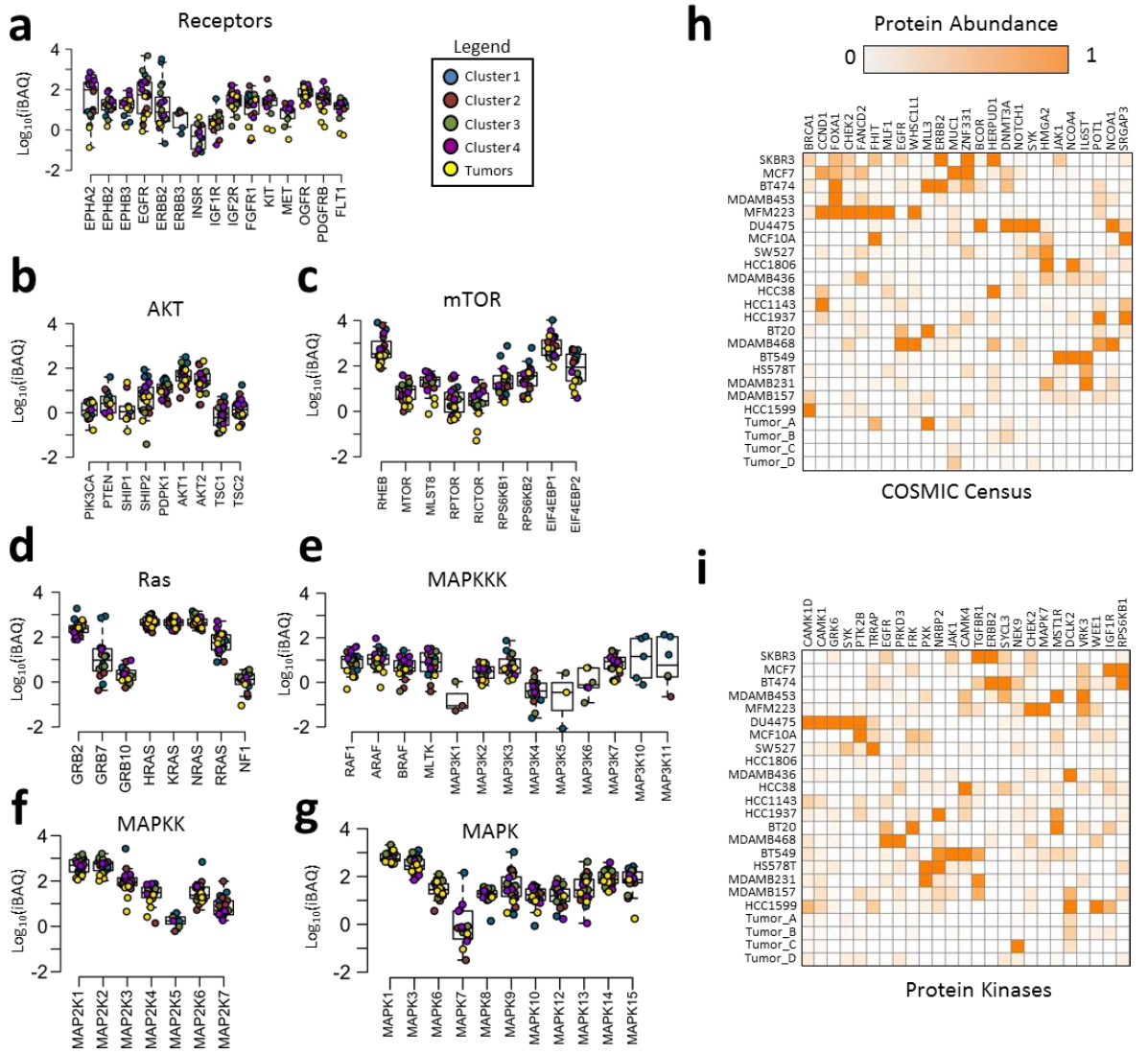


Figure 2.4 Expression of cancer signaling proteins

(a-g) Distribution of absolute abundance for each protein in the signaling network. Chart titles indicate subnetwork membership. Each data point represents a sample, color coded according to cluster membership from Figure 2.3a. **(h)** Top 25 most differentially expressed proteins (highest standard deviation between different samples) from the COSMIC gene census or **(i)** the protein kinase superfamily.

For example, ephrin type A receptors, which are involved in embryonic development and not normally present in adult tissues, were overexpressed by several orders of magnitude in many TNBC cell lines compared to luminal-like cells (Figure 2.4a). With the increasing availability of comprehensive quantitative proteomics datasets, protein expression should continue to be one of the most valuable parameters for drug development and clinical diagnostics.

2.3.4 Isoform-specific protein expression

The identification and quantification of protein isoforms resulting from alternative splicing is a significant challenge in proteomics, arising from the reduced number of isoform-specific peptides that are amenable to analysis by mass spectrometry. For this dataset, we first relied on isoform-specific peptides to unambiguously identify proteins mapping to the same gene in the Uniprot sequence database. This led to the identification of 1,860 protein isoforms that corresponded to 844 genes, 52 of which were members of the COSMIC census. Next, we examined the relative quantification of protein isoforms. Protein isoforms share long segments of identical sequence but are missing certain protein domains, resulting in altered signal intensity from those parts of the protein.

We relied on manual inspection to analyze the expression of isoforms for proteins involved in cancer progression. For most proteins, different isoforms were nearly perfectly correlated, indicating no difference in expression of specific isoforms, but there were notable exceptions. For example, we identified variants in the p65 subunit of the transcription factor NF-κB, the tumor antigen CD47, and focal adhesion kinase PTK2. The protein sequence of the NF-κB p65 variant is identical to the canonical sequence until proline 344, followed by the read-through translation of 33 amino acids and an early stop (Figure 2.5a). The alternative sequence lacks many important regulatory regions including the residues phosphorylated by IKKB that directly affect its transcriptional activity (Sakurai et al., 1999). The p65 variant was detected in two cell lines and

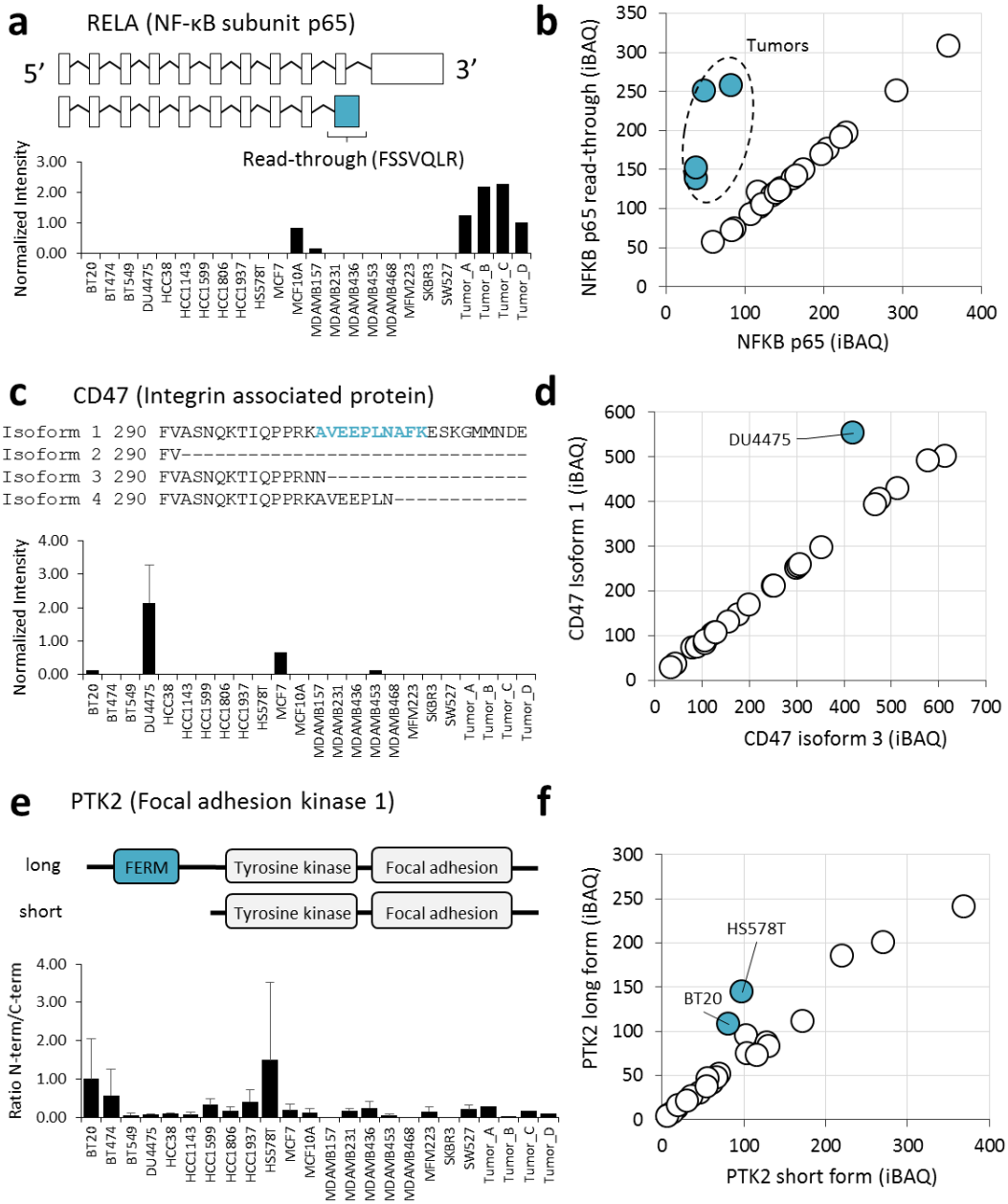


Figure 2.5 Differential expression of protein isoforms

(a) Schematic of RELA (NF-κB subunit p65) mRNA sequence variants and intensity-based quantification of the isoform-specific peptide FSSVQLR in each sample. Peptide intensity was divided by the total proteome intensity for normalization. The location of an exon read-through event is indicated. **(b)** Scatterplot of the full length NF-κB protein *versus* the read-through variant highlighting off-diagonal samples. **(c)** Four alternative splice variants encode the cytoplasmic tail of integrin associated protein CD47. The sequence of these variants is shown along with the quantification of the peptide specific to isoform 1, AVEEPLNAFK. **(d)** Scatterplot of CD47 isoform 1 *versus* isoform 3 highlighting off-diagonal samples. **(e)** Schematic of N-terminally truncated form of focal adhesion kinase PTK2 and quantification of N-terminal/C-terminal intensity in each sample. **(f)** Scatterplot of PTK2 long form *versus* short form highlighting off-diagonal samples.

was expressed at higher levels in all four tumor samples (Figure 2.5b). This result was confirmed by an isoform-specific peptide, FSSVQLR, which matched no other entry in the Uniprot protein sequence database (Figure 2.5a). This finding was especially interesting since the tumor proteomes were enriched in immuno-modulatory pathways. NF-κB modulates the inflammatory response and plays an important role in cancer by promoting metastasis (Huber et al., 2004; Luo et al., 2004).

CD47 is an atypical G-protein coupled receptor with five membrane spanning domains that participates in integrin signaling and is proposed to have many important roles in cancer (Sick et al., 2012). We detected two of the four known alternative splice variants which differentially encode the cytoplasmic tail. The cell line DU4475 expressed higher levels of the long isoform (Figure 2.5c-d), which is highly expressed in neurons (Brown and Frazier, 2001). While little is known about the functional differences between the isoforms, it is likely that this tail mediates intracellular signaling downstream of the receptor.

PTK2, or focal adhesion kinase 1, is a tyrosine protein kinase involved in cell migration (McLean et al., 2005). We confirmed the presence an N-terminally truncated form of this protein which lacks the FERM (4.1-Ezrin-Radixin-Moesin) domain (Figure 2.5e). The FERM domain regulates PTK2 localization and interaction with other proteins to affect its activity (Frame et al., 2010). Interestingly, the full-length form appeared to be expressed higher in HS578T and BT20 cells based on the relative intensity of N-terminal versus C-terminal peptides (Figure 2.5e-f). The differential expression of structural protein variants, many of which occur post-translationally, could be a significant regulatory mechanism in cancer. Further work will be necessary to systematically identify and accurately quantify these events.

2.3.5 Proteogenomic analysis identifies signatures of driver mutations

Genetic aberrations such as sequence mutations and amplifications, which typically occur in regulatory proteins, can have pleiotropic downstream effects on other proteins that more directly drive cancer phenotypes. We integrated publicly available exome sequence and gene copy number (CN) data from COSMIC (Forbes et al., 2011) with proteome profiles from 18 cell lines. Protein abundance trended positively with gene CN. The average expression of all proteins in each CN bin correlated strongly with CN (R = 0.96). However, it was more variable and correlated poorly on a pairwise basis (n = 56,579, R = 0.19) (Figure 2.6a). For example, the cancer census gene NDRG1 was not correlated with CN (R = -0.06) and was not highly expressed even when amplified (Figure 2.6b). This poor correlation is expected for proteins under high transcriptional, translational or proteasomal control.

Driver mutations occur frequently in regulatory proteins such as protein kinases, E3 ubiquitin ligases, and transcription factors which alter the physiology of the cell by modulating the abundance or activity of other proteins. For example, our data showed that DU4475, the cell line with an APC mutation, expressed more than 4-fold median levels of β-catenin (P = 3.3x10-4, heteroscedastic t-test), which APC normally targets for degradation. Initially we characterized cellular subtypes according to protein abundance profiles and asked whether frequent genetic mutations were associated with these subtypes (Figure 2.3). An alternative analysis approach is to group cell lines by their mutational status, and ask whether the abundance of specific proteins are associated with these mutations, as in the β-catenin and APC example.

We reasoned that mutations in certain driver genes, such as those in the same signaling pathway, would likely converge to regulate common effectors. To determine the global effects of driver genetic mutations on protein expression, we systematically evaluated gene-protein associations

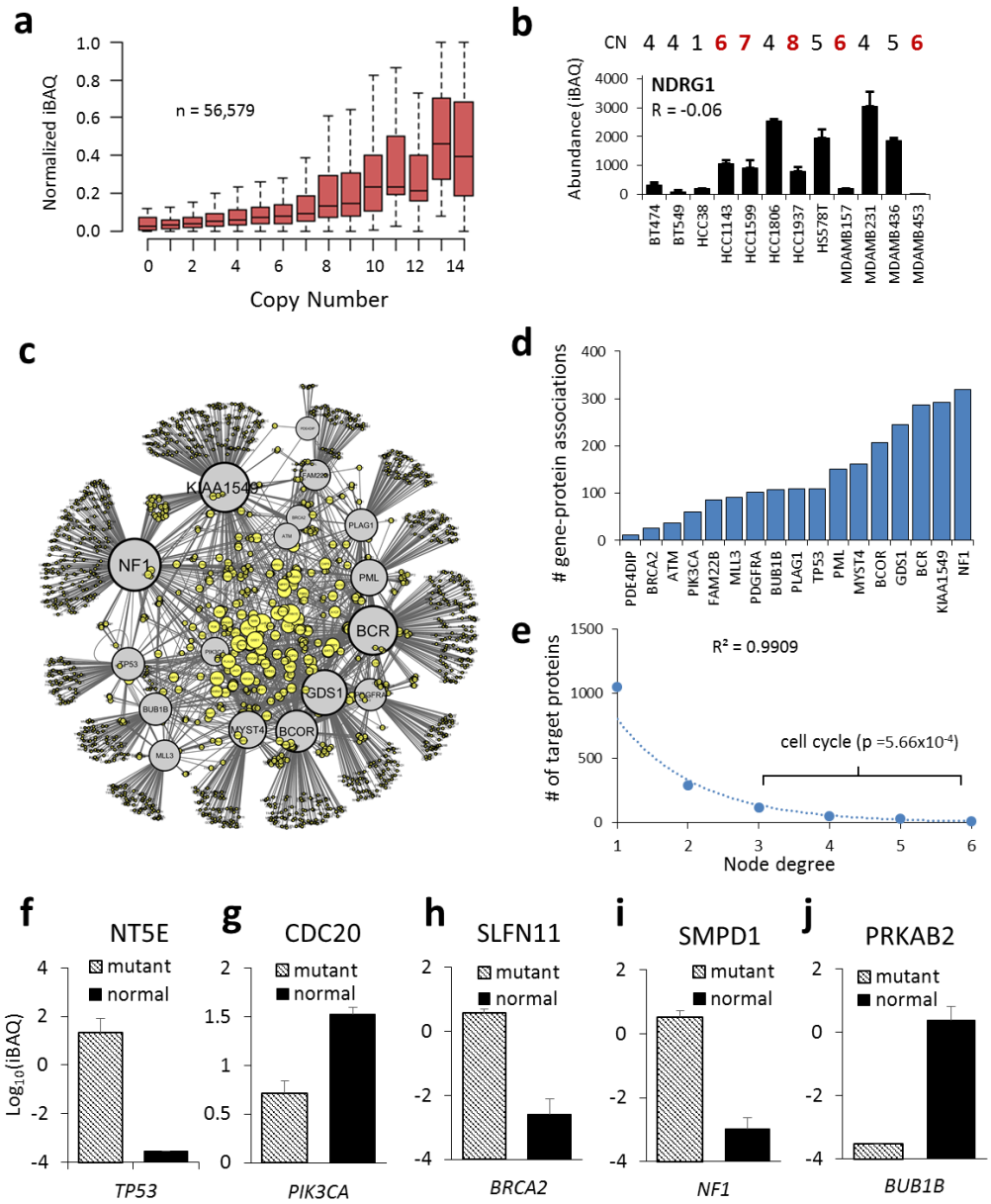


Figure 2.6 Proteogenomic associations

(a) Boxplot showing relationship of protein abundance *versus* gene copy number. Protein abundances were row-normalized to a scale of 0 to 1 to account for differences in absolute expression. **(b)** NDRG1 (N-myc downstream regulated gene 1), a representative protein that was not correlated with copy number. CN: copy number. CN>6 highlighted in red. R represents Pearson's correlation. Error bars represent S.D. between replicate measurements. **(c)** Network of gene-protein associations. Each edge represents an association ($p < 0.001$) between a mutated census gene (gray nodes) and protein expression (yellow nodes). Only genes from the COSMIC census mutated in at least 3 cell lines were analyzed. Node size represents the number of connections. The network was plotted in Cytoscape using 'edge-weighted spring embedded' layout so that genes with common associations cluster together. **(d)** Number of outgoing associations for each mutated gene in network. **(e)** Number of incoming associations for each target protein in network (node degree distribution). Cell cycle proteins were enriched among proteins with 3 or more associated genes ($p = 5.66 \times 10^{-4}$). **(f-j)** Representative gene-protein associations ($p < 0.001$) for common genetic lesions in breast cancer. Protein is indicated in chart title, and mutated gene shown in italics below plot.

for frequently mutated census genes (n ≥ 3 cell lines) by comparing the abundance of each protein in cell lines with versus without a mutation, and plotted this information as a network.

Driver genes and their protein targets formed clusters according to their shared associations (Figure 2.6c). The number of significant (P<0.001) associations for each gene ranged from 11 to 320 (Figure 2.6d). The network degree distribution fit an exponential function (R^2 = 0.99), revealing 233 'hub' proteins, each associated with 3 or more cancer census genes (Figure 2.6e). 'Cell cycle' was the only significantly enriched gene ontology term among hub proteins (P = 5.66x10-4). While not surprising, it demonstrates that dysregulation of cell cycle protein abundance may be a common effect of diverse genetic mutations.

On an individual basis, proteins regulated downstream of genetic lesions (e.g. TP53 loss-of-function) might represent more suitable therapeutic targets than the gene product itself. Several highly significant (P<0.001) gene-protein associations are shown (Figure 2.6f-j). In the case of TP53, nearly all of the significantly associated proteins were involved in DNA metabolism and repair. One such protein was ecto-5'-nucleotidase (NT5E or CD73), a GPI-anchored cell surface enzyme involved in the production of membrane-permeable nucleosides which can be used for nucleotide salvage (Zimmermann, 1992). Targeting it by siRNA or small molecule inhibition (using adenosine [(α,β)-methylene] diphosphate) arrested the cell cycle and triggered apoptosis in MDA-MB-231 breast cancer cells (Zhi et al., 2010). Monoclonal antibodies against NT5E were also demonstrated to block breast cancer metastasis in vivo (Stagg et al., 2010). NT5E may be an effective drug target specifically for cancers with TP53 mutations. In addition to the discovery of potential drug targets, these proteins could also be used as markers to infer whether or not a mutation is deleterious.

2.3.6 Proteomics of drug sensitivity

To generate a resource for drug sensitivity prediction, we screened the sixteen TNBC cell lines from our panel against a library of 160 compounds at eight different concentrations spanning four orders of magnitude. We used this data to determine the IC50, defined as the dose required to reach a 50% reduction in cell viability, for each drug in each cell line. Approximately three quarters (123/160) of the compounds elicited a measurable response in at least one cell line, and each cell line was sensitive to at least 5 compounds at sub-micromolar doses. The distribution of responses for each drug was diverse (Figure 2.7a). The IC50 distribution for most drugs spanned a wide range, 790-fold on average. Some drugs were very specific with few sensitive cell lines (e.g. everolimus, methotrexate, lapatinib), while other drugs were indiscriminate with few resistant cell lines (e.g. bortezomib, paclitaxel, MG132).

Next, we combined our pharmacological data set with publicly accessible data from the Genomics of Drug Sensitivity in Cancer (CRx) resource (Yang et al., 2013) and performed regression analysis against mass spectrometry-derived protein abundances to discover proteomic markers of drug sensitivity or resistance. We used hierarchical clustering to analyze global patterns among drug sensitivity-protein expression relationships, revealing many distinct clusters (Figure 2.7b). Drugs targeting proteins in the same pathway (e.g. BRAF and MEK inhibitors) showed similar correlation profiles. Interestingly, proteins that were part of the same pathways or complexes also clustered together, which did not occur using protein expression data alone (Figure 2.3a). The cluster that was highly enriched with mitochondrial proteins was associated with sensitivity to drugs that might depend on mitochondrial protein expression (belinostat, vorinostat, obatoclax). For example, since protein acetylation is known to be enriched within the mitochondrial space, cells with more mitochondria might be more sensitive to deacetylase inhibition. In a similar vein, the cluster that was enriched with translation factors was associated with increased sensitivity to proteasome inhibitors MG132 and bortezomib. These results show that integration of proteomics

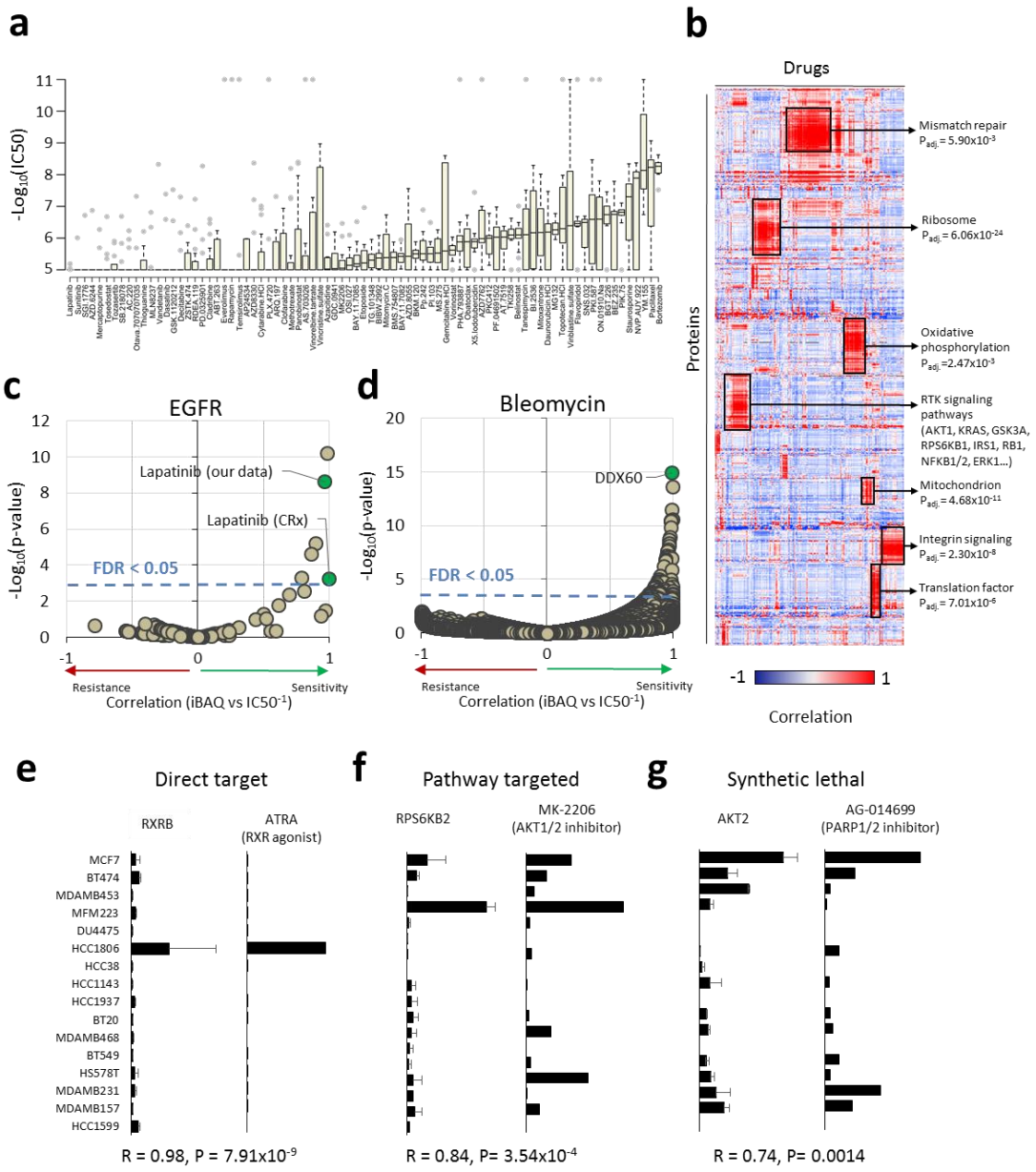


Figure 2.7 Protein expression and drug sensitivity

(a) Distribution of drug sensitivity (-log$_{10}$IC50) values across 16 TNBC cell lines for each drug in order of increasing median sensitivity. Drugs with sub-micromolar IC50 in at least one cell line are shown. Grey points represent outlier values (>1.5x interquartile range). **(b)** Hierarchical clustering of drug-protein associations. Pairwise Pearson's correlation was calculated systematically between drug sensitivity (inverted IC50) and protein abundance (iBAQ) values and clustered in both dimensions. Enriched gene ontology terms are shown for several clusters with Benjamini-Hochberg adjusted p-value. **(c)** Association of drug sensitivity with EGFR expression. The EGFR inhibitor lapatinib was significantly associated in both drug screen datasets (CRx: P = 6.2x10^{-4}, our data: P = 2.4x10^{-9}, FDR < 0.05). **(d)** Association of protein expression with bleomycin sensitivity. The protein DDX60 was significantly associated bleomycin sensitivity (P = 1.1x10^{-15}, FDR < 0.05). **(e-g)** Pairwise comparison of protein expression and drug sensitivity for three examples. Left panel: protein abundance (iBAQ) across cell lines. Right panel: drug sensitivity (inverse IC50, M^{-1}) across the same cell lines. RXRB: retinoid X receptor beta, RPS6KB2: ribosomal protein S6 kinase-2, AKT1: RAC-alpha serine/threonine-protein kinase. ATRA: RXR agonist all-trans retinoic acid, MK-2206: pan-isoform AKT inhibitor, AG-014699: poly-ADP ribose polymerase 1/2 inhibitor. Pearson's correlation and p-value is indicated below the plots. CRx: Data from (Yang et al., 2013). Panel A includes only data generated in this study. For panels b-g, data from the CRx was included. Missing IC50 values were not imputed.

and drug sensitivity data using regression analysis provides a rich resource to identify unexpected modes-of-action and to discover new features of target pathways.

We used the regression analysis to select the most effective and robust drugs for known targets. For example, EGFR expression was, as expected, strongly associated with sensitivity to the EGFR inhibitor lapatinib in both drug screens (our data: R = 0.96, P = 2.36 x 10-9; CRx: R = 0.99, P = 6.2 x 10-4) (Figure 2.7c). Proteomics data can also be used to uncover mechanisms of drug sensitivity. For example, several cell lines were hypersensitive to the drug bleomycin, an antibiotic used to treat plantar warts as well as many forms of cancer by inducing DNA damage.
Expression of DDX60, an antiviral RNA/DNA helicase that binds cytosolic DNA (Miyashita et al., 2011), was most significantly associated with sensitivity to bleomycin (R = 0.99, P = 1.1 x 10-15) (Figure 2.7d).

We curated these drug sensitivity results to ask whether drug sensitivity associated with (1) genetic mutations or protein expression of the drug target itself, (2) proteins in the same pathway as the target, or (3) other literature-supported 'synthetic lethal' interactions. Drug sensitivity associated strongly with both genomic and proteomic features of known targets. For example, we found that sensitivity to all-trans retinoic acid (ATRA) was correlated with the expression of its target protein RXRB (R = 0.98, P = 7.91 x 10-9). HCC1806 cells, which expressed the highest level of RXRB, were >200-fold more sensitive than the median cell line (Figure 2.7e). The cell line DU4475, which harbors the hyperactive BRAF-V600E mutation, was hypersensitive to both BRAF and MEK inhibitors (6,000-fold and 100,000-fold versus median, respectively) despite similar expression of the target proteins.

Another potential mechanism of drug sensitivity is synthetic lethality, in which the right combination of genetic, proteomic, or pharmacologic perturbations leads to cell death. Synthetic lethality tends to occur between proteins in the same pathway. For example, the AKT1/2 inhibitor

MK-2206 was not associated with expression of AKT isoforms, but was significantly associated with expression of RPS6KB2 (R = 0.84, P = 3.54x10-4) (Figure 2.7f), which lies downstream in the signaling pathway (Shaw and Cantley, 2006). Other drugs correlated with proteins that are not known to be in the same pathway, but have been previously proposed to be synthetic lethal relationships in genetic datasets. For example, poly-ADP ribose polymerase (PARP) inhibition disrupts DNA repair leading to genotoxic stress and cellular senescence, a process shown to be accelerated in overactive AKT signaling mutants (Chatterjee et al., 2013; Mendes- Pereira et al., 2009). In our data, AKT protein expression was also significantly correlated with sensitivity to PARP inhibition using AG-014699 (R = 0.74, P = 0.0014) (Figure 2.7g).

Finally, we explored how the differences in drug sensitivity and target expression between members of a signaling pathway relate to pathway structure. In the Akt-mTOR-S6K signaling pathway, ribosomal protein S6 kinases (RPS6KB1/2) are activated by mTOR. Curiously, despite its association with MK-2206 sensitivity, expression of either RPS6KB1 or RPS6KB2 was inversely correlated with the S6K inhibitor PF-4708671 in luminal breast cancer cells (R = -0.96, P = 0.04) (Figure A.2.5). This is consistent with the suggestion that S6K inhibition may amplify upstream cancer signaling due to the chronic ablation of a negative feedback loop (Carracedo et al., 2008; Manning, 2004). Thus, the tumorigenic action of this protein may be best targeted indirectly (Figure A.2.5). Unlike RPS6KB2, RPS6KB1 expression did not correlate with AKT1/2 inhibitor MK-2206 sensitivity but instead was most highly correlated with the p21-activated kinase (PAK) inhibitor IPA-3 (R = 0.99, P = 1.91×10^{-12}). Based on images from the Human Protein Atlas, RPS6KB1 and PAK2 are localized to the nucleus whereas RPS6KB2 and PAK1 are cytoplasmic (Uhlen et al., 2010). Thus, the reported activation of PAK1 downstream of S6K (Ishida et al., 2007) might be localized and isoform-specific. Together, these results demonstrate that integrated analysis of drug sensitivity and protein expression provides a useful strategy for drug selection,

finding diagnostic markers, and identifying potential mechanisms of cellular signaling. Further experimentation will be required to confirm these findings.

Finally, to demonstrate the potential clinical utility of these results, we asked how many proteins from the drug association analysis could be identified in primary tumors. We found that 73% (6,798/9,292) were quantifiable in the four clinical specimens we analyzed (Figure A.2.5). Of these, 494 were at least 5-fold more abundant than the average sample in at least 1 tumor. For example, the abundance of the protein kinase AKT2 was higher in one of the tumor samples than in any cell line analyzed in this study (Figure A.2.5).

2.4 Discussion

Despite the success of large-scale 'omic' studies in providing molecular targets for therapeutic intervention, these studies have been limited by the lack of comprehensive protein data. Mass spectrometry-based proteomics has advanced rapidly and it has become a routine to reproducibly quantify near-complete proteomes using this technology. Here we used mass spectrometry to interrogate the proteomes of TNBC. We then integrated proteomics, genomics and drug sensitivity data to study the effects of genomic aberrations in the proteome and build prediction models of drug response using proteomics.

This dataset is a useful resource to further explore the biology of TNBC. For example, many of the recently described metastatic stem cell pathways were highly expressed at the protein level in TNBC compared to luminal breast cells. The most invasive TNBC cells and solid tumors expressed low levels of proteins involved in cell proliferation and high levels of proteins involved in the epithelial-to-mesenchymal transition. Thus, the highly specialized nature of metastatic TNBC cells may be one reason they are so difficult to treat using conventional cytotoxic agents

that target highly proliferative cells. Precise knowledge of the proteomes of these cells can guide the development of new drugs to target the metastatic transition.

Machine learning has become a useful tool to capture the molecular features responsible for differences in drug sensitivity (Barretina et al., 2012; Costello et al., 2014; Weinstein et al., 1997; Yang et al., 2013). Statistically significant differences in drug sensitivity based on cellular subtype have been observed (Lehmann et al., 2011), but the effect sizes are small compared to treatment strategies directed towards precise molecular insults. Examples include ERBB2 amplification (trastuzumab), BCR-ABL fusion (imatinib), or BRAF-V600E mutation (vemurafenib), all of which result in orders-of-magnitude increases in drug sensitivity. In reality, large effect sizes are needed to make an impact in the clinic. In this study, drug sensitivity and the expression of cancer-related proteins was not generally attributable to subtypes derived by clustering global protein profiles. Considering these cells were all derived from the same tissue type (breast) and were cultured in the same conditions, the dynamic range and specificity of protein expression for established regulatory proteins and drug targets was surprising. Using regression and prior knowledge to interrogate mechanisms of protein expression in drug sensitivity, we found that in many cases, drug sensitivity was strongly correlated with the expression of the drug target itself (e.g. retinoic acid receptors, EGFR) or proteins in the same biological pathway (e.g. S6K expression as a marker for sensitivity to AKT inhibitors).

With the exception of drugs targeting proteins expressed from amplified genes, the importance of protein expression in drug efficacy might be underestimated. While it is evident that the target of a drug must be expressed at some level in order for the drug to take effect, many drugs are developed with the assumption that the target is expressed at similar levels in all cells. Even in the case of gene amplification, copy number does not fully account for differences in protein expression between specimens. In any case, quantitative analysis of drug targets and genetic abnormalities at the protein level might represent a useful addition to the current adjuvant therapy

selection algorithm. Indeed, this is already routine for estrogen, progesterone, and epidermal growth factor receptor-2. Larger panels of cell lines will be necessary to capture rare genetic events and to enable more robust machine learning approaches. This will facilitate the discovery of less obvious markers of drug sensitivity, such as synthetic lethal interactions. Proteomics could also provide an indispensable tool to rescue clinical trial results which do not improve patient outcomes in aggregate, but have many exceptional responses that might be due to underlying molecular features.

This study builds upon other deep proteomic characterizations of cancer (Geiger et al., 2012b; Gholami et al., 2013; Nagaraj et al., 2011; Zhang et al., 2014) and represents the first deep proteome characterization targeting triple-negative breast cancer. With the development of large "omics" approaches, personalized, predictive medicine is the prevailing direction of next-generation healthcare technology (Tian et al., 2012). Systematic, data-driven approaches are necessary to meet this goal. We anticipate that genome-scale nucleic acid sequencing and protein analysis will provide the basic molecular diagnostics toolbox for precision cancer medicine. Triple-negative breast cancer is one of many unmet clinical needs that will benefit from future research in this area.

2.5 Experimental Procedures

2.5.1 Sample preparation

Samples were lysed in denaturing buffer and centrifuged at 12,000 g for 10 min to pellet insoluble material. Protein extracts were reduced with 5mM DTT at 55°C and alkylated with 15mM iodoacetamide at room temperature in the dark. Extracts from each sample (25µg) were diluted and digested in solution overnight with either lysyl-endopeptidase (Lys-C) (Wako) or sequencing grade trypsin (Promega). Peptides were desalted and fractionated on StageTips (Rappsilber et

al., 2007) by basic reverse-phase using a step-wise gradient of increasing acetonitrile (5%, 10%, 15%, 25%, 80%) in 0.1% NH4OH. The resulting fractions were analyzed by LC-MS/MS.

2.5.2 LC-MS/MS

Peptide fractions were analyzed on an EASY-nLC-1000 (Thermo) coupled to a hybrid quadrupole-orbitrap Q-Exactive mass spectrometer (Thermo) configured for data dependent acquisition. Raw mass spectra were searched using Sequest (release 2012.01.0 of UW Sequest) against a concatenated forward and reverse version of the Uniprot human protein sequence database (v11/29/2012). Peptide spectral matches for all fractions corresponding to the same sample were filtered to reach a protein identification false discovery rate of less than 1%, resulting in an aggregate peptide-level FDR of less than 0.1% for the entire dataset. Protein quantifications were calculated using the intensity-based absolute quantitation (iBAQ) approach (Schwanhäusser et al., 2011).

2.5.3 Drug screen and curve fitting

Compounds were added to cells using the CyBi-Well Vario Workstation (CyBio) and incubated at 37°C, 5% CO2 for 96 hours. Cell viability was measured by luminescence using quantitation of ATP as an indicator of metabolically active cells. Measurements were corrected for background luminescence and percentage cell viability is reported as relative to the DMSO solvent control. Non-linear curve fitting was performed using MATLAB’s 'nlinfit' function. External drug sensitivity data (IC50) was downloaded from the “Genomics of Drug Sensitivity in Cancer” resource (Yang et al., 2013) release 2.0 (http://www.cancerrxgene.org).

2.5.4 Statistical analysis

Significance tests and correlation analysis were performed using built-in functions within Microsoft Office Excel 2013 or R statistical computing environment version 3.1.0. Gene enrichment significance testing was performed in DAVID version 6.7 using the EASE metric, a modified

Fisher's exact test (Huang et al., 2009). All error bars represent standard deviation unless otherwise noted.

Chapter 3. Exploring the range, topology, and dynamics of cellular signal transmission

This chapter is based on unpublished observations.

3.1 Summary

Many mechanisms for cellular information transfer have been described including protein-protein interaction, allosteric modulation, small molecule diffusion, and reversible covalent modification of proteins (e.g. phosphorylation, acetylation, ubiquitylation). Among those, protein phosphorylation is a ubiquitous but precise form of signal transduction that is utilized extensively by eukaryotic cells. Remarkably, approximately one out of every twenty proteins encoded by the human genome is involved directly in the reversible transfer of phosphate to protein serine, threonine, or tyrosine residues (including protein kinases, phosphatases, and their targeting subunits). Mass spectrometry enables systematic, multiplexed quantification of protein phosphorylation, and thus provides an aerial view of cellular signaling. The overarching goal of the following studies is to use proteomics to explore fundamental questions in signal transduction, and to discover new signaling processes that pose a significant threat to human health when dysregulated.

3.2 Introduction

Cells use networks of protein kinases and phosphatases to execute real-time decisions by coordinating the behavior of effector proteins (actin and myosin, for example). Cellular signaling networks are inherently complex, a property exemplified by the sheer number of protein kinases (>100) and target proteins (>10,000) expressed by a cell and the vast diversity of stimuli a cell might encounter in its local environment. The goal of the following studies is to “reverse-engineer” the phosphorylation-dependent control mechanisms of the cell using proteome-wide

measurements as a read-out of the cellular signaling state as it is perturbed by different stimuli and precisely targeted pharmacological agents. Proteomics is an ideal tool for systems biology applications in signal transduction. Signaling systems act globally within the cell, and proteomics often reveals important events that are entirely unexpected. Proteomic analysis is also attractive because although biological systems are complex, they are comprised of modular protein parts that have been described in detail during the post-genomic era. These parts include large globular domains that establish protein function, and short linear motifs within disordered regions that facilitate regulation by reversible covalent modifications. Many databases and tools have been assembled to make these protein modules and processes more amenable to data-driven applications such as proteomics. High-dimensional (many stimuli/perturbations) phosphoproteomics experiments have not been widely adopted to study the physiology within distinct human cell types. The studies here provide in-depth quantitative models of phosphorylation-dependent control in two human cell lines, HeLa cervical cancer cells and MCF7 breast cancer cells.

3.3 Results

3.3.1 Draft map of the HeLa cellular signaling network

The cellular signalsome represents the unique cadre of signaling modules available to a specific cell type, and encodes its distinct response to a variety of stimuli (Berridge, 2014). Cells respond to signals in their local intra- and extra-cellular environment by activating specific signaling pathways which lead to the precise and coordinated regulation of cellular physiology. Importantly, since many signals require the cell to produce similar (or even identical) outcomes, signaling pathways are wired to interact and converge on the same targets, a process referred to as signal integration (Cohen, 1992). Here, we exposed HeLa cells to 16 different local signaling environments and measured their effect on the phosphoproteome by quantitative mass spectrometry. HeLa S3 cells were selected as an initial model system due to their ubiquitous

usage in cell biology, and unlike most transformed human cell lines, they lack upstream mutations in signaling pathways (Adey et al., 2013). The stimuli were selected to cover a comprehensive swath of canonical signaling modules including hormone and growth factor receptors (insulin-like growth factor-1, epidermal growth factor, insulin, beta-adrenergic receptor), small molecule second messenger systems (phorbol esters, isobutyl-methylxanthine/forskolin), cytokines (tumor necrosis factor alpha, interferon gamma, transforming growth factor beta), and stress responses (staurosporine, pervanadate, cycloheximide, ultraviolet irradiation, sorbitol) (Figure 3.1a and Table B.2.1). Each experiment was performed in cells that had been deprived of serum for four hours and each stimulus was applied for twenty minutes, a point of peak activity for many phosphorylation events. These parameters were selected to construct a baseline model of the phosphoproteome, which can later be refined in diverse cell types along with dosage, temporal and topological constraints.

To obtain deep coverage of the phosphoproteome with accurate quantification, ten biological replicates of each experiment were performed and were processed using two phosphopeptide enrichment strategies – strong cation exchange fractionation followed by immobilized metal affinity chromatography (SCX/IMAC, 4 replicates) and single-shot immobilized metal affinity chromatography (single-shot, 10 replicates) (Figure 3.1b). To increase the throughput and reproducibility of phosphopeptide enrichment, a volatile SCX buffer system was developed for peptide fractionation (eliminating the need for solid-phase desalting after fraction collection) and IMAC was performed using an automated magnetic bead processor (Figure 3.1c). Each fraction was analyzed using an LTQ-Orbitrap mass spectrometer resulting in 24 hours of analysis time for each SCX/IMAC and 2 hours for each single-shot experiment. In total more than 7 million phosphopeptides were detected, and 102,638 phosphorylation sites on 10,567 proteins were characterized (Figure 3.1d). On average, we quantified 5,364 phosphosites in each single-shot experiment and 23,745 phosphosites in each SCX/IMAC experiment, and 17,733 sites were

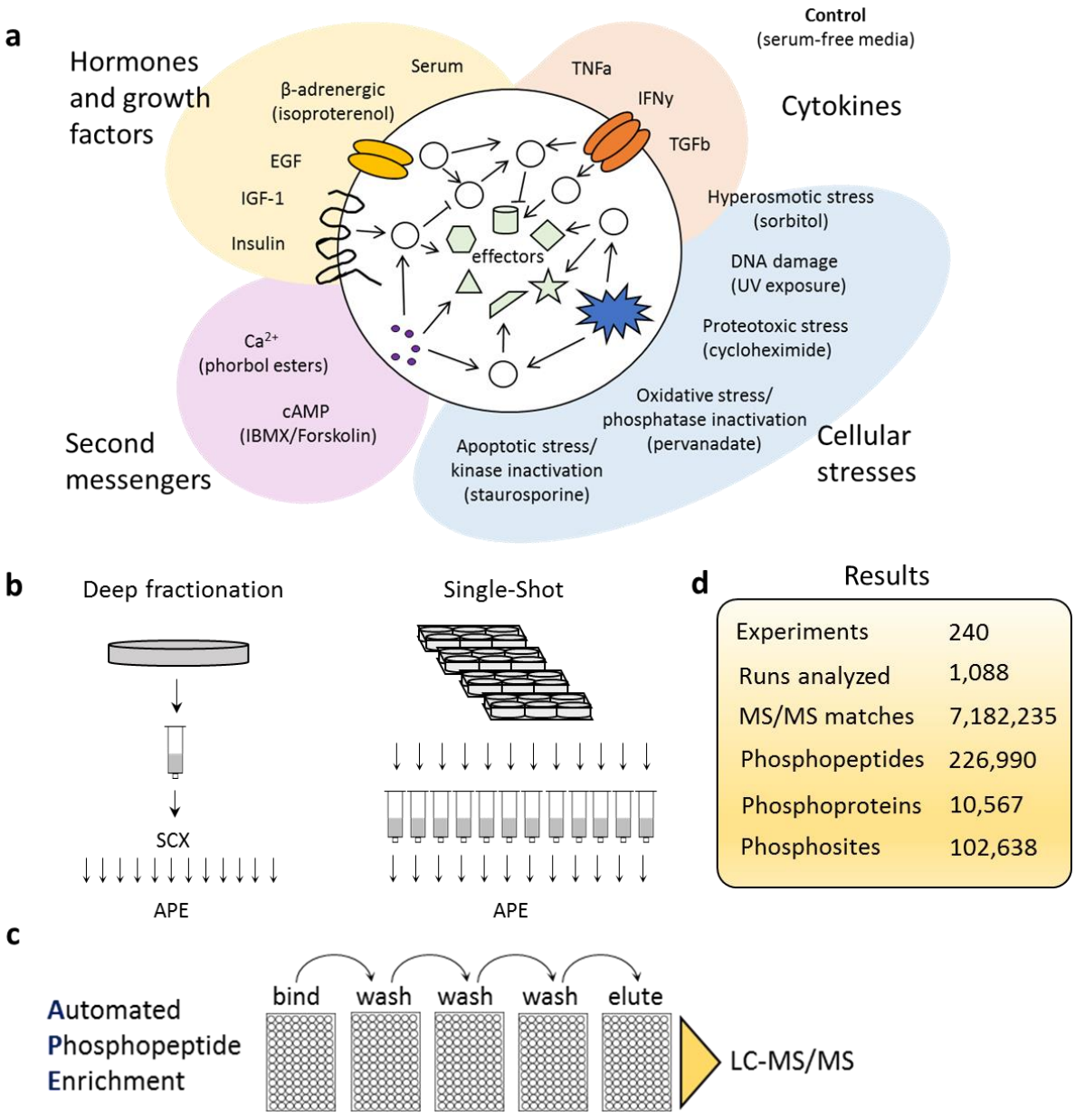


Figure 3.1 Analysis of the HeLa phosphoproteome by quantitative mass spectrometry
(a) Overview of stimuli panel. **(b) and (c)** Sample collection and **p**hosphopeptide enrichment workflow. SCX: strong cation exchange, APE: automated phosphopeptide enrichment. **(d) R**esults summary. MS/MS matches and corresponding quantifications were filtered to a false discovery rate (FDR) of <1%. The final data set of aggregate phosphopeptides and phosphosites were filtered to FDR<5%.

quantified across all sixteen experimental conditions (Figure 3.2a). Despite the depth of coverage obtained by using many replicates and fractionation strategies, stochastic peptide sampling in the mass spectrometer remains a challenging issue. The problem is accentuated by this analysis of diverse stimuli which can lead to dramatically different phosphopeptide abundance and composition.

We used hierarchical clustering analysis to determine which stimuli were most similar, and plotted them as a dendrogram (Figure 3.2b), revealing four clusters. The first cluster contained the control, interferon gamma (IFNy), transforming growth factor beta (TGFb), and insulin-like growth factor 1 (IGF-1). Compared to the other stimuli, these molecules did not have widespread effects on the HeLa phosphoproteome. Interestingly, IGF-1 stimulated many substrates in the canonical insulin pathway, but did did not cluster with insulin, which elicited a much broader phosphoproteomic response. Insulin and IGF-1 are often described interchangeably when speaking of their downstream signaling. Other phosphosites were specifically regulated by IGF1 but not insulin. The second cluster contained the classic cellular stresses sorbitol, cycloheximide, and ultraviolet irradiation, as well as the cAMP agonist and serum. Cyclic AMP signaling plays an important role in the stress response. Serum presumably contains a diverse array of signaling molecules, some of which may induce the stress response. The addition of 10% serum might also alter the osmotic pressure of the culture medium, which could induce stress signaling.

The third cluster contained phorbol myristate acetate (PMA), protein phosphatase inhibitors (PPI), and epidermal growth factor (EGF) as well as tumor necrosis factor alpha (TNF-α). PMA and EGF are known to be potent mitogens activating the MEK-ERK pathway. PMA is a diacylglycerol analog which activates Ca^{2+} ion channels, thus compared to stress signaling, mitogenic signaling is likely dependent on the Ca^{2+} as a second messenger. The potent activation of mitogenic signaling pathways by TNF-α was unexpected, but not undocumented. TNF-α plays a paradoxical role in cancer, promoting invasiveness as well as cell death (Wang and Lin, 2008), and was shown

to activate MEK-ERK signaling independently of Raf in mouse macrophages (Winston et al., 1995).

The final cluster containing staurosporine (STR), insulin (INS), and isoproterenol (ISO) was the most inconsistent with expected results. Staurosporine is a promiscuous protein kinase inhibitor, so it was surprising to observe any reproducible increases in phosphorylation in response to this molecule, however similar effects were recently reported in other cell types (Abelin et al., 2016). Similarly, the actions of catecholamines like isoproterenol are thought to specifically counteract the actions of insulin, but here they are regulating phosphorylation in the same direction. Again, it most likely depends on which specific actions are studied. In contrast to insulin, here isoproterenol clearly did not activate the AKT pathway, yet they had many other shared effects.

Signal integration was a dominant feature of the HeLa phosphoproteome, with a remarkable degree of overlap in the effects of different stimuli. To illustrate the cross-talk and breadth of effects between the sixteen signaling environments we plotted the clustered phosphorylation data as a heatmap (Figure 3.2c) and plotted the distribution of each stimulus versus control (Figure 3.2d). As expected, protein phosphatase inhibitors had the most widespread effects on the phosphoproteome *versus* control. Phosphorylation events in the canonical MEK-ERK signaling pathway were highly correlated (Figure 3.2e). However, cross-talk was evident at the effector level (MEF2D, ATF7, and ATF2). Compared to the activation sites on ERK1/2, each effector was more sensitive to both UV exposure and cycloheximide. MEF2D pS121 was sensitive to cAMP pathway activation, and ATF7 pT151 responded to IGF1 stimulation.

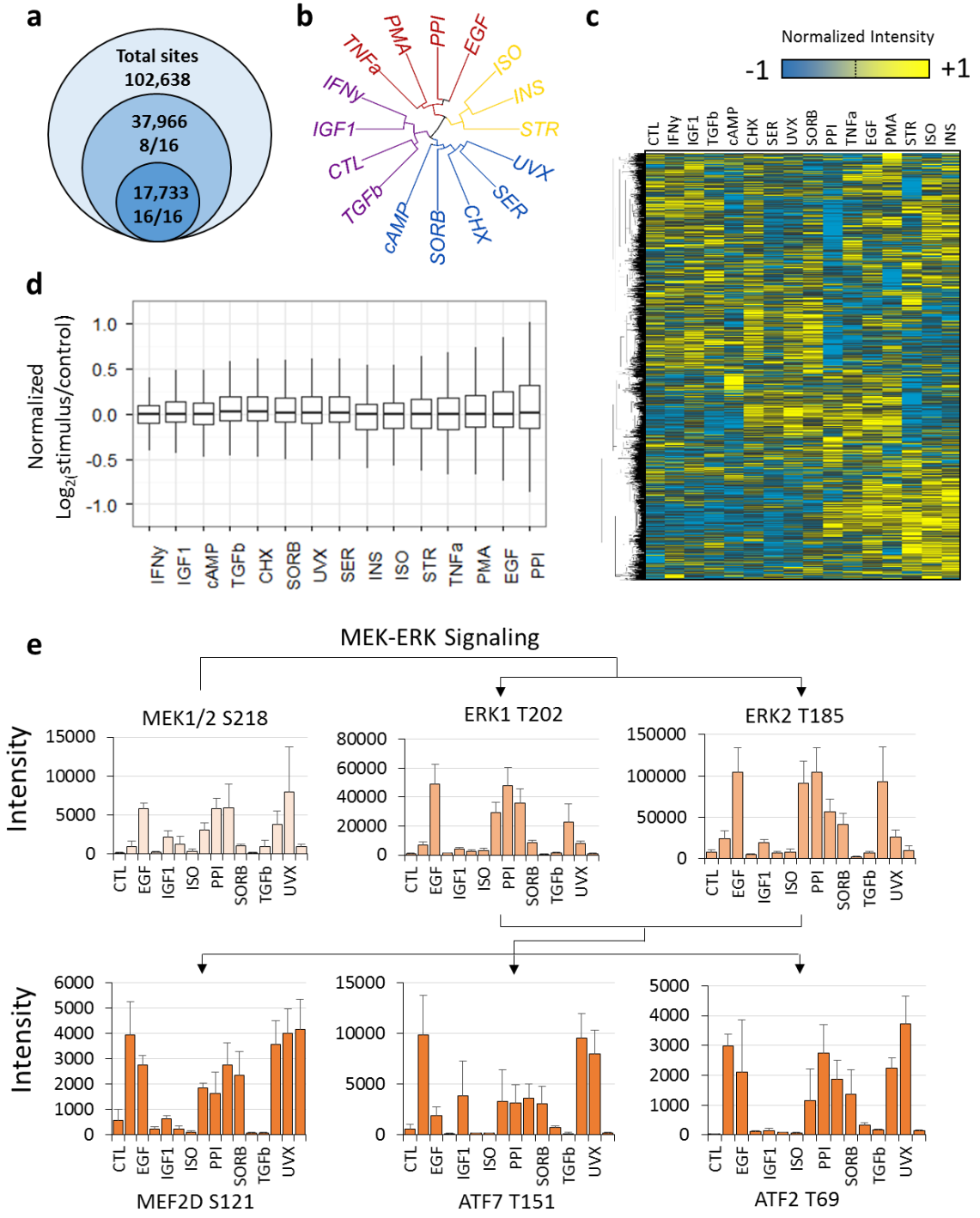


Figure 3.2 HeLa signal transmission and pathway cross-talk
(a) Total number of phosphosites detected, number of phosphosites detected in 8/16 experiments, and number of phosphosites detected in all 16 experiments. **(b)** Radial dendrogram computed using average linkage hierarchical clustering of Pearson's correlation. **(c)** Heatmap of conditional protein phosphorylation. Data were log-transformed, median-centered, and normalized such that the sum of squares for each row was equal to 1. **(d)** Distribution of values for each stimuli relative to control. **(e)** Depiction of signal transmission through the MEK-ERK pathway. Error bars represent standard error of the mean.

3.3.2 Systematic analysis of cancer signal integration

Signal integration is a ubiquitous phenomenon in signaling pathways. Large branches of the protein kinase superfamily share the same substrate sequence specificity, but they are activated and targeted towards their substrates by distinct mechanisms. At least 10 members of the AGC kinase family are directed towards peptide sequences harboring an Arg-X-Arg-X-X-[S|T] motif (Pearce et al., 2010), where 'X' represents any amino acid (Figure 3.3a). All of these kinases are direct modulators of effector function, and thus play an important role in cellular behavior. In this study we employed a systematic stimulation and perturbation strategy to determine the wiring of signaling pathways leading to the phosphorylation of arginine directed protein kinase substrates (Figure 3.3b). We used strong cation exchange followed by immobilized metal affinity chromatography (SCX/IMAC) and AGC kinase motif immuno-affinity purification (IAP) strategies followed by LC-MS/MS to characterize the phosphoproteome (Figure 3.3c). These strategies are complementary. SCX/IMAC provides a global survey of phosphorylated peptide species, while IAP targets peptides harboring AGC kinase motifs, which may be at significantly lower abundance in the cell.

We detected 44,083 phosphorylation sites on 7,012 proteins (Figure 3.3d). Despite the superior depth of SCX/IMAC, one third of phosphorylation sites detected by AGC IAP were not accessible using the global approach. Out of the sites quantified, more than 4,000 were regulated greater than 2-fold in at least one condition and accurately recapitulated known signaling events (Figure B.2.1). We subjected the regulated phosphoproteome to clustering analysis to identify patterns of pathway cross-talk (Figure 3.5a). Distinct modes of signal integration could be identified based on the response of each phosphorylation site to combinatorial stimulus and drug perturbation. For example FOXO3 pS284 was sensitive to both EGF and insulin stimulation. However, the insulin stimulated phosphosite responded only to the PI3K inhibitor, while the EGF stimulated phosphosite responded only to the MEK inhibitor (Figure 3.5b).

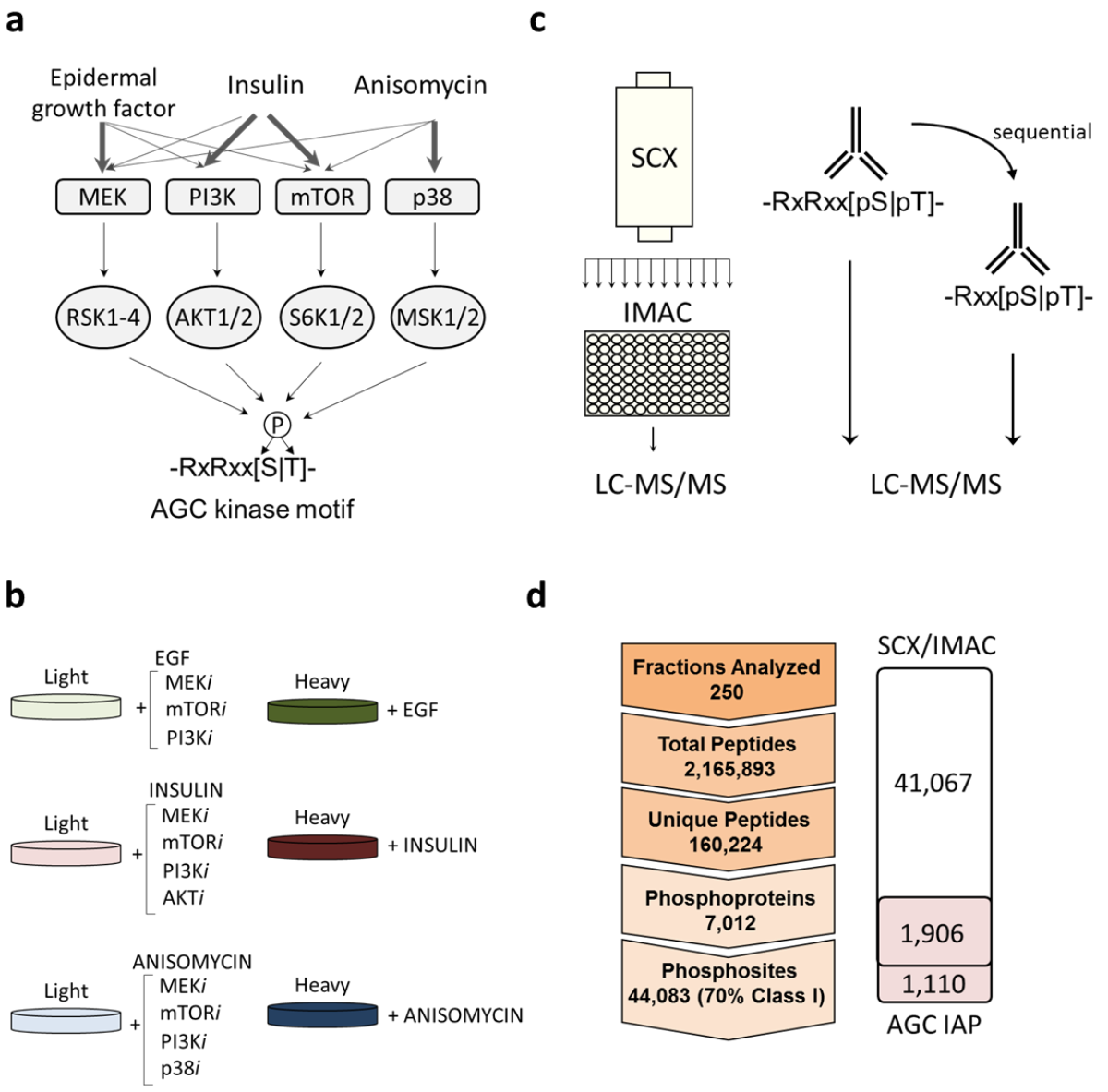


Figure 3.3 A proteomics approach to study AGC kinase signal integration
(a) Depiction of arginine-directed protein kinases downstream of epidermal growth factor, insulin, and anisomycin. **(b)** SILAC labeling and systematic perturbation strategy to dissect signal integration upstream of AGC kinase motifs. MEKi: U0126, mTORi: rapamycin, PI3Ki: wortmannin, AKTi: AKT VIII, p38i: SB20358 **(c)** Sample processing work-flow illustrating SCX/IMAC and AGC Motif IAP strategies for phosphopeptide enrichment. **(d)** Summary of results and overlap between AGC IAP and SCX/IMAC strategies. Total peptides and phosphosites were filtered to <1% FDR.

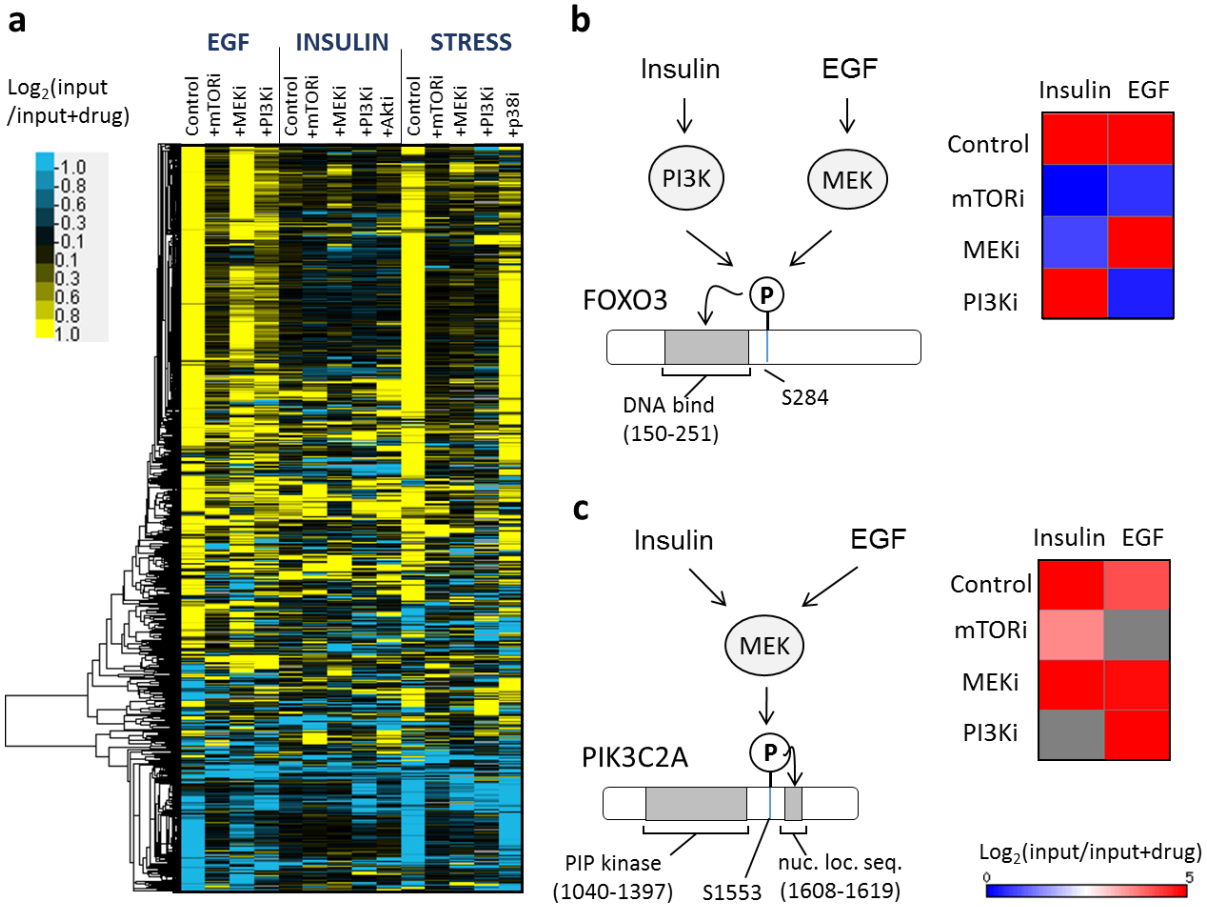


Figure 3.5 Signal integration is a widespread phenomenon
(a) Hierarchical clustering of Log_2(input/input+drug) ratios. **(b)** Example of downstream signal integration towards FOXO3 pS284. **(c)** Example of upstream signal integration towards PIK3C2A pS1553. Protein domain topology schematics are shown to indicate putative regulatory mechanisms.

Thus, two independent signaling pathways exist. PIK3C2A pS1553 was also sensitive to both EGF and insulin stimulation, but in both cases the response was blocked by a MEK inhibitor (Figure 3.5c). In this case, the signals are integrated somewhere upstream in the pathway between the receptors and MEK, and presumably the same downstream kinase(s) directly phosphorylate the site.

3.3.4 Identification of a second regulatory phosphorylation site on glycogen phosphorylase

We uncovered a novel phosphorylation event at serine 430 on glycogen phosphorylase (GPh), which was the first phosphoprotein characterized (Fischer and Krebs, 1955). GPh is a polymerase enzyme which is subjected to unique control mechanisms. Operating in the direction of glycogen synthesis, GPh requires the presence of its product (glycogen) to provide a template for chain extension, yet in the absence of regulation, glycogen would be degraded as it is produced, a futile cycle. GPh reversibly converts glucose-1-phosphate into glycogen in a manner that depends on both allosteric and covalent modification. The novel phosphorylation site lies within a short, disordered, and highly conserved segment within the glycogen binding domain (Figure 3.6a). It was stimulated by insulin and blocked by AKT inhibition (Figure 3.6b). It contains an AKT consensus motif (RxRxx[S|T]) that is conserved through all metazoans, consistent with a role in the insulin pathway (Figure 3.6c). More interestingly, this motif is subject to rare human genetic variation. The arginine in the -5 position is mutated with an allele frequency of almost 0.1% (Figure 3.6d). Thus, this control event is likely disrupted in several hundred thousand people in the United States alone.

Due to its potential relevance in diabetes, we performed follow-up experiments to confirm the functional relevance of this event. We validated that the site is a direct substrate of AKT *in vitro* (Figure 3.6e-f) and we set up an *in vitro* assay to test its enzymatic activity (Figure B.2.2), which

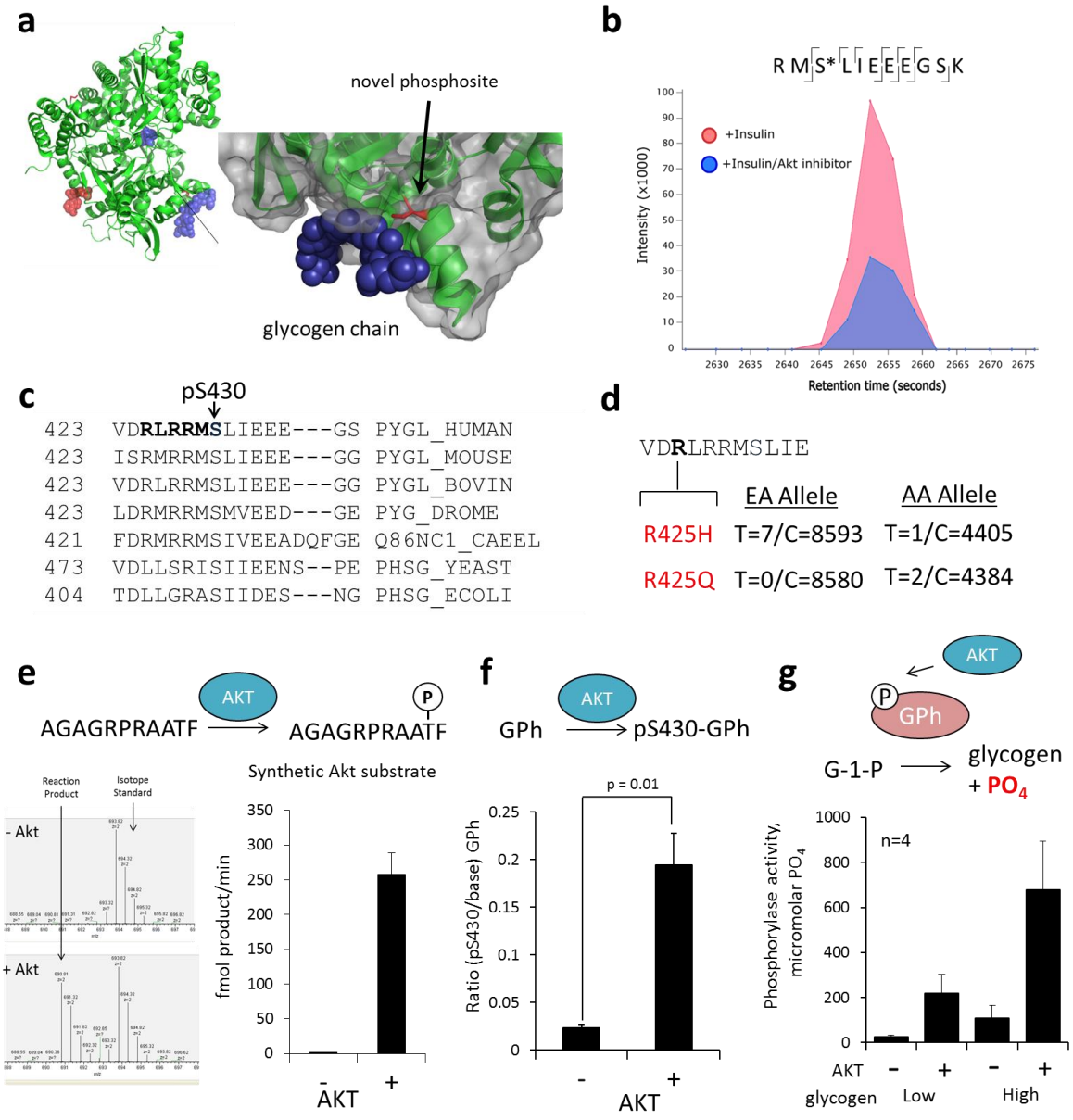


Figure 3.6 Identification of a second regulatory site on glycogen phosphorylase
(a) Localization of novel phosphorylation site within GPh glycogen binding domain. **(b)** GPh S430 is phosphorylated in response to insulin and blocked by AKT inhibition in HeLa cells **(c)** The serine is conserved across all species and the AKT consensus motif is conserved across all metazoans. **(d)** The arginine at the -5 position is mutated in approximately 0.1% of the human population. EA: European ancestry, AA: African ancestry. Source: Exome Variant Server, University of Washington. **(e)** AKT activity is measurable using mass spectrometry. **(f)** AKT phosphorylates pS430-GPh *in vitro.* **(g)** AKT-induced phosphorylation increases the activity of GPh. Statistical significance was assessed by unpaired t-test.

was increased more than 8-fold after being phosphorylated by AKT at low glycogen concentrations (0.05mg/ml), an effect which was significant but less pronounced at higher glycogen (5mg/ml)(Figure 3.6g). It is thought that the the GPh metabolic pathway is restricted to glycogen degradation, so it is wholly unexpected that it might also have a role in the insulin pathway, which promotes glycogen synthesis. However, before the glycogen synthase pathway was discovered, the leading hypothesis was that GPh accounted for both glycogen synthesis and degradation. Recent experiments have indicated an AKT-dependent, but GSK3-independent pathway for glycogen synthesis (Wan et al., 2013). Phosphorylation of the glycogen binding domain of GPh is a fair candidate for this missing link in the hormonal control of glycogen storage, and it the kinase-binding motif is genetically disrupted in the human population, thus it is likely important for human health. Further studies are necessary to confirm the effects of these mutations on enzymatic activity, glycogen binding, and glycogen metabolism.

3.3.5 A topologically and temporally resolved breast cancer phosphoproteome

The epidermal growth factor (EGF) and insulin-like growth factor 1 (IGF-1) signaling pathways are genetically dysregulated in a majority of human solid cancers and are particularly enriched in breast cancer (>90%) (Shaw and Cantley, 2006; The Cancer Genome Atlas Network, 2012). A precise mechanistic understanding of the topology and dynamics of these pathways is needed and will pave the way for new therapeutics directed towards the specific vulnerabilities of these cells. Here, we designed a high-throughput phosphoproteomics study to characterize the system-wide response of MCF7 breast cancer cells to both EGF and IGF1 over a 60 minute timecourse and in combination with a panel 10 different drugs targeted towards MEK-ERK and PI3K-AKT-MTOR signaling networks (Figure 3.7a). We utilized a high-throughput cell culture and sample processing workflow enabling automated phosphopeptide enrichment (PhosphoHT)(Figure 3.7b). Phosphopeptides were analyzed by mass spectrometry for peptide identification and label-free quantification (Figure 3.7c). PhosphoHT enabled the detection of 24,923 phosphorylation sites on

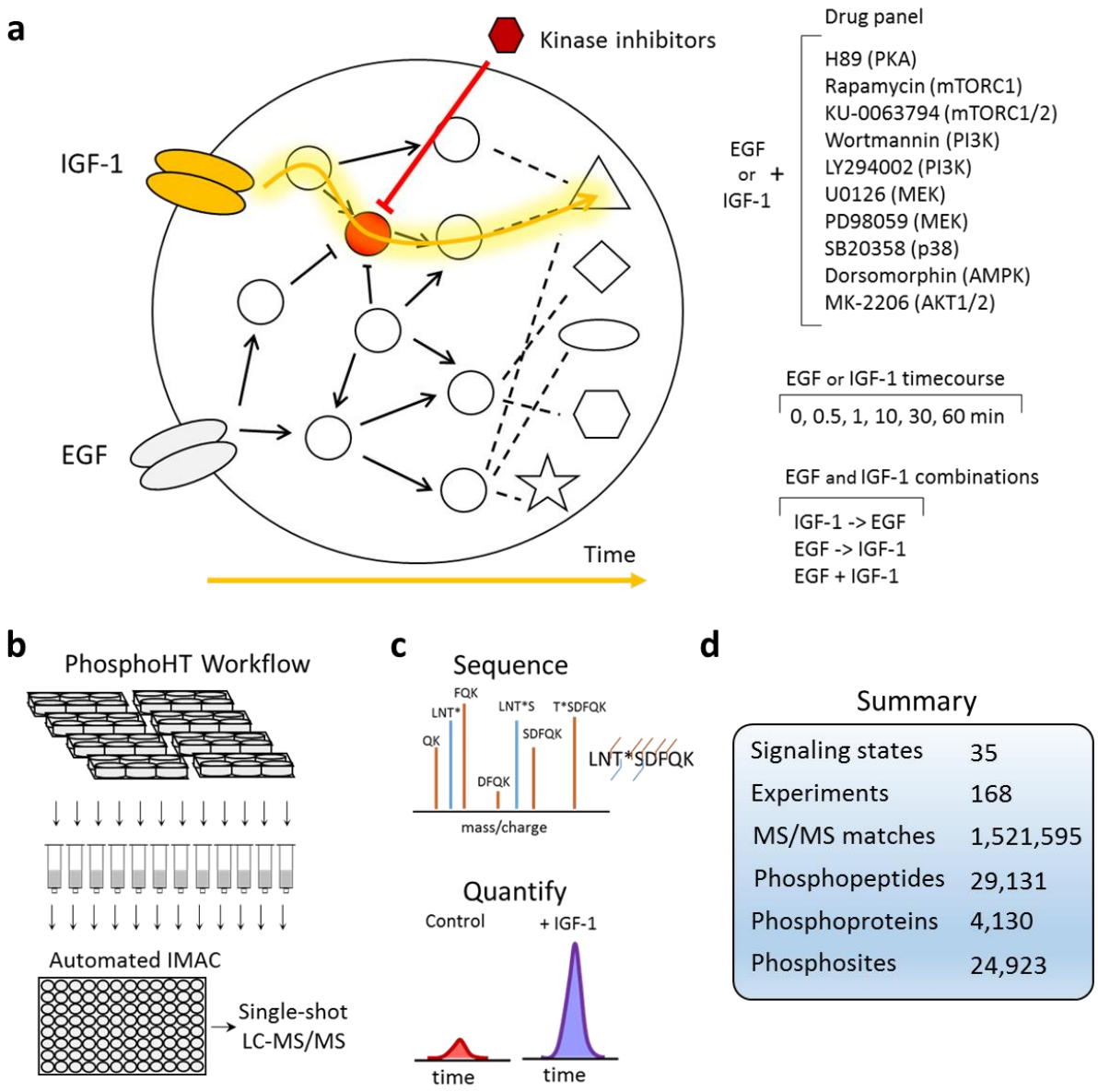


Figure 3.7 Dynamic topological characterization of the breast cancer phosphoproteome
(a) Schematic of EGF/IGF1 stimulation, timecourse, and perturbation strategies. **(b)** High-throughput phosphoproteomics workflow. Cells are cultured in 6-well plates, harvested, desalted and subjected to IMAC using a magnetic bead processor. **(c)** Illustration of how phosphopeptides are sequenced and quantified in the mass spectrometer. (d) Summary of results. MS/MS matches and phosphosites were filtered to FDR<1%.

4,130 proteins (Figure 3.7d), a comparable depth to recent phosphoproteomics studies performed on the same instruments using fractionation approaches. A key advantage of single-shot phosphoproteomics is that it allows up to twelve biological replicates to be performed in the same amount of time as a single fractionation experiment, which increases the accuracy of site quantification.

Phosphorylation sites across all the experiments were analyzed by principal component analysis, which clearly segregated the responses to EGF and IGF1 and the perturbation of these responses by targeted kinase inhibition (Figure 3.8a). PC1 accounted for the major differences between IGF1 and EGF stimulation. Stimulation along PC2 was similar for both EGF and IGF1 and was completely reversed by MEK-ERK inhibition. Stimulation along PC3 was only observed with IGF1 stimulation and with the addition of PI3K-AKT. Interestingly, none of the inhibitors elicited a difference along PC1, which would be necessary to fully reverse EGF/IGF1 signaling. Nonetheless, it is encouraging that the molecular tools at our disposal can control such a large proportion of the signaling within these pathways.

We next used hierarchical clustering to assess the relationships between signals and perturbation (Figure 3.8b). As expected, EGF and IGF1 responses formed distinct clusters, and the drug perturbations clustered according to their molecular targets (e.g. MEK inhibitors clustered together). We detected many PI3K-AKT-dependent and MEK-ERK dependent phosphorylation events (Figure 3.8c). These events could be quantified with high quantitative precision as is demonstrated with the activation site of ERK1 pY204 (Figure 3.8d).

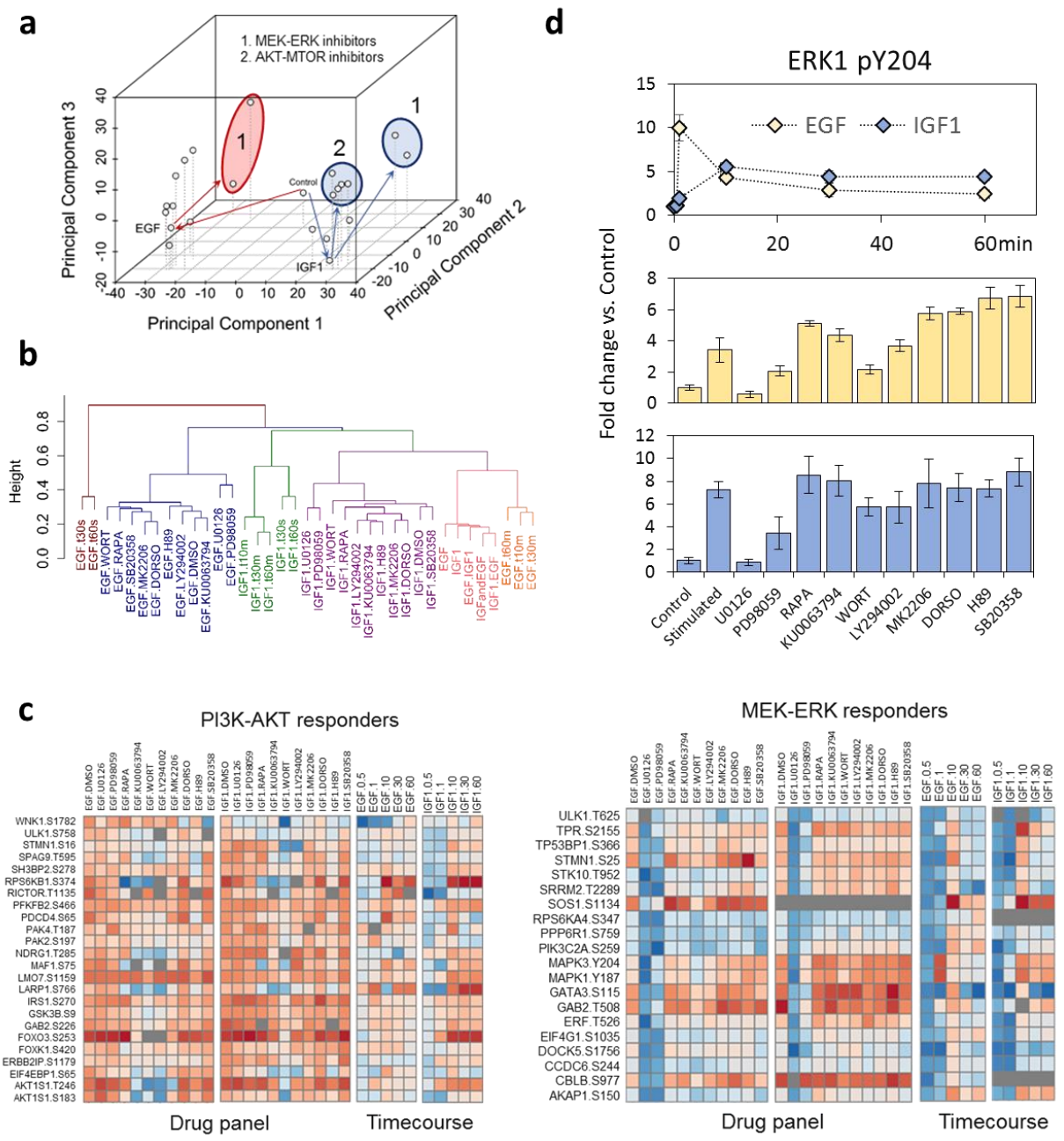


Figure 3.8 Analysis of signaling perturbations downstream of EGF and IGF-1
(a) Principal component analysis of drug perturbation data. **(b)** Hierarchical clustering was used to characterize signaling patterns. **(c)** Heatmap of phosphorylation sites responding to either PI3K-AKT inhibition or MEK-ERK inhibition. Data input for (a-c) were $\log_2$(perturbed/control) ratios. **(d)** Quantification of ERK1 pY204 across both time courses and drug screens.

3.4 Discussion

Global ‘top-down’ analysis of the phosphoproteome response to diverse stimuli and perturbations revealed unexpected twists and turns to well-characterized signal transduction systems. In the first study, we demonstrate that cross-talk is a ubiquitous phenomenon, and that when examining signaling from the ‘top-down’, many new relationships between signals emerge. In HeLa cells, IGF1 elicited an AKT signaling response but not an EGF response, and was not similar on a global level to the insulin response. TNF-alpha elicited a response most similar to mitogens such as EGF and PMA. Insulin shared many phosphorylation events with isoproterenol. Based on a more limited subset of data, as found in the majority of signaling studies (i.e. <10 phosphorylation sites) one would most likely draw completely different conclusions. In the second study, we use perturbations to confirm that in fact, distinct phosphorylation pathways can lead to the same phosphorylation site depending on which stimulus was applied. Both kinases and their substrates are promiscuous.

In MCF7 cells, we demonstrated that IGF1 and EGF signaling were highly similar. This is different from what was observed in HeLa, where IGF1 had a less pronounced effect on mitogenic pathways than EGF. EGF is thought to more strongly affect the MEK-ERK pathway and only slightly activate PI3K-AKT signaling, and vice versa for IGF1. However, we found the MEK-ERK pathway was strongly activated in both cases, and the PI3K-AKT pathway was only slightly more sensitive to IGF1 stimulation.

3.5 Experimental Procedures

<u>HeLa cellular signaling network</u>

HeLa cells were passaged in DMEM with 4.5 g/L glucose, penicillin-streptomycin, 10% FBS at 37°C, 5% CO_2. For deep fractionation experiments, cells were split into 15cm plates and one plate was used for each experiment. For single-shot experiments, cells were split into 6-well plates,

and 3 experiments were performed per plate (technical duplicate stimulations). For all experiments, cells were serum-deprived for 4 hours prior to stimulation according to Table B.2.1. Proteins were lysed in urea buffer, digested with trypsin and desalted on tC18 SepPaks. From here, single-shot samples were enriched directly by IMAC or subjected to SCX using a volatile buffer system and dried prior to IMAC. All samples were analyzed on an EASY nLC-II coupled to a Velos-Orbitrap mass spectrometer with a data dependent acquisition strategy. Raw files were searched with Comet against the human SwissProt database allowing for phosphorylation of S, T, or Y, filtered using percolator, site-localized using Ascore, and quantified using an in-house peak area integration algorithm.

AGC protein kinase signal integration

HeLa cells were passaged in DMEM (-Arg,-Lys) with penicillin-streptomycin, 10% dialyzed FBS at 37°C, 5% CO_2, supplemented with either normal L-lysine and L-arginine (light K0, R0) or $^{13}C_6$,$^{15}N_2$-lysine and $^{13}C_6$,$^{15}N_4$-arginine (heavy K8, R10). Both populations of cells were deprived of serum overnight. 'Light' labeled cells were treated with inhibitor for 30 min prior to stimulation with for an additional 30 min. 'Heavy' labeled cells were stimulated for 30 min. Cells were lysed in ice-cold urea buffer. Protein concentration was assayed using the BCA method and lysates from 'light' and 'heavy' cultures were mixed in a 1:1 ratio and digested with Lys-C. For SCX/IMAC peptides were additionally digested with trypsin and separated using a volatile buffer system. For IAP, Lys-C peptides were enriched sequentially using RxRxx[pS|pT] resin followed by Rxx[pS|pT] resin and digested with trypsin. All samples were analyzed on an EASY nLC-II coupled to a Velos-Orbitrap mass spectrometer with a data dependent acquisition strategy.

Dynamic MCF7 phosphoproteome

MCF7 cells were passaged in DMEM with 4.5 g/L glucose, penicillin-streptomycin, 10% FBS at 37°C, 5% CO_2. For the PhosphoHT workflow, cells were split into 6-well plates, and 3 experiments

were performed per plate (technical duplicate stimulations). For all experiments, cells were serum-deprived for 4 hours prior to stimulation. For the timecourse experiments, cells were stimulated with 100ng/ml IGF1 or EGF for 30 sec, 1min, 10min, 30min, or 60min. For the drug panel experiments, control cells were treated with 1.5ul DMSO, and 10,000x stocks of drugs prepared for a final dosage of 10x manufacturer reported IC50 values. Proteins were lysed in urea buffer, digested with trypsin and desalted on tC18 SepPaks. From here, single-shot samples were IMAC-enriched in multiplex using a magnetic bead processing robot (Thermo KingFisher). All samples were analyzed on an EASY nLC-II coupled to a Velos-Orbitrap mass spectrometer with a data dependent acquisition strategy.

Additional methods can be found in the accompanying supplementary materials (B.1).

Chapter 4. Targeting the phosphoproteome

This chapter is based on the following published article:

Robert T Lawrence, Brian C Searle, Ariadna Llovet and Judit Villén. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nature Methods. 2016.

4.1 Summary

Systematic approaches to study cellular signaling require phosphoproteomic techniques that reproducibly measure the same phosphopeptides across multiple replicates, conditions, and time points. Here we present a method to mine information from large-scale, heterogeneous phosphoproteomics datasets to rapidly generate robust targeted mass spectrometry assays. We demonstrate the performance of our method by interrogating the IGF-1/AKT signaling pathway, showing that even rarely observed phosphorylation events can be consistently detected and precisely quantified.

4.2 Introduction

Each human cell harbors a signaling landscape likely spanning hundreds of thousands of phosphorylated residues (Ubersax and Ferrell, 2007). Investigating how these residues are dynamically engaged to control cell behavior in the context of time, environment, cellular identity, and genetic variation requires systematic phosphoproteome analysis using high-throughput assays that are accurate, sensitive and reproducible. Measuring phosphorylation events in a targeted manner presents many hurdles. To date, it has been only achievable after tedious assay optimization and reliance on synthetic peptide standards (Gerber et al., 2003; Soste et al., 2014; de Graaf et al., 2015), which impedes assay versatility and limits widespread adoption of the technique by researchers outside the proteomics community. Our goal was to develop the capability to easily generate 1-hour “plug-and-play” targeted phosphoprotein assays equivalent in

sensitivity to prolonged deep fractionation experiments (>12-hr analysis time) with more reproducible sampling and quantification.

Much of the work on cellular signaling using mass spectrometry based proteomics has focused on phosphorylation site discovery, generating vast catalogues of novel phosphorylation events and their regulation (Huttlin et al., 2010; Lundby et al., 2012; Rikova et al., 2007; Sharma et al., 2014). This workflow generally employs a data-dependent acquisition (DDA) strategy, in which the most abundant features in each full MS scan are selected for MS/MS fragmentation and identification. However, one of the major problems that has plagued quantitative proteomics using DDA is stochastic sampling, which leads to extensive but sparse datasets that have many missing values across different experimental conditions (Bantscheff et al., 2007). Analysis of phosphopeptide-enriched samples is further complicated by their high dynamic range, limiting the sensitivity and reproducibility of DDA. Recently, more systematic and sensitive data acquisition strategies have emerged to meet these analytical challenges, including data independent acquisition (DIA) and parallel reaction monitoring (PRM). In DIA, MS/MS scans are acquired across the full mass range each duty cycle (Gillet et al., 2012; Venable et al., 2004). In PRM, MS/MS scans are targeted towards narrow prespecified mass and time windows corresponding to analytes of interest (Peterson et al., 2012). Compared to selected-reaction monitoring (SRM), which has been the workhorse of targeted proteomics, PRM simplifies the targeted mass spectrometry workflow. All one needs to specify to configure an assay is the precursor mass-to-charge ratio (m/z) and the expected retention time, but no optimization is required *a priori*. Potential interferences can be identified and fragment ions quantified *post-hoc*.

The promise of targeted quantitative phosphoproteomic analysis has been demonstrated in several recent studies(Soste et al., 2014; de Graaf et al., 2015; Parker et al., 2015). However, selecting the best peptide sequence and charge state to monitor for phosphorylation sites still represents a major obstacle to targeted analysis. And because protein phosphorylation is site-

specific, selection of MS-compatible peptide sequences is limited by the local sequence composition and protease enzyme used for digestion. Notably, phosphorylation alters the local charge distribution, which interferes with routinely used enzymes like Lys-C and trypsin, further hampering peptide selection (Dickhut et al., 2014; Giansanti et al., 2015). Thus, the preferred peptide cleavage and charge state are difficult to predict *a priori*.

4.3 Results

Here, we instead relied upon a large-scale database of previously observed human phosphopeptide sequences. We assembled this database by searching nearly 1,000 LC-MS/MS runs from human label-free trypsin-digested phospho-enriched samples. The samples were derived from a variety of human cell lines exposed to many different stimuli and processed using different phosphopeptide enrichment methods and single-shot as well as deep offline fractionation techniques. More than two thirds of the data (727 runs) were collected in-house. We additionally searched 262 LC-MS/MS runs from three other groups (Sharma et al., 2014; de Graaf et al., 2014). Overall, we identified more than 7.5 million phosphopeptide spectral matches (PSM-level FDR < 1%) corresponding to 109,611 phosphorylation sites (90,103 localized $p < 0.05$) on 11,428 proteins (phosphosite-adjusted FDR < 5%), commensurate with the human phosphoproteome coverage provided in resources such as PhosphoSitePlus™ (Hornbeck et al., 2015) (Figure 3.1a, Figure C.3.1).

We used the database to quantify several key parameters of data-dependent phosphoproteome analysis that we hypothesize can be addressed with targeted analysis. Without fractionation, DDA is limited in sensitivity. In line with our expectations, 64% of phosphopeptides we identified were only observable in experiments that used extensive fractionation prior to phosphopeptide enrichment (Figure 3.1b). Phosphopeptide sampling stochasticity is detrimental in both single-

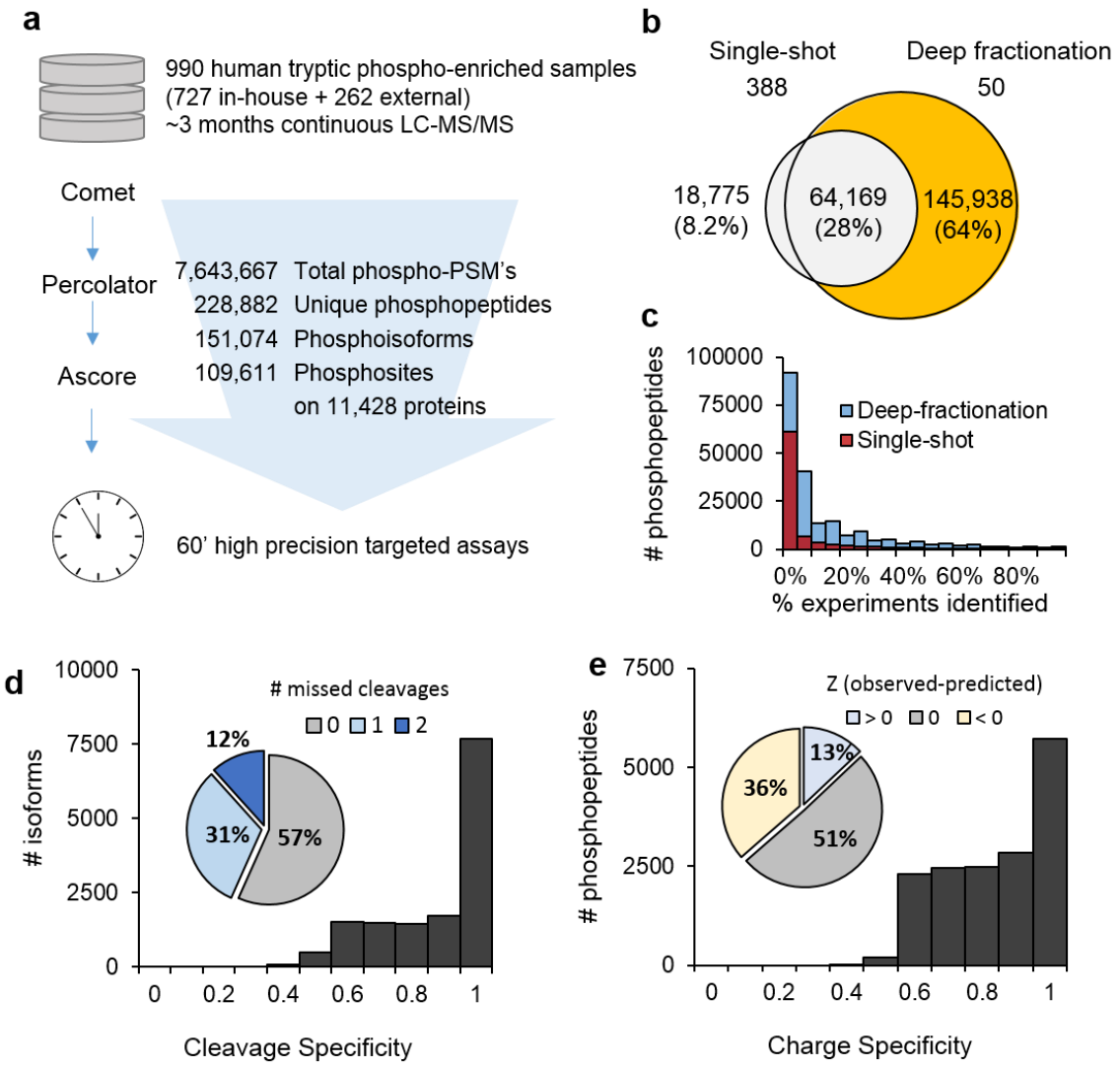


Figure 4.1 A database for targeted human phosphoproteome analysis.
(**a**) Data analysis pipeline (Online Methods) and summary statistics. Phosphopeptide spectral matches (PSM's) were filtered to FDR < 1%. Phosphoisoforms refer to unique protein phosphorylation states (i.e. phosphopeptides representing multisite phosphorylation reported independently from singly phosphorylated species). Phosphosites refer to total unique protein phosphorylation sites (90,103 were confidently localized, Ascore ≥ 13). To account for data aggregation, unique phosphopeptides, phosphoisoforms, and phosphosites were each additionally filtered to reach an aggregate FDR < 5%. (**b**) Comparison of phosphopeptides identified in single-shot experiments (n=388) versus deep fractionation experiments (n=50). (**c**) Reproducibility of phosphopeptide sampling across experiments. (**d**) Cleavage form specificity (counts of most frequently observed cleavage state/total counts) and distribution. For specificity calculation, only phosphoisoforms observed at least 100 times were analyzed (n=14,480). The pie chart represents the distribution of the most frequently observed number of miscleavages for each phosphoisoform in the database. (**e**) Charge state specificity (counts of most frequently observed charge state/total counts) and distribution. For specificity calculation, only phosphopeptide sequences observed at least 100 times were analyzed (n=15,985). The pie chart represents the distribution of the most frequently observed charge state versus predicted (positively charged amino acids + 1) for each phosphopeptide in the database.

shot and deep fractionation routines. Sample fractionation increases the depth of phosphoproteome coverage (on average, 5,834 unique phosphopeptides were identified per single-shot experiment versus 32,311 per fractionation experiment), but the overlap between samples was still lower than expected. For single shot experiments only 2,544/82,944 (3%) phosphopeptides identified were observable in at least 50% of experiments compared with 16,677/210,107 (8%) for fractionation experiments (Figure 3.1c).

Next, we examined phosphopeptides sequenced with deep coverage (identified a minimum of 100 times) to ascertain the distribution of charge and cleavage state specificity (preferred state/total observations). We found that most phosphosites were predominantly observed in only one cleavage state (Figure 3.1d), and that charge state was moderately specific (Figure 3.1e). The preferred phosphopeptide forms were both fully cleaved and of the expected charge state only 35% of the time. As expected, we observed a significant number of missed cleavage sites (Figure C.3.2), a small fraction (10%) of which rescued peptides that would otherwise be too short for analysis. We also identified ~9,000 phosphoisoforms mapping to protein N-terminal clipping and/or acetylation. Prediction of the most frequent charge was wrong 50% of the time (Figure 3.1e), with many of the ions predicted by heuristics falling outside the optimal mass range of the mass spectrometer (Figure C.3.2). Lastly, we found that the preferred peptide sequence was the same between at least 3 of the 4 datasets 87% of the time (Figure C.3.2), suggesting that the preferred phosphopeptide sequences provided in our database should be compatible with most laboratory trypsin digestion and phosphopeptide enrichment protocols.

We hypothesized that leveraging a large-scale database of previously observed phosphopeptides would enable rapid assay deployment with higher success rates than traditional approaches. We evaluated the precision of the retention time scheduling by using DDA analysis of BSA or phospho-enriched tryptic digest to predict the retention times of independent phosphopeptides in a subsequent DDA run, and established that when using a complex phosphopeptide mixture for

assay calibration, 5-minute retention time windows are sufficient to capture 95% of the targets (Figure 3.2a). Next, we targeted pairs of phosphopeptides in which the sequence and charge state selected using heuristics differed from the database selection. Our data-driven selection approach outperformed heuristics (Figure 3.2b, Figure C.3.3) by magnitudes to similar to what was predicted by our previous analysis of sequence and charge specificity (Figure 3.1d,e).

We further benchmarked our method by selecting 101 phosphopeptides spanning a wide range of detectability and configuring a 1-h parallel reaction monitoring assay to detect those peptides in a phospho-enriched tryptic digest of MCF7 breast cancer cells treated with a cocktail of insulin-like growth factor-1 (IGF-1), epidermal growth factor, and pervanadate. We analyzed the same sample using PRM, DIA, and DDA strategies in technical quadruplicate. The PRM and DIA results were analyzed in a targeted 'peptide-centric' manner querying specifically for the target phosphopeptides as well as in a 'spectrum-centric' manner using a database search pipeline. Using the PRM method, we readily detected several species that were only sparingly detected in our phosphopeptide database, such as the peptide corresponding to the activation site of tyrosine protein kinase SYK (Figure 3.2c). Measured retention times correlated well with the retention times in the database (R^2>0.99) enabling efficient assay scheduling and interpretation (Figure C.3.4a). Out of the 101 targets, PRM was superior to DIA and DDA 1-h assays in terms of the number of peptides detected and sampling reproducibility (Figure 3.2c, Figure C.3.4). It has been suggested that 'peptide-centric' targeted analysis might offer advantages over the traditional 'spectrum-centric' approach since it more directly evaluates the evidence for a given peptide (Röst et al., 2014; Ting et al., 2015). In our analysis, peptide-centric analysis was more sensitive than database searching for both PRM and DIA, and the signal (if measurable) was consistently detectable across all 4 runs (Figure 3.2c and Figure C.3.4).

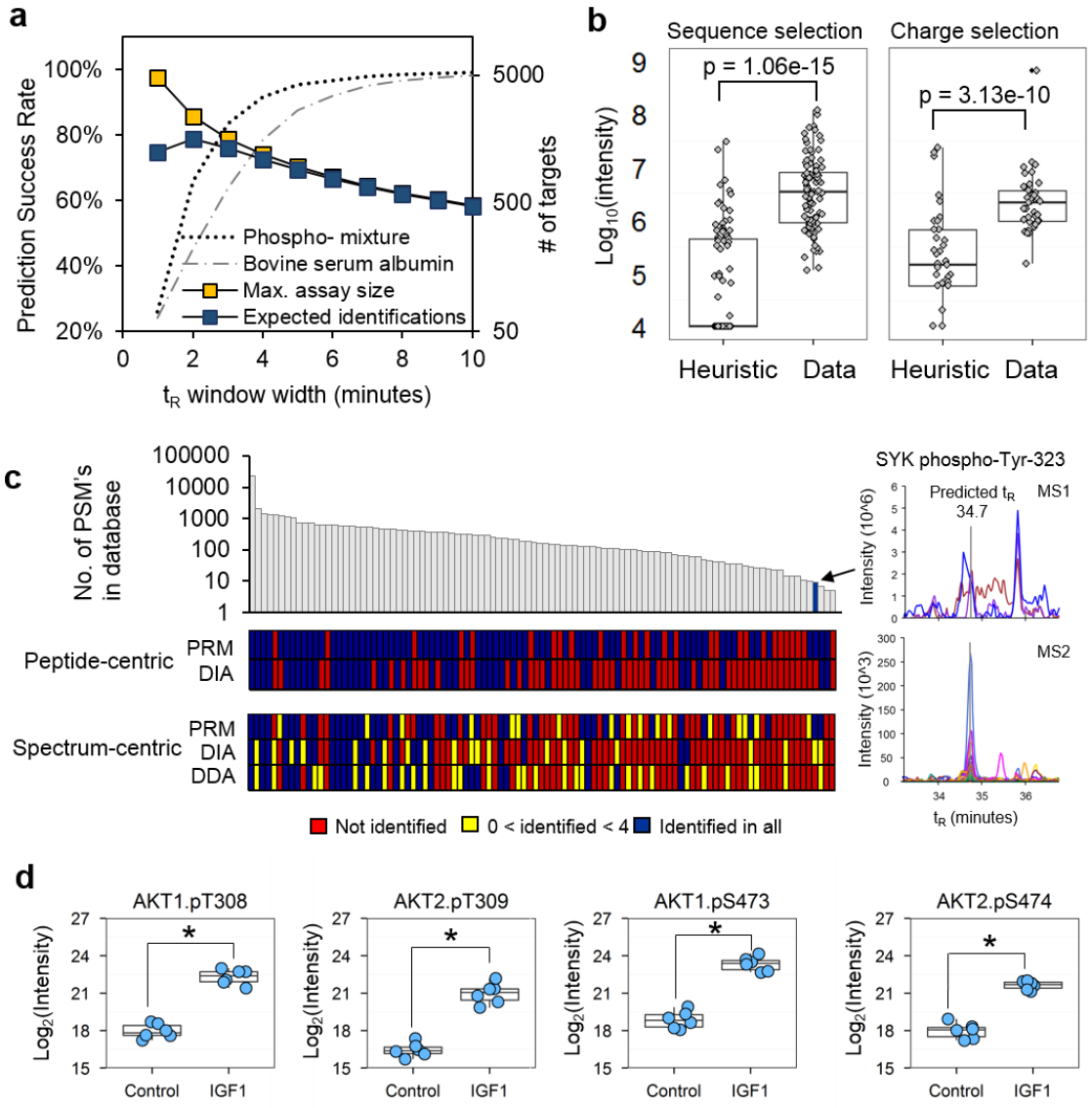


Figure 4.2 Plug-and-play assay performance.
(**a**) Retention time prediction performance. “Prediction Success Rate” refers to the percentage of phosphopeptides in the validation set that were identified within the predicted retention time windows. “Max. assay size” is based on 80 concurrent targets scheduled optimally over 60 minutes. “Expected identifications” represents the success rate using the “phospho-mixture” multiplied by the max. assay size. (**b**) Phosphopeptide sequence (n=100) and charge state (n=50) selection performance. The reported intensity represents the summed peak areas of all identified y-ions and unidentified targets are imputed at noise level (1e4). Significance was assessed using a Wilcoxon signed rank (paired non-parametric) test. (**c**) PRM, DIA, and DDA analysis of 101 targets in phosphopeptides purified from IGF-1/EGF/pervanadate stimulated MCF7 cells. Peptides are ordered from most to least frequently observed in the database and PSMs are indicated in the y-axis. Peptide centric analysis refers to targeted analysis using Skyline. Spectrum centric analysis refers to a database search using a Comet-Percolator-Ascore pipeline. Right panel shows PRM analysis of a phosphopeptide (R.QESTVSFNPY*EPELAPWAADKGPQR.E) rarely observed by DDA. The top right panel shows the MS chromatogram of the 3 precursor isotopes and the bottom right panel shows the MS/MS chromatograms of all identified fragment ions. (Y*: phosphotyrosine). (**d**) Targeted site and isoform specific quantification of AKT1/2 phosphorylation at T308/309 and S473/474 (n=6). *$p < 1x10^{-6}$, unpaired t-test. For boxplots in (b) and (d), the lower and upper edges of the box correspond to the boundaries of the first and third quartiles. The "whiskers" extend to the most extreme value within 1.5 x interquartile range.

Lastly, we designed and implemented a targeted assay to quantify phosphorylation sites on proteins within the IGF-1/AKT signaling pathway in MCF7 cells before and after stimulation with IGF-1. Our assay enabled reproducible isoform-specific quantification of protein kinase AKT1/2 activation via phosphorylation at T308/309 and S473/474 (Figure 3.2d). The specific isoforms of AKT are thought to have distinct roles in cellular signaling, but the respective kinase activation sites T308/T309 are not distinguishable using specific antibodies due to nearly identical local sequence composition. They are sparingly detectable by DDA even after deep fractionation but were reproducibly detected using PRM in the single stage enrichment protocol used here.

In addition to its advantages of sensitivity and reproducibility, PRM has the capability to monitor isobaric peptide species, which are ubiquitous in phosphorylated proteins (Figure C.2.5). These positional isomers can often be resolved by retention time, allowing for more accurate quantification (Figure C.2.5).

4.4 Discussion

Overall, we demonstrate the potential of label-free PRM assays for robust, high-throughput, targeted phosphoproteome analysis. In order to facilitate “plug-and-play” assay development, we created a web-based application that queries our database for optimal peptide selection and retention time scheduling (https://phosphopedia.gs.washington.edu). This application provides several tools for assay development, including pre-curated lists of phosphosites, information for sequence and charge state selection, an MS/MS spectra viewer, retention time calibration, automated variable window assignment for positional isomers, and dynamic schedule visualization and optimization. Using this tool, targeted phosphoproteomic assays are convenient to configure, sensitive enough to detect low-abundance analytes without sample fractionation, and more reproducible than DDA. The use of label-free targeted quantification in conjunction with data-driven peptide selection enables rapid deployment of assays to measure virtually any known

phosphorylation event in human specimens. These qualities make the method suitable for interrogating the diverse dimensions of the cellular signaling landscape with high throughput and versatility.

Accession codes

Raw MS data for the experiments performed in this study are available at MassIVE (MSV000079423) and ProteomeXchange (PXD003344).

Chapter 5. Towards a comprehensive understanding of signal transduction systems

In this final chapter, I discuss the signal transduction paradigm in the context of the work presented in this dissertation, how it has changed, and how it must continue to change to reflect the experimental observations now possible using data-driven systems biology approaches. The observations and methods presented in this dissertation offer only a suggestion of what is to come, but clearly demonstrate the need to continue working in this direction.

Initially, my work was motivated by two general questions. First, what is the diversity of signal transduction networks *between* cells? To begin to answer this question I measured the proteomes of a panel of cell lines all originating from human breast, which should have relatively similar levels of protein expression compared to cells from different tissues of origin. Even in these cells, differences in protein expression were remarkable, typically varying by greater than ten-fold. In the signaling space, some pathways had more variable expression than others. For example, the canonical Ras-Raf-MEK-ERK pathway was invariant and expressed at similar levels across the panel. However, expression within the PI3K-AKT-MTOR pathway and others was extremely variable. Cell surface receptor expression also varied dramatically. Interestingly, receptors often have affinity for many ligands (e.g. EGFR), and different isoforms of the receptor that binds those ligands have different downstream signaling (e.g. EGFR vs. ERBB2, ERBB3). The combination of receptor and kinome expression diversity allows cells to elicit very different signaling responses to the same stimuli and likewise to elicit a similar response to very different stimuli.

Second, what is the diversity of signal transduction networks *within* cells? To begin to answer this I designed a large-scale study of the phosphoproteome of HeLa cells exposed to sixteen different "environments." Nearly every protein (>10,000) that has been identified in HeLa was found to be phosphorylated. Some stimuli had more widespread effects on protein phosphorylation than

others, but one overarching principle was that the majority of phosphorylation sites were activated by multiple stimuli. How are these signals integrated to reach the same protein positions? Signal integration allows the transmission of signals that require similar functional outcomes to flow through common pathways. The protein kinase superfamily encodes around 500 enzymes, and the large branches of the family all have similar target sequence specificity. Based on this fact alone, it is unlikely that any given phosphorylation site has just one upstream kinase, but signal integration can occur anywhere between the sensor and the target. I demonstrated this principle using a systematic stimulation and perturbation strategy in HeLa cells. I elaborated on this by investigating cross-talk between IGF1 and EGF signaling in MCF7 cells.

Many implicit narratives have been constructed around the signaling ‘pathways’ and ‘cascades’ downstream of hormones, growth factors, and cytokines. While the studies in this dissertation did recapitulate known protein expression profiles and phosphorylation events, these concepts have probably led the field of signal transduction astray in certain ways. In light of systems biology, the taxonomy of signaling could use some refinement. First, signaling pathways ought not to be named for their receptor. The ‘insulin pathway’ in one cell type might be the ‘EGF pathway’ in another cell type, and the downstream substrates are likely different depending on which effectors are expressed by the cell. Second, phosphorylation sites should not be considered as exclusive kinase substrates. Virtually no amount of evidence is enough to confirm such a conclusion, because the upstream kinase can be different in another cell type or even a different stimulus.

The signal transduction paradigm is now changing. A ‘top-down’ approach to study signaling (nowadays referred to as systems biology) was attractive from the beginning, but, as outlined previously, lacked appropriate technologies. Thus, our current body of knowledge must be somewhat biased by ‘bottom-up’ analytical strategies. On one hand, ‘bottom-up’ approaches were indispensable, leading us to discover the key principles of signal transduction. One the other, these early experiments were performed in a limited number of contexts, neglecting cell-type and

stimulus-specific effects. System level data, e.g. phosphoproteomics, brings unexpected signaling events into the light. In a traditional, hypothesis-driven experiment, these events would never be measured and in many cases it would be impossible to do so due to the lack of appropriate reagents. Proteome-wide analysis of protein phosphorylation is now becoming more routine, with mass spectrometry-based technologies in command. It is now possible to rapidly characterize cellular signaling systems with high-throughput. I will briefly outline several areas that must be addressed going forward.

I. Phospho-regulatory diversity from cells to organisms

Comparative phosphoproteomics is a particularly exciting area. The way that our diverse cellular communities talk to each other is key to what makes us human. How do the thousands of different cell types respond to stimuli? Much recent work, including the studies presented in this dissertation, have focused on the wiring of protein kinases and their substrates. Within the next couple of years, I predict these networks will be characterized in extensive detail through large-scale experimentation involving systematic combinations of different cell types, organisms, stimuli, pharmacological/genetic perturbations. Some of these experiments are already underway.

II. The uncharacterized phosphoproteome

Out of the >100,000 phosphorylation sites we have observed, less than 5% have been functionally characterized, speaking liberally. More realistically, only a few hundred are really well understood. The function of many proteins remains unknown, let alone their phosphorylation sites. If the goal is to understand how complex cellular behavior is controlled, it is essential to understand how phosphorylation sites are linked to phenotype. That will pave the way for therapeutic modulation or even engineering synthetic signal transduction circuits. High-throughput strategies for site-directed mutagenesis and functional profiling will likely be helpful for this task.

III. The complexity of cellular microenvironments

All of the experiments presented in this dissertation were performed in simplified experimental conditions, but such is not the reality of the local cellular environment. *In vivo*, cells are exposed to many different signaling molecules simultaneously as they converse with other cell types both near (paracrine) and far (endocrine). For example, how does EGF action depend on the presence of cytokines, and vice-versa? How do the concentrations of these molecules vary and how are the sensitivity and dynamics of signal transduction systems programmed to respond to these differences? Furthermore, there are hundreds of uncharacterized signals including secreted proteins, peptides, and small molecules. The function of many of these molecules is completely unknown.

5.5 Concluding remarks

This dissertation contributes to a rich history of signaling research, but many questions still remain. Advances in mass spectrometry-based proteomics have revolutionized the study of signal transduction in recent years. As the broader community begins to harness these advances, a more comprehensive understanding of phosphorylation-dependent signaling networks seems within reach. The data and methods presented here should be a valuable resource for future work in this area.

Appendix A. Supplementary Material for Chapter 2

A.1 Supplementary Experimental Procedures for Chapter 2

A.1.1 Cell Culture

Triple negative breast cancer cell lines were purchased from ATCC (American Type Culture Collection, Manassus, VA). MCF7, SKBR3, BT474, and MCF10A cells were obtained from Dr. Hanna Irie (Mt. Sinai Hospital). MCF10A were grown in DMEM-F12 (Gibco) with addition of 5% horse serum (Gibco), 20ng/mL EGF (Peprotech), 0.5mg/mL hydrocortisone (Sigma), 100ng/mL cholera toxin (Sigma), 10µg/mL insulin (Sigma). All other cell lines were grown in RPMI-1640 (Gibco) with addition of 10% fetal bovine serum (Gibco) and penicillin-streptomycin-glutamine (Gibco). Patient tumor specimens were purchased from Indivumed GmbH.

A.1.2 Sample preparation

Cultured cells were washed 3 times quickly with ice cold phosphate buffered saline and flash frozen on liquid nitrogen. They were scraped directly into chilled denaturing buffer containing 50mM Tris pH 8.2, 75mM NaCl, 9M urea, complete EDTA-free protease inhibitor cocktail (Roche), and phosphatase inhibitors (50mM sodium fluoride, 1mM sodium orthovanadate, 10mM sodium pyrophosphate, 50mM β-glycerophosphate) and sonicated on ice for two cycles of 30s each. Tumor tissues were dounce homogenized on ice in the same lysis buffer above prior to sonication. All lysates were centrifuged at 12,000 g for 10 min to pellet insoluble material, the supernatant assayed for protein content using the bicinchoninic acid method and saved for analysis at -80°C. Protein extracts were reduced with 5mM DTT at 55°C and alkylated with 15mM iodoacetamide at room temperature in the dark. Extracts from each sample (25µg) were diluted and digested in solution overnight with either lysyl-endopeptidase (Lys-C) (Wako) or sequencing grade trypsin (Promega). Digestion products were acidified to pH ~2 and loaded directly onto pre-equilibrated stop-and-go-extraction tips constructed in-house from SDB-XC Empore wafers (3M)(Rappsilber et al., 2007). Peptides were desalted and fractionated on the tips by basic reverse-phase using a

step-wise gradient of increasing acetonitrile (5%, 10%, 15%, 25%, 80%) in 0.1% NH_4OH. Finally, fractions were dried by vacuum centrifugation and resuspended in 3% MeCN, 4% formic acid for analysis by LC-MS/MS.

A.1.3 LC-MS/MS

Peptide fractions were injected onto a 40cm x 100μm column packed in-house with 1.9μm Reprosil C18 reverse phase material (Dr. Maisch GmbH), separated by liquid chromatography gradient on an EASY-nLC-1000 (Thermo) equipped with column oven set to 50°C, and analyzed online by tandem mass spectrometry in a hybrid quadrupole-orbitrap Q-Exactive mass spectrometer (Thermo). Mass spectra were acquired in centroid mode using a data dependent acquisition strategy where the twenty most intense precursors were selected for fragmentation, and fragmented ions were excluded from further selection during 40s. Full MS scans were acquired from 300 to 2000 m/z at 70,000 FWHM resolution with a maximum injection time of 100ms and fill target of 3e6 ions. MS/MS fragmentation spectra were collected at 17,500 FWHM with maximum injection time of 50ms using a 2.0 m/z precursor isolation window and fill target of 5e4 ions. Acquisition time for each fraction was 90min, and included column wash and equilibration.

A.1.4 Data processing

Raw spectra were converted to the mzXML open data format and searched using Sequest (release 2012.01.0 of UW Sequest) against a concatenated forward and reverse version of the Uniprot human protein sequence database (v11/29/2012), allowing for up to two missed cleavages, methionine oxidation (+15.9949 Da), and protein N-terminal acetylation (+42.0105 Da). Cysteine carbamidomethylation (+57.0214 Da) was set as a fixed modification. Precursor mass tolerance was set to 50ppm and fragment ion tolerance set to 0.01 Da. Peptide spectral matches for all fractions corresponding to the same sample were filtered to reach a protein identification false discovery rate of less than 1%, resulting in an aggregate peptide-level FDR of less than 0.1%. Peptides were assembled into proteins using parsimony principles (Nesvizhskii

and Aebersold, 2005). Integrated MS1 intensity over time peak areas for identified peptides were calculated using an in-house peptide quantification algorithm. Protein quantifications were calculated using the intensity-based absolute quantitation (iBAQ) approach (Schwanhäusser et al., 2011). For each sample (including all fractions), summed peak areas for all peptides matching to the same protein were divided by the maximum number of observable peptides, and normalized to the total sum intensity of the observed proteome. Common contaminants (e.g. keratins, serum proteins) were excluded prior to normalization. Mass spectrometry data was also analyzed using the standalone MaxQuant platform (Cox and Mann, 2008), for which we obtained similar number of peptide and protein identifications as well as quantification values (data not shown).

A.1.5 Drug screen and curve fitting

Cells were added to 384-well plates at a density of 2,000 cells per well in 50µL of RPMI 1640 containing penicillin-streptomycin using a Matrix WellMate liquid handler (Thermo Scientific), and incubated overnight to allow attachment. Compounds were added (50nL ranging from 5pM to 100µM) to cells using the CyBi-Well Vario Workstation (CyBio) and incubated at 37°C, 5% CO_2 for 96 hours. The final solvent (DMSO) concentration in the assay was 0.1%. Cell viability was measured by luminescence using quantitation of ATP as an indicator of metabolically active cells. CellTiter-Glo reagent (Promega) was dispensed into individual wells with the WellMate following the manufacturer's recommended procedures and, following 20 minutes incubation on an orbital shaker, luminescence was measured on an EnVision Multi-label plate reader (Perkin Elmer). Measurements were corrected for background luminescence and percentage cell viability is reported as relative to the DMSO solvent control. Non-linear curve fitting was performed using MATLAB's 'nlinfit' function. After curve fitting, IC50 values were extracted based on the curve fits, similar to the Cancer Cell Line Encyclopedia (CCLE)(Barretina et al., 2012). External drug sensitivity data (IC50) was downloaded from the "Genomics of Drug Sensitivity in Cancer" resource (Yang et al., 2013), release 2.0 (http://www.cancerrxgene.org).

A.1.6 Bioinformatics

Hierarchical clustering and PCA (principal component analysis) were performed using Cluster 3.0. Average iBAQ values were normalized to a scale from 0 to 1 and filtered for presence in at least 25% of samples and differentially regulated proteins (by S.D.). Unweighted clustering was performed in both dimensions using centered correlation as the similarity metric with centroid linkage. Gene ontology (GO) mapping and enrichment analysis was performed using DAVID 6.7 (Database for Annotation, Visualization and Integrated Discovery)(Huang et al., 2009). The cancer gene census, copy number, and exome data were downloaded from COSMIC (catalog of somatic mutations in cancer)(Forbes et al., 2011). For the cell line SKBR3, mutational data was acquired from CCLE (Barretina et al., 2012). Associations between census gene mutations and protein expression were assessed by using a heteroscedastic unpaired t-test on $\log_{10}$ transformed protein expression values. To generate a network of common gene-protein relationships, we applied a significance threshold ($P < 0.001$) to differences in protein expression that were associated with cancer census gene mutations. These were plotted using Cytoscape version 3.1.0 (Shannon et al., 2003) with a spring-embedded layout. Drug sensitivity associations were assessed by pairwise Pearson's correlation of protein abundance versus inverted IC50 using the 'cor' function in R. Correlation significance was assessed using default settings of the 'cor.test' function in R (a Fisher's Z transformation), and corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

A.1.7 Statistical analysis

Significance tests and correlation analysis were performed using built-in functions within Microsoft Office Excel 2013 or R statistical computing environment version 3.1.0. Gene enrichment significance testing was performed in DAVID version 6.7 using the EASE metric, a modified Fisher's exact test (Huang et al., 2009). All error bars represent standard deviation unless otherwise noted.

A.2 Supplementary Figures for Chapter 2

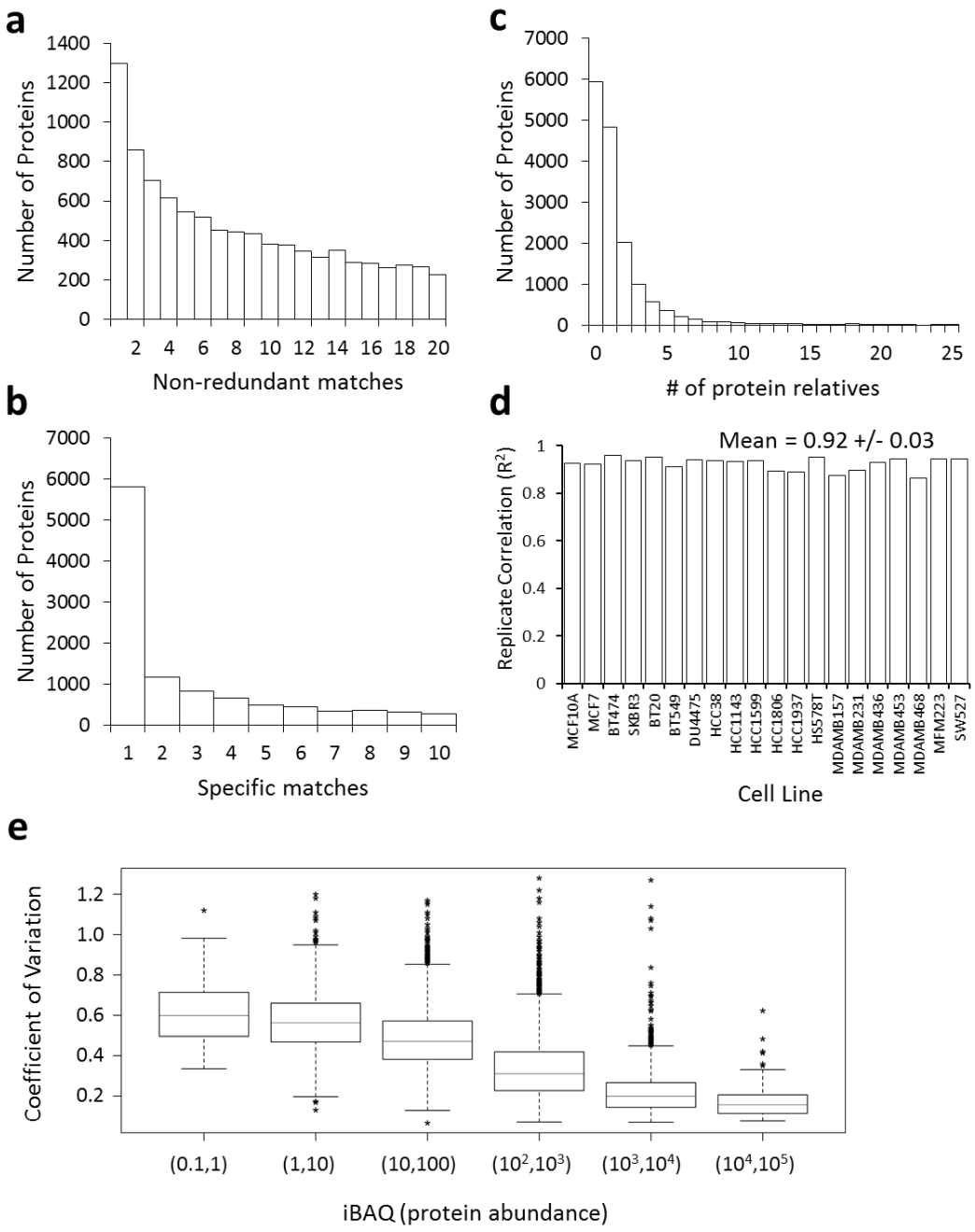


Figure A.2.1 Statistics of protein identification and quantification
(a) Distribution of the number of non-redundant peptide matches (peptides with a unique amino acid sequence) per protein. **(b)** Distribution of the number of specific peptide matches (peptides matching to no other protein in the dataset) per protein. **(c)** Distribution of the number of protein relatives (proteins sharing at least one peptide with one other protein in the dataset) per protein. **(d)** Correlation coefficient between replicates for each cell line. **(e)** Coefficient of variation *versus* absolute protein abundance. Protein abundances were binned as indicated in the x-axis.

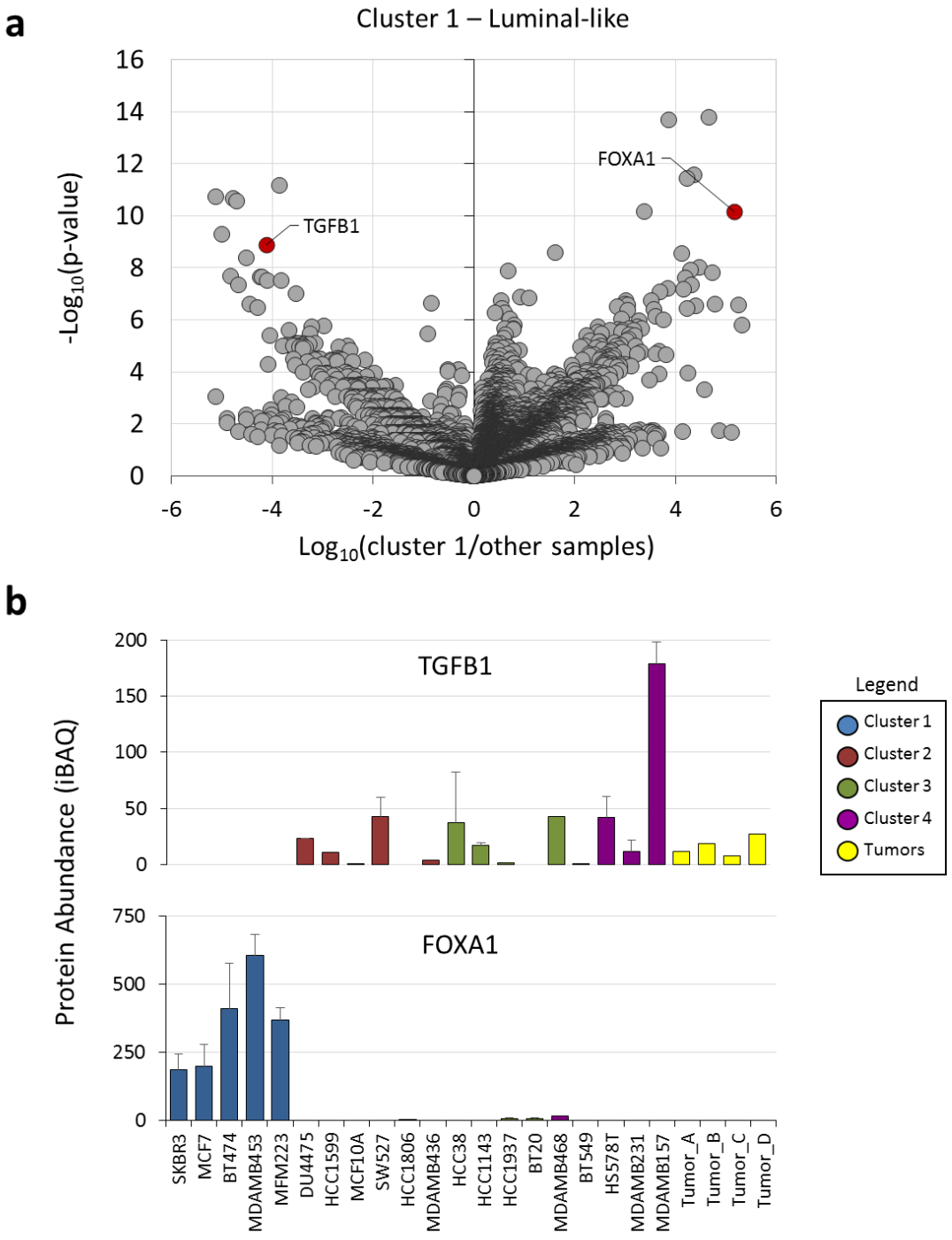


Figure A.2.2 Identification of subtype-associated proteins using volcano plots
(a) Volcano plot enables selection of proteins significantly underexpressed (top left, e.g. TGFB1) or overexpressed (top right, e.g. FOXA1) in the samples from subtype cluster 1 *versus* other samples. **(b)** Protein expression of transforming growth factor beta 1 was only expressed in TNBC samples. Expression of Forkhead box A1 (FOXA), also known as hepatocyte nuclear factor 3-alpha, was almost exclusive to samples in cluster 1. Sample labels are shown in the bottom panel. Color corresponds to cluster assignment from Figure 2.3a. Error bars represent S.D.

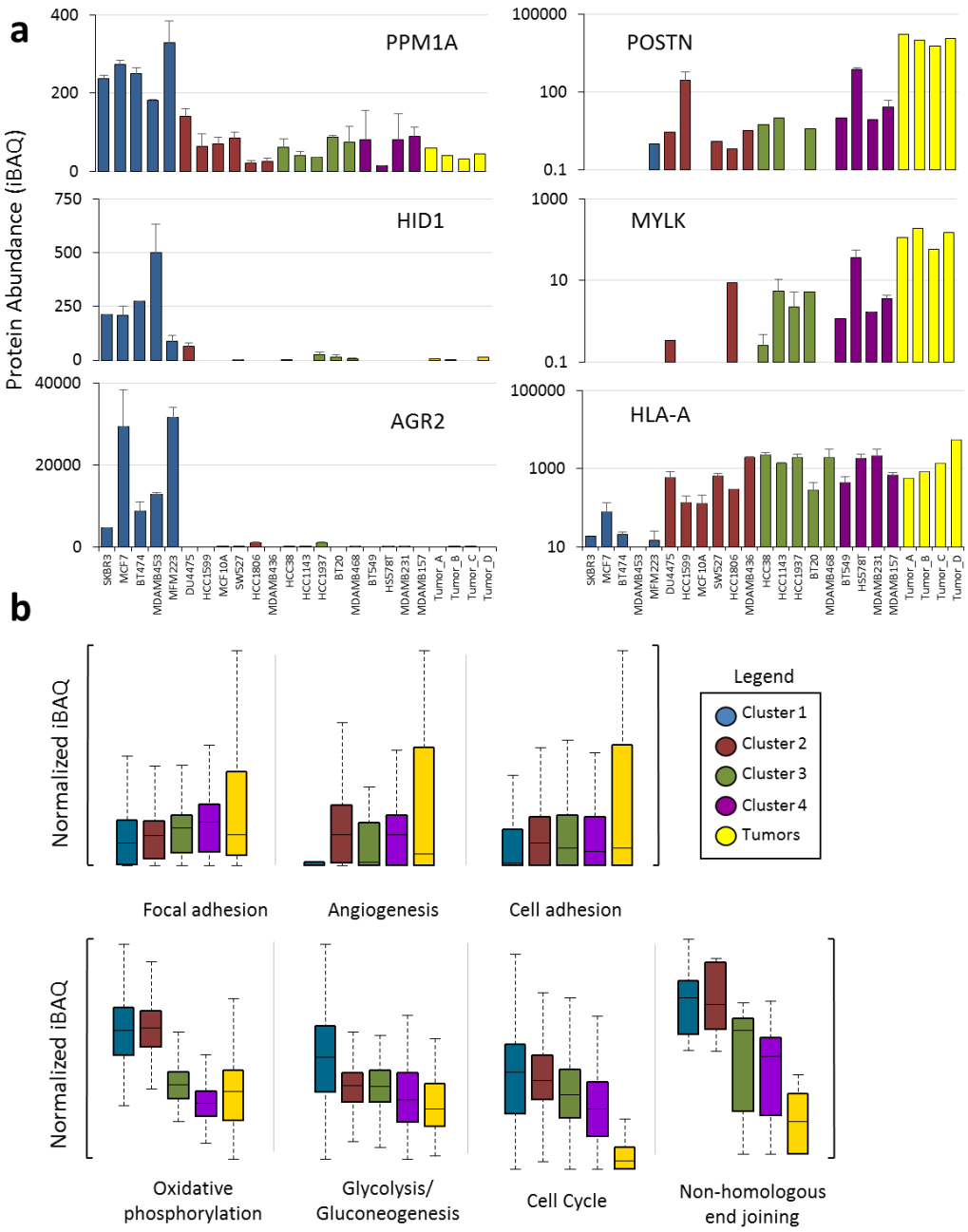


Figure A.2.3 Additional subtype-associated proteins and pathways
(a) Proteins associated with cluster 1, representing luminal-like breast cancer *versus* triple-negative breast cancer. Protein phosphatase 1A (PPM1A), down-regulated in multiple cancers 1 (HID1), and anterior gradient protein 2 (AGR2) expression were significantly decreased in triple-negative breast cancer. Periostin (POSTN), myosin light chain kinase (MYLK), and MHC class I antigen A (HLA-A) were increased in TNBC. Error bars represent S.D. **(b)** Boxplot distribution of protein abundance for each cluster within gene ontology pathways as indicated. Color corresponds to cluster assignment from Figure 2.3a.

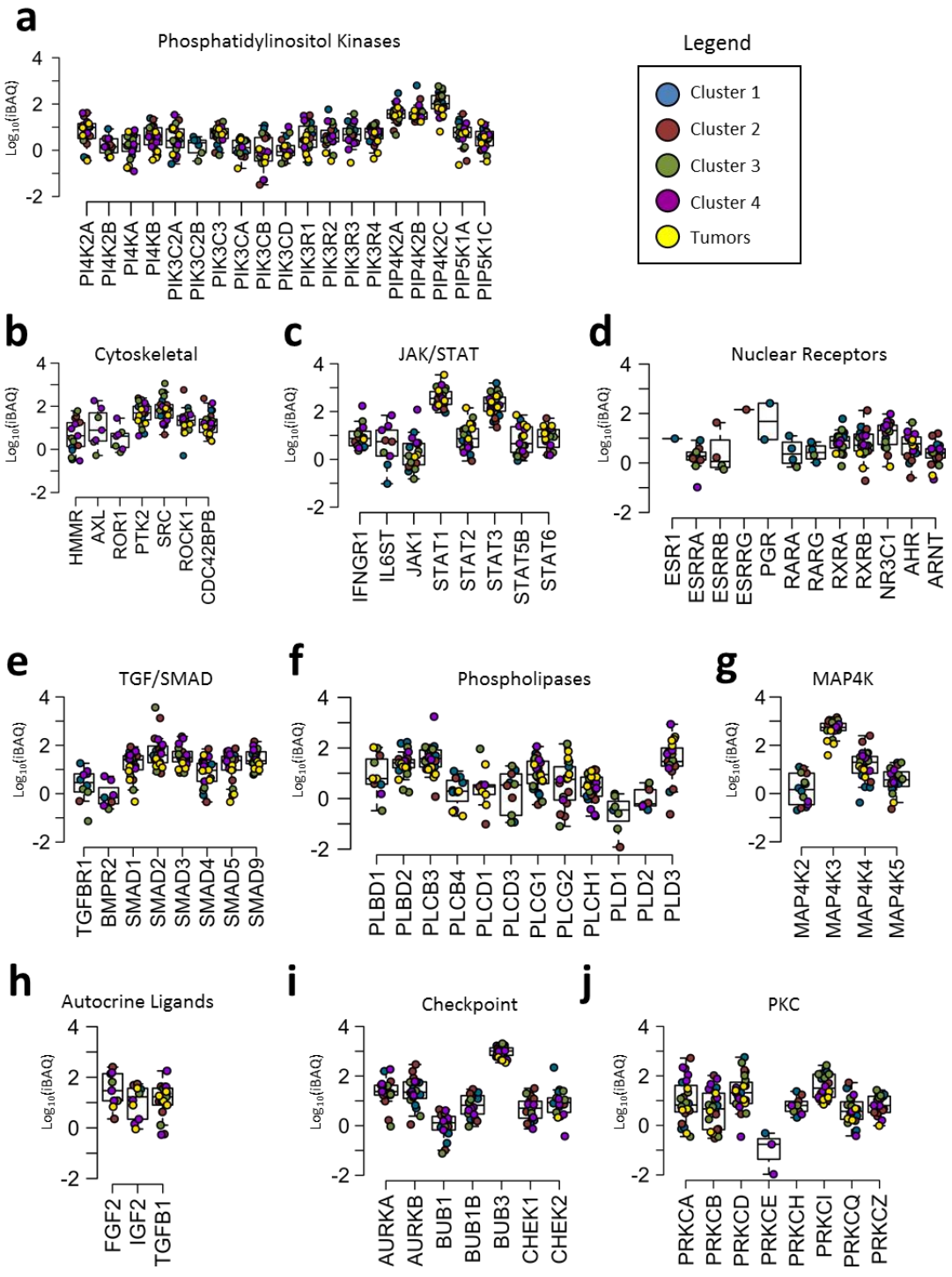


Figure A.2.4 Absolute abundance distribution of additional signaling components
Distribution of absolute abundance across samples for each protein in the indicated signaling networks. Each point represents a sample, color coded according to cluster assignment from Figure 2.3a.

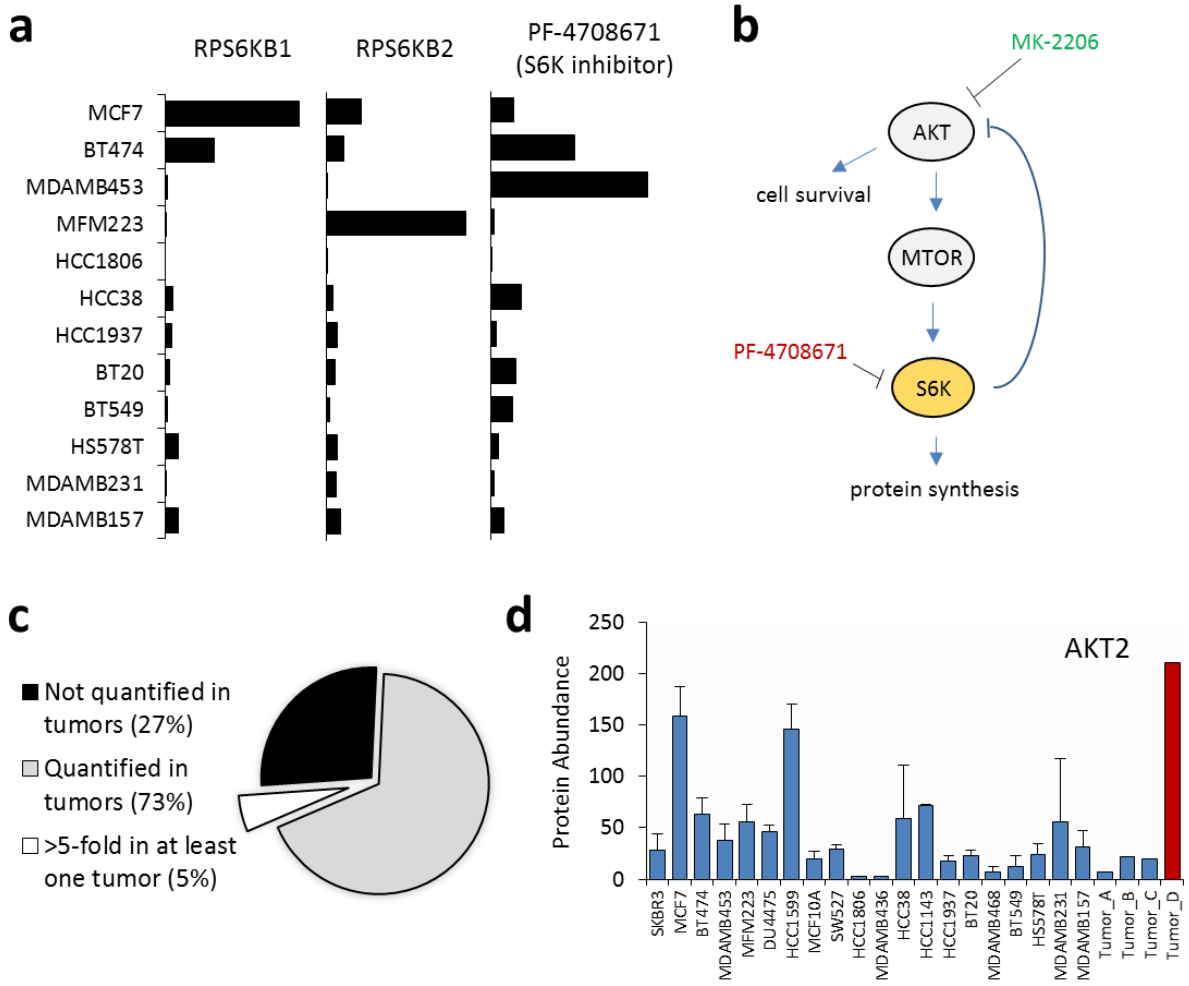


Figure A.2.5 Correlation of signaling components and drug sensitivity
(a) S6K expression was inversely correlated with sensitivity to an S6K inhibitor. Left two panels: protein abundance (iBAQ) across cell lines of RPS6KB1 and RPS6KB2. Right panel: drug sensitivity (inverse IC50, M^{-1}) across the same cell lines of S6K inhibitor (PF-4708671). **(b)** Schematic of negative feedback signaling between S6K and AKT. S6K expression is associated with sensitivity to AKT inhibitor (MK-2206) but inversely associated with S6K inhibitor (PF-4708671). **(c)** Percentage of proteins used in the drug sensitivity analysis that were also quantifiable in tumors. **(d)** Protein kinase AKT2 was most highly expressed in a tumor sample.

Appendix B. Supplementary Material for Chapter 3

B.1 Supplementary Experimental Procedures for Chapter 3

B.1.1 Common procedures

Protein extraction and digestion

Cells were scraped into ice-cold urea buffer (8M urea, 75 mM NaCl, 50 mM Tris HCl pH 8.2, complete protease inhibitor cocktail (Roche), 50 mM sodium fluoride, 50 mM beta-glycerophosphate, 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate). Protein concentration was assayed using the BCA method and lysates from 'light' and 'heavy' cultures were mixed in a 1:1 ratio. Protein lysates were reduced with 5 mM DTT for 30 min at 55°C, alkylated with 10 mM iodoacetamide for 15 min at room temperature, and quenched with 10 mM DTT. Proteins were diluted 5-fold with 50 mM Tris pH 8.8 and digested with trypsin (Promega) overnight at 37C. The resulting peptides were desalted over a tC18 SepPak cartridge (Waters) and dried by lyophilization.

Mass spectrometry

Phosphopeptide-enriched samples were resuspended in 4% formic acid, 3% MeCN and subjected to liquid chromatography on an EASY-nLC II system equipped with a 100 µm inner diameter x 40 cm column packed in-house with Reprosil C18 1.9 µm particles (Dr. Maisch GmbH) and column oven set to 50°C. Separations were performed using gradients of 9% to 32% MeCN in 0.125% formic acid ranging in length from 55-105 min and were coupled directly with a LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher) configured to conduct a full MS scan (60k resolution, 3e6 AGC target, 500 ms maximum injection time, 300 to 1500 m/z) followed by up to 20 data-dependent MS/MS acquisitions on the top 20 most intense precursor ions (3e3 AGC target, 100 ms maximum injection time, 35% normalized collision energy, 40 sec dynamic exclusion).

B.1.2 Specific procedures for draft map of the HeLa cellular signaling network

HeLa cells were passaged in DMEM with 4.5 g/L glucose, penicillin-streptomycin, 10% FBS at 37°C, 5% CO_2. For deep fractionation experiments, cells were split into 15cm plates and one plate was used for each experiment. For single-shot experiments, cells were split into 6-well plates, and 3 experiments were performed per plate (technical duplicate stimulations). For all experiments, cells were serum-deprived for 4 hours prior to stimulation according to Table B.2.1. Proteins were lysed in urea buffer, digested with trypsin and desalted on tC18 SepPaks. From here, single-shot samples were enriched directly by IMAC or subjected to SCX using a volatile buffer system and dried prior to IMAC. All samples were analyzed on an EASY nLC-II coupled to a Velos-Orbitrap mass spectrometer with a data dependent acquisition strategy. Raw files were searched with Comet against the human Swissprot database allowing for phosphorylation of S, T, or Y, filtered using percolator, site-localized using Ascore, and quantified using an in-house peak area integration algorithm.

Raw data files were converted to mzXML and searched using Comet version 2015.01 against the human Swissprot database including reviewed isoforms (April 2015; 42,121 entries) allowing for binary (all or none) labeling of lysine (+8.0142) and arginine (+10.0083), and variable oxidation of methionine, protein N-terminal acetylation, and phosphorylation of serine, threonine, and tyrosine residues. Carbamidomethylation of cysteines was set as a fixed modification. Trypsin (KR|P) fully digested was selected allowing for up to 2 missed cleavages. Precursor mass tolerance was set to 50 ppm, and fragment ion tolerance to 1.0005 Daltons. Search results were filtered using Percolator to reach a 1% false discovery rate at the PSM level. Peak areas were calculated using an in-house quantification algorithm. Phosphosite assignment was performed using an in-house implementation of Ascore, and sites with Ascore ≥ 13 were considered localized (p=0.05). Phosphopeptides in the database with multiple non-localized instances spanning the same sequence were only considered to correspond to the minimum

number of phosphosites that explain the data. Finally, the dataset was additionally filtered to reach a site-adjusted false discovery rate of 1%.

B.1.3 Specific procedures for systematic analysis of cancer signal integration

HeLa cells were passaged in DMEM (-Arg,-Lys) with penicillin-streptomycin, 10% dialyzed FBS at 37°C, 5% CO_2, supplemented with either normal L-lysine and L-arginine (light K0, R0) or $^{13}C_6$,$^{15}N_2$-lysine and $^{13}C_6$,$^{15}N_4$-arginine (heavy K8, R10). Both populations of cells were deprived of serum overnight. 'Light' labeled cells were treated with inhibitor (rapamycin, U0126, wortmannin, SB20358, AKTVIII) for 30 min prior to stimulation (insulin, anisomycin, epidermal growth factor) for an additional 30 min. 'Heavy' labeled cells were stimulated with for 30 min. At the time of harvest, cells were washed three times with ice cold PBS and flash frozen over liquid nitrogen.

Approximately 3 mg of peptides were resuspended in 50 mM Tris pH 8.2 and further digested with trypsin (Promega) overnight at 37°C. The resulting tryptic peptides were desalted over a tC18 SepPak cartridge (Waters) and dried by vacuum centrifugation. They were separated by strong cation exchange into 12 fractions using a volatile binary solvent system (A: 10 mM NH_4HCO_2 + 25% MeCN + 0.05% FA, B: 500 mM NH_4HCO_2 +25% MeCN+ 0.05% FA). Fractions were dried and desalted by vacuum centrifugation. Fractions were resuspended in 100 µl IMAC loading solution (80% MeCN+ 0.1% TFA). To prepare IMAC slurry, Ni-NTA magnetic agarose (Qiagen) was stripped with 40 mM EDTA for 30 min, reloaded with 10 mM $FeCl_3$ for 30 min, washed 3 times and resuspended in IMAC loading solution. To enrich phosphopeptides, 50 µl of 5% bead slurry was added to each fraction and incubated with rotation for 30 min at room temperature, washed 3 times with 150 µl 80% MeCN, 0.1% TFA, and eluted with 60 µl 1:1 MeCN:1% NH_4OH. The eluates were acidified with 10% FA and dried by vacuum centrifugation for LC-MS/MS.

Approximately 15 mg of peptides were resuspended in 1.4 ml IAP buffer (50 mM MOPS, 10 mM Na_2HPO_4, 50 mM NaCl, pH 7.2) and added to 20 ul phospho-AKT substrate RxRxxS*|T* antibody

pre-conjugated to sepharose beads (Cell Signaling Technologies). The mixture was incubated for 2 hours rotating at 4°C. The flow-through was collected and subjected to a sequential IAP using phospho-AKT substrate RxxS*|T* antibody. Beads were washed 3x with cold IAP buffer and 2x with cold H_2O and eluted with 100 µl 0.15% TFA. Eluates were dried by vacuum centrifugation and resuspended in 50 µl of 50 mM NH_4HCO_3 with 5% MeCN and further digested using 500 ng trypsin (Promega) at 37°C for 4 hours. The digestion product was acidified, concentrated using a C18 StageTip, and dried by vacuum centrifugation for LC-MS/MS.

Raw data files were processed using MaxQuant v1.4 against the human Swissprot database including reviewed isoforms (April 2015; 42,121 entries) allowing for binary (all or none) labeling of lysine (+8.0142) and arginine (+10.0083), and variable oxidation of methionine, protein N-terminal acetylation, and phosphorylation of serine, threonine, and tyrosine residues. Carbamidomethylation of cysteines was set as a fixed modification. Trypsin (KR|P) fully digested was selected allowing for up to 2 missed cleavages. The dataset was filtered to reach a site-adjusted false discovery rate of 1%.

B.1.4 Specific procedures for dynamic MCF7 phosphoproteome

MCF7 cells were passaged in DMEM with 4.5 g/L glucose, penicillin-streptomycin, 10% FBS at 37°C, 5% CO_2. For the PhosphoHT workflow, cells were split into 6-well plates, and 3 experiments were performed per plate (technical duplicate stimulations). For all experiments, cells were serum-deprived for 4 hours prior to stimulation. For the timecourse experiments, cells were stimulated with 100ng/ml IGF1 or EGF for 30 sec, 1min, 10min, 30min, or 60min. For the drug panel experiments, control cells were treated with 1.5ul DMSO, and 10,000x stocks of drugs prepared for a final dosage of 10x manufacturer reported IC50 values. Proteins were lysed in urea buffer, digested with trypsin and desalted on tC18 SepPaks. From here, single-shot samples were IMAC-enriched in multiplex using a magnetic bead processing robot (Thermo KingFisher). All samples were analyzed on an EASY nLC-II coupled to a Velos-Orbitrap mass spectrometer with a data

dependent acquisition strategy. Raw data files were processed using MaxQuant v1.4 against the human Swissprot database including reviewed isoforms (April 2015; 42,121 entries) allowing for variable oxidation of methionine, protein N-terminal acetylation, and phosphorylation of serine, threonine, and tyrosine residues. Carbamidomethylation of cysteines was set as a fixed modification. Trypsin (KR|P) fully digested was selected allowing for up to 2 missed cleavages. The dataset was filtered to reach a site-adjusted false discovery rate of 1%.

B.2 Supplementary Figures and Tables for Chapter 3

Table B.2.1 Panel of local signaling environments to map the HeLa cellular signaling network

Signal	**Abbreviation**	**Description**
Control	CTL	Serum-free media
Insulin	INS	+ 1uM insulin
Serum	SER	+ 10% fetal bovine serum
cAMP signaling	cAMP	+ 100uM isobutyl-methylxanthine + 20uM forskolin
Ca^{2+} signaling	PMA	+ 100nM phorbol myristate acetate
Insulin-like growth factor 1	IGF1	+ 100ng/ml IGF1
Epidermal growth factor	EGF	+ 100ng/ml EGF
Beta-adrenergic signaling	ISO	+ 10uM isoproterenol
Ultraviolet irradiation	UVX	+ 40 J/m^2 UVC
Tumor necrosis factor alpha	TNFa	+ 50ng/ml TNF-alpha
Transforming growth factor beta	TGFb	+ 5ng/ml TGF-beta
Interferon gamma	IFNy	+ 20ng/ml IFN-gamma
Osmotic stress	SORB	+ 200mM sorbitol
Proteotoxic stress	CHX	+ 50ug/ml cycloheximide
Phosphatase inactivation	PPI	+ 50nM calyculin A, 100nM okadaic acid, 1mM pervanadate, 10% serum
Kinase inactivation	STR	+ 100nM staurosporine

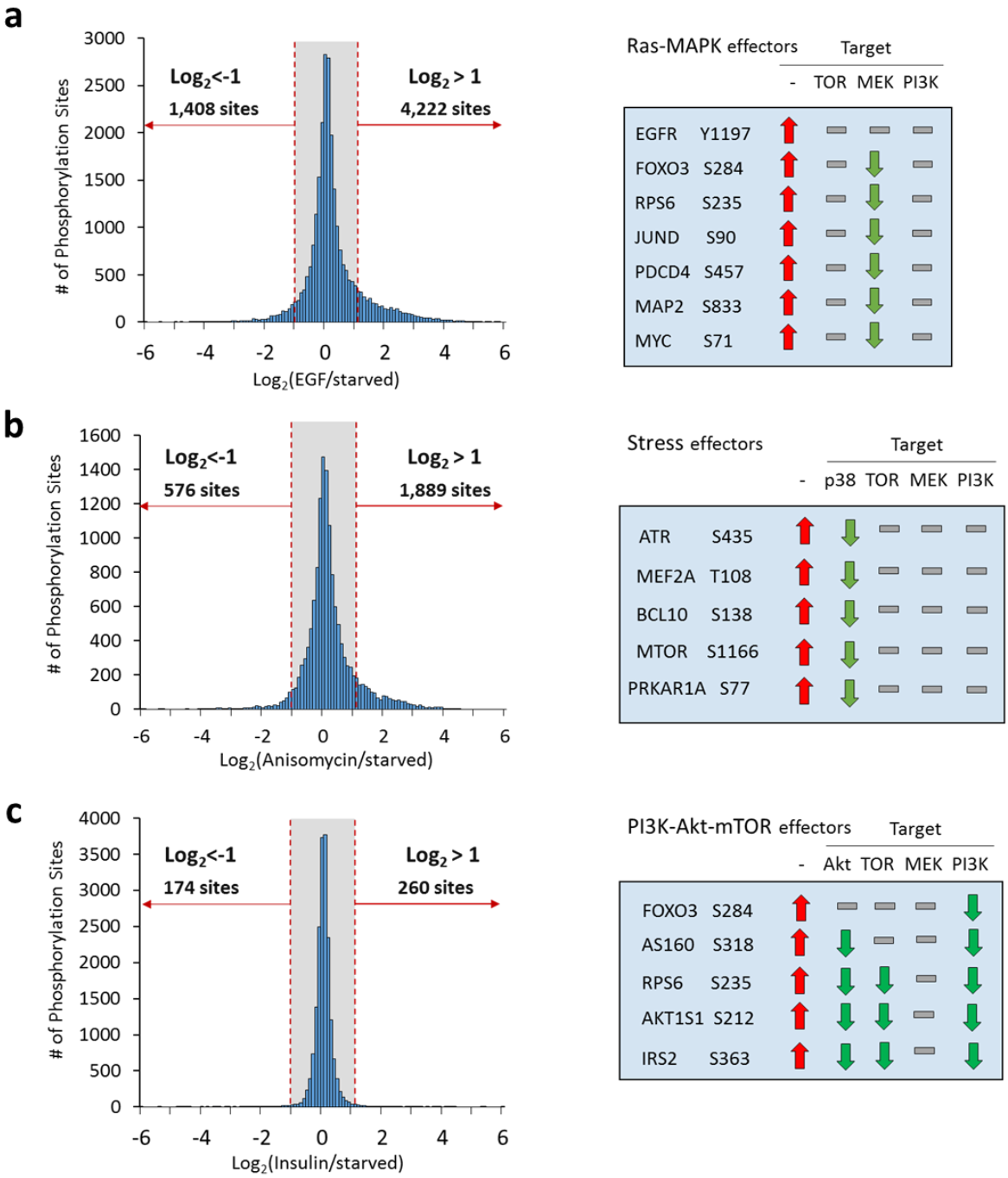


Figure B.2.1 Quantitative overview of cellular responses to stimulation
(a) Distribution of EGF-stimulated phosphorylation sites and effector responses. **(b)** Distribution of anisomycin-stimulated phosphorylation sites and effector responses. **(c)** Distribution of insulin-stimulated phosphorylation sites and effector responses.

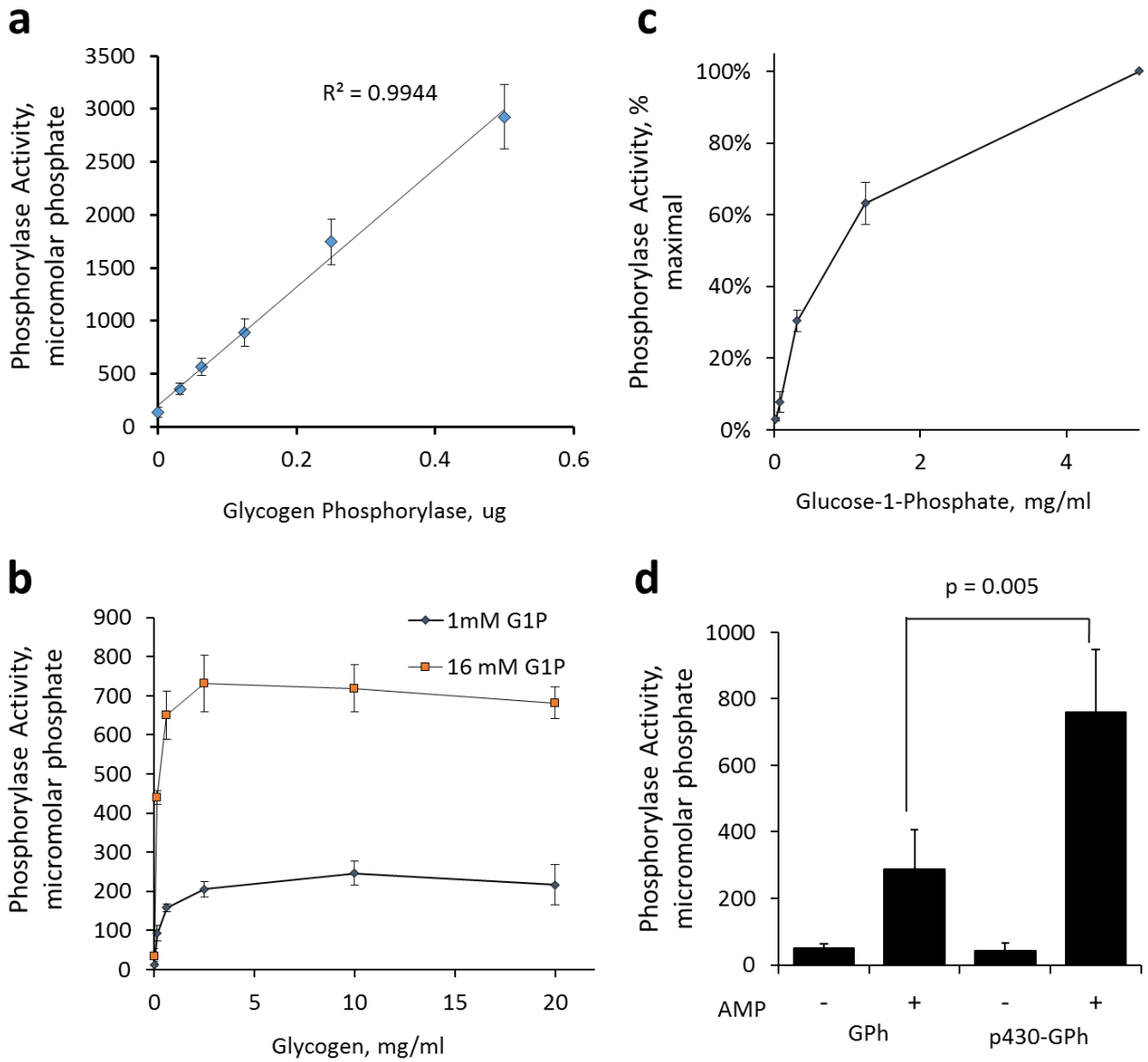


Figure B.2.2 Glycogen phosphorylase assay validation
(a) Linearity of the malachite green phosphorylase activity assay. **(b)** Glycogen synthesis kinetics with varying glycogen concentration. **(c)** Glycogen synthesis kinetics with varying glucose-1-phosphate. **(d)** Dependence of GPh and pS430-GPh activity on the presence of AMP.

Appendix C. Supplementary Material for Chapter 4

C.1 Supplementary Experimental Procedures for Chapter 4

C.1.1 Cell culture

MCF7 breast cancer cells were obtained from ATCC, and tested biannually for the presence of mycoplasma. MCF7 cells were cultured at 37°C in 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with 4.5 g/L glucose, L-glutamine, and 10% fetal bovine serum. To generate bulk phosphopeptides for method comparisons, cells were incubated in serum-free medium for 4 hours prior to treatment with IGF-1 (100 ng/ml), EGF (100 ng/ml), and pervanadate (1 mM) for 15 minutes. For IGF-1 experiments, cells were incubated in serum-free medium for 4 hours and stimulated with or without IGF-1 (100 ng/ml) for 15 minutes (n=6). At the time of harvest, cells were rinsed 3 times quickly with ice-cold phosphate-buffered saline and flash frozen on liquid nitrogen.

C.1.2 Sample preparation

Cell lysis was performed in 9 M urea, 50 mM Tris pH 8.2, 75 mM NaCl with protease inhibitors (Roche) and phosphatase inhibitors (50 mM beta-glycerophosphate, 50 mM sodium fluoride, 10 mM sodium pyrophosphate, 1 mM sodium orthovanadate). Cells were scraped off of plates directly into ice cold lysis buffer and subjected to 20 seconds of probe sonication, incubated on ice for 20 minutes to solubilize proteins and spun at 12,000 g for 10 min, and protein content was assayed using the bicinchoninic acid method (Pierce). Proteins were reduced with 5 mM dithiothreitol for 30 min at 55°C, alkylated with 10 mM iodoacetamide for 15 min at room temperature, and quenched with an additional 10 mM dithiothreitol. Protein extracts were diluted 5-fold with 50 mM Tris pH 8.2 and digested overnight at 37°C with sequencing grade trypsin (Promega) in a 1:200 enzyme/substrate ratio. Following digestion, the reactions were quenched with 10% TFA to pH ~2, desalted on tC18 SepPak cartridges (Waters), and dried by vacuum centrifugation. For bulk phosphopeptide preparation, 5 mg of tryptic peptides were resuspended

in immobilized metal affinity chromatography (IMAC) loading solution (80% MeCN, 0.1% TFA) and divided into 12 x 150 µl aliquots. To prepare IMAC slurry, Ni-NTA magnetic agarose (Qiagen) was stripped with 40 mM EDTA for 30 min, reloaded with 10 mM $FeCl_3$ for 30 min, washed 3 times and resuspended in IMAC loading solution. Phosphopeptide enrichment was performed using a KingFisher Flex robot (Thermo Scientific) programmed to incubate peptides with 150 µl 5% bead slurry for 30 minutes, wash 3 times with 150 µl 80% MeCN, 0.1% TFA, and elute with 60 µl 1:1 MeCN:1% NH_4OH. The eluates were acidified with 10% formic acid, pooled, and dried by vacuum centrifugation. For IGF-1 experiments, 350 µg tryptic peptides were enriched for each sample. To control variability in phosphopeptide enrichment and mass spectrometry, we used a spike-in standard of bovine serum albumin tryptic peptides previously subjected to *in vitro* phosphorylation by serum-stimulated HeLa cell lysate in the presence of 2.5 mM ATP for 60 min at 30°C. The resulting peptides were purified by solid phase extraction on a tC18 SepPak cartridge and spiked in prior to IMAC at a mass ratio of 1:50.

C.1.3 Mass spectrometry

Phosphopeptide-enriched samples were resuspended in 4% formic acid, 3% MeCN and subjected to liquid chromatography on an EASY-nLC 1000 system equipped with a 100 µm inner diameter x 25 cm column packed in-house with Reprosil C18 1.9 µm particles (Dr. Maisch GmbH) and column oven set to 50°C. All separations were performed using a gradient 9% to 32% MeCN in 0.15% formic acid over 44 min (60 min total method length) at a flow rate of 500 nl/min. The HPLC was coupled directly with a Q-Exactive mass spectrometer. The DDA method consisted of a full MS scan (70k resolution, 3e6 automatic gain control (AGC) target, 240 ms maximum injection time, 400 to 1200 m/z, centroid mode) followed by up to 20 data-dependent MS/MS acquisitions on the top 20 most intense precursor ions (35k resolution, 5e4 AGC target, 120 ms maximum injection time, 2 m/z isolation window, 27% normalized collision energy, centroid mode). The DIA method consisted of a full MS scan configured as above followed by 33 data-

independent MS/MS acquisitions configured using an inclusion list with 25 m/z overlapping windows (12.5 m/z with deconvolution) covering the 400 to 1200 m/z mass range (35k resolution, 5e5 AGC target, 120 ms maximum injection time, 25 m/z isolation window, 27% normalized collision energy, centroid mode). The PRM method consisted of a full MS scan configured as above followed by up to 20 targeted MS/MS scans as defined by a time-scheduled inclusion list (35k resolution, 5e5 AGC target, 120 ms maximum injection time, 2 m/z isolation window, 27% normalized collision energy, centroid mode). To prevent systematic bias, the order of acquisition for “Control” and “IGF-1” samples was randomized. Benchmarking experiments for retention time, sequence and charge selection were performed on a nanoACQUITY liquid chromatography system coupled to a Q-Exactive Plus mass spectrometer with the following modifications to the above parameters: flow rate was set to 400 nl/min and for the PRM assays 25 unscheduled targeted MS/MS scans using 50 ms maximum injection time and 17.5k resolution were collected after each full MS scan. For DDA and DIA, AGC targets were optimized for speed with a goal of reaching the target before reaching maximum injection time. For PRM, the AGC target was selected for enhanced sensitivity and dynamic range with a goal of reaching the maximum injection time before reaching the target. PRM assay scheduling was performed within Skyline (MacLean et al., 2010) (version 3.1.0.7382). To calibrate the schedule, an initial pilot run was conducted with 10min wide acquisition windows, and aligned to the normalized retention database to build a retention time predictor. Subsequently, a scheduled isolation list with refined 6min windows was exported from Skyline as a .csv file and imported directly into the instrument PRM method as an inclusion list. Any group of peptides from the database may be used as retention time calibrators, including any of the phosphopeptides, the Peptide Retention Time Calibration (PRTC) mixture (Pierce), or tryptic peptides from BSA. An equivalent retention time scheduling tool with other capabilities such as variable windows for positional isomeric phosphopeptides is available with the web portal that accompanies this manuscript.

C.1.4 DDA data processing and analysis

Raw DDA data files were converted to mzXML and searched using Comet (Eng et al., 2013) (version 2015.01) against the human Swissprot database including reviewed isoforms (April 2015; 42,121 entries) allowing for variable oxidation of methionine, protein N-terminal acetylation, and phosphorylation of serine, threonine, and tyrosine residues. Carbamidomethylation of cysteines was set as a fixed modification. Trypsin (KR|P) fully digested was selected allowing for up to 2 missed cleavages. Precursor mass tolerance was set to 50 ppm, and fragment ion tolerance to 0.02 Daltons. Search results were filtered using Percolator (Käll et al., 2007) to reach a 1% false discovery rate at the PSM level. Phosphosite assignment was performed using an in-house implementation of Ascore (Beausoleil et al., 2006), and sites with Ascore ≥ 13 were considered localized (p=0.05). To construct the large-scale phosphopeptide database we imposed additional filters to prevent accumulation of false hits associated with data aggregation. First, phosphopeptides in the database with multiple non-localized instances spanning the same sequence were only considered to correspond to the minimum number of phosphosites that explain the data. Second, we carried forward the best posterior error probability for each phosphopeptide spectral match, phosphoisoform, and phosphosite in order to compute an adjusted FDR at each level. A phosphoisoform represents multiple peptide sequences containing the same combination of phosphorylation sites. Multiple peptides with different degrees of phosphorylation or cleavage may represent the same phosphosite. Without imposing additional filtering beyond peptide spectral matches 196,744 phosphosites were identified, but the phosphosite-level false discovery rate after data aggregation was 29.2% and the adjusted posterior error probability for phosphosites identified by only a single MS/MS scan was 44% (Figure C.2.1b). Accordingly, we suspect that phosphorylation site databases that aggregate large volumes of spectral data without imposing additional filters are also likely to aggregate false

discoveries. Lastly, spectral libraries were constructed from aggregate phosphopeptide search results and assembled into a normalized retention time database using Skyline.

C.1.5 DIA and PRM data processing and analysis

For spectrum-centric analysis of DIA and PRM mass spectrometry results, we used DIA-Umpire (Tsou et al., 2015) version 1.4 with default parameters to assemble pseudo-MS/MS spectra for the database search pipeline described above. Peptide-centric analysis was performed using Skyline. Signal extraction was performed on +2, +3, +4 precursors and +1, +2 b and y fragment ions. Full MS resolving power was set to 70,000, and MS/MS resolving power set to 17,500. After importing an initial run, extracted ion chromatograms were aligned to the retention time library to generate a predictor and all results were re-imported using retention time filtering to within 5 minutes of predicted RT. Peptide identifications were further refined by manual interpretation using several criteria including product ion mass accuracy, correlation of precursor and fragment ion peak shapes, and signal-to-noise ratios. Specifically, we required at least three highly resolved fragment ions without interference to consider a peptide identified. To consider a peptide localized, we required at least 1 site-diagnostic ion. For IGF-1 experiments, integrated peak areas were measured for each peptide in Skyline and exported for analysis. Values were normalized to the average peak areas of 3 spiked-in phosphorylated bovine serum albumin peptides and log2 transformed. Statistical significance was assessed using a two-sample unpaired t-test.

C.1.6 Benchmarking experiments

Retention times from bovine serum albumin digest (110 peptides) or phosphopeptide-enriched MCF7 digest (phospho-mixture) were used to predict the retention times of a subsequent analysis of the same MCF7 phospho-mixture. For the phospho-mixture prediction, half of the identifications (2,357 phosphopeptides) were randomly selected as “training” data and used to predict the retention times of an independent set of phosphopeptides in the subsequent run. Unscheduled

PRM experiments were used to evaluate disagreements in data-driven versus heuristic peptide sequence and charge state selection. For sequence selection, the heuristic was the fully cleaved phosphopeptide (i.e. cleavage after all lysines and arginines except when residue at +1 is proline), and 100 phosphosites were analyzed (200 targets). For charge selection, the heuristic was positively charged amino acids +1, and 50 phosphopeptides were analyzed (100 targets).

C.1.7 Statistical analysis

Statistics regarding peptide and phosphorylation site identification are discussed in the above section entitled “Data processing and analysis”. Sample sizes necessary for non-parametric comparisons are not easily predictable since the expected shape of the distribution is unknown. For the comparison of peptide sequence and charge state selection we used n=100 and n=50 respectively, which we predicted would be sufficient to detect a roughly 10-fold difference in the median peak area intensity. Peptide sequence selection implicitly also requires charge state selection for each sequence, hence the sample size was increased for that experiment to account for additional variability. A Wilcoxon signed rank test (paired non-parametric) was used because the intensity of different peptides representing the same phosphorylation site in the same sample are related but the intensity of peptides arising from different phosphorylation sites in the sample are unrelated and not normally distributed. Similarly, sample size necessary for high-throughput measurements is also difficult to predict, since different analytes in the assay each have different expected effect sizes and precision. For the quantitative analysis of IGF-1 stimulation *versus* control, we assumed that replicate measures of a typical phosphopeptide would follow a normal distribution after log transformation with coefficient of variation of approximately 20%. Under these assumptions we used a sample size of 6, which we predicted would be sufficient to detect a 2-fold change in most targets. Unpaired t-tests were used to assess significant differences between control and IGF-1 treated samples. Replicates were from independent treatments of the same source of MCF7 cells.

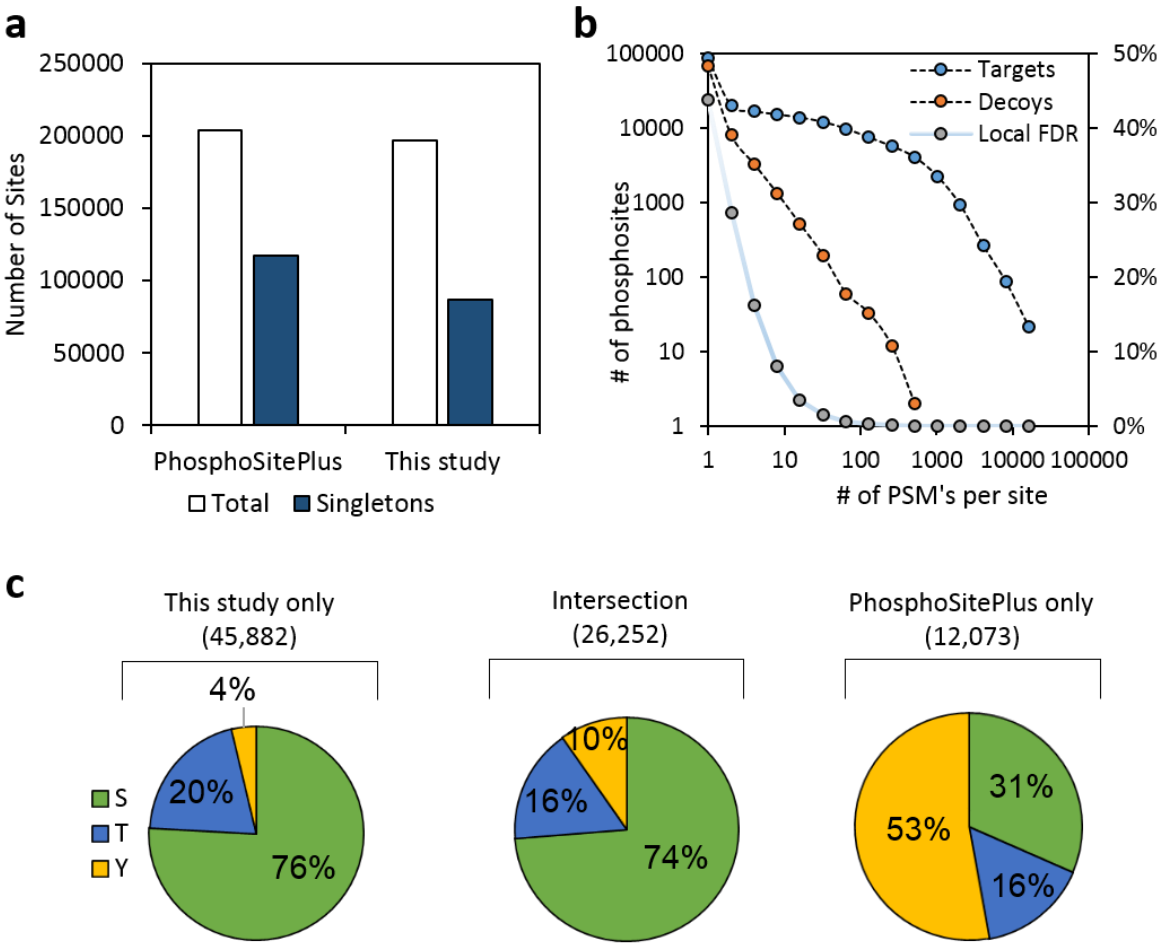


Figure C.2.1 Database coverage comparison
(a) Number of phosphosites detected, PSM-level FDR 1%, before correcting for data aggregation. (b) Effect of data aggregation on phosphosite FDR. Since PSM-level target-decoy data were not accessible from PhosphoSitePlus, we estimated site-level "local FDR" from our data using the number of phosphosite PSM's as the scoring metric. Sites only observed one time (singletons) are likely to be observed by random chance in aggregated data (> 40% in this case). (c) Database composition compared to PhosphoSitePlus. To control FDR, we compared the overlap of sites observed at least 5 times in each database.

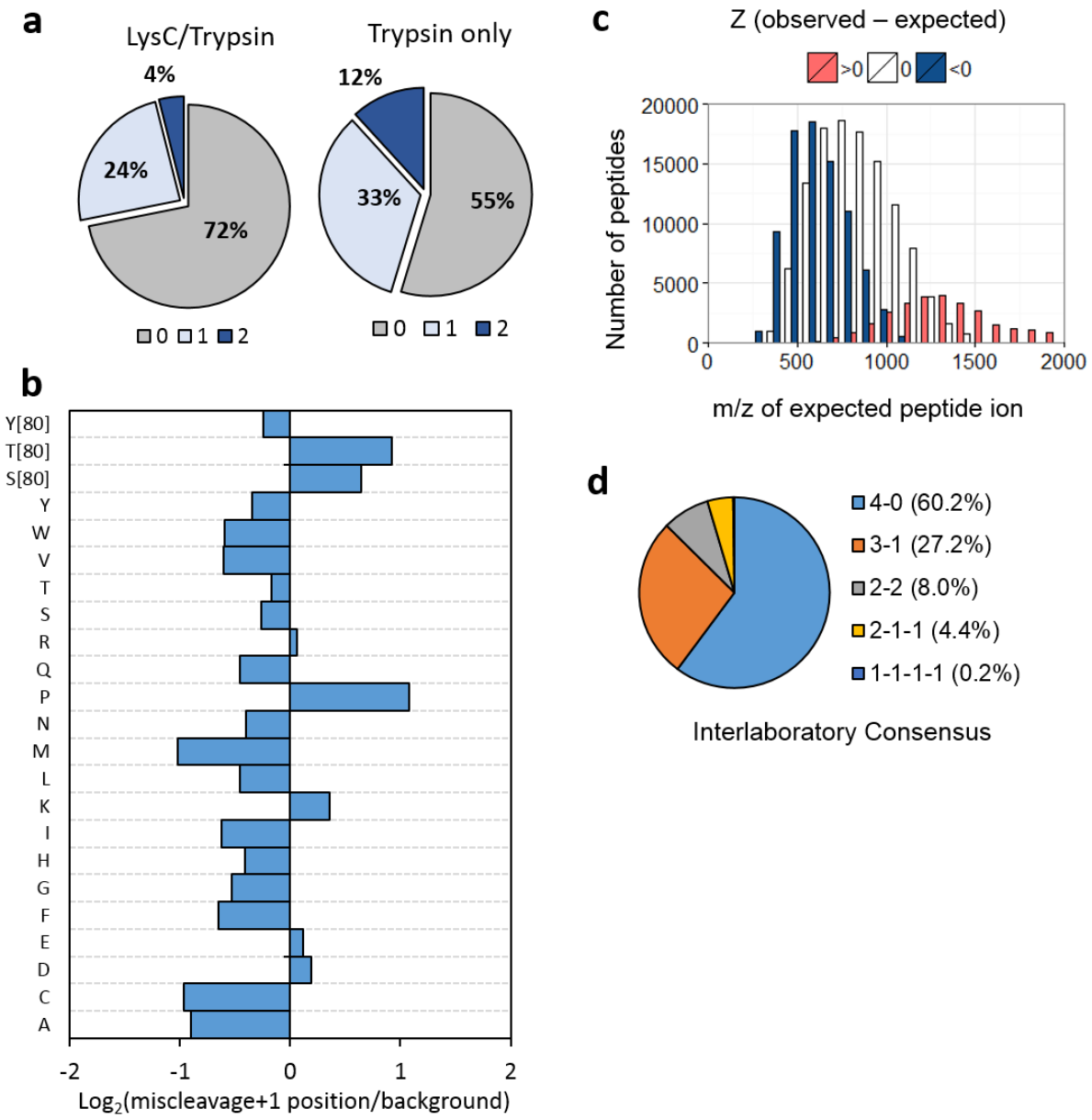


Figure C.2.2 Analysis of most frequently observed phosphopeptide forms
(a) Distribution of preferred cleavage states with sequential LysC/Trypsin digest or Trypsin only. (b) Enrichment versus background of amino acids in the miscleavage+1 position of preferred peptides. (c) Mass-to-charge distributions of predicted peptide charge states grouped by lower than observed, equal to observed or greater than observed. (d) Inter-laboratory cleavage form consensus. Phosphoisoforms observed at least once by 4 different laboratories were analyzed (n=7,897). The most frequently observed peptide representing a unique phosphoisoform for each study was considered the preferred sequence. The preferred sequence for each of the 4 studies was compared. For example “4-0” indicates that the preferred sequence was the same for all 4 studies.

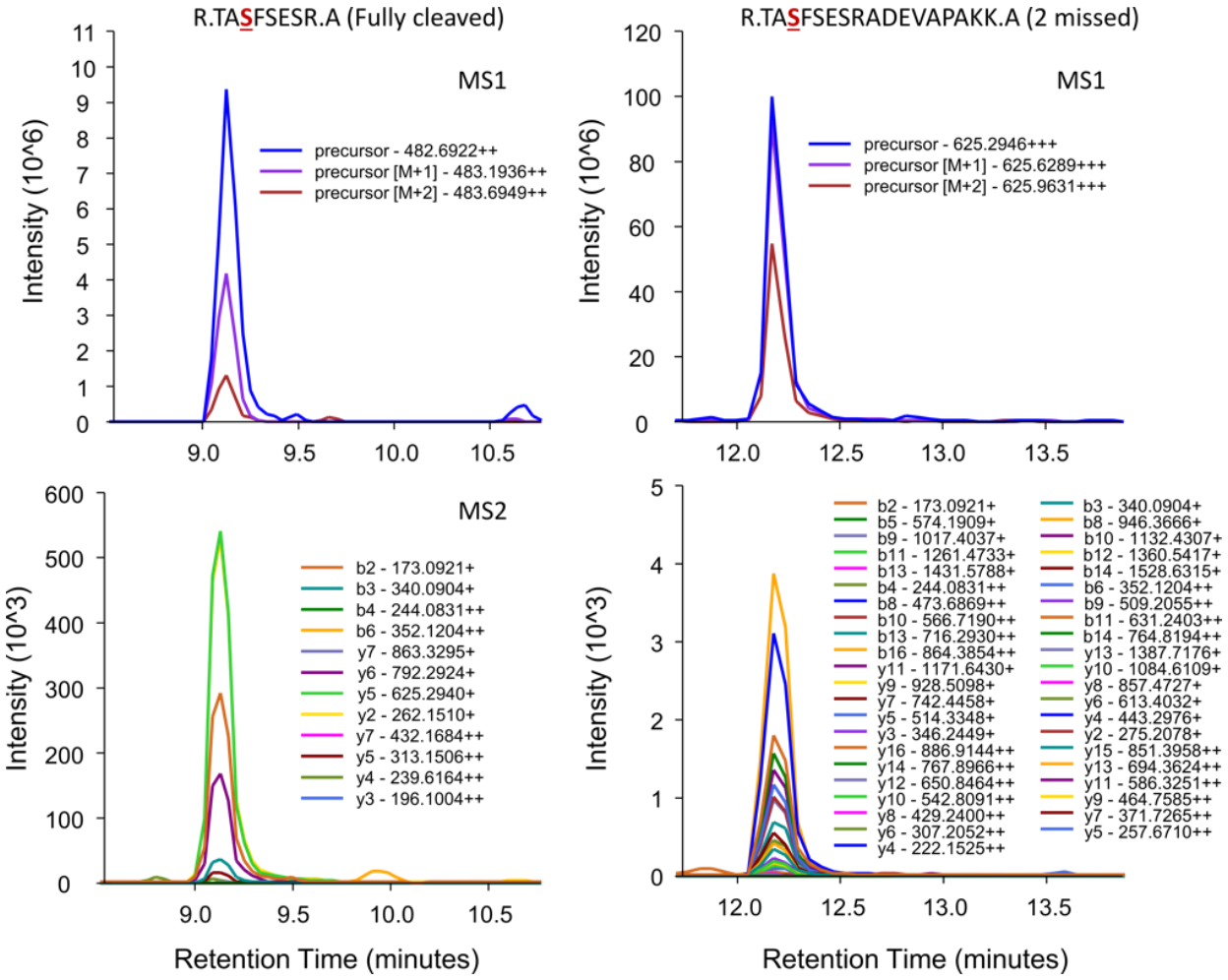


Figure C.2.3 Targeted comparison of two cleavage states of ACLY phosphorylation at S455
The isoform with two miscleavages is observed with greater than 10-fold higher intensity in than the fully cleaved isoform.

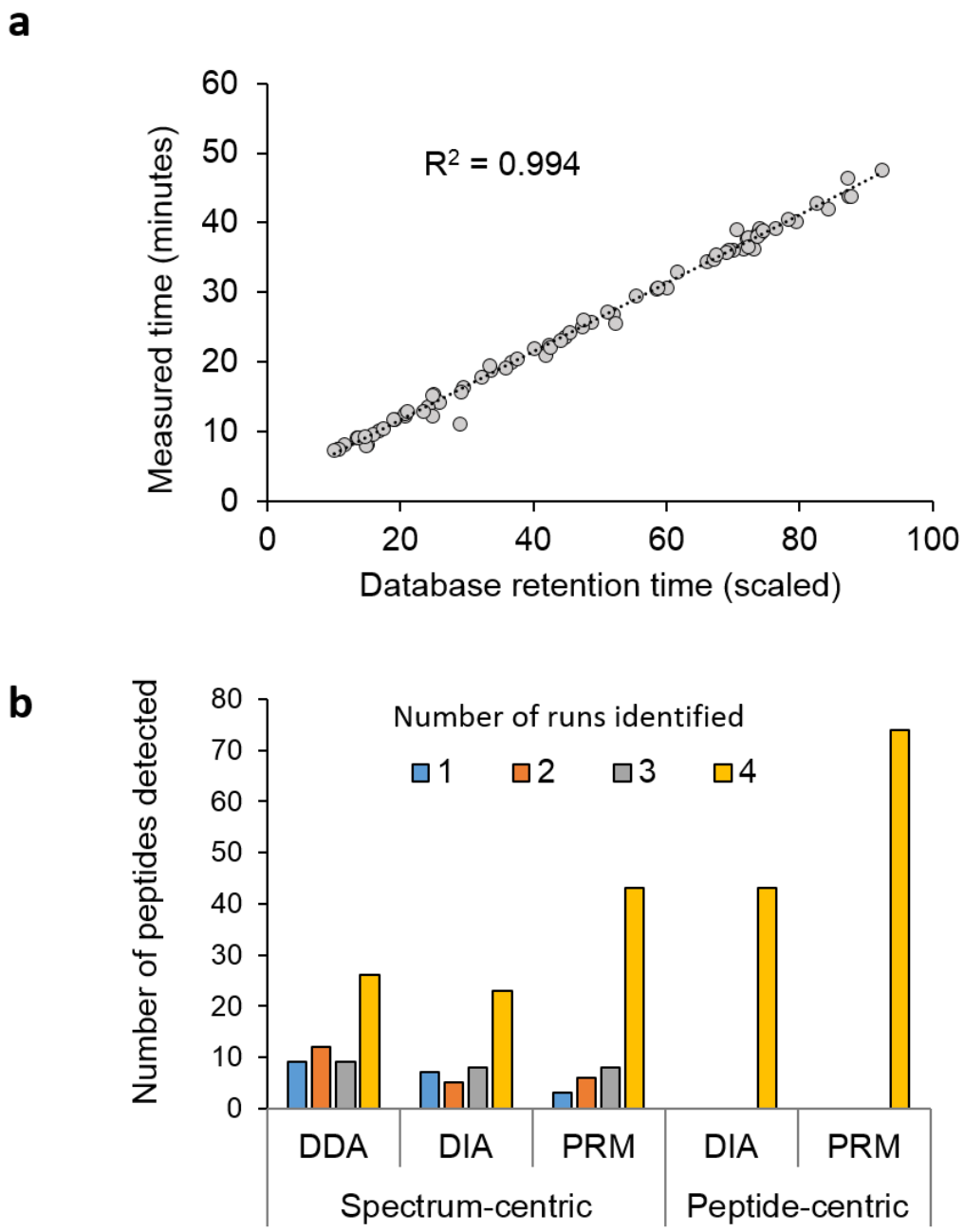


Figure C.2.4 Benchmarking PRM versus DIA and DDA
(a) Correlation of observed retention times to database retention times (normalized to a 0-100 scale in Skyline). (b) Comparison of sampling efficiency of DDA, DIA, and PRM using either spectrum-centric (database searching) or peptide-centric (targeted signal extraction) approaches. DDA was not analyzed using the peptide-centric approach (e.g. by MS1 filtering).

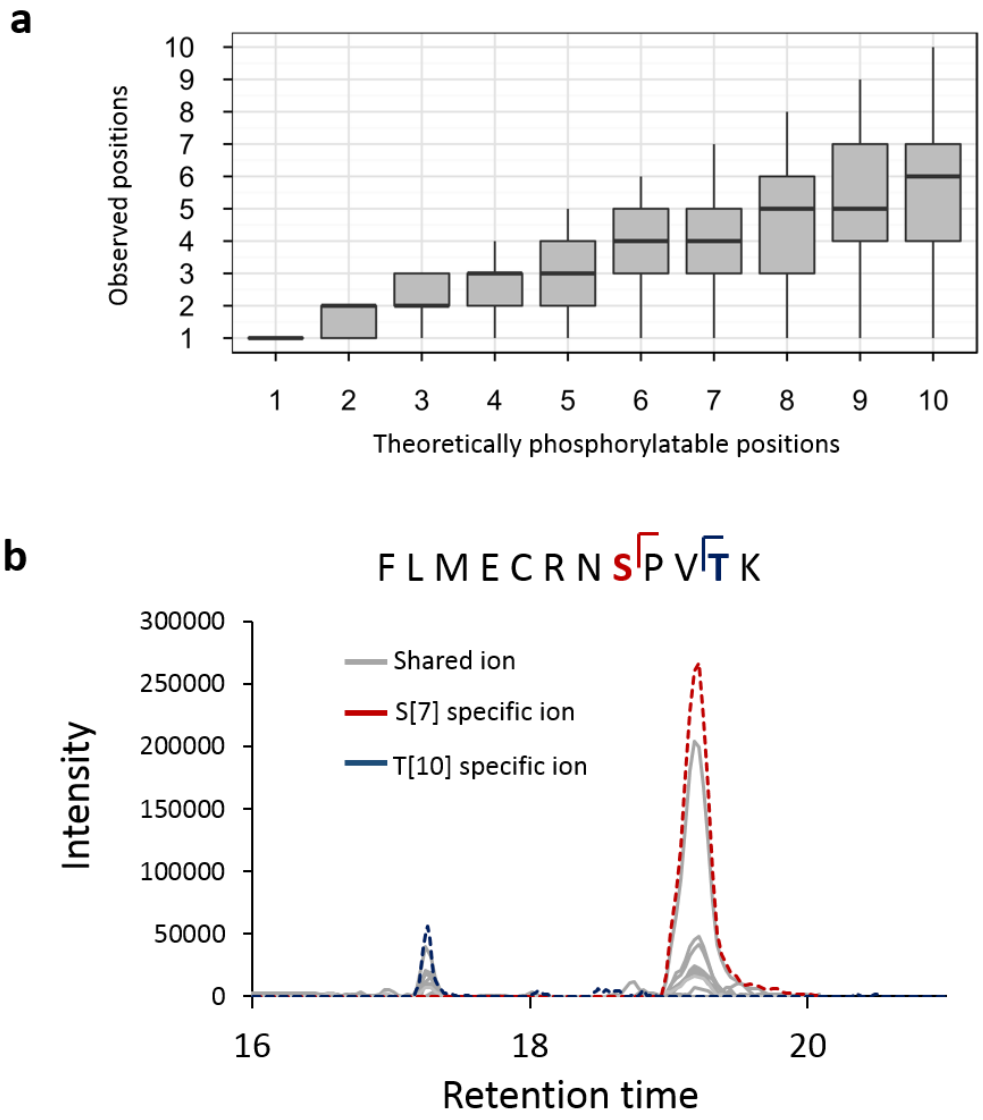


Figure C.2.5 Analysis of phosphosite positional isomers
(a) The number of observed versus theoretical positions by bin. In peptides with multiple phosphorylatable residues, typically more than half of the possible sites have been observed. Only singly-modified underlying phospho-peptide sequences observed at least 100x were analyzed (n=12,357). (b) Extracted fragment ion chromatogram of two singly phosphorylated isobaric peptides (FLMECRNS*PVTK and FLMECRNSPVT*K) resolvable by retention time.

C.3 Phosphopedia User Manual

C.3.1 Introduction

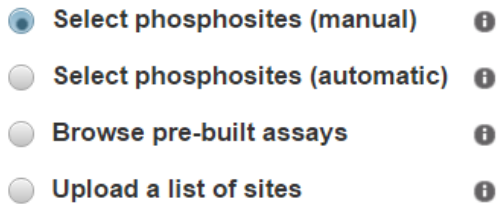
This portal provides access to an extensive and growing database of human phosphorylation sites (>100,000) and corresponding phosphopeptides that have been observed across hundreds of mass spectrometry-based phosphoproteomics discovery experiments. In addition to cataloging these sites we provide a tool for the design and implementation of targeted phosphoproteomic assays. You can mix-and-match phosphosites from pre-assembled lists, individually select the sites you are interested in, and/or upload a list of phosphosites. By adding retention time calibration data from a sample run on your LC-MS/MS system, you can automatically schedule and optimize your targeted assay. Using this method, highly precise and sensitive assays to screen hundreds of phosphopeptide targets can be developed rapidly.

For more information please refer to the publication:
Robert T. Lawrence, Brian C. Searle, Ariadna Llovet, Judit Villén. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nature Methods. 2016.

Note:
ⓘ click for more information anywhere this icon appears on the website.

C.3.2 Selecting targets

Several different mechanisms are provided for adding phosphorylation sites to an assay. They can be used in combination to rapidly generate a highly customized experiment.

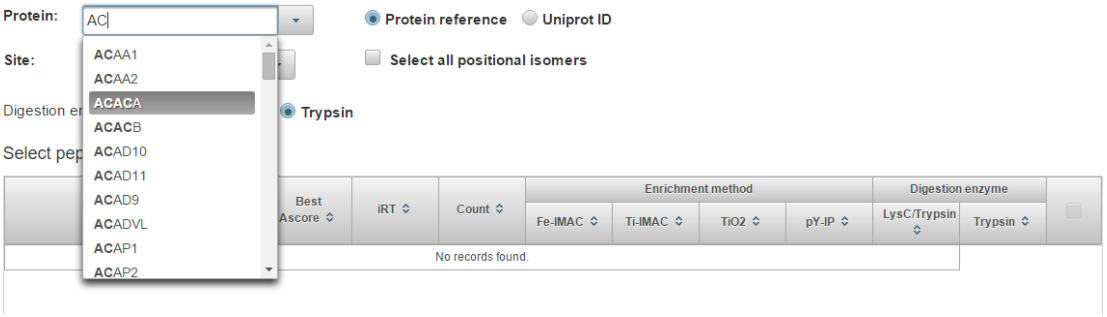


For each option, select the digestion configuration used in the experiment (LysC/Trypsin or Trypsin) to automatically choose as default the most frequently observed peptide cleavage state for that scheme. In most cases these will be the same, but miscleavage events are enriched around acidic residues, sequential cleavage sites, and phosphorylated serine/threonine.

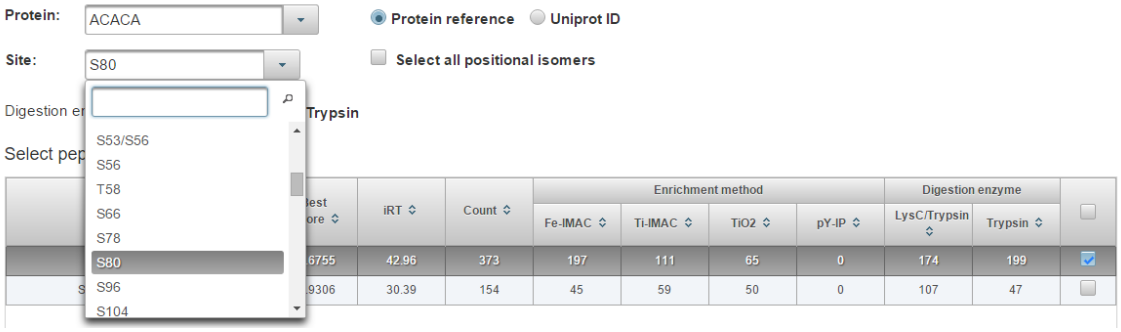
i) Select phosphosites (manual)

This is the most detailed and flexible option for assay design. It allows for manual selection of each phosphopeptide sequence and charge state to be measured and direct inspection of representative MS/MS spectra. It is the best way to directly evaluate data supporting an individual phosphorylation site. Targeting multiple sequences and charge states provides the best chances of detecting the site.

1. Start typing a protein identifier (HGNC gene symbol or Uniprot/Swiss-Prot accession are supported) and choose from the drop-down list.

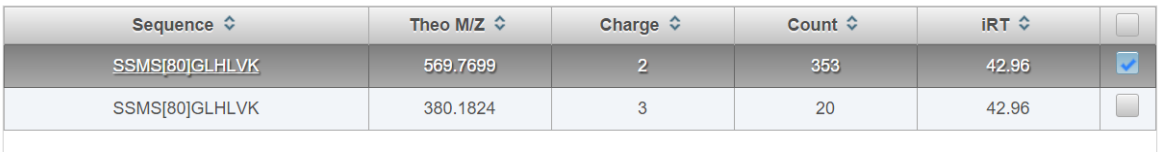


2. Select the site (position) on the protein. All peptide species that have been observed for this phosphosite will populate the table. Check the box next to the peptide(s) you would like to add or select all using the box in the top right. The most frequently observed sequence is selected by default. Choose “Select all positional isomers” to automatically schedule other known phosphosites that are localized to the same peptide (e.g. THFPQFS[80]YSASIRE and THFPQFSYSAS[80]IRE).

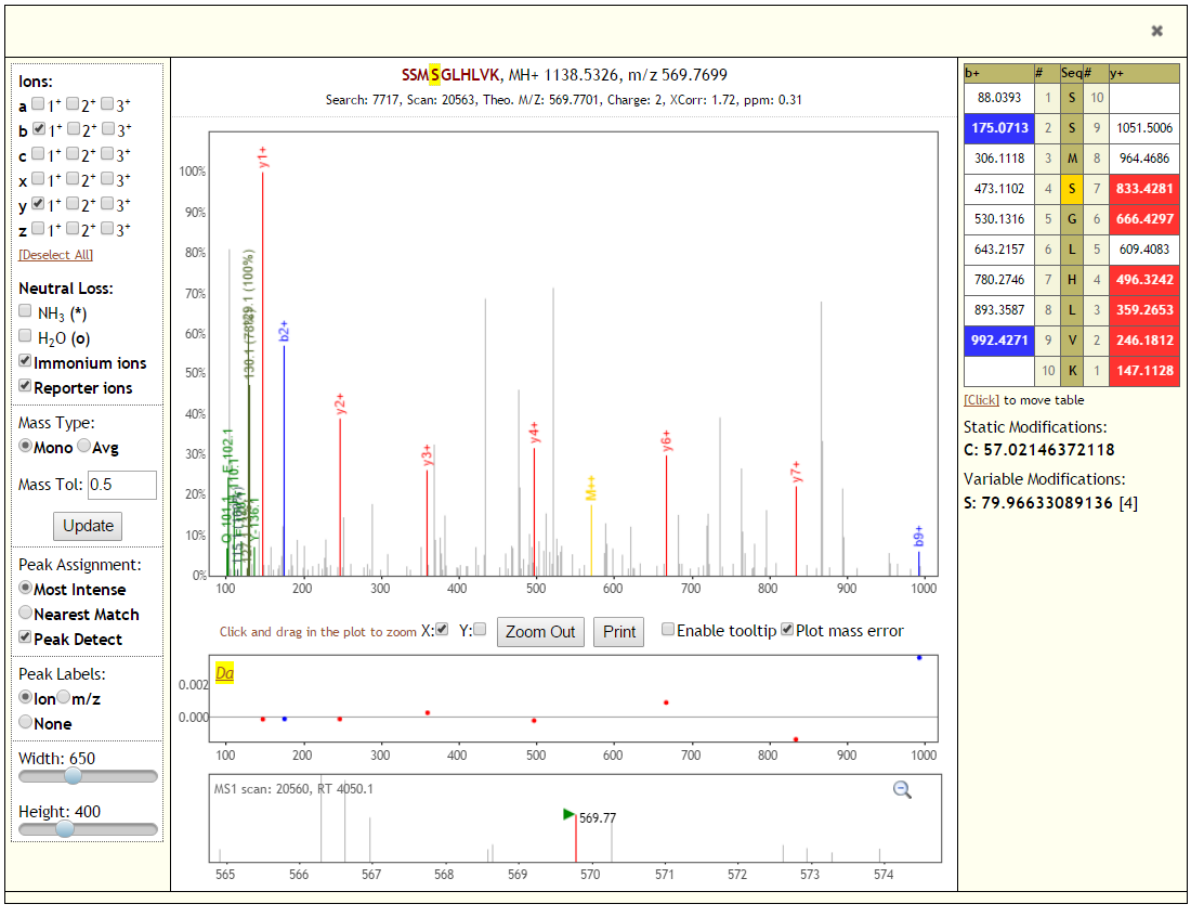


3. Choose “Select Peptides”. A list of precursor ions that have been observed for each peptide should appear.

4. Click on the peptide sequence to see its MS/MS spectrum.



MS/MS spectra are visualized using Lorikeet (http://uwpr.github.io/Lorikeet/)



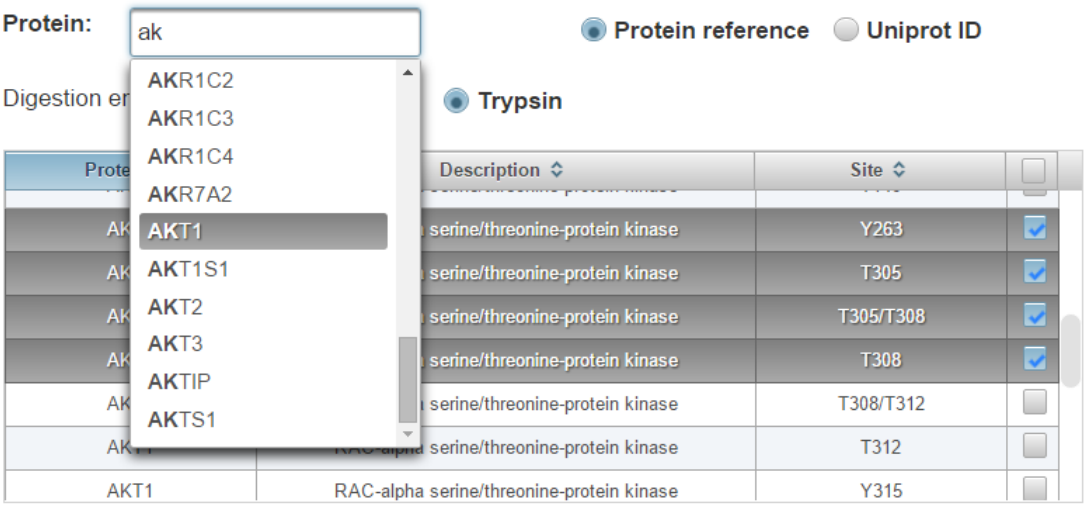
5. Check the box next to the precursor you would like to add or top right box to select all. The most frequently observed charge state is selected by default. Click “Add selected peptides”. The selected phosphopeptide precursor ions will be added to the current assay schedule which can be visualized at the bottom of the page (See section entitled “Assay Refinement”)

6. Repeat steps 1-5 to add more targets.

ii) Select phosphosites (automatic)

This option is similar as above, but precursor selection is performed automatically. It is the fastest way to browse the database to add individual phosphosites to an assay.

1. Start typing a protein identifier (HGNC gene symbol or Uniprot/Swiss-Prot accession are supported) and choose from the drop-down list. All phosphorylation site isoforms that have been observed on the protein will populate the table.
2. Check the boxes next to the phosphorylation sites you would like to add to the current assay or the top right box to select all.

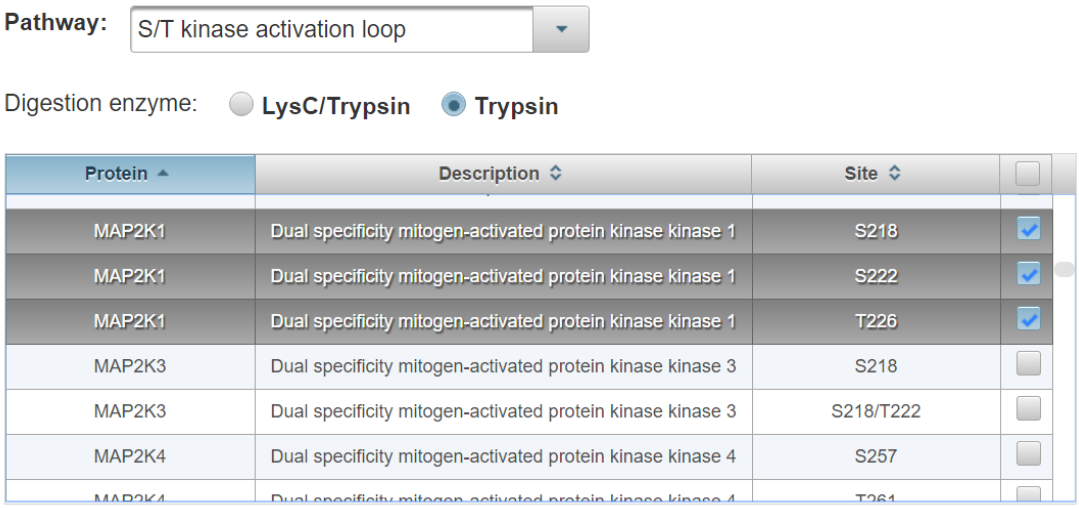


3. Click “Add Selected Sites”. The most frequently observed peptide sequence and charge state will be automatically added to the current assay schedule which can be visualized at the bottom of the page (See section entitled “Assay refinement”)

iii) Browse pre-built assays

This option provides a menu of pre-built lists of phosphorylation sites curated from phosphospecific antibody catalogs, pathway diagrams, or derived on the basis of protein structural features. It is the easiest way to assemble a large assay containing phosphosites of known function. These lists are provided as a starting point to simplify assay design. They are not comprehensive nor are they necessarily generalizable to the unique signaling biology of specific human cell types.

1. Select a pathway from the drop-down menu. Check the boxes next to the phosphorylation sites you would like to add to the current assay or top right box to select all.



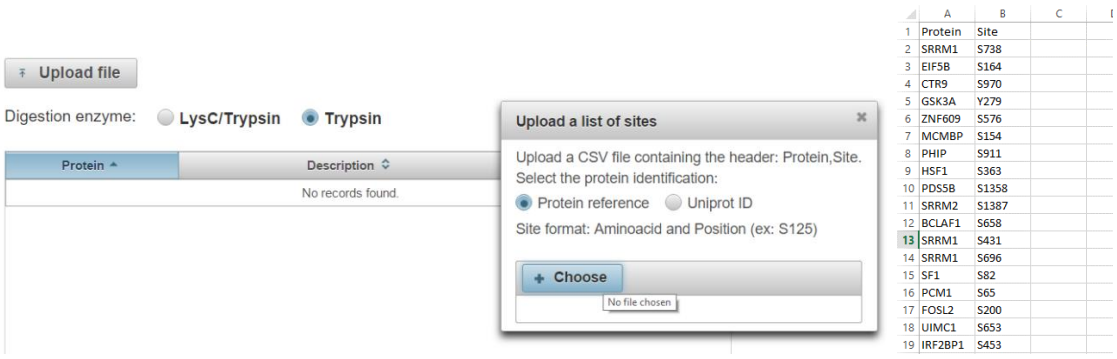
2. Click “Add Selected Sites”. The most frequently observed peptide sequence and charge state will be automatically added to the current assay schedule which can be visualized at the bottom of the page (See section entitled “Assay Refinement”)

3. Add more phosphosites from other pathways and proteins to create a customized assay for your experiment.

iv) Upload a list of sites

This option allows you to upload a custom list of phosphorylation sites for automatic precursor selection. It is the fastest way to generate a large assay for example to validate or expand on results from discovery experiments.

1. Click “Upload file” to generate a dialog for .csv file upload. The file must be in the format shown below, using HGNC gene symbols or Uniprot ID.



2. Check the boxes next to the phosphorylation sites you would like to add to the current assay or the top right box to select all.

Protein	Description	Site	
AXL	Tyrosine-protein kinase receptor UFO	Y702	☐
BABAM1	BRISC and BRCA1-A complex member 1	S29	☑
BAD	Bcl2-associated agonist of cell death	S134	☑
BAP1	Ubiquitin carboxyl-terminal hydrolase BAP1	S592	☑
BAP18	Chromatin complexes subunit BAP18	S96	☑
BCAR1	Breast cancer anti-estrogen resistance protein 1	Y249	☐
BCAR1	Breast cancer anti-estrogen resistance protein 1	Y410	☐

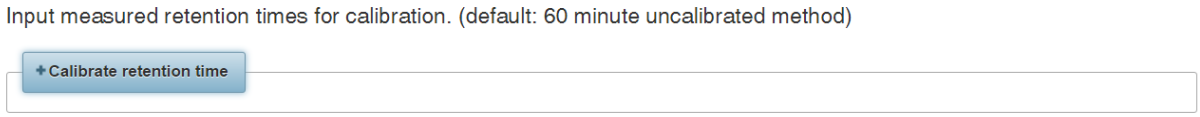
3. Click “Add Selected Sites”. The most frequently observed peptide sequence and charge state will be automatically added to the current assay schedule which can be visualized at the bottom of the page (See section entitled “Assay refinement”)

C.3.3 Retention time calibration

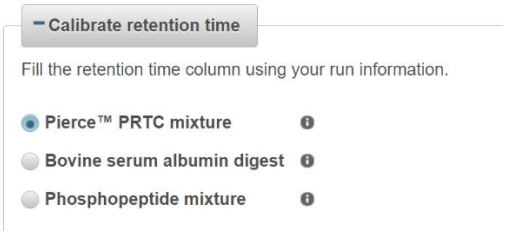
Retention times in the database are stored as iRT values, a dimensionless unit of relative hydrophobicity. For accurate retention time scheduling the assay must first be calibrated by analyzing a set of peptides measured using the same liquid chromatography configuration to be used for the experiment (i.e. the same pre-column, column, gradient, and flow rate).

For more information:
Escher C, Reiter L, MacLean B, Ossola R, Herzog F, Chilton J, MacCoss MJ, Rinner O. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics. 2012.

Click the “Calibrate retention time” button to expand the RT calibration panel.

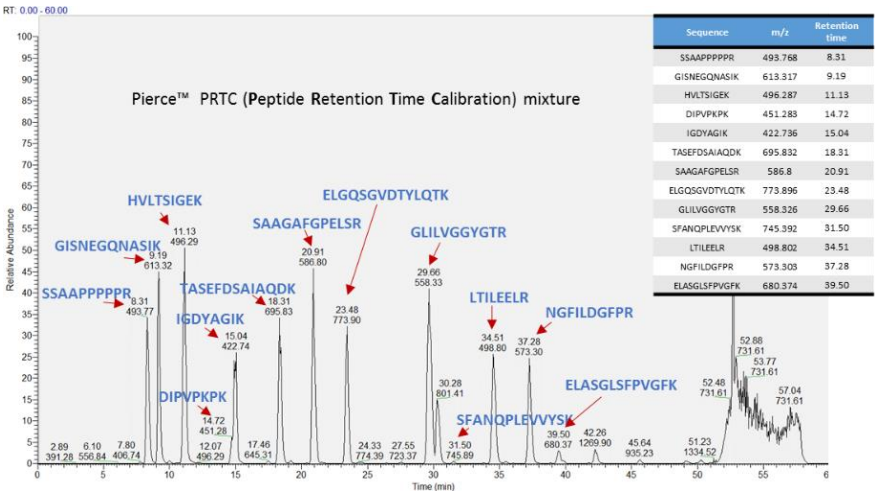


Currently there are three options for retention time calibration. Using more calibration points enables more accurate prediction. We achieved the greatest accuracy by using data-dependent acquisition (DDA) to measure a complex mixture of human phosphopeptides.



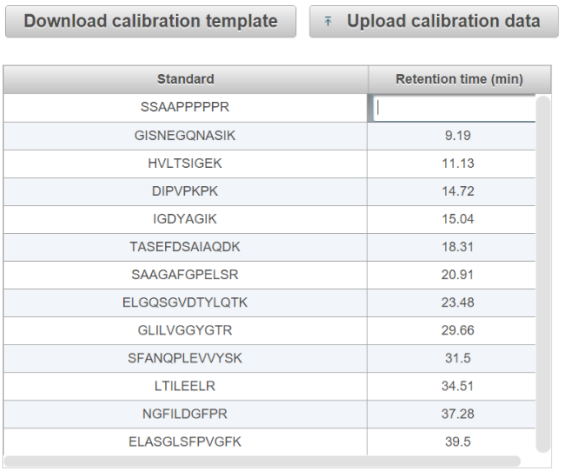
i) Pierce™ PRTC mixture (cat#. 88320)

1. Analyze the PRTC peptide mixture using your LC-MS/MS system.



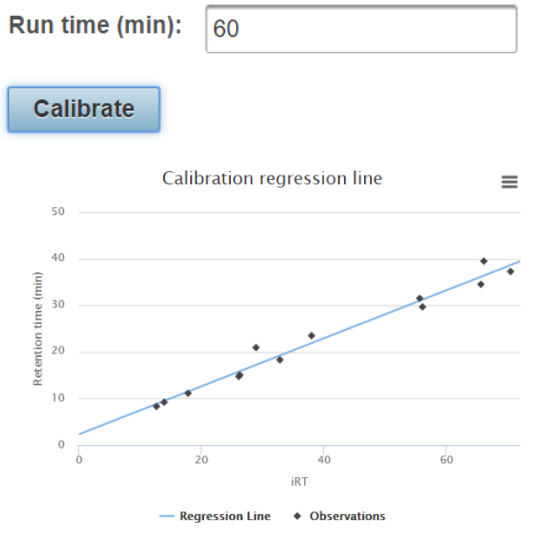
Sequence	m/z	Retention time
SSAAPPPPPR	493.768	8.31
GISNEGQNASIK	613.317	9.19
HVLTSIGEK	496.287	11.13
DIPVPKPK	451.283	14.72
IGDYAGIK	422.736	15.04
TASEFDSAIAQDK	695.832	18.31
SAAGAFGPELSR	586.8	20.91
ELGQSGVDTYLQTK	773.896	23.48
GLILVGGYGTR	558.326	29.66
SFANQPLEVVYSK	745.392	31.50
LTILEELR	498.802	34.51
NGFILDGFPR	573.303	37.28
ELASGLSFPVGFK	680.374	39.50

2. Enter the empirical retention times of each peptide directly in the table as below OR download the template .csv file, edit the table, and re-upload.



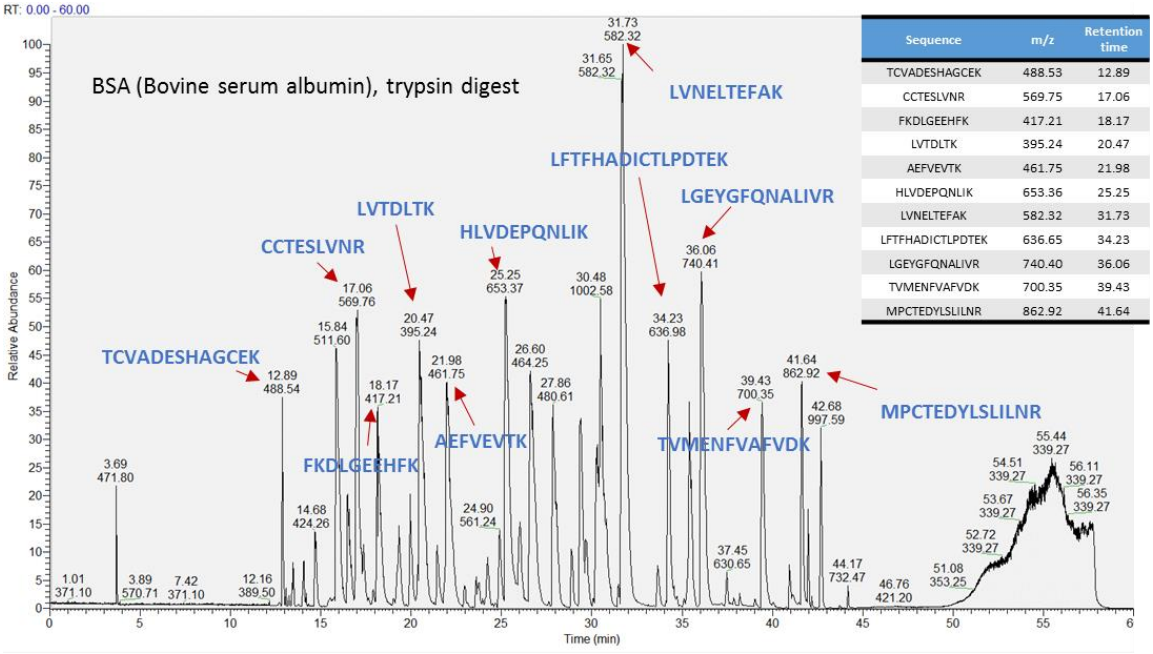
Standard	Retention time (min)
SSAAPPPPPR	
GISNEGQNASIK	9.19
HVLTSIGEK	11.13
DIPVPKPK	14.72
IGDYAGIK	15.04
TASEFDSAIAQDK	18.31
SAAGAFGPELSR	20.91
ELGQSGVDTYLQTK	23.48
GLILVGGYGTR	29.66
SFANQPLEVVYSK	31.5
LTILEELR	34.51
NGFILDGFPR	37.28
ELASGLSFPVGFK	39.5

3. Enter the total length of the method in minutes in the box labeled “Run time” and click “Calibrate”. The retention times of the current assay will be updated automatically.



ii) Bovine serum albumin (BSA) digest

1. Prepare a BSA peptide standard by digesting BSA protein using trypsin. Analyze the BSA peptide mixture using your LC-MS/MS system. Some representative peptides are illustrated below, but more than 100 BSA-derived peptides are included in the retention time database. For this reason we recommend analyzing the sample using DDA to identify as many peptides as possible.



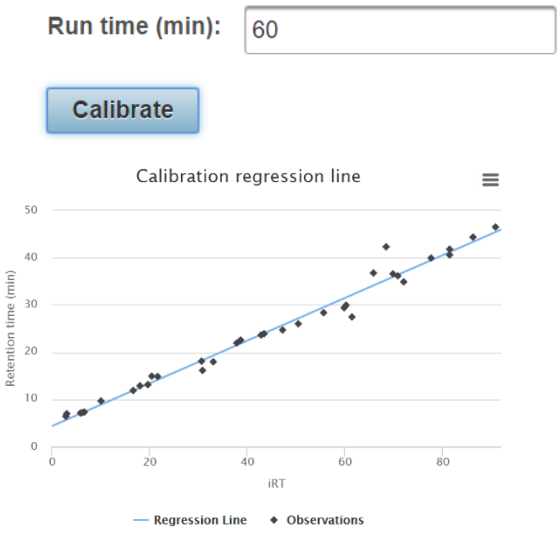
Sequence	m/z	Retention time
TCVADESHAGCEK	488.53	12.89
CCTESLVNR	569.75	17.06
FKDLGEEHFK	417.21	18.17
LVTDLTK	395.24	20.47
AEFVEVTK	461.75	21.98
HLVDEPQNLIK	653.36	25.25
LVNELTEFAK	582.32	31.73
LFTFHADICTLPDTEK	636.65	34.23
LGEYGFQNALIVR	740.40	36.06
TVMENFVAFVDK	700.35	39.43
MPCTEDYLSLILNR	862.92	41.64

2. Enter the empirical retention times of each peptide directly in the table as below (not recommended) OR download the template .csv file and upload a complete list of BSA peptides identified using DDA (recommended).

Download calibration template Upload calibration data

Standard	Retention time (min)
DTHKSEIAHR	6.42
HKPKATEEQLK	6.96
TPVSEKVTK	7.12
SLGKVGTR	7.2
ATEEQLK	7.3
LSQKFPK	9.67
KFWGK	11.86
DLGEEHFK	12.87
DDSPDLPK	13.12
LVTDLTKVHK	14.84
FKDLGEEHFK	14.93
LVTDLTK	16.12
AEFVEVTK	17.95

3. Enter the total length of the method in minutes in the box labeled “Run time” and click “Calibrate”. The retention times of the current assay will be updated automatically.

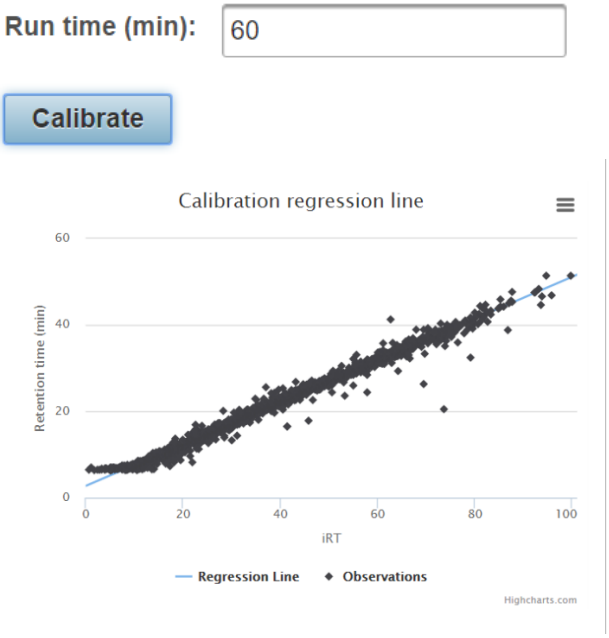


iii) Phosphopeptide mixture (human)

1. Prepare an aliquot of phosphopeptides from a sample you will be using for your experiment or prepare a human phosphopeptide sample similar to those to be analyzed in your experiment (e.g. use a HeLa lysate if your experiment is performed in HeLa cells). Analyze by DDA using your LC-MS/MS system.

2. Upload the peptide sequences and retention times using the template provided.

3. Enter the total length of the method in minutes in the box labeled “Run time” and click “Calibrate”. The retention times of the current assay will be updated automatically.



B.3.4 Assay refinement

Once you have calibrated retention times you are ready to optimize your assay for export. The number of targets you can measure concurrently depends on the scanning rate of the mass spectrometer, chromatographic peak width, and the desired number of points along the curve. For example, if the peak-width is 20 seconds and you would like to measure 10 points, then the maximum duty cycle is 2 seconds. If the mass spectrometer scans at a rate of 10 Hz, it is feasible to measure up to 20 targets every 2 seconds.

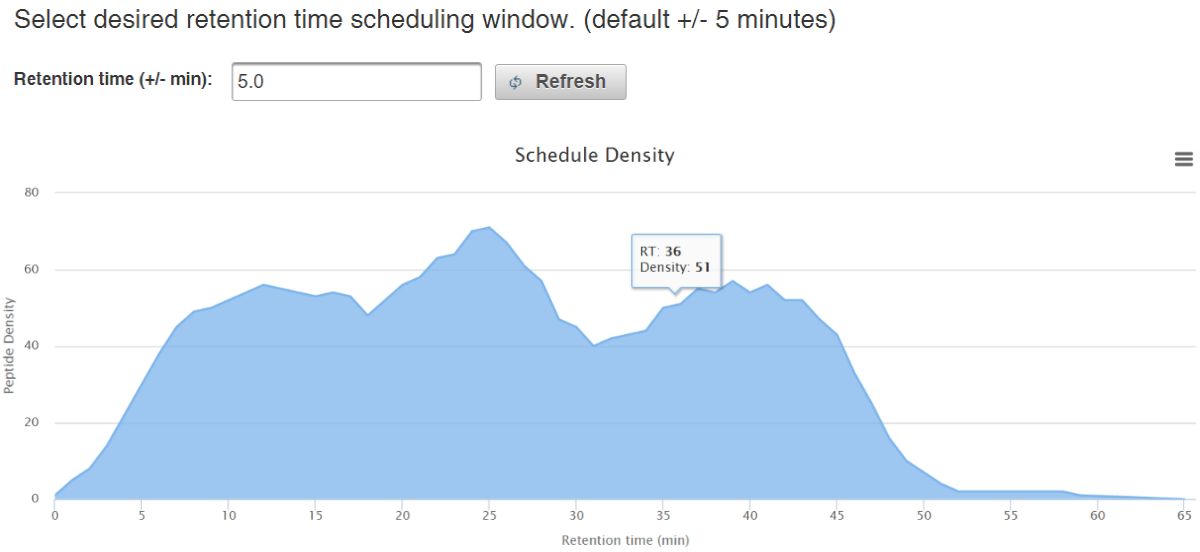
Recommended strategy: screen, optimize, deploy

By using wide retention time windows and a lower peak sampling rate, it is feasible to screen several hundred targets simultaneously. Use a “screening” approach to empirically pinpoint retention times and exclude targets that cannot be detected. Then, narrow the window enough to accommodate intraday chromatographic variability (typically <30s in a 60min method) and increase the maximum injection time to boost sensitivity until the desired duty cycle is reached. The assay is ready to deploy.

The UWPR provides a helpful overview of some considerations for targeted assay design: http://proteomicsresource.washington.edu/tools/PRM.php

Using the portal:

1. Enter the desired retention time tolerance window in the box. This represents the length of time before and after the predicted retention time to monitor for each phosphopeptide. The default tolerance is +/- 5 minutes, which was sufficient to capture >99% of targets using DDA analysis of a phosphopeptide mixture for calibration. Click refresh to update the schedule density plot.



2. [Optional] Continue to add or remove phosphopeptide precursors to reach the desired number of concurrent precursors (precursor density). Click the (x) next to a peptide to remove it from the assay.

Current Assay

Mass [m/z]	CS [z]	Start [min]	End [min]	Comment	
569.77	2	21.01255	26.01255	ACACA:S80 SSMS[80]GLHLVK	✖
873.391	2	13.128245	18.128245	ACLY:S455 TAS[80]FSESRADEVAPAK	✖
733.987	3	20.205606	25.205606	AKT1:S129 SGSPSDNS[80]GAEEMEVSLAKPK	✖
578.254	3	25.379217	30.379217	AKT1:S473 RPHFPQFS[80]YSASGTA	✖
1263.53	2	38.208796	43.208796	AKT1:T308 T[80]FC[57]GTPEYLAPEVLEDNDYGR	✖
698.357	2	44.948465	49.948465	AKT1S1:S183 S[80]LPVSVPVWGFK	✖
516.724	2	9.495322	14.495322	AKT1S1:T246 LNT[80]SDFQK	✖
583.926	3	28.581859	33.581859	AKT2:S474 THFPQFS[80]YSASIRE	✖
1172.505	3	38.683474	43.683474	AKT2:T309 EGISDGATMKT[80]FC[57]GTPEYLAPEVLEDNDYGR	✖
1200.226	3	45.879206	50.879206	AKT2:T451 YFDDEFTAQSITIT[80]PPDRYDSLGLLELDQR	✖
728.346	3	23.45652	28.45652	AP2M1:T156 EEQSQITSQVT[80]GQIGWRR	✖
950.46	3	36.247166	41.247166	ARAF:S299 NLGYRDS[80]GYYWEVPPSEVQLLKR	✖

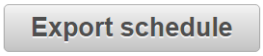
3. Click “Upload schedule” to upload an assay previously generated using this tool. This is helpful for adding new targets and recalibrating the retention times of old assays.

4. Click “Merge positional isomers” (e.g. THFPQFS[80]YSASIRE and THFPQFSYSAS[80]IRE) to merge peptides with the same precursor m/z. Since these precursors often have overlapping windows, this eliminates redundant sampling.

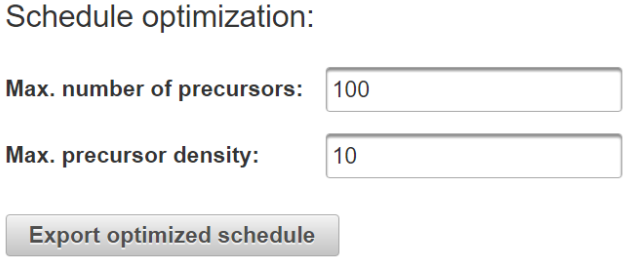
C.3.5 Assay export

Two options are provided for exporting the current assay.

1. Click “Export schedule” to download the current assay schedule above.



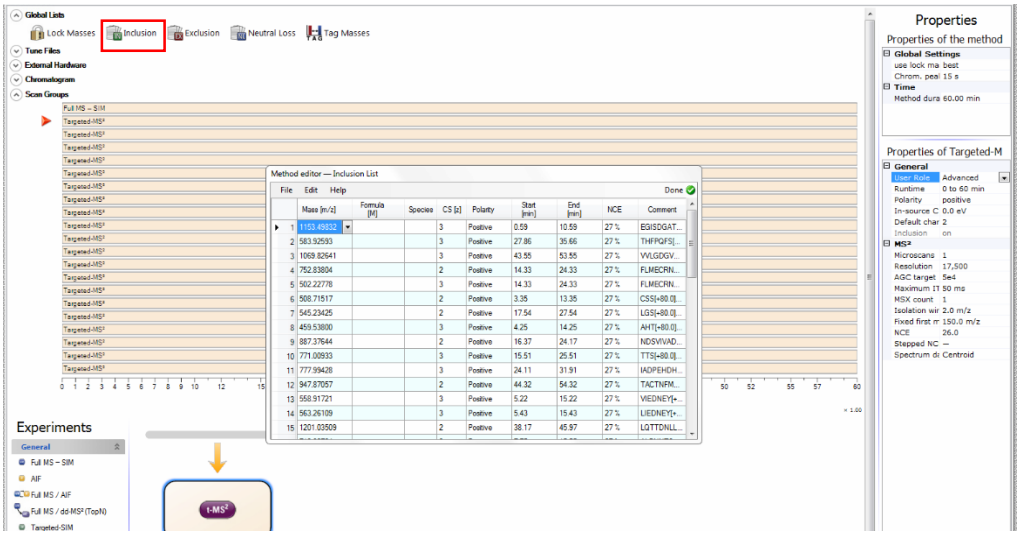
2. Enter the maximum number of precursors and maximum precursor density as desired. Click “Export optimized schedule” to automatically divide the assay into multiple runs.



3. The exported schedule is a csv file directly compatible with Thermo Q-Exactive mass spectrometers and can be modified to work with virtually any platform.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Mass [m/z]	Formula [I	Species	CS [z]	Polarity	Start [min	End [min]	NCE	Comment				
2	569.77			2	Positive	19.463	29.463	27	ACACA:S80 SSMS[80]GLHLVK				
3	873.391			2	Positive	11.0711	21.0711	27	ACLY:S455 TAS[80]FSESRADEVAPAK				
4	733.987			3	Positive	18.6041	28.6041	27	AKT1:S129 SGSPSDNS[80]GAEEMEVSLAKPK				
5	578.254			3	Positive	24.1108	34.1108	27	AKT1:S473 RPHFPQFS[80]YSASGTA				
6	889.897			4	Positive	34.1984	44.1984	27	AKT1:T308 EGIKDGATMKT[80]FC[57]GTPEYLAPEVLED				
7	698.357			2	Positive	44.9398	54.9398	27	AKT1S1:S183 S[80]LPVSVPVWGFK				
8	516.724			2	Positive	7.20433	17.2043	27	AKT1S1:T246 LNT[80]SDFQK				
9	583.926			3	Positive	27.5196	37.5196	27	AKT2:S474 THFPQFS[80]YSASIRE				
10	1172.51			3	Positive	38.2715	48.2715	27	AKT2:T309 EGISDGATMKT[80]FC[57]GTPEYLAPEVLEDI				
11	1200.23			3	Positive	45.9305	55.9305	27	AKT2:T451 YFDDEFTAQSITIT[80]PPDRYDSLGLLELDQR				
12	1013.97			2	Positive	26.5598	36.5598	27	AP2M1:T156 EEQSQITSQVT[80]GQIGWR				
13	950.46			3	Positive	35.6784	45.6784	27	ARAF:S299 NLGYRDS[80]GYYWEVPPSEVQLLKR				
14	808.072			3	Positive	30.4732	40.4732	27	ARHGEF2:S886 RRS[80]LPAGDALYLSFNPPQPSR				
15	810.336			3	Positive	5.03824	15.0382	27	ARRB1:S412 GMKDDKEEEEDGTGS[80]PQLNNR				
16	847.393			2	Positive	14.2996	24.6617	27	ATF2:T69 NDSVIVADQT[80]PTPTR - ATF2:T71 NDSVIV				

4. To import the assay into a Thermo Q-Exactive method, click “Global Lists”-> “Inclusion” -> “File” -> “Import”



References

Abelin, J.G., Patel, J., Lu, X., Feeney, C.M., Fagbami, L., Creech, A.L., Hu, R., Lam, D., Davison, D., Pino, L., et al. (2016). Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. Mol. Cell. Proteomics MCP.

Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature *500*, 207–211.

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Arteaga, C.L., and Baselga, J. (2012). Impact of Genomics on Personalized Cancer Medicine. Clin. Cancer Res. *18*, 612–618.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., et al. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405–409.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. *389*, 1017–1031.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol. *24*, 1285–1292.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549.

Berridge, M.J. (2014). Cell Signalling Biology.

Brown, E.J., and Frazier, W.A. (2001). Integrin-associated protein (CD47) and its ligands. Trends Cell Biol. *11*, 130–135.

Carpenter, G., King, L., and Cohen, S. (1978). Epidermal growth factor stimulates phosphorylation in membrane preparations in vitro. Nature *276*, 409–410.

Carracedo, A., Ma, L., Teruya-Feldstein, J., Rojo, F., Salmena, L., Alimonti, A., Egia, A., Sasaki, A.T., Thomas, G., Kozma, S.C., et al. (2008). Inhibition of mTORC1 leads to MAPK pathway activation through a PI3K-dependent feedback loop in human cancer. J. Clin. Invest. *118*, 3065–3074.

Chatterjee, P., Choudhary, G.S., Sharma, A., Singh, K., Heston, W.D., Ciezki, J., Klein, E.A., and Almasan, A. (2013). PARP Inhibition Sensitizes to Low Dose-Rate Radiation TMPRSS2-

ERG Fusion Gene-Expressing and PTEN-Deficient Prostate Cancer Cells. PLoS ONE *8*, e60408.

Cohen, P. (1992). Signal integration at the level of protein kinases, protein phosphatases and their substrates. Trends Biochem. Sci. *17*, 408–413.

Cohen, P. (2002a). The origins of protein phosphorylation. Nat. Cell Biol. *4*, E127–E130.

Cohen, P. (2002b). Protein kinases--the major drug targets of the twenty-first century? Nat. Rev. Drug Discov. *1*, 309–315.

Cori, C.F., and Cori, G.T. (1928). The mechanism of epinephrine action II. The influence of epinephrine and insulin on the carbohydrate metabolism of rats in the postabsorptive state. J. Biol. Chem. *79*, 321–341.

Cori, G.T., and Green, A.A. (1943). Crystalline muscle phosphorylase II. Prosthetic group. J. Biol. Chem. *151*, 31–38.

Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Dickhut, C., Feldmann, I., Lambert, J., and Zahedi, R.P. (2014). Impact of digestion conditions on phosphoproteomics. J. Proteome Res. *13*, 2761–2770.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989.

Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics *13*, 22–24.

Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. (2002). Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nat. Biotechnol. *20*, 301–305.

Fischer, E.H., and Krebs, E.G. (1955). Conversion of phosphorylase b to phosphorylase a in muscle extracts. J. Biol. Chem. *216*, 121–132.

Fischer, E.H., Graves, D.J., Crittenden, E.R.S., and Krebs, E.G. (1959). Structure of the site phosphorylated in the phosphorylase b to a reaction. J. Biol. Chem. *234*, 1698–1704.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

Frame, M.C., Patel, H., Serrels, B., Lietha, D., and Eck, M.J. (2010). The FERM domain: organizing the structure and function of FAK. Nat. Rev. Mol. Cell Biol. *11*, 802–814.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Geiger, T., Madden, S.F., Gallagher, W.M., Cox, J., and Mann, M. (2012a). Proteomic portrait of human breast cancer progression identifies novel prognostic markers. Cancer Res. *72*, 2428–2439.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012b). Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. Mol. Cell. Proteomics *11*, M111.014050.

Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc. Natl. Acad. Sci. U. S. A. *100*, 6940–6945.

Gholami, A.M., Hahne, H., Wu, Z., Auer, F.J., Meng, C., Wilhelm, M., and Kuster, B. (2013). Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Rep. *4*, 609–620.

Giansanti, P., Aye, T.T., van den Toorn, H., Peng, M., van Breukelen, B., and Heck, A.J.R. (2015). An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. Cell Rep. *11*, 1834–1843.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics MCP *11*, O111.016717.

de Graaf, E.L., Giansanti, P., Altelaar, A.F.M., and Heck, A.J.R. (2014). Single-step enrichment by Ti4+-IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. Mol. Cell. Proteomics MCP *13*, 2426–2434.

de Graaf, E.L., Kaplon, J., Mohammed, S., Vereijken, L.A.M., Duarte, D.P., Redondo Gallego, L., Heck, A.J.R., Peeper, D.S., and Altelaar, A.F.M. (2015). Signal Transduction Reaction Monitoring Deciphers Site-Specific PI3K-mTOR/MAPK Pathway Dynamics in Oncogene-Induced Senescence. J. Proteome Res. *14*, 2906–2914.

Graves, J.D., and Krebs, E.G. (1999). Protein phosphorylation and signal transduction. Pharmacol. Ther. *82*, 111–121.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19*, 1720–1730.

Hoehn, K.L., Hohnen-Behrens, C., Cederberg, A., Wu, L.E., Turner, N., Yuasa, T., Ebina, Y., and James, D.E. (2008). IRS1-independent defects define major nodes of insulin resistance. Cell Metab. *7*, 421–433.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.

Howlader, N., Altekruse, S.F., Li, C.I., Chen, V.W., Clarke, C.A., Ries, L.A.G., and Cronin, K.A. (2014). US Incidence of Breast Cancer Subtypes Defined by Joint Hormone Receptor and HER2 Status. J. Natl. Cancer Inst. *106*, dju055.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huber, M.A., Azoitei, N., Baumann, B., Grünert, S., Sommer, A., Pehamberger, H., Kraut, N., Beug, H., and Wirth, T. (2004). NF-kappaB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. J. Clin. Invest. *114*, 569–581.

Hudis, C.A., and Gianni, L. (2011). Triple-Negative Breast Cancer: An Unmet Medical Need. The Oncologist *16*, 1–11.

Humphrey, S.J., Yang, G., Yang, P., Fazakerley, D.J., Stöckli, J., Yang, J.Y., and James, D.E. (2013). Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. Cell Metab. *17*, 1009–1020.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Ishida, H., Li, K., Yi, M., and Lemon, S.M. (2007). p21-activated kinase 1 is activated through the mammalian target of rapamycin/p70 S6 kinase pathway and regulates the replication of hepatitis C virus in human hepatoma cells. J. Biol. Chem. *282*, 11836–11848.

Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925.

Kennedy, J.J., Abbatiello, S.E., Kim, K., Yan, P., Whiteaker, J.R., Lin, C., Kim, J.S., Zhang, Y., Wang, X., Ivey, R.G., et al. (2014). Demonstrating the feasibility of large-scale development of standardized assays to quantify human proteins. Nat. Methods *11*, 149–155.

Kenny, P.A., Lee, G.Y., Myers, C.A., Neve, R.M., Semeiks, J.R., Spellman, P.T., Lorenz, K., Lee, E.H., Barcellos-Hoff, M.H., Petersen, O.W., et al. (2007). The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. Mol. Oncol. *1*, 84–96.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Krebs, E.G., and Fischer, E.H. (1956). The phosphorylase b to a converting enzyme of rabbit skeletal muscle. Biochim. Biophys. Acta *20*, 150–157.

Lawrence, J.C. (1992). Signal transduction and protein phosphorylation in the regulation of cellular metabolism by insulin. Annu. Rev. Physiol. *54*, 177–193.

Lawrence, R.T., Perez, E.M., Hernández, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.-I., Blau, C.A., and Villén, J. (2015). The proteomic landscape of triple-negative breast cancer. Cell Rep. *11*, 630–644.

Lawrence, R.T., Searle, B.C., Llovet, A., and Villén, J. (2016). Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nat. Methods.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*, 2750–2767.

Lin, X., Duan, X., Liang, Y.-Y., Su, Y., Wrighton, K.H., Long, J., Hu, M., Davis, C.M., Wang, J., Brunicardi, F.C., et al. (2006). PPM1A functions as a Smad phosphatase to terminate TGFbeta signaling. Cell *125*, 915–928.

Lundby, A., Secher, A., Lage, K., Nordsborg, N.B., Dmytriyev, A., Lundby, C., and Olsen, J.V. (2012). Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. Nat. Commun. *3*, 876.

Luo, J.-L., Maeda, S., Hsu, L.-C., Yagita, H., and Karin, M. (2004). Inhibition of NF-κB in cancer cells converts inflammation- induced tumor growth mediated by TNFα to TRAIL-mediated tumor regression. Cancer Cell *6*, 297–305.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinforma. Oxf. Engl. *26*, 966–968.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. *583*, 3966–3973.

Manning, B.D. (2004). Balancing Akt with S6K: implications for both metabolic diseases and tumorigenesis. J. Cell Biol. *167*, 399–403.

McLean, G.W., Carragher, N.O., Avizienyte, E., Evans, J., Brunton, V.G., and Frame, M.C. (2005). The role of focal-adhesion kinase in cancer - a new therapeutic opportunity. Nat. Rev. Cancer *5*, 505–515.

Mendes-Pereira, A.M., Martin, S.A., Brough, R., McCarthy, A., Taylor, J.R., Kim, J.-S., Waldman, T., Lord, C.J., and Ashworth, A. (2009). Synthetic lethal targeting of PTEN mutant cells with PARP inhibitors. EMBO Mol. Med. *1*, 315–322.

Miyashita, M., Oshiumi, H., Matsumoto, M., and Seya, T. (2011). DDX60, a DEXD/H box helicase, is a novel antiviral factor promoting RIG-I-like receptor-mediated signaling. Mol. Cell. Biol. *31*, 3802–3819.

Moritz, A., Li, Y., Guo, A., Villén, J., Wang, Y., MacNeill, J., Kornhauser, J., Sprott, K., Zhou, J., Possemato, A., et al. (2010). Akt-RSK-S6 kinase signaling networks activated by oncogenic receptor tyrosine kinases. Sci. Signal. *3*, ra64.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics MCP *4*, 1419–1440.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515–527.

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell *127*, 635–648.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., et al. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci. Signal. *3*, ra3.

Parker, B.L., Yang, G., Humphrey, S.J., Chaudhuri, R., Ma, X., Peterman, S., and James, D.E. (2015). Targeted phosphoproteomics of insulin signaling using data-independent acquisition mass spectrometry. Sci. Signal. *8*, rs6.

Pearce, L.R., Komander, D., and Alessi, D.R. (2010). The nuts and bolts of AGC protein kinases. Nat. Rev. Mol. Cell Biol. *11*, 9–22.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol. Cell. Proteomics MCP *11*, 1475–1488.

Petruzzelli, L.M., Ganguly, S., Smith, C.J., Cobb, M.H., Rubin, C.S., and Rosen, O.M. (1982). Insulin activates a tyrosine-specific protein kinase in extracts of 3T3-L1 adipocytes and human placenta. Proc. Natl. Acad. Sci. U. S. A. *79*, 6792–6796.

Pinna, L.A., and Ruzzene, M. (1996). How do protein kinases recognize their substrates? Biochim. Biophys. Acta *1314*, 191–225.

Prat, A., and Perou, C.M. (2011). Deconstructing the molecular portraits of breast cancer. Mol. Oncol. *5*, 5–23.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. *12*, R68.

Quail, D.F., and Joyce, J.A. (2013). Microenvironmental regulation of tumor progression and metastasis. Nat. Med. *19*, 1423–1437.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. *2*, 1896–1906.

Ray, L.B., and Sturgill, T.W. (1987). Rapid stimulation by insulin of a serine/threonine kinase in 3T3-L1 adipocytes that phosphorylates microtubule-associated protein 2 in vitro. Proc. Natl. Acad. Sci. U. S. A. *84*, 1502–1506.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., et al. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell *131*, 1190–1203.

Roach, P.J. (1991). Multisite and hierarchal protein phosphorylation. J. Biol. Chem. *266*, 14139–14142.

Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223.

Sakurai, H., Chiba, H., Miyoshi, H., Sugita, T., and Toriumi, W. (1999). IkappaB kinases phosphorylate NF-kappaB p65 subunit on serine 536 in the transactivation domain. J. Biol. Chem. *274*, 30353–30356.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Sharma, K., D'Souza, R.C.J., Tyanova, S., Schaab, C., Wiśniewski, J.R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep. *8*, 1583–1594.

Shaw, R.J., and Cantley, L.C. (2006). Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature *441*, 424–430.

Sick, E., Jeanne, A., Schneider, C., Dedieu, S., Takeda, K., and Martiny, L. (2012). CD47 update: a multifaceted actor in the tumour microenvironment of potential therapeutic interest. Br. J. Pharmacol. *167*, 1415–1430.

Smith, C.J., Wejksnora, P.J., Warner, J.R., Rubin, C.S., and Rosen, O.M. (1979). Insulin-stimulated protein phosphorylation in 3T3-L1 preadipocytes. Proc. Natl. Acad. Sci. U. S. A. *76*, 2725–2729.

Soste, M., Hrabakova, R., Wanka, S., Melnik, A., Boersema, P., Maiolica, A., Wernas, T., Tognetti, M., von Mering, C., and Picotti, P. (2014). A sentinel protein assay for simultaneously quantifying cellular processes. Nat. Methods *11*, 1045–1048.

Stagg, J., Divisekera, U., McLaughlin, N., Sharkey, J., Pommey, S., Denoyer, D., Dwyer, K.M., and Smyth, M.J. (2010). Anti-CD73 antibody therapy inhibits breast tumor growth and metastasis. Proc. Natl. Acad. Sci. U. S. A. *107*, 1547–1552.

Subik, K., Lee, J.-F., Baxter, L., Strzepek, T., Costello, D., Crowley, P., Xing, L., Hung, M.-C., Bonfiglio, T., Hicks, D.G., et al. (2010). The Expression Patterns of ER, PR, HER2, CK5/6, EGFR, Ki-67 and AR by Immunohistochemical Analysis in Breast Cancer Cell Lines. Breast Cancer Basic Clin. Res. *4*, 35–41.

Sutherland, E.W., and Cori, C.F. (1951). Effect of hyperglycemic-glycogenolytic factor and epinephrine on liver phosphorylase. J. Biol. Chem. *188*, 531–543.

Terfve, C.D.A., Wilkes, E.H., Casado, P., Cutillas, P.R., and Saez-Rodriguez, J. (2015). Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. Nat. Commun. *6*, 8033.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Tian, Q., Price, N.D., and Hood, L. (2012). Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J. Intern. Med. *271*, 111–121.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., and Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol. Cancer Ther. *5*, 2512–2521.

Ting, Y.S., Egertson, J.D., Payne, S.H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R.D., Noble, W.S., and MacCoss, M.J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Mol. Cell. Proteomics MCP *14*, 2301–2307.

Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A.I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods *12*, 258–264, 7 p following 264.

Ubersax, J.A., and Ferrell, J.E. (2007). Mechanisms of specificity in protein phosphorylation. Nat. Rev. Mol. Cell Biol. *8*, 530–541.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

Untch, M., Konecny, G.E., Paepke, S., and von Minckwitz, G. (2014). Current and future role of neoadjuvant therapy for breast cancer. Breast *23*, 526–537.

Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods *1*, 39–45.

Vidal, M., Chan, D.W., Gerstein, M., Mann, M., Omenn, G.S., Tagle, D., Sechi, S., and Workshop Participants (2012). The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. Clin. Proteomics *9*, 6.

Villén, J., and Gygi, S.P. (2008). The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc. *3*, 1630–1638.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science *339*, 1546–1558.

Vranic, S., Gatalica, Z., and Wang, Z.-Y. (2011). Update on the molecular profile of the MDA-MB-453 cell line as a model for apocrine breast carcinoma studies. Oncol. Lett. *2*, 1131–1137.

Wan, M., Leavens, K.F., Hunter, R.W., Koren, S., von Wilamowitz-Moellendorff, A., Lu, M., Satapati, S., Chu, Q., Sakamoto, K., Burgess, S.C., et al. (2013). A noncanonical, GSK3-independent pathway controls postprandial hepatic glycogen deposition. Cell Metab. *18*, 99–105.

Wang, X., and Lin, Y. (2008). Tumor necrosis factor and cancer, buddies or foes? Acta Pharmacol. Sin. *29*, 1275–1288.

Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., et al. (1997). An Information-Intensive Approach to the Molecular Pharmacology of Cancer. Science *275*, 343–349.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Winston, B.W., Lange-Carter, C.A., Gardner, A.M., Johnson, G.L., and Riches, D.W. (1995). Tumor necrosis factor alpha rapidly activates the mitogen-activated protein kinase (MAPK) cascade in a MAPK kinase kinase-dependent, c-Raf-1-independent fashion in mouse macrophages. Proc. Natl. Acad. Sci. U. S. A. *92*, 1614–1618.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.

Yuan, Y., Van Allen, E.M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L.A., Xu, Y., Hess, K.R., Diao, L., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. *32*, 644–652.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhi, X., Wang, Y., Zhou, X., Yu, J., Jian, R., Tang, S., Yin, L., and Zhou, P. (2010). RNAi-mediated CD73 suppression induces apoptosis and cell-cycle arrest in human breast cancer cells. Cancer Sci. *101*, 2561–2569.

Zimmermann, H. (1992). 5'-Nucleotidase: molecular structure and functional aspects. Biochem. J. *285 (Pt 2)*, 345–365.

VITA

Robert Lawrence graduated with a bachelor of science in biomedical engineering and economics in 2009 from the University of Virginia, where he became interested in data-driven approaches to systems biology. He was selected for a postgraduate fellowship to study insulin signaling in the laboratory of Dr. David James at the Garvan Institute in Sydney, Australia. Inspired by the work of his colleagues down under, Rob found his calling for proteomics research. After returning to the USA, Rob continued to work on insulin signaling in the laboratory of Dr. Kyle Hoehn and joined the Molecular and Cellular Biology program at the University of Washington in 2011. He carried out his doctoral studies in the laboratory of Dr. Judit Villén in the Department of Genome Sciences. Outside the lab, Rob is an avid rower and outdoorsman with an aptitude for choosing dark and rainy weekends to go hiking.