5-2017

# Genome-scale Precision Proteomics Identifies Cancer Signaling Networks and Therapeutic Vulnerabilities

Hong Wang
*University of Tennessee Health Science Center*

# Genome-scale Precision Proteomics Identifies Cancer Signaling Networks and Therapeutic Vulnerabilities



**ORCID**
http://orcid.org/0000-0002-6215-6348

**DOI**
10.21007/etd.cghs.2017.0441

# Genome-scale Precision Proteomics Identifies Cancer Signaling Networks and Therapeutic Vulnerabilities

A Dissertation
Presented for
The Graduate Studies Council
The University of Tennessee
Health Science Center

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy
From The University of Tennessee

By
Hong Wang
May 2017

## DEDICATION

This dissertation is dedicated to the people in my life who have been supportive along the way, especially my parents and wife, who are the heroes of my life and the source of my inspiration.

## ACKNOWLEDGEMENTS

It is my great honor to work with my mentor Dr. Junmin Peng. This work would not have been possible without him. He has always been a brilliant and humble scientist as well as a great friend. His guidance, mentorship, and encouragement were invaluable for my scientific career. It is his passion for science, stringency in research, and care of others that made me understand how to become a great scientist. I am proud to say that I can walk away from his lab a better person and for that he deserves the credit.

I would like to acknowledge the Peng lab and the St. Jude proteomics core as a whole for providing an extremely supportive and friendly environment for me. It has been my family for the last five years. I appreciate every one of them for their utmost help, advice and discussions. Special thanks to Zhiping Wu, Haiyan Tan, Bing Bai, Tim Shaw, Xusheng Wang, Yuxin Li, Ji-Hoon Cho, Ping-chung Chen, Anthony High, and Jeffery Sifford for their help and advice.

I thank my collaborators Dr. Michael Dyer group and Dr. Suzanne Baker group including Elizabeth Stewart, Cori L Bradley, Ocarz, Monica, Alexander K. Diaz, and Barbara S. Paugh for being so supportive and helpful. It is my privilege to work with Mike and Suzy, they are always generous and willing to help, thanks for your invaluable insights to my work. I am sure that these wonderful experiences will benefit the rest of science career.

My thanks also goes to the faculty members of my committee, Drs. Suzanne Baker, Michael Dyer, David Nelson, Stanley Pounds, and Jinghui Zhang for their guidance along the progress of my Ph.D projects. I really appreciate the time and effort they invested in my work and Ph.D training.

I also thank the MIB track of IPBS program at UTHSC and the Structural Biology department at St. Jude Children's Research Hospital for providing an extraordinary training environment.

Lastly, I would like to thank my wife Mingming Niu for bringing happiness and love to my life, also thank my parents, parents-in-law and my friends for their ultimate support.

## ABSTRACT

Mass spectrometry (MS) based-proteomics technology has been emerging as an indispensable tool for biomedical research. But the highly diverse physical and chemical properties of the protein building blocks and the dramatic human proteome complexity largely limited proteomic profiling depth. Moreover, there was a lack of high-throughput quantitative strategies that were both precise and parallel to in-depth proteomic techniques. To solve these grand challenges, a high resolution liquid chromatography (LC) system that coupled with an advanced mass spectrometer was developed to allow genome-scale human proteome identification. Using the combination of pre-MS peptide fractionation, MS2-based interference detection and post-MS computational interference correction, we enabled precise proteome quantification with isobaric labeling. We then applied these advanced proteomics tools for cancer proteome analyses on high grade gliomas (HGG) and rhabdomyosarcomas (RMS). Using systems biology approaches, we demonstrated that these newly developed proteomic analysis pipelines are able to (i) define human proteotypes that link oncogenotypes to cancer phenotypes in HGG and to (ii) identify therapeutic vulnerabilities in RMS.

Development of high resolution liquid chromatography is essential for improving the sensitivity and throughput of mass spectrometry-based proteomics to genome-scale. Here we present systematic optimization of a long gradient LC-MS/MS platform to enhance protein identification from a complex mixture. The platform employed an in-house fabricated, reverse phase long column (100 μm x 150 cm, 5 μm C18 beads) coupled with Q Exactive MS. The column was capable of achieving a peak capacity of approximately 700 in a 720 min gradient of 10-45% acetonitrile. The optimal loading amount was about 6 micrograms of peptides, although the column allowed loading as many as 20 micrograms. Gas phase fractionation of peptide ions further increased the number of peptides identified by ~10%. Moreover, the combination of basic pH LC pre-fractionation with the long gradient LC-MS/MS platform enabled the identification of 96,127 peptides and 10,544 proteins at 1% protein false discovery rate in a postmortem brain sample of Alzheimer's disease. As deep RNA sequencing of the same specimen suggested that ~16,000 genes were expressed, current analysis covered more than 60% of the expressed proteome.

Isobaric labeling quantification by mass spectrometry has emerged as a powerful technology for multiplexed large-scale protein profiling, but measurement accuracy in complex mixtures is confounded by the interference from co-isolated ions, resulting in ratio compression. Here we report that the ratio compression can be essentially resolved by the combination of pre-MS peptide fractionation, MS2-based interference detection and post-MS computational interference correction. To recapitulate the complexity of biological samples, we pooled tandem mass tag (TMT) labeled *E. coli* peptides at 1 : 3 : 10 ratios, and added in ~20-fold more rat peptides as background, followed by the analysis of two dimensional liquid chromatography-MS/MS. Systematic investigation indicated that the quantitative interference was impacted by LC fractionation depth, MS isolation window and peptide loading amount. Exhaustive fractionation (320 x 4 h) can

nearly eliminate the interference and achieve results comparable to the MS3-based method. Importantly, the interference in MS2 scans can be estimated by the intensity of contaminated *y*1 product ions, and we thus developed an algorithm to correct reporter ion ratios of tryptic peptides. Our data indicated that intermediate fractionation (40 x 2 h) and *y*1 ion-based correction allowed accurate and deep TMT protein profiling, which represents a straightforward and affordable strategy in isobaric labeling proteomics

High throughput omics approaches provide an unprecedented opportunity for dissecting molecular mechanisms in cancer biology. Here we present deep profiling of whole proteome, phosphoproteome and transcriptome in two high-grade glioma mouse models driven by mutated receptor tyrosine kinase (RTK) oncogenes, platelet-derived growth factor receptor alpha (*PDGFRA*) and neurotrophic receptor tyrosine kinase 1 (*NTRK1*), analyzing 13,860 proteins (11,941 genes) and 30,431 phosphosites by mass spectrometry. Systems biology approaches identified numerous functional modules and master regulators, including 41 kinases and 26 transcription factors. Pathway activity computation and mouse survival curves indicate the NTRK1 mutation induces a higher activation of AKT targets, drives a positive feedback loop to up-regulate multiple other RTKs, and shows higher oncogenic potency than the PDGFRA mutation. Further integration of the mouse data with human HGG transcriptome data determines shared regulators of invasion and stemness. Thus, multi-omics integrative profiling is a powerful avenue to characterize oncogenic activity.

There is growing emphasis on personalizing cancer therapy based on somatic mutations identified in patient's tumors. Among pediatric solid tumors, RAS pathway mutations in rhabdomyosarcoma are the most common potentially actionable lesions. Recent success targeting CDK4/6 and MEK in RAS mutant adult cancers led our collaborator Dr. Dyer's group to test this approach for rhabdomyosarcoma. They achieved synergistic killing of RAS mutant rhabdomyosarcoma tumor cells by combining MEK and CDK4/6 inhibitors in culture but failed to achieve efficacy *in vivo* using orthotopic patient derived xenografts (O-PDXs). To determine how rhabdomyosarcomas evade targeting of CDK4/6 and MEK, we collaborated to perform large-scale deep proteomic, phosphoproteomic, and epigenomic profiling of RMS tumors. Integrative analysis of these omics data detected that RMS tumor cells rapidly compensate and overcome CDK4/6 and MEK combination therapy through 6 myogenic signal transduction pathways including WNT, HH, BMP, Adenyl Cyclase, P38/MAPK and PI3K. While it is not feasible to target each of these signal transduction pathways simultaneously in RMS, we discovered that they require the HSP90 chaperone to sustain the complex developmental signal transduction milieu. We achieved specific and synergistic killing of RMS cells using sub-therapeutic concentrations of an HSP90 inhibitor (ganetespib) in combination with conventional chemotherapy used for recurrent RMS. These effects were seen in the most aggressive recurrent RMS orthotopic patient derived xenografts irrespective of RAS pathway perturbations, histologic or molecular classification. Thus, multi-omics integrative cancer profiling using our newly developed tools is powerful to identify core signaling transduction networks, tumor vulnerability (master regulators) for novel cancer therapy.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AD | Alzheimer's Disease |
| ANOVA | Analysis of Variance |
| ARMS | Alveolar Rhabdomyosarcoma |
| CID | Collision Induced Dissociation |
| CNV | Copy Number Variations |
| CSTN | Childhood Solid Tumor Network |
| DE | Differential Expression |
| DiLeu | N-dimethyl Leucine Isobaric Tags |
| DMSO | Dimethyl Sulfoxide |
| DNA | Deoxyribonucleic Acid |
| ERMS | Embryonal Rhabdomyosarcoma |
| FDR | False Discovery Rate |
| FPKM | Kilobase of Exon per Million Fragments |
| FWHM | Full Width at Half Maximum |
| GFP | Green Fluorescent Protein |
| GPF | Gas Phase Fractionation |
| GSP | Ganetespib |
| HCD | Higher-energy Collisional Dissociation |
| HDB | High-dimensional Biology |
| HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic Acid |
| HGG | High Grade Glioma |
| HPLC | High-pressure Liquid Chromatography |

| | |
|---|---|
| IHC | Immunohistochemistry |
| IRN | Irinotecan |
| iTRAQ | Isobaric Tags for Relative and Absolute Quantification |
| LC | Liquid Chromatography |
| LTQ | Linear Ion Trap |
| MED | Murine Equivalent Dose |
| MS | Mass Spectrometry |
| NTRK | Neurotrophic Tyrosine Receptor Kinase |
| O-PDX | Orthotopic Patient Derived Xenograft |
| PAGE | Polyacrylamide Gel Electrophoresis |
| PDGFRA | Platelet Derived Growth Factor Receptor Alpha |
| PMD | Partially Methylated Domain |
| PPI | Percentage of Precursor Intensity |
| PSM | Peptide Spectrum Match |
| PTM | Post-translational Modifications |
| RMS | Rhabdomyosarcoma |
| RNA | Ribonucleic Acid |
| RTK | Receptor Tyrosine Kinase |
| SAX | Strong Anion Exchange |
| SCX | Strong Cation Exchange |
| SDS | Sodium Dodecyl Sulfate |
| SILAC | Stable Isotope Labeling With Amino Acids in Cell Culture |
| SNV | Single Nucleotide Variation |

| | |
|---|---|
| SV | Structural Variations |
| TF | Transcription Factor |
| TMT | Tandem Mass Tags |
| UPR | Unfolded Protein Response |
| VCR | Vincristine |
| WES | Whole Exome Sequencing |
| WGBS | Whole Genome Bisulfite Sequencing |
| WGCNA | Weighted Gene Co-Expression Network Analysis |
| WGS | Whole Genome Sequencing |

# CHAPTER 1. INTRODUCTION

## Mass Spectrometry Based Proteomics

The era of whole genomic-sequencing has made 'proteomics' research possible. As most of biological functions are executed and transmitted through proteins, proteomics studies provide complementary knowledge to genomics, shedding new light on understanding biology and disease. Development of diverse technologies has allowed the exploration of different aspects of protein function and structure. Numerous proteomic methods encompassing protein microarray[1], yeast two-hybrid assay[2] and high-throughput protein crystallization[3] have made great impact on our understanding of protein structure, interaction, activity and function. Among these technologies, mass spectrometry has been emerging as a mainstream tool for analyzing protein production, modification and function in distinct biological systems[4,5].

### Principles

The main purpose of mass spectrometry-based proteomics has been focusing on the systematic identification and characterization of protein sequences and protein posttranslational modifications and on the quantification of proteins in a biological system[6]. The most well-established proteomic platform relies on tandem mass spectrometry, a method also known as bottom-up proteomics[6]. During bottom-up proteomics analysis, proteins are first extracted from tissues or cells and then digested into peptides using proteinases such as trypsin. Peptides in samples are then separated by different peptide separation strategies, often by liquid chromatography and then are ionized and loaded on a mass spectrometer, where peptide ion spectra are scanned and further fragmented into peptide fragment ions. Fragment ion spectra are then assigned to peptide sequences to infer proteins. Fragment ion spectra can also be used to identify amino acids with modifications and to localize the modification sites. Numerous tools and methods have been developed for each step of the procedure[6], these including sample preparation, fractionation strategy, MS settings, quantification and data analysis. Individual options can be combined into distinct MS workflows from the numerous available choices for each procedure step to address different types of biological questions. The most widely used approach is known as shotgun proteomics or discovery proteomics. During shotgun proteomics, precursor ions are scanned and selected automatically via a strategy known as data-dependent analysis. There, top abundant precursor peptides are selected, collected and further broken into fragment peptide ions sequentially. This allows identification of a vast amount of proteins and also enables quantitative comparison between samples, either with stable isotope labeling, isobaric labelling or without labeling, an approach referred to as label free quantification.

## Towards genome-scale proteomic profiling

The central dogma of molecular biology (i.e. the flow of information from DNA to RNA to proteins) is fundamental in biological studies. Proteomic analysis is indispensable for understanding how cells function, however, collecting information at the global proteome level has proven to be challenging compared to data collection at the whole genome and transcriptome level[7]. One intrinsic limitation of proteomic analysis is the highly diverse physical and chemical properties of amino acids, the building blocks of proteins, in contrast to the linear array of genomic information[7]. Another inherent limitation of human proteome study concerns the dramatic proteome complexity and the large dynamic protein abundance range, which spans more than eight orders of magnitude in human tissues and cells[7,8]. MS technologies have enjoyed rapid development in the last decade; modern mass spectrometers are sensitive enough to detect the lowest abundant proteins present in human cells[7,8]. However, when a complex human proteome is loaded on a mass spectrometer, the signal of low abundant proteins can easily be buried in the dominant signals of highly abundant proteins[9]. Thus approaches to improve the separation capacity of high-performance liquid chromatography to reduce the sample complexity before MS analysis became a prerequisite for achieving genome-scale proteomic profiling.

## Towards accurate quantification with high-throughput

One central component beyond protein identification for proteomic analysis is quantification. MS-based proteomics can determine the relative change of proteins in two or more conditions, this is referred to as relative quantification. It can also study the absolute amount of each protein in a mixture sample, also known as absolute quantification. Often it is the relative changes of protein amounts upon a specific perturbation that are of more biological interest[10]. Traditionally, relative quantitative information can be obtained from a labelling free strategy. Labelling free quantifications based on spectral counting and /or intensity of detected peptides seem to be the easiest approaches to obtain quantitative information, however large variations can easily be introduced during sample handling. Moreover, it works in a sequential manner which shows much lower throughput compared to highly parallel genomics profiling. More recently, isobaric labeling methods, such as isobaric tags for relative and absolute quantitation (iTRAQ)[11], tandem mass tags (TMT)[12] and DiLeu isobaric tags[13], are emerging as mainstream strategies for quantitative proteomic analysis largely due to the multiplexed capacity of processing up to 12 samples[14]. For example, isobaric labeling enables the analysis of hundreds of mammalian samples in tens of batches, detecting a total of more than 15K proteins (from 12K genes) and 60K phosphosites in mammalian samples[15-17]. Despite the advances of isobaric labeling, the method often suffers from high noise levels due to co-eluted interfering ions, leading to quantitative ratio compression that underestimates the difference, particularly in complex protein samples[18-22]. Thus strategies to eliminate this effect become essential for accurate quantification of high throughput proteomics data[22].

## Protein phosphorylation

Our understanding of the players in different signal transduction pathways has been rapidly accumulated in recent decades, largely via high-throughput approaches such as microarray analysis[23]. Many downstream transcriptional responses following diverse stimuli have been elucidated through these studies[23]. However, protein posttranslational modifications rather than transcriptional alterations have been demonstrated to mediate critical events in distinct cellular responses through controlling enzymatic activity, protein-protein interaction, protein confirmation and cellular localization[23]. It is estimated that protein phosphorylation, the most widely studied posttranslational modification, affects one-third of all proteins[24]. Unfortunately, only a small portion of total *in vivo* phosphorylation sites have been revealed so far. Traditionally, phosphorylation has been analyzed mainly by *in vitro* assays, such as protein chip array[25]. Similarly, synthetic peptides have been used as kinase substrates to explore consensus kinase motifs[26]. However, Kinases are in general much less specific *in vitro* than *in vivo*, necessitating additional approaches[23]. Therefore, the development of a global and quantitative strategy for elucidating phosphorylation events becomes essential for systematic understanding of cellular behaviors.

# Systems Proteomics in Cancer

## Cancer proteomics

Advances in genomic sequencing technologies in the last decade enable large-scale genomic profiling of cancers. Comprehensive genomic landscapes have been generated for numerous types of cancers via genomic sequencing projects such as The Cancer Genome Atlas (TCGA) and The Pediatric Cancer Genome Project (PCGP), resulting in hundreds of cancer driver genes. Despite these extraordinary achievements, how these genomic alterations lead to deregulation of proteins, which result in cancers remains largely unclear. As genomic landscapes are accomplished for major cancer types today, large-scale proteomic profiling becomes essential for the next phase of cancer omics study, which is to decipher the oncogenesis processes for filling the gap between genotypes and phenotypes.

The quest to explore global protein changes in cancers started half a century ago. However, the large dynamic range of protein abundance, a plethora of protein isoforms, and paradigmatic sample and disease heterogeneity have been the grand challenges that largely limited the outcomes of early cancer proteomics studies[27]. Fortunately, recent advances in analytical techniques and strategies to de-complex the cancer proteome is making the genome-scale proteomic profiling for cancers possible[6,7,10]. The massive data that can be collected with the advanced proteomics workflows are expected to reveal the signaling networks downstream of cancer drivers, guiding the discovery of tumor vulnerability and new drug targets. Recent successes on large-scale cancer proteomics

projects[28,29] demonstrate that, with the constant technique advancements, MS-based proteomics are ready to carry the torch to push cancer research forward.

## Integration of proteomics data with other omics

Traditionally, scientists solve biomedical problems based on reductionism or "divide and conquer". That is, to learn a complex system by dividing it to simpler, smaller and hence more manageable units of the whole. This approach has been successful and will continue to be crucial for biomedical research[30]. However, it cannot deliver a comprehensive view of biological processes with complex disorders such as cancers. Meanwhile, High-dimensional biology (HDB), also refered to as the simultaneous analysis of changes in genome, transcriptome, proteome, or metabolome in a sample, has emerged as a powerful alternative paradigm. The main rationale for this approach is that omics data are complementary to each other. For example, multiple processes beyond transcript concentration (e.g. the spatial and temporal variations of mRNA levels, the availability of local resources for protein biosynthesis) strongly impact the relationship between protein amounts and their coding transcripts level[31-34]. As a result, the transcript levels by themselves are not adequate to accurately infer protein amounts in many scenarios[35,36]. Also, sequencing of DNA and RNA does not capture information on post transcriptional regulations including protein translation, protein degradation, posttranslational modifications (e.g. phosphorylation)[37]. Indeed, numerous protein functions including enzymatic activity, protein-protein interaction, protein confirmation and cellular localization[23] are mediated by protein PTMs rather than protein expression alone.

As mentioned in the previous section, high-throughput, quantitative MS-based proteomics is becoming as powerful as other well established omics technologies. Collecting large-scale and high quality proteomics data has never been easier than today. Obviously, combination of proteomic with other omics data will provide the largest benefit for a systematic view of cancer cells. Indeed, as proteins are intrinsic carriers of biological functions, proteomic data provides an ideal scaffold for integrating multidimensional data[10] to define the inter-relationships of all the components in cancers.

Numerous strategies to integrate multiple omics data are emerging in recent years[28,29,38]. However, when compared to the remarkable advancements of omics sequencing technologies, it is still left largely behind. Indeed, most biomedical researchers only utilize a small portion of large-scale omics data. As a result, the majority of information collected in omics data are wasted. With the rapid development of technologies and accumulation of big omics datasets, it is obvious that the lack of bioinformatics tools to make full use of the massive omics data will become the next bottleneck to limit the outcomes of biomedical research. For complex diseases with systematic reprogramming of genome and proteome such as cancers, it is essential to develop novel bioinformatics pipelines to prioritize master regulators and core signaling networks from thousands of passenger changes for illuminating the complex oncogenic processes and potential therapeutic vulnerabilities.

## High-grade Glioma

HGGs are the most prevalent malignant tumors that develop in the central nervous system, and confer devastating mortality[39,40]. Histologically, gliomas display similarities to glial cells (e.g. astrocytes and oligodendrocytes) and can be classified to astrocytomas, oligodendrogliomas, oligoastrocytomas, *etc*. More than 50% of gliomas are grade IV astrocytoma (glioblastoma multiforme), one of the most aggressive human cancers[41]. Despite advances in treatment strategies including radiation, chemotherapy and surgery, the overall five year survival rate of glioblastoma multiforme remains < 5%, making glioma an urgent focus of cancer research. Significant efforts in glioma genome sequencing including the TGCA projects[42,43] have already revealed comprehensive HGG genomic landscapes[39,40,43-48], these include three core signaling pathways (i.e. the RTK pathway, the P53 pathway and the RB pathway) that are frequently activated in HGGs[42,43]. Genetic alterations in all three pathways are often presented in most of HGG tumors, promoting enhanced cell proliferation and survival while aiding tumor cells to escape apoptosis, senescence and cell-cycle checkpoints. Nevertheless, how genomic mutations in these core pathways lead to dysregulation of particular master regulators and specific pathways remains unclear. On the other hand, HGG proteomic and phosphoproteomic studies have extended our understanding of HGG signaling[44,49]. Yet, most of these attempts have used proteomic approaches of relatively shallow depth. There is essentially no deep HGG proteomic landscape available for the cancer research community. Therefore, comprehensive profiling of the HGG proteome becomes essential to define the global HGG proteotypes, filling the gap between genotypes and phenotypes.

Laboratory mice display extensive physiological and molecular similarities compared to human[40]. Dr. Suzanne Baker's group reported a comprehensive genomic analysis on a mouse HGG model and unveiled an astounding similarity of gene copy number and molecular subtypes in HGG mouse models and those found in human HGGs. While the paradigmatic inter- and intratumoral heterogeneity of HGG largely confounded the power of surgical tumor samples to dissect the global proteome and signaling network[49], HGG mouse models with a much clearer genomic background become valuable alternatives for deep proteomic profiling[40].

## Rhabdomyosarcoma

Rhabdomyosarcoma (RMS) is a childhood solid tumor with molecular and cellular features of skeletal muscle. Approximately 75% of patients with localized disease are cured with conventional multimodal therapy but patients with recurrent or metastatic disease have overall survival rates of 17% and 30%, respectively[50]. Historically, RMS was divided into two histopathologic subtypes: embryonal rhabdomyosarcoma (ERMS), which accounts for about 60% of all RMS patients, and alveolar rhabdomyosarcoma (ARMS), which accounts for 25% of patients[51]. The majority of tumors classified as ARMS by histopathology, also have a translocation between the FOXO1 gene on chromosome 13q14 and either PAX3 on chromosome 2q35 or PAX7 on chromosome 1p36[52,53]. However, a subset of ARMS tumors are translocation negative and have

molecular features more similar to ERMS despite their ARMS histopathologic classification[54]. This is important because patients with ERMS have a relatively good prognosis[55,56] and many of the current clinical trials are focused on stratifying molecular subgroups (translocation positive and negative) rather than the ARMS/ERMS histologic subgroups.

Previous genomic characterization showed that the molecular differences between translocation positive and negative rhabdomyosarcomas extend beyond the FOXO1-PAX3/7 translocation. Translocation negative ERMS tumors have a loss of heterozygosity at the 11p15 locus, a high rate of structural variations (SVs), copy number variations (CNVs) and single nucleotide variations (SNVs) leading to recurrent genetic lesions in oncogenes and tumor suppressor genes[57,58]. In contrast, translocation positive ARMS tumors have relatively few SVs, CNVs and SNVs and few if any recurrent mutations in cancer consensus genes[57,58]. The most commonly mutated genes in translocation negative ERMS are in the RAS pathway[57,58]. In some adult RAS mutant cancer types, targeting kinases in the RAS and PI3K pathways has shown promise in laboratory studies[59-61]. However, efforts to target the RAS pathway alone or in combination with PI3 kinase pathway inhibitors in RMS have been unsuccessful to date[58,60,61]. Moreover, it is not known if translocation positive ARMS tumors have a completely different molecular profile that requires a unique treatment approach or if they have the same fundamental vulnerabilities as translocation negative ERMS tumors. Another important consideration is disease recurrence. Previous analyses of clonal evolution in RMS patients demonstrated that rare clones of cells in the diagnostic tumor can survive and contribute to recurrent rhabdomyosarcoma. However, it is not known how tumor evolution impacts sensitivity to molecularly targeted therapy.

## Research Aims

### Aim 1: to develop a genome-scale proteomic analysis platform

Although significant progress has been achieved to identify the deep mammalian proteome[7,62], there is no systematic report on the adjustment of parameters for ultra-long LC-MS/MS runs to optimize protein identification at a genome wide scale. To obtain an in-depth coverage of the mammalian proteome and provide a comprehensive genome-scale proteomic analysis workflow, we determined to engineer a 150 cm LC column to couple with Q Exactive MS and further optimize key steps and systematically tune shotgun proteomics parameters in the LC-MS/MS platform, following our previous optimization work using a regular short column (75 μm x 12 cm)[63]. Finally, we will use the optimized LC/LC-MS/MS platform to process a human brain specimen and to examine the proteome profiling depth by comparing it to the transcriptome of the same sample.

**Aim 2: to develop an accurate quantitative strategy by isobaric labelling and mass spectrometry for high-throughput genome-scale proteomic analysis**

Quantitative proteomics has been an essential tool in biomedical research[4,64] and shows high potential for clinical application[65]. Despite the advances of isobaric labeling, the method often suffers from high noise levels due to co-eluted interfering ions, leading to quantitative ratio compression that underestimates the difference, particularly in complex protein samples[18-21]. We seek to address the ratio compression issue by extensive high resolution fractionation and a novel *y*1 ion-based interference correction method. To mimic real biological samples, we will mix TMT-labeled *E. coli* proteins at known ratios, in the presence of a 20-fold excess amount of background peptides from rat proteins. The mix will be analyzed under multiple LC-MS/MS conditions by adjusting key parameters, including fraction number collected in the offline pre-fractionation, MS2 isolation window, peptide loading amount and online RP fractionation depth (gradient length). We will also develop a computational method that uses the known *E. coli* protein ratios to estimate interference levels from rat proteins. Finally, we will try to eliminate the interference by pre-MS fractionation, optimization of MS parameters, and post-MS *y*1 ion-based correction and introduce a general pipeline for accurate isobaric labeling quantification.

**Aim 3: to apply the techniques developed above to define the proteotypes to link the genotypes and phenotypes in HGG mouse models**

A central gap in cancer biology concerns how oncogenes drive the reprogramming of molecular signaling networks to execute phenotypic changes[28,29,38]. Significant efforts in glioma sequencing have unveiled comprehensive genome-wide mutation landscapes[39,40,43-48]. These include mutations and/or amplifications of PDGFRA and fusion genes of the NTRK family of neurotrophin receptors identified in pediatric and adult HGG[39,46,48,66,67]. However, a complete understanding of how these genomic alterations lead to dysregulation of particular master regulators and specific pathways remains unclear. Here we seek to perform genome-scale proteomic and phosphoproteomic profiling on two HGG mouse models driven by PDGFRA mutations or NTRK fusions using pipelines developed above to first evaluate the strength of the novel pipeline. Following that we will use global proteomic data to identify functional modules and master signaling networks reprogrammed in HGGs, validate these discoveries through *in vitro* and *in vivo* experiments to further explore phenotype differences driven by these two genotypes. Finally, we will assess how the global proteotype can explain the phenotype differences driven by different genotypes.

**Aim 4: to develop a bioinformatics pipeline to prioritize master regulators in cancer through integrating multi-omics data**

With the rapid development of technologies and accumulation of big omics datasets, there is a growing request for novel bioinformatics pipelines to handle these data

and to prioritize master regulators and core signaling networks from thousands of passenger changes for illuminating the complex oncogenic processes and potential therapeutic vulnerabilities. To address this, we seek to develop approaches to integrate proteome, phosphoproteome, and transcriptome data collected in Aim 3 with human HGG transcriptome and numerous publicly available databases to prioritize master regulators including kinases and transcription factors in cancers. Kinase activity will be inferred from phosphorylation of substrates using a machine learning approach. Transcription factor activity will be derived from target gene expression and will be validated by proteome and phosphoproteome data. We will further define a core kinase to transcription factor network in cancer by examining the expression patterns of master regulators in reported kinase to transcription factor relationships in databases. Finally we will perform cross-species analysis to match the changes prioritized in mouse omics data to corresponding human HGG data to search for consensus master regulators.

**Aim 5: to apply the techniques and bioinformatics tools developed above to identify therapeutic vulnerabilities in rhabdomyosarcoma**

Recent success targeting CDK4/6 and MEK in RAS mutant adult cancers led us to test this approach for rhabdomyosarcoma[68-71]. We achieved synergistic killing of RAS mutant rhabdomyosarcoma tumor cells by combining MEK and CDK4/6 inhibitors in culture but failed to achieve efficacy *in vivo* using orthotopic patient derived xenografts (O-PDXs). In this section, we seek to apply the pipelines, methodologies and bioinformatics tools developed above to perform large-scale profiling on rhabdomyosarcoma (RMS) using orthotopic patient derived xenografts. A cohort consists of two subtypes of RMS (i.e. ERMS and ARMS), normal myoblasts and myotube will be profiled in 3 TMT batches (total 30 samples) to provide a deep proteomic and phosphoproteomic landscape for RMS. Through integrative analysis of genomic and epigenomic data, we aim to determine the pathways that enable the rhabdomyosarcoma cells to evade the targeting of CDK4/6 and MEK and to identify a therapeutic vulnerability in rhabdomyosarcoma.

# CHAPTER 2. DEVELOPMENT OF A TMT-LC/LC-MS/MS PLATFORM FOR GENOME-SCALE PROTEOMIC ANALYSIS WITH HIGH-THROUGHPUT AND ACCURATE QUANTIFICATION*

## A Genome-scale Proteomic Analysis Platform

### Introduction

In the post genomic era, next generation sequencing technology is now widely used to characterize the alterations of the genome and transcriptome in the context of human diseases[72,73]. Although gene expression can be analyzed by transcriptomic profiling, transcriptomic data do not always correlate well with protein expression in biological samples and often lack the information of protein posttranslational modifications. Thus, development of proteomics platforms for deep proteome coverage becomes an urgent task to provide systematic and comparable protein expression information complementary to DNA and RNA data.

Mass spectrometry (MS)-based shotgun proteomics is predominantly used for complex proteome analysis[4,5,74]. In a typical shotgun experiment, complex protein samples extracted from cells or tissues are digested with protease(s) and the resulting peptide mixtures are fractionated by organic gradient on HPLC columns followed by tandem mass spectrometry analysis. The MS/MS spectra are then searched against protein databases for the identification of proteins and posttranslational modifications. With the advent of high resolution MS and the improvement of LC performance, current platforms of shotgun proteomics can routinely identify thousands of proteins in mammalian cells in a single LC-MS/MS analysis. One of the key measurements of LC performance is the peak capacity that is defined as the number of peaks separated within a resolution of unity in a given LC gradient time[75]. Peak capacity is estimated to be proportional to the root square of LC column length and inversely proportional to the root square of LC particle size[76]. Several reports demonstrated the benefits of small particles (<2 μm) with ultra-high pressure solvent delivery (up to 70 000 psi)[77-80]. High values of peak capacities were obtained on these columns (i.e. 75 μm x 50 cm) depending on the

---

* Reprinted with permissions from ACS under the ACS AuthorChoice license, further permission requests should be directed to the ACS. Wang, H. *et al*. Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. *J. Proteome Res.* 2015, **14** (2), pp 829–838. http://pubs.acs.org/doi/abs/10.1021%2Fpr500882h;
Niu, M. *et al*. Extensive peptide fractionation and *y*1 ion-based interference detection enable accurate quantification by isobaric labeling and mass spectrometry. Analytical Chemistry, doi:10.1021/acs.analchem.6b04415 (2017). http://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b04415

gradient length[80-86]. However, column heating and ultra-high system pressure (> 10,000 psi) are usually required for running long columns packed with sub 2 μm beads, compromising the robustness of the system. Alternatively, when HPLC time is not a limiting factor, longer LC columns improve resolving power but with a higher backpressure. For instance, several reports have shown comparable peak capacity using 5 μm C18 particles and up to a 1 meter long column within regular HPLC pressure limits[87,88]. When long LC was coupled with Q Exactive MS, it resulted in more than 4,000 identified proteins in the human proteome under optimized conditions[16,89]. However, due to the large (>$10^7$) dynamic range of proteins in mammalian cells, additional pre-fractionation step(s) (e.g. SDS-PAGE, strong anion exchange (SAX), strong cation exchange (SCX), basic pH LC and isofocusing) were applied to reduce peptide complexity and deepen the mammalian proteomic analysis[7,62,90-92]. Further peptide separation was also achieved through gas phase fractionation (GPF) through MS1 ion selection on mass spectrometer[93,94].

To date, a few studies lead to the detection of more than 10,000 proteins in several human cancer cell lines using SAX fractionation and analysis on LTQ Orbitrap Velos MS and about a month of instrument time[95,96]. The Marto group identified 11,352 mouse genes-derived proteins using LTQ XL MS and Triple TOF 5600 MS in 8 days from murine embryonic stem cells[62]. The Lehtio group reported the identification of 13,078 human proteins and 10,637 mouse proteins from cancer cell lines using high resolution isofocusing fractionation and LTQ Orbitrap Velos MS with ~15 days of instrument time[89,92]. While we were preparing this manuscript, Mann's group reported the identification of ~10,000 proteins on Q Exactive MS using 4 day instrument time and a long column coupled with UPLC system[96]. Most recently, drafts of the entire human proteome (identifications of ~18,000 gene products in varieties of human tissues and hematopoietic cells) were completed from ~2,000 LC-MS/MS runs using several months of MS instrument time by two research groups[97,98].

Although significant progress has been achieved to identify the deep mammalian proteome, there is no systematic report on the adjustment of parameters for ultra-long LC-MS/MS runs to optimize protein identification at a genome wide scale. To obtain an in-depth coverage of the mammalian proteome, we determined to further optimize key steps in the LC-MS/MS platform, following our previous optimization work using a regular short column (75 μm x 12 cm)[63]. In this study, we described a step-wise analysis to tune shotgun proteomics parameters using an in-house manufactured 150 cm LC column coupled with Q Exactive MS. The optimization process consisted of more than 30 LC-MS/MS runs of analyzing mammalian tissue (e.g. rat brain). Finally, we used the optimized LC/LC-MS/MS platform to process a human brain specimen of Alzheimer's disease (AD) and identified more than ten thousand proteins, covering more than 60% of the expressed proteome.

## Methods and materials

**Construction of 100 µm × 150 cm analytical columns.** Capillary column of 150 cm in length and 100 m inner diameter (ID) was packed in house following the previously reported protocol with modifications[88]. This column consisted of two segments, namely, one 110 cm long blunt end capillary column and one 40 cm long capillary column with a 15 µm opening tip. To make the blunt end column, 100 µm ID fused silica tubing was dipped into the activated silicate solution (Next Advance Inc., NY) briefly followed by heating to 100ºC on a heater plate for one min before the ejection of excess silicate solution. Then the frit was further heated for another hour at 100ºC and cut to 2 mm in length. The capillary tubing was washed with methanol thoroughly. The blunt end column was then packed a with slurry of Magic C18 AQ 200 beads (5 µm) at a concentration of 30 mg/ml in methanol. Bed length of 110 cm was obtained after 6 h of continuous packing at 2,800 psi using a Pressure Injection Cell system (Next Advance Inc, NY). The second segment of the capillary column was packed similarly to 40 cm in length using Self-Pack PicoFrit column (New Objective, 15 µm tip opening, 100 µm inner diameter, cat # PF360-100-N-5). Finally two columns were connected through a metal union with zero dead volume (Upchurch Scientific, NY).

**Protein extraction and digestion from the rat brain and AD brain.** Human tissues of prefrontal cortical regions were provided by the Brain and Body Donation Program at Banner Sun Health Research Institute. The AD case with a short post-mortem interval (< 3 h) was clinically and pathologically characterized in accordance with established criteria[99]. This study was approved by Banner Sun Health Research Institute. Adult rat brains were purchased from Pel Freez Biologicals, and rat brain peptides were prepared as previously described[100]. The cerebral cortex of AD brain was homogenized in 100 µl of lysis buffer (0.1 M Tris, pH 8.5, 8 M urea, 0.15% sodium deoxycholate) at 4ºC using 0.5 mm glass beads for 5 min in a Bullet Blender instrument (Next Advance Inc.)[100,101]. The entire cell lysate without clarification of the insoluble materials was digested with Lys-C (Wako, 200:1 by weight) at room temperature for half hour in the lysis buffer followed by trypsin digestion (Promega, 200:1 by weight) in 2 M urea, 0.1 M Tris-HCl , pH 8.5 at room temperature overnight. The peptides were then acidified with 0.15% TFA, pre-cleared by centrifugation, desalted with Sep-Pak C18 SPE column (Waters), and eluted with 40% acetonitrile (ACN) plus 0.1%TFA. The eluent was dried and stored at -80 ºC for further usage[86]. Protein quantification was carried out by short SDS-gel based staining and BCA method[63].

**Basic pH LC fractionation of peptides.** The desalted peptides from AD brain were re-suspended in 10 mM ammonium formate pH 8 at a concentration of 10 mg/ml. Basic pH HPLC was performed on a 4.6 mm x 250 mm Xbridge C18 column (Waters, 3.5 µm bead size) using an Agilent 1270 HPLC instrument. About 400 µg peptides were loaded on the column and HPLC gradient started at 90% solvent A (10 mM ammonium formate, pH 8.0) for 5 min and went up to 50% solvent B (90% acetoniltrile, 10 mM ammonium formate, pH 8.0) during a 50 min time period followed by a steep increase to 90% B within 5 min at a flow rate of 0.4 ml/min. The eluted peptides were collected into 60 fractions and every 6 fractions were combined into ten sub-fractions in a concatenated

pattern to ensure that each sub-fraction contained similar complexity of hydrophilic and hydrophobic peptides[102-104]. The sub-fractions were then dried and stored at -80°C for further analysis.

**Protein identification by LC−MS/MS.** Dried peptides were dissolved in 5% formic acid and 0.1% TFA. Peptides were loaded on a 100 μm x 150 cm column using nano ACQUITY UHPLC (Waters) system which was interfaced to a Q Exactive MS (Thermo Fisher Scientific) through a nanoelectrospray ion source[105]. Peptides were separated by a designed gradient as indicated (solvent A: 0.2% formic acid, and solvent B: 70% acetonitrile, 0.2% formic acid). The peak capacity at each gradient time was calculated using formula $p = 1 + tg/w$, where tg is the time of the gradient and w is the average peak width across entire LC runs[63]. The peak width of an individual LC run was estimated by averaging the chromatographic peak width ($4\sigma$, where $2\sigma$ is defined as FWHM of the corresponding extracted ion chromatograms) of major peptide ions. Peptides in the ten basic pH LC sub-fractions were resolved similarly on this long column using a 540 min, 15-65% buffer B linear gradient. The Q Exactive was operated in a data dependent mode switching between full scan MS and up to 20 MS/MS acquisitions. The survey scans with a m/z range of 300-1600 were acquired in the Orbitrap with 35,000 resolution at m/z = 200 and a predicted AGC value of 1 x 106 with maximal ion time of 60 ms. The ions detected in survey scans were then sequentially isolated and fragmented by HCD at normalized collision energy of 28 eV. The maximal ion injection time for MS/MS was set to 60 ms at a resolution of 17,500 or 128 ms with a resolution of 35,000. Isolation of precursor ions was performed at 1.6 m/z window. Different dynamic exclusion times were evaluated to maximize peptide identification including 10 s, 20 s, 40 s and 60 s. At last 20 s was chosen for AD brain samples. For the GPF method, the operation of Q Exactive MS was similar to the non GPF method with minor modifications. The entire m/z range for MS1 was 300–1600 but divided into multiple m/z subsections which were described in the results section. Each m/z subsection had 10 m/z overlapping with adjacent subsections[94,106]. For data acquisition of GPF, the cycle started at the first m/z subsection of MS1 acquisition and its data dependent MS/MS followed by the 2nd m/z subsection of MS1 acquisition and its data dependent MS/MS until the full m/z range in MS1 was covered.

**Database search and analysis.** The acquired raw MS data were processed with an in-house data processing pipeline as previously reported[63]. Briefly, the MS raw data were converted to mzXML format using ReAdW software. Up to six precursor ions were selected for a mixed MS/MS spectrum. The search was performed by the SEQUEST algorithm (version 28 revision 13)[107] against a composite target / decoy human or rat protein database[108,109]. The target human protein database was generated from Uniprot (combined Swissprot and Tremble) human database containing 71,809 protein entries. The target rat protein database contained 35,570 protein entries. Spectra were searched with ±10 ppm for precursor ion mass tolerance, ±0.02 Da for fragment ion mass tolerance, fully tryptic restriction, dynamic mass shift for oxidized Met (+15.9949), two maximal missed cleavages, and three maximal modification sites. Only a, b and y ions were considered during the search. The peptide spectrum matches (PSMs) were first filtered by the length of matched peptides (removal of PSMs with 6 or less amino acids)

and then by mass accuracy. The survival PSMs were further filtered by matching scores to achieve unique protein identification (grouped using a parsimony algorithm) at 1% FDR. To perform integrative analysis with RNAseq data, UniProt IDs were converted to official gene symbols according to UCSC annotation (downloaded on 01/23/14). For each gene, the number of accepted PSMs was calculated and further normalized by gene length.

**RNA-seq analysis.** Total RNA was extracted from ~20 mg inferior frontal cortex of the same AD brain for proteomics study using the RNeasy mini kit (Qiagen)[110]. On-column DNA digestion was performed to eliminate the endogenous genomic DNA contaminants. The mRNA samples were purified by poly(dT) beads and then fragmented before reverse transcription. The paired end adaptors were used to ligate the processed double stranded cDNA fragments. The sequencing was carried out on the Illumina Genome Analyzer IIx platform. Using BWA (0.5.10) aligner, RNAseq reads were aligned to multiple databases, including human genome (GRCh37), human transcriptome (RefSeq and AceView), and all possible combinations of RefSeq exons. Finally, the reads mapped to the transcriptome were converted to genomic mapping, and merged together in the final output BAM files.

## Results and discussion

**Installation of a long gradient LC-MS/MS platform.** We packed a 100 μm x 150 cm nano LC column using 5 μm C18 beads and interfaced this column with a Q Exactive MS for deep shotgun proteomic analysis of a mammalian proteome (**Figure 2-1A**). Recently, the Marto group[62,88] has shown that nano LC columns packed with large beads (e.g. 5 μm) in extended length (up to 1 meter) performs as efficiently as nano HPLC columns packed with sub 2 μm C18 beads for separation of mammalian protein digest complexes, but the one meter column was operated at a flow rate of 5-10 nl/min under 1,500 psi with a regular HPLC system. Although the extremely low flow rate may improve ionization sensitivity, it is not optimal for resolving peptides on the majority of nano LC-MS/MS platforms that typically run in the range of 150 to 300 nl/min[111,112]. The current long LC system normally flowed at 300 nl/min with backpressure of 7,500 psi. When heating the column to 60ºC with lower flow rate of 150 nl/min, this backpressure was reduced to ~3,000 psi. To achieve stable electrospray ionization of the eluted peptides, the column was split into two portions (110 cm and 40 cm) and connected by a metal zero dead volume union where the voltage was applied.

To evaluate the reproducibility of this system, we examined the run to run variation by repeated LC-MS/MS analyses. The rat brain tryptic peptide mixture was used for the optimization of the system because of similar compositions and dynamic ranges between human and rat brain proteomes. The rat brain peptide mixture (~1 μg) was analyzed three times on this column during a 4 h run. Base peak profiles for the replicates were almost identical (**Figure 2-1B**) with the retention time shifts of less than 1 min. After database search and filtering, the relative standard deviations of accepted peptide-spectrum matches (PSMs), unique peptides, and proteins were 2.5%, 2.1% and

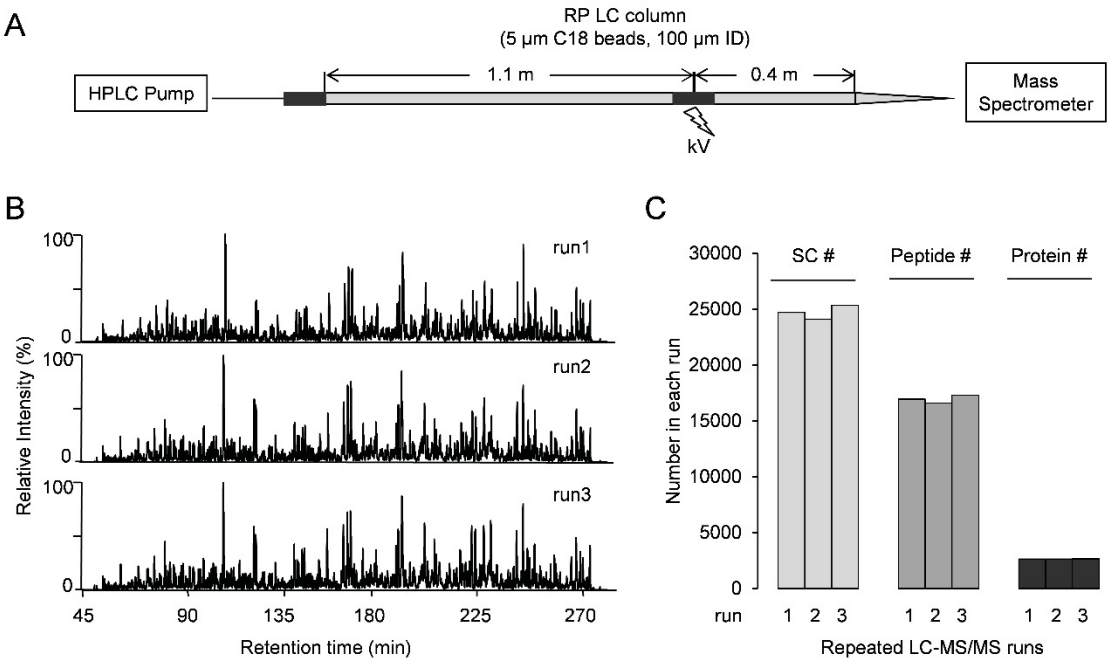



**Figure 2-1. Evaluation of the reproducibility of long LC column coupled with Q Exactive MS.**

(A) Illustration of the setup of long LC column (100 µm x 150 cm, 5 µm C18 particles) coupled with Q Exactive MS. (B) Base peak chromatographs of three technically repeated runs. About 1 µg of rat brain tryptic peptide mixture was loaded on the column and then eluted in a 10-45% acetonitrile gradient over 4 h. (C) Comparison of accepted peptide-spectrum matches (PSMs), peptide and protein identifications.

0.6%, respectively (**Figure 2-1C**). This result strongly indicated high reproducibility of the LC-MS/MS platform. The same column was used for the entire optimization process, comprising of more than 100 runs, and no obvious column deterioration was observed.

**Optimization of LC parameters.** Increasing LC loading capacity is one of the leading approaches to maximize peptide detection in shotgun proteomics analysis[113]. We examined the effect of peptide loading amount on peptide and protein identifications using this ultra-long capillary LC column. When the loading amount of rat brain peptides were increased from 0.2 μg to 6 μg, the identified peptides and proteins were increased by 60.6% (from 12,159 to 19,529), and 39.9% (from 2,105 to 2,948), respectively (**Figure 2-2A, B**). However, further addition of loading amount to 20 μg resulted in only 1.4% gain of peptides and 0.9% gain of proteins. Consistently, the ion intensities of peptides exemplified by one 14-3-3 peptide and one TBB3 peptide were increased by ~10 fold or 3.8 fold, respectively, when the loading amount was increased from 0.6 μg to 6 μg (**Figure 2-3**). However, further increasing the loading to 20 μg did not lead to stronger ion intensity. This result suggested that the optimal loading amount of peptides for the current system was ~6 μg which was 6 times higher than the optimal loading amount on a regular capillary LC column (e.g. 75 μm x 12 cm) and twice as much as the regular loading amount reported on other long LC columns[63,91,95].

Next we examined the impact of increased peptide loading amount on the LC peak width (**Figure 2-2C**). In general, the average peak width only increased ~20% (from 0.65 min to 0.77 min) when the loading amount varied from 0.6 to 20 μg, indicating that this column has high loading capacity and reasonable performance during chromatography. Interestingly, when loading 20 μg of peptides, we found that a fraction of strong peaks showed significant peak broadening (**Figure 2-2D**), which may result in ion suppression of adjacent weak peptides. This observation may also contribute to no gain of identified peptides at 20 μg loading. To balance the benefit of peak intensity and disadvantage of peak broadening, we selected ~6 μg peptides as a standard loading level on this LC-MS/MS platform.

To utilize the MS instrument efficiently in the shotgun proteomics platform, it is desirable to select a LC gradient range in which the number of identified peptides in unit time across the entire LC gradient region is similar [63]. We evaluated the LC gradient for the long column and found that the optimal gradient was in a linear gradient range of 10-45% acetonitrile (**Figure 2-4**). Over 98% of the identified peptides were eluted within this gradient range during a 4 h run and about 78.3 ± 21.6 peptides were identified per min. Interestingly, the reported optimal LC gradient range for mammalian cellular tryptic peptide mixture was about 10-30% of acetonitrile for both a regular 12 cm column and a long LC column (up to 50 cm )[63,79,84,85]. However, only about half of the peptides were eluted at 30% of acetonitrile on this extra-long column, suggesting that higher organic gradient was required for efficient elution of mammalian cellular peptide complexes on ultra-long C18 LC columns[76]. This observation may be explained by the increasing interaction between peptides and C18 beads created by the long distance through which peptides have to travel.

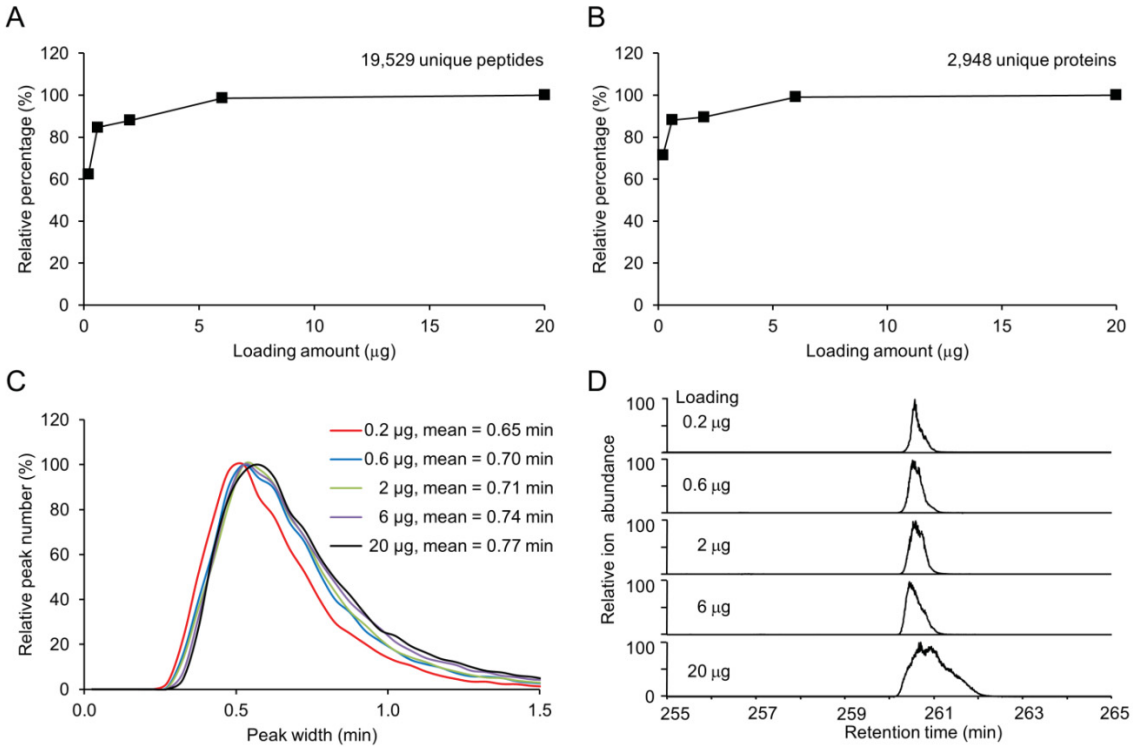



**Figure 2-2. Optimization of the loading amount of rat brain peptides for LC-MS/MS identification.**
Various amounts of rat brain peptides were loaded on the long column and analyzed by a 4 h gradient. (A) The number of detected peptides with different loading levels. (B) Protein identification with different loading levels. (C) The effect of different peptide loading amount on the global distribution of peak width for major peptide ions.

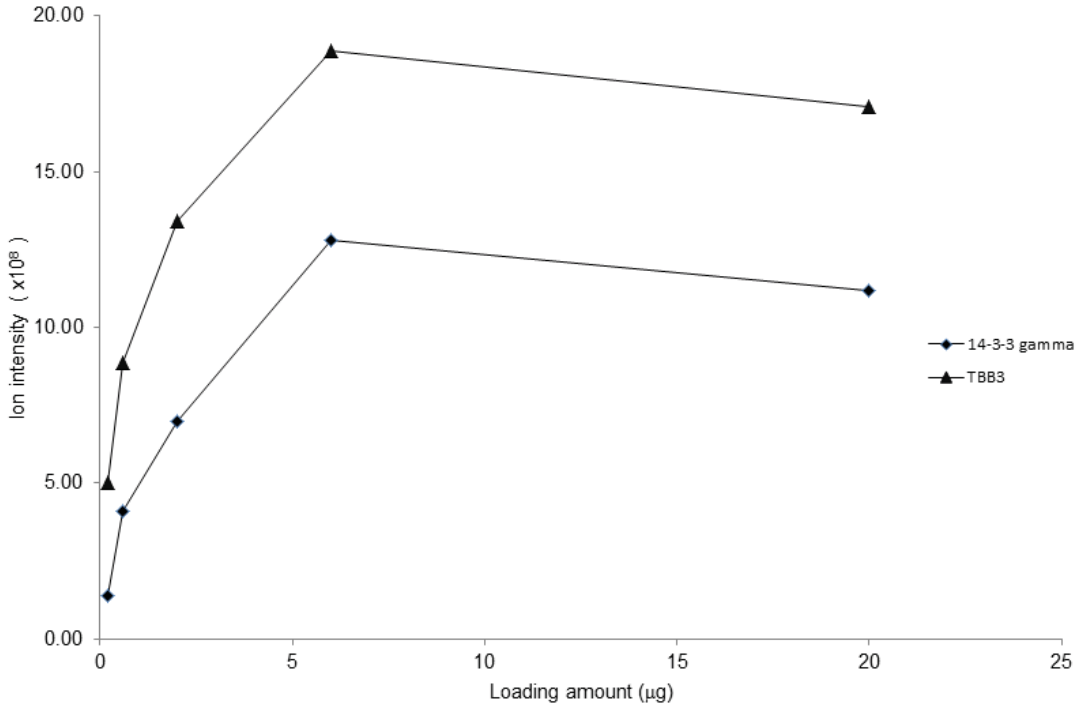



**Figure 2-3. Extracted base peak ion intensity of peptide LAEQAER of 14-3-3 γ protein and NSSYFVEWIPNNVK of TBB3 protein on different loading amounts.**

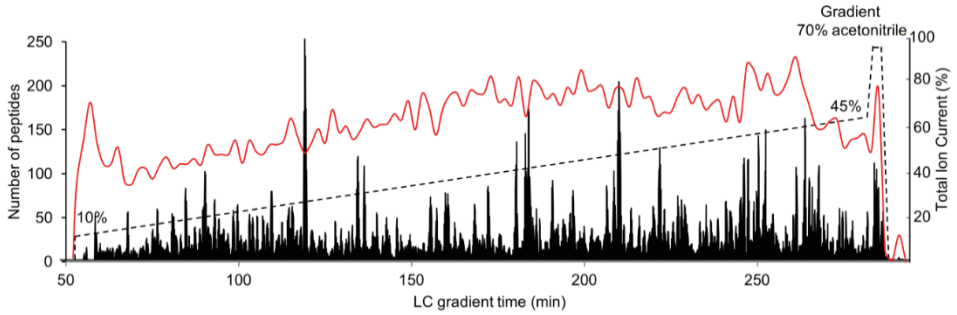



**Figure 2-4. Optimization of the LC gradient buffer for peptide elution.**
~2 μg of peptides was loaded on the long column and eluted in a 10-45% gradient of acetonitrile over 4 h. The LC elution profile was represented by total ion current (solid black line) along with the gradient (dotted black line). The number of identified peptides every two min was plotted (solid red line). About 157 ± 42 peptides were identified in every two min.

To determine optimal gradient time for peptide and protein identifications on this system, we tested various LC gradients ranging from 2 h to 12 h. We first calculated the average peptide peak width of major peptide ions across the entire elution, and then derived the corresponding peak capacity for each gradient (**Figure 2-5A**). When the peak capacities were plotted as a function of gradient time, a positive correlation was observed between peak capacity and gradient time. The peak capacity reached its maximum of 730, similar to the reported peak capacities of other nano LC columns used for in-depth proteomics analysis[84,88]. Next, we investigated whether the increased peak capacities can lead to more peptide identifications. As expected, the number of PSMs was increased proportionally to the extended gradient time while the number of identified peptides and proteins also followed this trend (**Figure 2-5B**). The number of detected peptides and proteins almost reached plateau at 12 h gradient with the identification of 23,884 peptides and 3,484 proteins from 46,711 PSMs. Interestingly, there was a linear correlation (R2 = 0.985) between the peak capacity and the number of identified peptides (**Figure 2-5C**), supporting the notion that peak capacity is a major factor for optimizing LC-MS/MS based peptide identification[84].

**Evaluation of MS parameters.** One interesting finding was that MS sequencing efficiency was reduced when LC gradient time was extended on the long column, evidenced by a steady decline of the ratios of summed MS2 scans versus MS1 scans (**Figure 2-6A**). This result suggested that there was not sufficient number of ions detected in survey MS1 scans to trigger MS2 scans. Since GPF is capable of detecting weak sample ions within a narrow m/z range but it takes multiple MS1 scans to cover a full scan region[93,94], we assessed the function of GPF to improve the MS sequencing efficiency. The m/z subsections of GPF were determined experimentally to contain the same number of PSMs in each subsection of m/z windows using rat brain peptides as a testing sample. During a 4 h LC-MS/MS analysis, one, two, three and four m/z subsections in a full m/z range of MS1 were tested. Compared to no GPF, the implementation of GPF of three subsections exhibited the highest ratio of MS2/MS1 scans (**Figure 2-6B**) and led to 11.3% and 15.4% increase in the number of identified peptides and proteins, respectively (**Figure 2-6C**). Thus, the GFP of three subsections was chosen for later experiments.

To further optimize the sequencing efficiency of MS, we evaluated the effect of different dynamic exclusion time of MS on the identification of peptides and proteins. In a 4 h LC gradient on the long LC column, the number of PSMs, peptides and proteins was the highest at 20 s dynamic exclusion time (**Figure 2-6**). Since the calculated average peak width was approximately 40 s for the 4 h LC gradient, each m/z ion would be analyzed about twice. Reduction of the dynamic exclusion time from 20 s to 10 s leads to 22% and 15 % drop in the number of peptide and protein identifications respectively, even though the MS2/MS1 ratio reached the highest number of 10. This result clearly showed the redundant sampling at 10 s dynamic exclusion time because of repetitive sequencing of the same peptide ions. We observed 1.9% decrease of the number of peptide and protein identifications and more than 23% dropping of PSMs at 40 s dynamic exclusion time. Therefore the dynamic exclusion time was set at 20 s for the 4 h LC gradient.

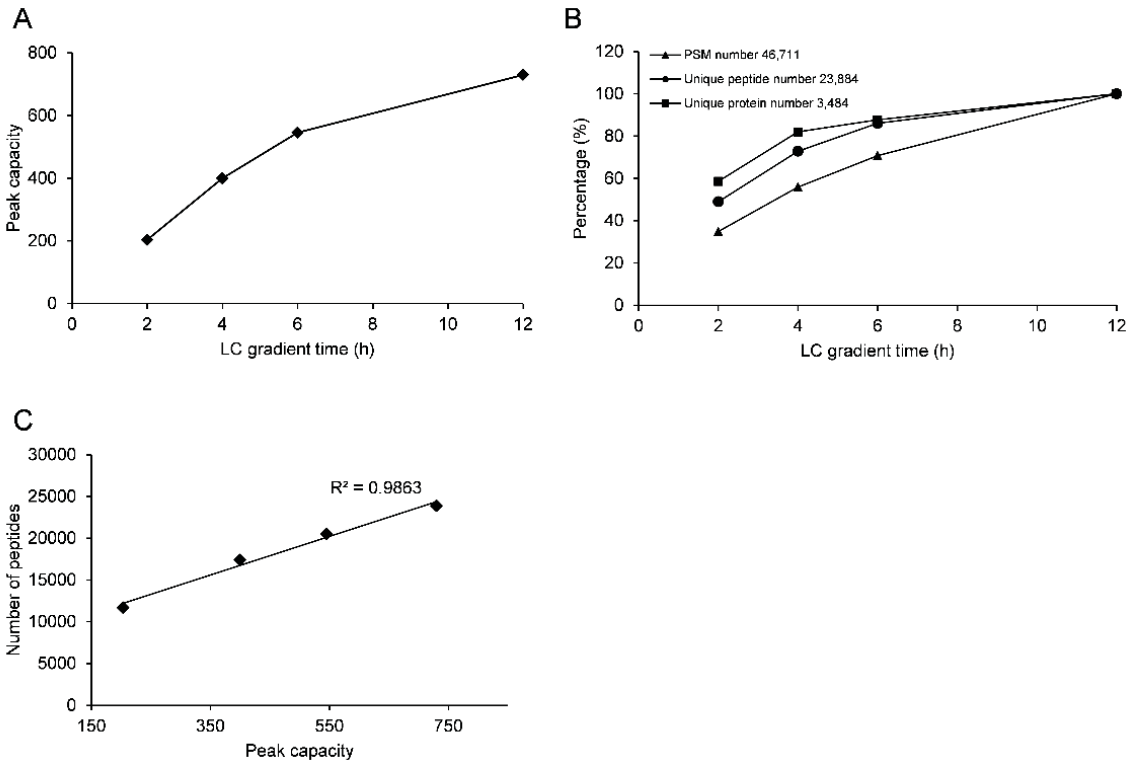



**Figure 2-5. Optimization of the LC gradient time for peptide elution.**
(A) Peak capacities plotted against gradient time. Peak capacities were calculated by dividing the average peak width of major peptide ions in a LC run over entire gradient time. (B) The correlation between the number of identified peptides/proteins and gradient time. (C) The number of detected peptides was in a linear relationship with the peak capacity.

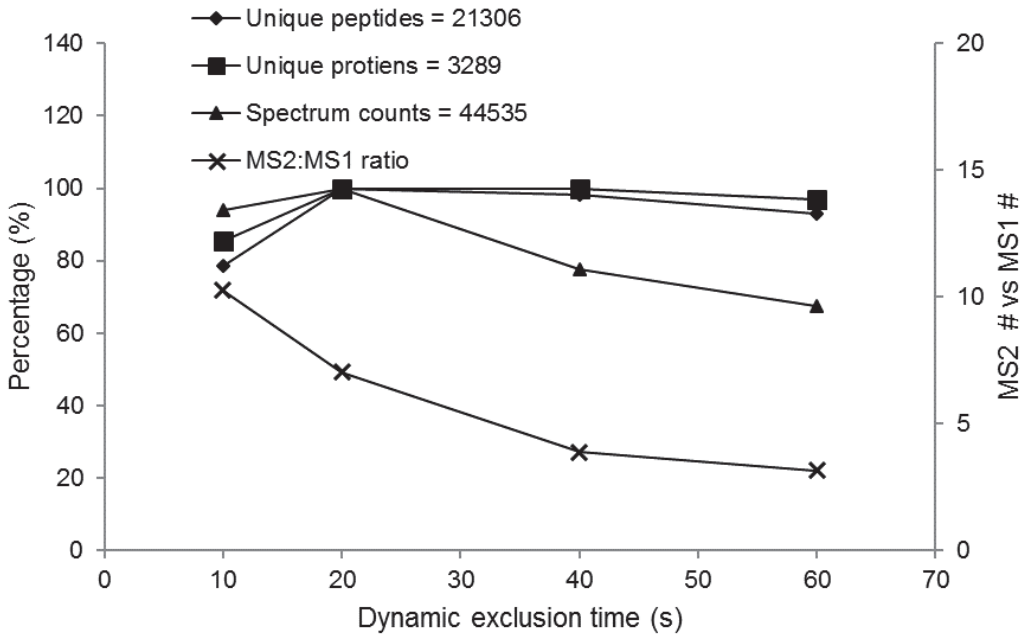



**Figure 2-6. Impact of dynamic exclusion time on the number of protein and peptide identifications.**
~ 6 µg of peptides was loaded on the long column and eluted in a 10% to 45% gradient of acetonitrile for 4 hr gradient time.

**Deep proteomic analysis of AD brain.** We then used the optimized conditions to explore the entire proteome of AD brain (**Figure 2-7A**). About 400 µg of protein was extracted from the tissue and subjected to Lys-C and trypsin digestion. We digested the entire cell lysate without removal of cell debris to increase the coverage of membrane and nucleus proteins as a recent study suggested[86]. Basic pH reverse phase LC was performed to pre-fractionate AD brain peptides because it provides better resolution and loading capacity than other methods (e.g. SCX, HILIC) and good orthogonality to acidic reverse phase LC (**Figure 2-7B**)[91,102]. We collected 10 basic pH LC fractions and analyzed each fraction on the acidic pH LC-MS/MS system in a 9 h gradient time. Total MS instrument time was about four days which is comparable to other reported instrument times required for in-depth proteomes analysis varying from one to two weeks[62,95,114]. A total of 1,695,626 high resolution MS/MS spectra was acquired, identifying 629, 747 PSMs (37% successful rate), 96,127 peptides and 10,544 proteins when protein FDR was controlled at 1%. On average, each peptide was identified by MS for about seven times. In each basic pH LC fraction, the average number of detected peptides and proteins were 11,9303 ± 651 and 4701 ± 119, respectively (**Figure 2-7C**). Nearly 80% of peptides were solely identified in one fraction and about 95% of peptides were only found in one or two fractions, suggesting high partitioning of peptides within each fraction (**Figure 2-7D**).

To evaluate the depth of AD brain proteome analyzed in this pilot study, we performed deep RNA-seq analysis of the same sample, and compared the proteome data with transcriptome results. We identified 16,670 protein coding genes by RNA expression, similar to the result in previous transcriptomic analysis of human brain[115]. The abundance of each transcript was calculated as reads in fragment per kilobase of exon per million fragments mapped (FPKM). A total of 10,161 human genes were detected in AD brain proteome, corresponding to 61% of the expressed genes (**Figure 2-8A**). Next we investigated the correlation between transcript and protein levels in our study. The protein level was indicated by a spectral counting based method[116], in which the total number of spectral counts for every protein was summed and normalized by the length of protein sequence (spectral counts per thousand amino acids) to adjust the bias created by protein size. We observed a modest correlation between the RNA and protein levels (Spearman correlation = 0.62, **Figure 2-8B**), which was consistent with the conclusions of other studies (Spearman correlation = 0.4-0.6)[100]. Taken together, our data suggested that the utilization of the current optimized LC/LC-MS/MS platform covers the majority of the AD brain proteome.

The multidimensional LC-MS/MS system presented here was robust with no instrument down time during the entire process of deep proteomic analysis. By heating the LC column to 60ºC, this LC/MS/MS system can be operated under a regular pressure limit (~3,000 psi with 0.15 µl/min flow rate), reducing potential problem of overpressure. To further enhance the identification of extremely low-abundance proteins, it is conceivable that extensive pre-fractionation of peptides during basic pH LC separation would further reduce sample complexity and improve dynamic range in the pre-fractionated pools. At last, combination of our long column LC/LC-MS/MS platform with the newly introduced Orbitrap Fusion Tribrid mass spectrometer would also allow considerably deeper proteomics analysis, due to its higher scan rate and peptide

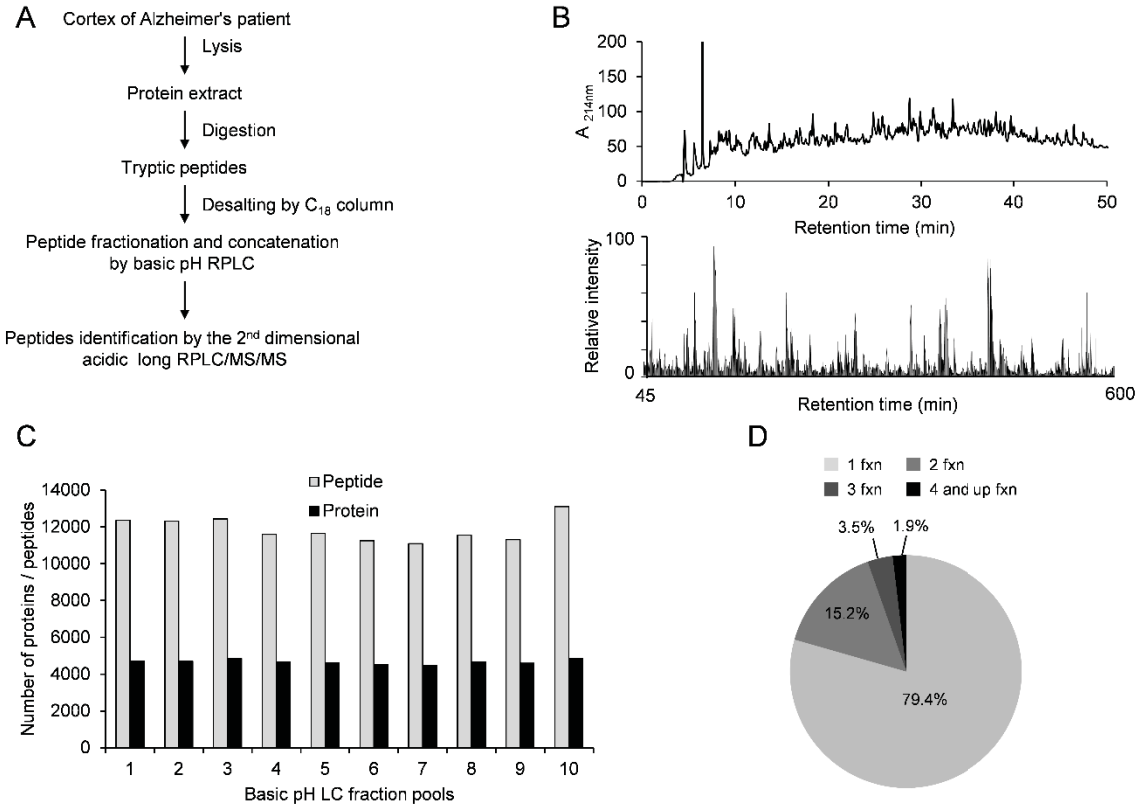



**Figure 2-7. Deep proteomics analysis of AD brain tissue.**
(A) Flow chart of the procedure. (B) Chromatograph of basic pH RPLC pre-fractionation of peptides (upper panel) monitored at 214 nm and an example base peak chromatograph of acidic pH long gradient PRLC-MS/MS (lower panel).(C) Basic pH RPLC fractionation yielded even partitioning of peptides which led to similar number of identified proteins in concatenated, pooled fractions. (D) Majority of the peptides was solely identified in one fraction.

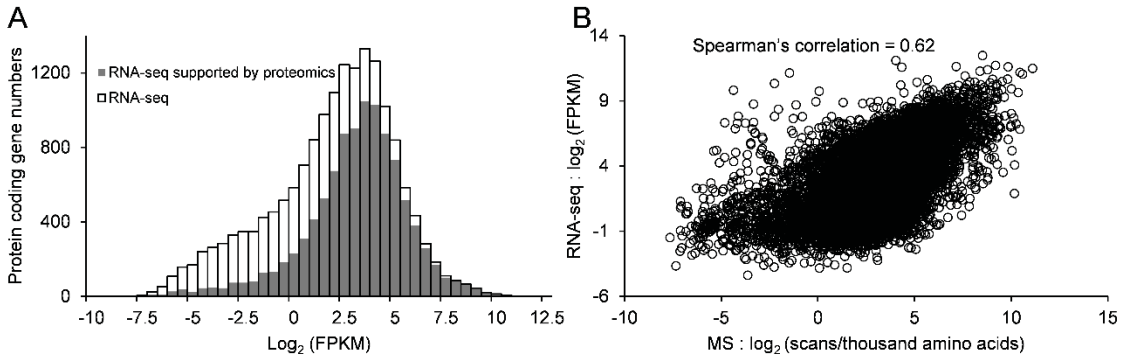



**Figure 2-8. Comparison of deep proteomics and RNA-seq data from the same AD brain tissue.**

(A) Histogram of FPKM distribution of RNA-seq and proteomics data. The open bar represents the distribution of protein coding gene numbers detected by RNAseq and the grey bar indicates the distribution of protein coding gene numbers validated by MS with different FPKM values. (B) Scatter plot of spectra counts per thousand amino acid of proteomic data versus FPKM of RNA-seq data.

identification efficiency than the Q Exactive MS instrument used in this study[82,117].

## Strategies to Enable Accurate Quantification by Isobaric Labelling and Mass Spectrometry for High-throughput Genome-scale Proteomic Analysis

This portion of chapter 2 will be the publication on the strategies to enable accurate quantification by isobaric labelling and mass spectrometry. Quoted text was taken from the manuscript. This study was led by two co-first authors Mingming Niu and Dr. Ji-Hoon Cho. Mingming Niu performed the proteomics experiments and data analysis. Ji-Hoon Cho developed the algorithm for *y*1 ion based interference correction. I contributed to the proteomics experiments and data analysis.

### Introduction

"Quantitative proteomics has been becoming an essential tool in biomedical research[4,64] and shows high potential for clinical application[65]. The integration of liquid chromatography and tandem mass spectrometry (LC-MS/MS) is the mainstream approach for global measurement of proteins and posttranslational modifications. Numerous MS strategies have been developed for large-scale profiling, including label free quantification and stable isotope labeling technologies[118]. More recently, isobaric labeling methods, such as isobaric tags for relative and absolute quantitation (iTRAQ)[11], tandem mass tags (TMT)[12] and DiLeu isobaric tags[13], gain popularity largely due to multiplexed capacity of processing up to 12 samples[14]. For example, isobaric labeling enables the analysis of hundreds of mammalian samples in tens of batches, detecting a total of more than 15K proteins (from 12K genes) and 60K phosphosites in mammalian samples[15-17].

Despite the advances of isobaric labeling, the method often suffers from high noise levels due to co-eluted interfering ions, leading to quantitative ratio compression that underestimates the difference, particularly in complex protein samples[18-21]. This drawback is ameliorated by some proposed approaches, which may be classified into three categories: pre-MS fractionation, MS setting modification, and post-MS correction. While pre-MS fractionation (e.g. 2D LC) partially reduced the co-elution of interfering peptides[119], 3D LC of basic pH reversed-phase (RP), strong anion exchange and acidic pH RPLC[120], yielded a more efficient platform for peptide separation. However, this platform involves complex LC settings that are not commonly used in other groups. The co-elution ions can also be reduced by a narrow isolation window[121], gas-phase purification[122] and complement reporter ion cluster quantification during MS analysis[123], usually at the expense of decreased information output. Some post-MS corrections were also reported by subtracting interference to enhance quantitation[124-126]. Moreover, a multistage MS3-based technique was developed to nearly eliminate the ratio compression, but has low sensitivity to detect weak peptide ions and requires expensive MS instrumentation[21,127,128]. Although all of these approaches improve quantitative

accuracy, the application is still limited by instrument dependency, time consumption, and computer algorithm availability.

In this study, we attempted to address the ratio compression issue by extensive high resolution fractionation and a novel *y1* ion-based interference correction method. To mimic real biological samples, we mixed TMT-labeled *E. coli* proteins at known ratios, in the presence of a 20-fold excess amount of background peptides from rat proteins. The mix was analyzed under multiple LC-MS/MS conditions by adjusting key parameters, including the fraction number collected in the 1st LC, MS isolation window, peptide loading amount and the 2nd LC fractionation depth (gradient length). We also developed a computational method that uses the known *E. coli* protein ratios to estimate the interference levels from rat proteins. Finally, the interference can be essentially eliminated by pre-MS fractionation, optimization of MS parameters, and post-MS *y1* ion-based correction, leading to a general pipeline for accurate isobaric labeling quantification.

## Methods and materials

**Preparation of *E. coli* and rat protein samples.** Proteins in *E. coli* cells or adult rat brain were extracted and digested as previously described[129]. Protein concentration was measured by the BCA method (Thermo Scientific); and the desalted tryptic peptides were resuspended in HEPES buffer (50 mM, pH 8.5) and labeled by individual 10-plex MT reagents (Thermo Scientific, *E. coli* peptides by 10 channels, and rat peptides by 8 channels from 126 to 130N). These peptides were pooled as specified (**Figure 2-9**).

**Basic pH LC prefractionation.** The pooled TMT-labeled sample was fractionated on a long reverse phase column (concatenated two Waters 4.6 mm × 25 cm Xbridge C18 columns, totaling 50 cm, 3.5 μm beads, Agilent 1270 HPLC, flow rate of ~0.4 ml/min). The gradient included 5 min of 95% buffer A (10 mM ammonium, pH 8.0), 215 min of 13%-35% buffer B (10 mM ammonium and 90% acetonitrile, pH 8.0), 90 min of 35-55% buffer B, and 10 min of 55-95% buffer B. A total of 320 fractions (one min each) were collected, dried and stored at -80ºC for further analysis.

**Acidic pH LC-MS/MS analysis.** Dried peptides were dissolved in 0.2% formic acid for LC-MS/MS analysis on an optimized platform[130,131] with modifications. Peptides were separated on a 75 μm x ~50 cm column (1.9 μm C18 beads, Dr. Maisch GmbH) and operated at 70ºC to reduce back pressure (solvent A: 0.2% formic acid, solvent B: 0.2% formic acid and 70% acetonitrile, 240 min gradient from 12-65% solvent B unless specified). The analysis used an Ultimate 3000 RSLC nano system coupled with an Orbitrap Fusion mass spectrometer (Thermo Scientific). The Orbitrap Fusion acquired data in a data-dependent manner alternating between full scan MS and MS/MS scans. The MS spectra (400-1600 m/z) were collected with 60,000 resolution, AGC of 2 x 105 and 50 ms maximal injection time. Selected ions were sequentially fragmented in a 3 sec cycle by HCD with 38% normalized collision energy, specified isolated windows (0.4-1.6 m/z, 0.3 m/z offset), 60,000 resolution. AGC of 1 x 105 and 150 ms maximal injection

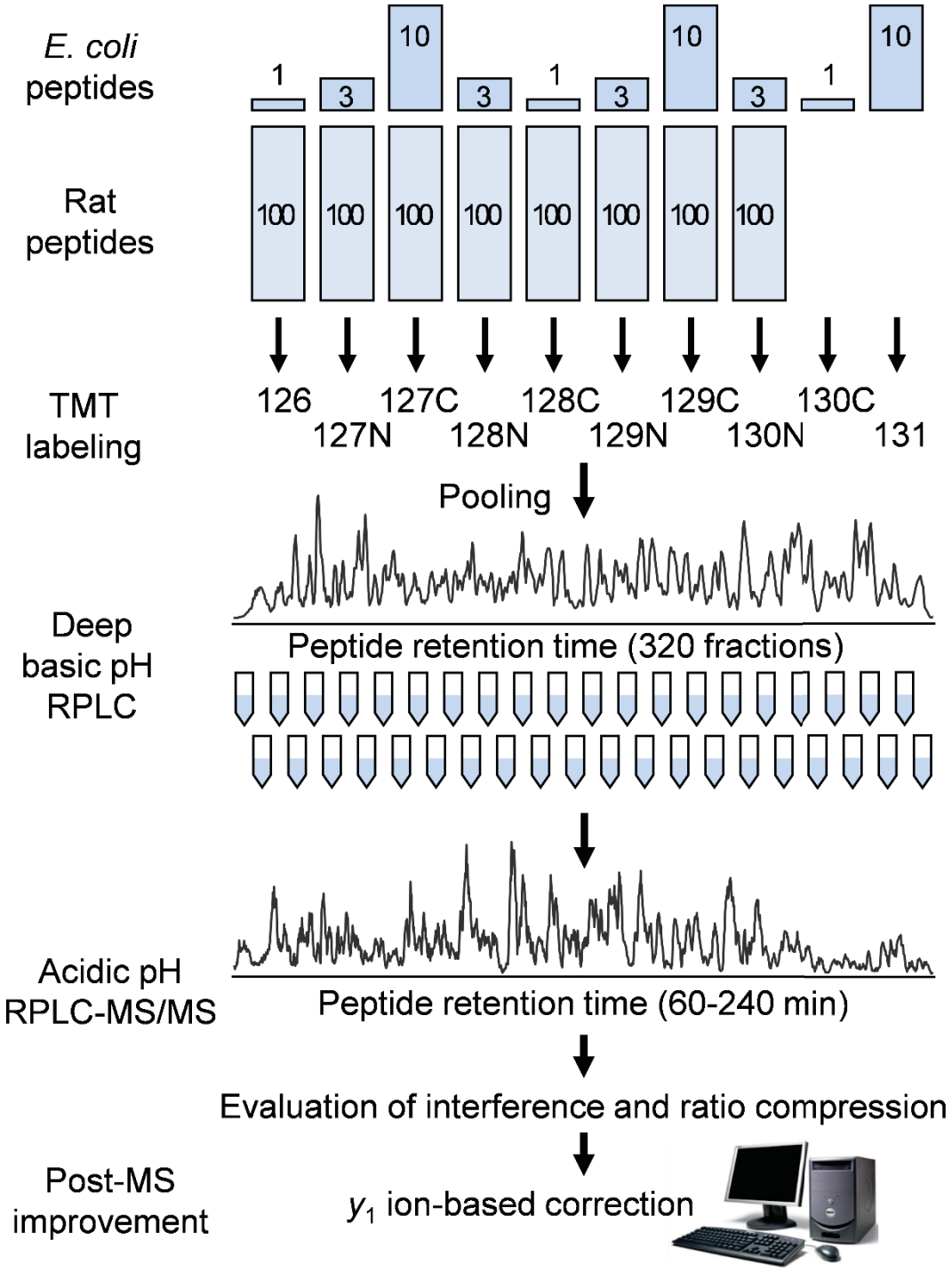



**Figure 2-9. Experimental design and procedures for evaluating TMT analysis.** Digested rat and *E. coli* peptides were TMT-labeled and mixed at known ratios, fractionated by basic pH RPLC, and analyzed by acidic pH RPLC-MS/MS. The interference of the TMT analysis was assessed by computational approaches, including a novel y1 ion-based correction method.

time. Dynamic exclusion was set to 20 sec.

For MS3 analysis, the precursors for MS2 analysis were isolated with a 1.6 m/z window (0.3 m/z offset). The CID-MS2 spectra were acquired in the ion trap with AGC of 1 x 104, 50 ms maximal injection time and 35% normalized collision energy. For HCD-MS3 mode, the setting were 65% normalized collision energy, 2 m/z isolated windows (0.3 m/z offset), 60,000 resolution. AGC of 1 x 105 and 500 ms maximal injection time. Dynamic exclusion was set to 30 sec.

**Protein/peptide identification and quantification.** The analysis was performed by our recently developed JUMP engine to improve sensitivity and specificity, which combines the advantages of pattern matching and de novo sequencing during a database search[132,133]. The JUMP hybrid algorithm was used to process numerous published large datasets[134-136]. RAW files were converted to mzXML format and MS2 spectra were searched against rat and *E. coli* target-decoy Uniprot databases to estimate the false discovery rate (FDR)[137,138]. Search parameters included precursor and product ion mass tolerance (6 ppm), fully tryptic restriction, two maximal missed cleavages, static TMT modification (+229.162932 Da on N-termini and Lys residues), dynamic Met oxidation (+15.99492 Da), and three maximal dynamic modification sites. Only a, b, and y ions were considered during the search. Peptide-spectrum matches (PSMs) were filtered by 7 minimal peptide length, mass accuracy (~2.5 ppm) and matching scores to achieve 1% protein FDR. For each accepted PSM, the peaks of TMT reporter ions were extracted for quantification

**Quantitative data analysis and post-MS computational correction approach.** To evaluate the levels of interference, TMT reporter ion intensities of each PSM were converted into relative intensities. For rat peptides that were equally mixed, the relative intensities were calculated by dividing individual reporter ion intensity by the mean intensity of eight reporters (126-130N). For *E. coli* peptides of three groups with known ratio, (126, 128C):(127N, 128N, 129N, 130N):(127C, 129C) to be 1:3:10 (**Figure 2-9**), the relative intensities were converted by dividing each channel intensity by the mean intensity of 126 and 128C. Then the relative intensity of each group was averaged in two steps:

For each PSM,

$$m_{1,i} = \frac{m_{126,i} + m_{128C,i}}{2},\ m_{2,i} = \frac{m_{127N,i} + m_{128N,i} + m_{129N,i} + m_{130N,i}}{4},\text{ and}$$

$$m_{3,i} = \frac{m_{127C,i} + m_{129C,i}}{2},$$ where $m_{reporter,i}$ and $m_{g,i}$ represent relative intensities of a reporter channel and the *g*-th group, respectively, for the *i*-th PSM.

Then for each group,

$\overline{m}_g = \sum_{i=1}^{N} m_{g,i}$, where $\overline{m}_g$ is the mean by averaging all PSM relative intensities, and *N* is the total number of PSMs.

Finally, the group mean was used to calculate interference level by compared to the expected ratio of the three groups ($r_1$ = 1, $r_2$ = 3 and $r_3$ = 10)

We also developed a post-MS computational approach to correct interference based on $y_1$ ion in MS2 scans. As K-TMT- and R-C-terminal tryptic peptides generate different $y_1$ ions (376.27574 Da and 175.11895 Da, respectively). If only one $y_1$ ion is detected and consistent with the identified peptide, the MS2 is termed a clean scan. If both $y_1$ ions are detected, the MS2 is termed a noisy scan. Assuming that the $y_1$ ion intensity is proportional to the reporter ion intensity, we computed their linear relationship from the "clean" scans, and then used the contaminated $y_1$ ion intensity in the "noisy" scans to derive the interference level.

## Results and discussion

**Generation of a cross-species peptide mix to mimic complex biological samples.** As ratio compression in isobaric labeling is largely influenced by sample complexity, we attempted to replicate the complexity of real biological samples by mixing cross-species peptides from *E. coli* and rat brains in a 10-plex TMT experiment (**Figure 2-9**). The *E. coli* peptides were labeled by 10 TMT reagents with known ratios (1 : 3 : 10 : 3 : 1 : 3 : 10 : 3 : 1 :10), while rat peptides were labeled in 8 channels and equally mixed as background, leaving 2 channels without the interference.

In the vast majority of proteomic comparison, proteins of high abundance are usually expressed from house-keeping genes and do not alter under experimental conditions, whereas changed proteins are likely to play regulatory roles and exist at low abundance. To simulate this scenario, we markedly increased the levels of background rat peptides, approximately 20-fold more than the targeted *E. coli* peptides, although in many of previous reports[127,139], the background and targeted peptides were pooled at comparable amounts.

This cross-species peptide mix was used for systematically dissecting the effect on ratio compression in three major steps, including pre-MS fractionation, MS settings and post-MS correction. The pre-MS fractionation was carried out by the combination of basic pH RPLC and acidic pH RPLC. During post-MS analysis, peptides shared between *E. coli* and rat were removed, and only species-specific peptides were considered (**Figure 2-9**).

**Confirmation of ratio compression and interference computation by known *E. coli* peptide ratios.** We initially analyzed the pooled peptide sample by one dimensional (1D) LC-MS/MS (**Figure 2-10**). As expected, rat peptide intensities in the 8 channels were almost equal, suggesting that these rat peptide measurements were not significantly affected by the *E. coli* peptides of low abundance (**Figure 2-10A**). In contrast, a clear ratio compression was observed for the *E. coli* peptides in the 8 channels in the presence of ~20-fold background rat peptides, but not in the 2 channels without rat peptides (i.e. 130C and 131, **Figure 2-10B**). Averaged ratios of 1 : 3 : 10 *E. coli* channels

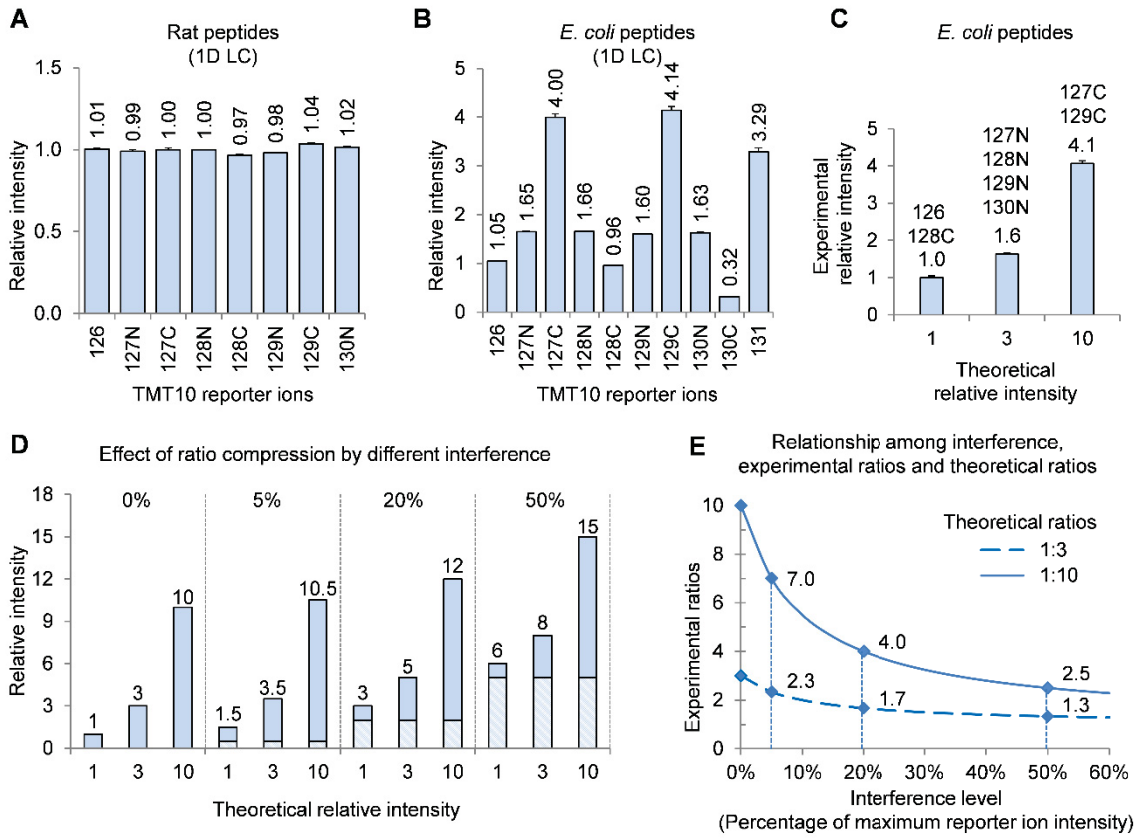



**Figure 2-10. Interference analysis of TMT data based on known peptide ratios.** (A) The averaged relative intensities of rat peptides in each TMT channel. (B) The averaged relative intensities of *E. coli* peptides in all TMT channels. (C) The summed relative intensities of *E. coli* peptides in three groups: the lowest level (126 and 128C), the medium level (127N, 128N, 129N and 130N) and the highest level (127C and 129C). (D) Schematic diagram showing the interference definition and its effect on ratio compression. The interference is defined as the proportion of maximal reporter intensity. If the interference is equal to 5% of maximal intensity (5% x 10 = 0.5) in each TMT channel, the intensities are elevated, resulting in ratio compression. The cases of interference of 20% and 50% are also shown. (E) Given theoretical ratios between reporters, the interference can be inferred by experimental ratios. For example, when theoretical and measured ratios are 1:10 and 1:4 respectively, the interference should be ~20% of maximal reporter intensity

were found to be 1 : 1.6 : 4.1 (**Figure 2-10C**), consistently with previously reported effects of ratio compression[19,21,121].

We then developed a method to calculate the level of interference based on known peptide ratios. Assuming the level of interference was stable among channels, defined as the percentage of maximum reporter ion intensity (e.g. 10, **Figure 2-10D**), if there was no interference, the relative intensities were 1, 3 and 10; if the interference was 20% (10 x 20% = 2), the relative intensities increased to 3, 5 and 12, resulting in compressed ratios (1 : 1.7 : 4.0); and so on. Such analyses were performed over various interference levels to generate a standard curve, describing the relationship among interference, experimental ratios and theoretical ratios (**Figure 2-10E**). With this standard curve, experimental and theoretical ratios, the interference could be calculated. For example, when experimental 1 : 4 ratio was detected for theoretical 1 : 10, we concluded that the interference was about 20% of maximum reporter ion intensity. If there were multiple experimental and theoretical ratios, the interference could be derived by minimizing errors with generalized equations

**Interference level affected by core LC/LC-MS/MS parameters.** With the interference computation method, we examined the effects of a number of core parameters in LC/LC-MS/MS, including the 1st LC resolution, MS2 isolation window, the 2nd LC loading amount, and the 2nd LC resolution. To change the resolution during the 1st LC, we separated the complex mix into 320 fractions, and then combined some of the fractions together to adjust the separation power. For instance, combination of every 4 adjacent fractions would yield 80 fractions, and so on. Thus, the 1st LC resolution was reflected by different number of collected fractions (1, 5, 10, 20, 40, 80, and 320), under which the interference levels decreased gradually from 16.4% to 2.8%, suggesting that extensive fractionation in the 1st LC alleviated co-eluted peptides but could not totally remove the interference (**Figure 2-11A**).

As to the 2nd LC-MS/MS analysis, we examined the effect of MS2 isolation window and found that the interferences were almost proportional to the size of the isolation window (**Figure 2-11B**), in agreement with previous studies[121,124]. For example, 4-fold difference of window size (1.6 to 0.4 Da) resulted in 4-fold difference of inference level (14.4% to 3.7%). We also observed a visible impact of peptide loading amount on the interference. When the loading level decreased from 900 ng to 100 ng, the interference reduced from 9.4% to 3.3%, implying that high loading led to peak broadening[140] and therefore raised the interference (**Figure 2-11C**). Finally, we adjusted the 2nd LC resolution by gradient length (1, 2, and 4 h) on a long column. The 4 h gradient nearly eliminated the interference (down to 0.4%) and gave the best result (averaged ratios of 1 : 2.8 : 9.9, **Figure 2-11D**). This result was comparable with that of the accurate MS3 strategy15 (the interference of 0.6%, averaged ratios of 1 : 2.8 : 9.7, **Figure 2-11E**), although the MS3 analysis usually has limited sensitivity of analyzing weak peptide ions, requires more MS acquisition time and uses low-resolution MS2 for protein identification[64]. Taken together, our data demonstrated that the combination of extreme fractionation (320 x 4 h = 1,280 h, 53 days) and narrow isolation window was able to solve the issue of ratio compression.

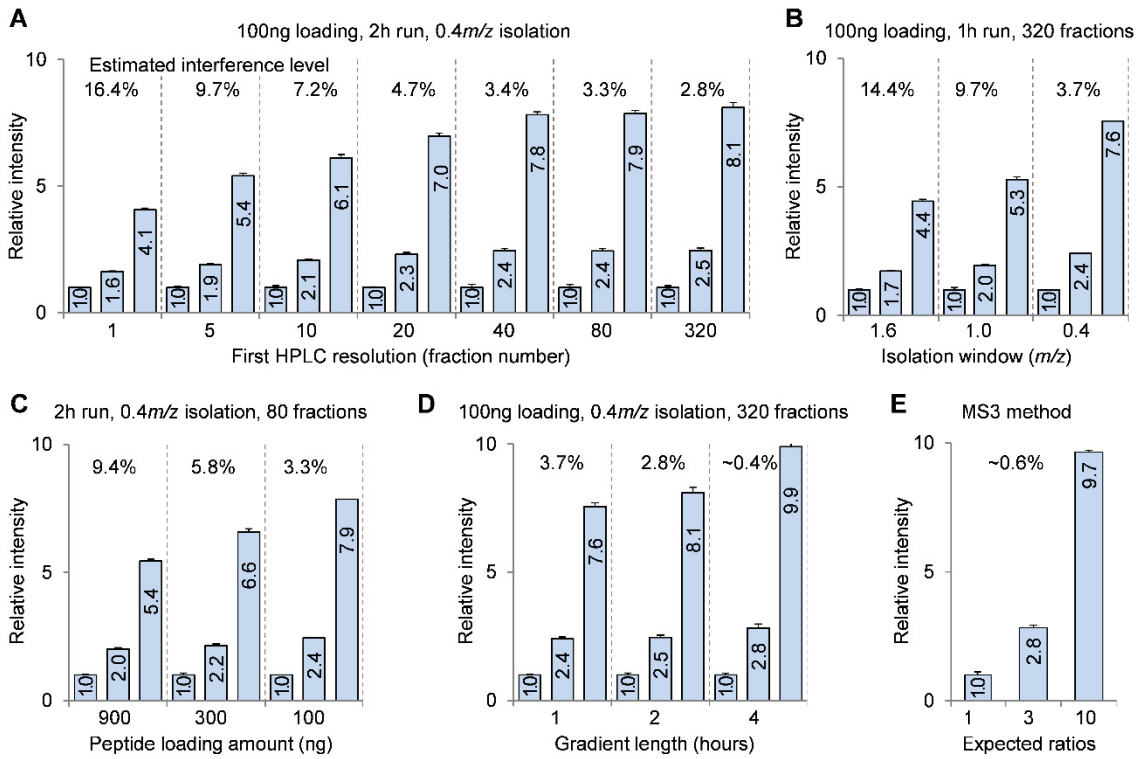



**Figure 2-11. Effects of LC-MS parameters on interference in the TMT analysis.** The summed relative intensities of *E. coli* peptides in three groups under different conditions of (A) the first LC, (B) isolation window, (C) peptide loading, (D) the 2nd LC gradient, and (E) MS3 setting. The interference level under each condition was computed from the known 1 : 3 : 10 ratios.

We also analyzed the effects of these core parameters on peptide/protein identification in this analysis. As anticipated, pre-MS fractionation reduced sample complexity, leading to low peptide/protein identification per fraction. However, if all fractions were analyzed, the combined number of identifications would be improved[120,131]. Interestingly, the reduction of MS2 isolation window had minor effect on peptide/protein identification, which may be due to high isolation efficiency of quadrupole mass filters that were installed in newly developed MS instruments (e.g. Q Exactive HF)[141]. The effect of sample loading was also relatively small, as even 100 ng of fractionated peptides (one of 80 fractions, **Figure 2-11C**) contained concentrated peptides from the initial 8,000 ng total peptides (80 x 100 ng). When some of the 320 fractions were analyzed by different gradient time, detected peptides were similar from 1 h to 2 h, but decreased in 4 h, because the long gradient might result in peak broadening and decreased the sensitivity of identifying weak peptides[140].

In summary, we could eliminate ratio compression by utilizing extreme separation power (LC/LC fractionation and narrow isolation window), but this strategy required >50 days instrument time. To compromise the effects of these core parameters on the interference and protein identification, we recommended the final setting of a narrow isolation window (0.4 Da), medium fractionation (~40 x 2 h, 3.3 days) and ~100 ng of fraction sample loading, which resulting in a low level of interference (3.4%, **Figure 2-11A**).

**Interference correction by *y1* ion-based post-MS method.** We devised a post-MS computational approach to calculate and remove the interference based on the information in MS2 scans, in contrast to MS1-based correction methods[125,126]. In the MS1-based strategies, the intensity of co-eluted peptides is estimated from the precursor isolation window in MS1 scans, but these MS1 survey scans are acquired at different time points from the MS2 scans during real time LC-MS/MS analysis. Therefore, the exact co-eluted peptides in MS2 are not directly measured.

In our method, the level of interference was directly analyzed in the same MS2 scan for reporter ion quantification. When examining TMT MS2 scans of tryptic peptides (**Figure 2-12A**), we found that some MS2 scans (clean scans) displayed only one *y1* ion (K-TMT or R residue) consistent with the matched peptide sequences. The other MS2 scans (noisy scans) had the two *y1* ions, indicative of contaminated peptides. In the clean scans, the *y1* ion intensity tended to be proportional to the measured TMT reporter intensities because they originated from the same precursor ions. Therefore, the relationship between intensities of either K-TMT- or R-*y1* ion and TMT reporter ions was modeled as a linear form. This linear relationship enabled the calculation and correction of interference level in the noisy scans (**Figure 2-12A**).

The performance of this post-MS correction approach was evaluated in the dataset of varying LC resolution (e.g. different fractions in the 1st LC), and showed unanimous improvement in precision. For example, under our recommended condition of 40 fractions, the interference level decreased from 3.4% to 1.3%. The measured ratio was 1:2.7:9.1, nearly reaching the expected ratio of 1:3:10 (**Figure 2-12B**). The analysis

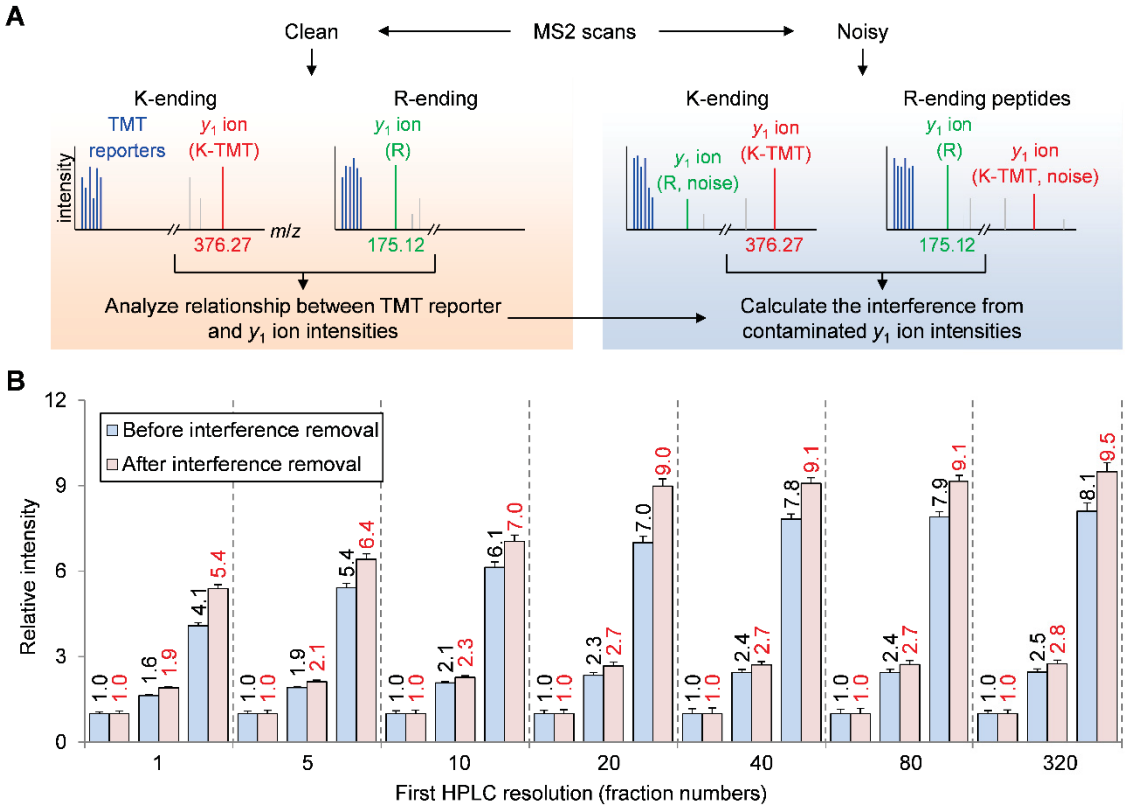



**Figure 2-12. Post-MS computational approach for interference removal.** (A) Schematic diagram. MS2 scans of tryptic peptides can be divided into clean and noisy scans based on the detected y1 ions. The relationship between y1 ions and TMT reporter ion intensities is modeled by the clean scans. Then the relationship is used to calculate the interference in the noisy scans. (B) The summed relative intensities of *E. coli* peptides in the three groups, before and after the post-MS computational correction.

indicated that the use of our optimized LC-MS parameters and post-MS correction almost eliminated ratio compression commonly observed in the TMT-based quantification.”

# CHAPTER 3. DEEP MULTI-OMICS PROFILING OF BRAIN TUMORS TO IDENTIFY SIGNALING NETWORKS DOWNSTREAM OF CANCER DRIVER GENES

## Introduction

A central gap in cancer biology concerns how oncogenes drive the rewiring of molecular signaling networks to execute phenotypic changes[28,29,38]. Initial attempts to decode the molecular networks were through proteomic characterization via an antibody-based approach (e.g. the reverse phase protein array)[142,143]. However, this targeted approach is restricted by profiling breadth and depth, largely due to antibody availability and specificity. The signaling networks are highly regulated by protein posttranslational modifications, such as phosphorylation, and thus phosphoproteomic measurement is indispensable for studying cancer signaling[144]. Recently, mass spectrometry (MS)-based proteomics technology has been emerging as the mainstream strategy for unbiased analysis of the genome-wide proteome and phosphoproteome. Together with advanced DNA sequencing, these methodologies provide an unprecedented opportunity of deep omics analysis. It is now possible to integrate transcriptome, deep proteome and phosphoproteome to dissect oncogenic signaling networks, broadening our understanding of cancer biology[28,29,38,145].

High-grade gliomas (HGGs) are the most prevalent malignant brain tumors, and confer devastating mortality[39,40]. Although significant efforts in glioma sequencing have unveiled comprehensive genome-wide mutation landscapes[39,40,43-48], a complete understanding of how genomic alterations lead to dysregulation of particular master regulators and specific pathways remains unclear. Previous HGG proteomic and phosphoproteomic studies extend our understanding of HGG signaling[44,49], but most of these attempts have used proteomic approaches of relatively shallow depth. There is essentially no deep HGG proteomic landscape available for the cancer research community. Here, for the first time, we present a new paradigm of identifying ~12,000 gene products (proteins) and >30,000 phosphosites for dissecting HGG cancer biology.

In the present paper, we have compared two HGG mouse models driven by oncogenic receptor tyrosine kinases (RTKs: PDGFRA D842V and TPM3-NTRK1 fusion) respectively, using integrative systems biology analyses of proteome, phosphoproteome and transcriptome. Mutations and/or amplifications of PDGFRA and fusion genes of the NTRK family of neurotrophin receptors have been identified in pediatric and adult HGG[39,46,48,66,67]. Human surgical HGG specimens exhibit paradigmatic inter- and intra-tumoral heterogeneity, dramatically undermining the power to dissect the global proteome and signaling networks[49,142,146,147]. To improve sample quality, we have used engineered mouse HGGs expressing the mutated RTKs in the same cell type to minimize confounding factors[40,47]. With a novel bioinformatics pipeline, we identify various functional modules and master regulators rewired in HGGs and demonstrate that the TPM3-NTRK1 oncogene upregulates multiple other RTKs to form a

positive feedback loop within the PI3K-AKT pathway, driving more rapid tumor development compared with the PDGFRA-driven HGG.

## Methods and Materials[1]

### Mutated RTK driven HGG mouse models and tissue collection

Mouse experiments were approved by Institutional Animal Care and Use Committee that are in compliance with national and institutional guidelines. Mouse HGGs were generated as previously described[46,48]. Briefly, a pooled population of p53-deficient primary mouse astrocytes was transduced with retrovirus expressing either human TPM3-NTRK1 fusion or PDGFRA D842V mutation along with an IRES-GFP, and then $2 \times 10^6$ cells per mouse were intracranially implanted into athymic nude mice. Mice were anesthetized and perfused with PBS on the manifestation of brain tumor symptoms. GFP-labeled HGG tumors were dissected with fluorescent dissecting microscope followed by snap freezing for proteome and transcriptome analyses.

### Antibodies and other reagents

Antibodies against the following proteins were used for Western blotting: FLAG (Sigma-Alrdrich, F1804), Tubulin (Santa Cruz, 23948), pEphA2 (Cell Signaling 6347), EphA2 (Cell Signaling 6997), p-C-Myc (Abcam, 32029), C-Myc (Cell Signaling, 9402), PDGFRA (Santa Cruz, 338). PhosStop phosphatase inhibitor (Roche); Lys-C (Wako, peptides: Lys-C = 100:1); Trypsin (Promega, peptides: Trypsin = 100:1); TiO2 beads (GL sciences, TiO2: peptides = 4:1); and $C_{18}$ 1.9 μm resin (Dr. Maisch GmbH)

### RNAseq analysis

RNAs were extracted by Trizol (Invitrogen) from about 20 mg aliquots of the same tumor samples for proteomics analyses. The mRNA samples were purified by poly(dT) beads. Paired end adaptors were used for ligation. RNAseq reads were aligned to multiple databases encompass human genome (GRCh37), human transcriptome (RefSeq and AceView), and all other possible combinations of RefSeq exons. The reads mapped to the transcriptome were converted to genomic mapping and merged in the final BAM files.

---

[1] I would like to acknowledge my collaborators from Dr. Suzanne Baker's group for their support on materials and experiments. Alex Diaz and others from Dr. Baker's group provided the mouse models and RNAseq, performed IHC and immunoblot experiments.

**Deep proteomics profiling by two-dimensional reverse phase LC-MS/MS**

Whole proteome and phosphoproteome analyses were processed similarly as previously described[148]. Tissue samples (~10 mg each) were homogenized at 4°C in 0.3 ml of lysis buffer (50 mM HEPES, pH 8.5, 8 M urea, 0.5% sodium deoxycholate, 1 x PhosStop Phosphatase Inhibitor). Cell lysate including insoluble debris was digested with Lys-C followed by trypsin overnight at room temperature. Peptides from each sample were labelled with TMT10-plex reagents and then pooled together with equal amount. Pooled peptides were pre-fractionated with a 2 hour gradient basic pH liquid chromatography. A nano UHPLC (Waters) system was applied to load peptides on a heated 75 μm × 110 cm column packed in house with $C_{18}$ 1.9 μm resin (Dr. Maisch GmbH) and interfaced to a Q Exactive MS. A long gradient of up to 9 hours for peptides separation was carried out in a buffer system with 5% DMSO added. Q Exactive was operated with m/z range 420−1600, MS1 resolution 70,000 at m/z = 200. MS2 resolution was set at 35,000 and a predicted AGC of 2 × 105 with maximal ion time of 128 ms, top 10, isolation window of 1.6 m/z, NCE of 31 or 33, dynamic exclusion of 45s or 60s. MS data was processed using our in house developed tool JUMP suites as previously described[149,150]. Briefly raw MS files were analyzed by JUMP version 12.1.0, peptide lists were searched against the database downloaded from Uniprot mouse database (52,490 protein entries) with methionine oxidations as dynamic modifications. FDR was set to 0.02 at protein level with a minimum amino acids length 7 and was determined by searching a reverse database. Initial precursor and fragment mass tolerance was set to 6 PPM and 10 PPM respectively. Minimal percentage of precursor peak intensity (PPI) was set to 70%; minimum and median TMT channel intensities were set to 2,000 and 10,000 respectively to guarantee only high quality PSMs were applied for quantification.

**Phosphoproteome analysis with an additional step of phosphopeptide enrichment**

95% of peptides were applied for phosphoproteome analysis with similar parameters as the whole proteome method. An extra step of phosphopeptide enrichment was performed with $TiO_2$ using our refined phosphopeptide enrichment strategy[151]. Briefly, peptides were incubated with 0.5 mM $KH_2PO_4$ non-phosphopeptides competitor and $TiO_2$ beads with a peptides-to-beads ratio of 1:4 for 20min to allow efficient and specific phosphopeptide enrichment. Stepped NCE was applied at 30±15%. For data processing, Minimal percentage of PPI was set to 50%; minimum and median TMT channel intensities were set to 1,000 and 5,000 respectively.

It is often problematic to pinpoint the location of phosphosites when there are presence of consecutive serine (S), threonine (T) or tyrosine (Y)[152]. As a result, randomness of phosphosites assignment occurs which often also results in inflation of total amount of phosphosites reported. In this study, to define a reasonable biological base for prioritization of sites that are essentially indistinguishable to reduce randomness of phosphosites assignment, we borrowed information from sites that already confidently assigned to the same protein to guide the assignment of indistinguishable sites. The algorithm of phosphoRS[152] was used to calculate phosphorylation site scores (between

0~100) in a peptide. We assigned the phosphorylation site when its peptide-level score was higher than the second-highest site by at least 10, which prevented the situation that the identical sites identified by multiple scans (peptides) were differently localized because of subtle differences of scores. In addition, we newly introduced a protein-level site score defined as the highest value of peptide-level scores corresponding to the site in a protein. Since, for each site, the protein-level score consolidated the scores from peptides, it was particularly useful when the peptide-level score of a site was not distinguishable from others. Nevertheless, phosphorylation sites in a peptide might not be determined when multiple sites had the exact same score in both peptide and protein-level. To address the problem, a heuristic priority was given to the amino acid residue of a site in the following order: SP-motif > S > T > Y.

**Evaluation of proteomic profiling depth**

Theoretically observable peptides were determined by three main factors including peptide mass, peptide hydrophobicity and corresponding minimal RNA FPKM value. In silico digestion of peptides was performed. Distributions of the mass, hydrophobicity and FPKM of detected peptides against all in silico digested peptides were evaluated sequentially to determine the cutoff of mass range, hydrophobicity range, and minimal FPMK for theoretically observable peptides. As a result, mass range was determined as 800 Da – 3200 Da, Hydrophobicity was set as -52 to 17, and minimal FPKM was set as 0.005. After applying these cutoffs, distribution of percentage of detected peptides against in silico digested peptides passed filters was plotted to determine coverage of theoretically observable peptides for each protein. As a result, a mean coverage of 42% of theoretically observable protein sequences was achieved through our deep proteomic profiling. It is often challenging to evaluate depth of phosphoproteome because the total amount of phosphorylation events in one specific cell or tissue is unknown. To provide a reasonable estimation of phosphoproteome depth, we compared our data to all mouse phosphosites reported previously in the PhosphoSitePlus database[153]. Phosphosites with at least two independent evidences of MS-based identifications in the database were accepted to evaluate the coverage of the amount of our phosphosites against the total reported mouse phosphosites.

**Differential expression analyses of whole proteome and phosphoproteome**

ANOVA was performed for differential expression analysis comparing cortex, NTRK HGG, and PDGFRA HGG with a P value determined by permutation and then adjusted by Benjamin Hochberg method to handle relative small sample size ($n$ = 10). A cutoff of P value 0.05 was applied. Qualified proteins were further filtered by fold change of 1.5 in at least one comparison among 3 sample groups; a final FDR was estimated by permutation using resulted differential expression (DE) genes. DE phosphosites were identified by the same procedure with a fold change of 2. Principal component analyses and hierarchical clustering analyses detected a NTRK HGG sample outlier

(**Figure 3-1e, f**) which was also supported by the evidence of low transduced TPM3-NTRK1 expression (**Figure 3-1d**). This was probably caused by normal brain tissue contamination. To avoid possible false discoveries, this sample outlier was excluded from differential expression analyses.

**Pathway and network module analysis**

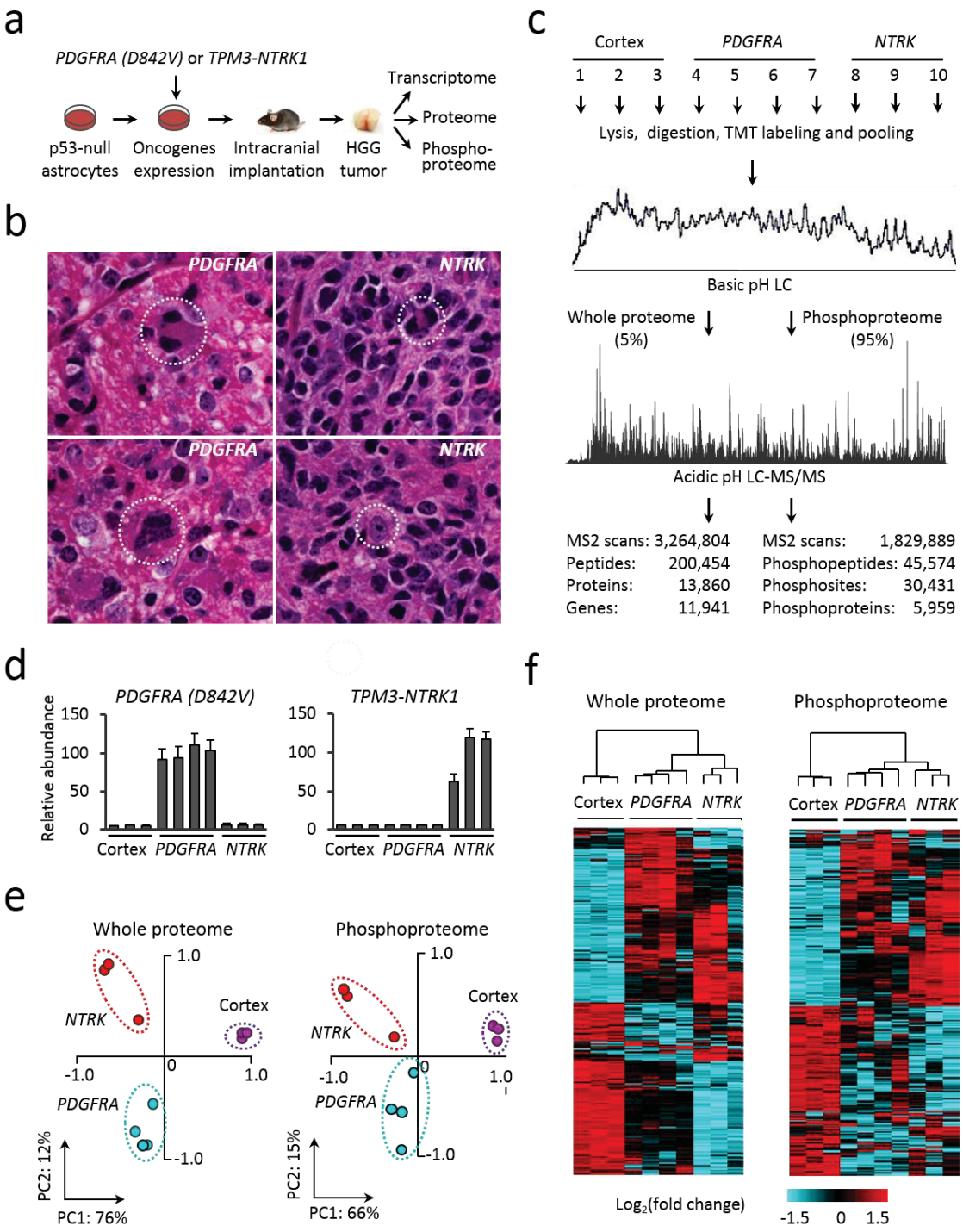
Following DE analysis, WGCNA R package[154] was utilized to perform weighted DE gene co-expression clustering analysis. Briefly, a Pearson correlation matrix was calculated using all samples, correlation type was set to "signed" to allow only positive correlation. A hybrid dynamic tree-cutting method with a minimum height for merging modules at 0.15 was applied to define co-expression clusters. The first principal component, also known as eigengene, was calculated as a consensus trend for each co-expression cluster. DE Proteins were assigned to the co-expression cluster with the highest correlation based on Pearson R value. Pathway and network module analyses were carried out using ClueGO[155], a software package based on cytoscape[156]. ClueGO can detect pathways utilizing diverse statistical tests and then apply Kappa statistics to link deregulated pathways to construct network modules according to their connections. DE proteins or phosphosites from each co-expression cluster were applied for this analysis. Pathway analysis was performed using a right-sided hypergeometric test, a Benjamin Hochberg corrected P value of 0.05 was set as cutoff. Pathways from KEGG, WikiPathways, and Reactome databases were combined to construct network modules. Kappa score cutoff was set to 0.5 to ensure stringent network module construction.

**Kinase activity analysis based on whole proteome normalized phosphoproteome using IKAP**

Kinase activity analysis was carried out using IKAP, a heuristic machine learning algorithm that infers the activities of kinases from substrate phosphorylation. Kinase-substrate relationship was extracted from the PhosphoSitePlus database[157]. With our deep phosphoproteome data, phosphosites instead of phosphopeptides can be applied as data input to increase the accuracy of IKAP analysis. The phosphorylation level in each tumor sample was first normalized against normal cortex samples, and then these normalized samples were normalized against the whole proteome to retain only the phosphosite changes driven by phosphorylation (Remove the contribution of protein expression change). We repeated the simulation process 10 times and assessed the solution's variation to overcome limitation of gradient descent optimization algorithm that could get stuck in a local minimum. And then, we applied a cutoff of 0.2 SD to filter results that failed to converge into a stable solution. Kinase activities derived from substrate size <3 were filtered out except ones that were supported by upstream kinases with co-activation patterns. Finally, we applied a cutoff of B.H. adjusted P value 0.05 to determine kinase with altered activity. A kinase-to-kinase network was constructed based on kinase co-activation patterns (e.g. according to database, PRKCE phosphorylates AKT1 on S473 and this phosphorylation activates AKT1 enzymatic activity; PRKCE shows activity

**Figure 3-1. Work flow of MS-based proteomic analyses and data quality evaluation.**

(a) Overview of HGG mouse models. 2 × 106 p53-null primary mouse astrocytes transduced with human TPM3-NTRK1 fusion or PDGFRA D842V mutation genes were implanted intracranially into athymic nude mice to generate HGG tumors for proteome, phosphoproteome and transcriptome analyses. (b) Mouse gliomas have high-grade features. 160 x H&E images highlight PDGFRA-driven HGG with mitotic figures (top) and a multinucleated giant cell (bottom); NTRK-driven HGG with mitotic figures (top) and tumor invasion of normal parenchyma as evidenced by entrapped native neurons (bottom). (c) Proteomic analysis work flow. 3 normal mouse cortex samples, 4 PDGFRA-driven HGG tumors, and 3 NTRK-driven HGG tumors were applied for whole proteome and phosphoproteome analyses using combination of TMT 10-plex labeling, extensive 2D-LC peptides separation and Q Exactive MS analysis. (d) Validation of cancer driver genes expression. MS-based quantification of human-specific peptides expression agreed with the HGG genotypes. (e) Principal component analyses (PCA) separate sample groups by genotypes. Graphs show PCA analyses results (PC1 and PC2) for whole proteome and phosphoproteome. (f) Unsupervised hierarchical clustering analyses cluster samples by genotypes. Heatmap shows hierarchical clustering analyses using top 1,000 most variable proteins and top 3,000 most variable phosphosites respectively.


a
PDGFRA (D842V) or TPM3-NTRK1
Transcriptome
Proteome
Phospho-proteome
p53-null astrocytes
Oncogenes expression
Intracranial implantation
HGG tumor
b
PDGFRA
NTRK
PDGFRA
NTRK
c
Cortex
PDGFRA
NTRK
1 2 3 4 5 6 7 8 9 10
Lysis, digestion, TMT labeling and pooling
Basic pH LC
Whole proteome (5%)
Phosphoproteome (95%)
Acidic pH LC-MS/MS
MS2 scans: 3,264,804
Peptides: 200,454
Proteins: 13,860
Genes: 11,941
MS2 scans: 1,829,889
Phosphopeptides: 45,574
Phosphosites: 30,431
Phosphoproteins: 5,959
d
PDGFRA (D842V)
TPM3-NTRK1
Relative abundance
150
100
50
0
Cortex PDGFRA NTRK
e
Whole proteome
Phosphoproteome
1.0
0
-1.0
Cortex
NTRK
PDGFRA
PC2: 12%
PC1: 76%
PC2: 15%
PC1: 66%
f
Whole proteome
Phosphoproteome
Cortex PDGFRA NTRK
Log2(fold change)
-1.5 0 1.5

pattern of NTRK > PDGFRA >cortex according to IKAP. We accept this kinase-to-kinase relationship if IKAP inferred AKT1 activity shows the same pattern as PRKCE. Moreover, since AKT1 S473 phosphorylation is detected in our data, we further require the measurement of phosphorylation on S473 follow the same DE pattern).

**TF activity inference by integrative analysis of transcriptome, proteome and phosphoproteome**

TF activity was derived from target gene expression in transcriptome and whole proteome clusters and was further validated by the measurements of TF whole protein expression and phosphorylation. Proteins from each WP-C were first overlapped with targets of TFs according to publicly available TF-target relationship from the Encode database[158] to search for TFs with differential activity among samples. Fisher exact test was performed to determine the significance of this overlapping, which was followed by B.H. FDR correction. P value cutoff was set to 0.05. Similarly, we overlapped the known target genes of TFs with differentially expressed genes determined by student T test comparing HGG samples with normal cortex in the transcriptome data with the same criteria. We only accepted TFs that passed cutoffs in both transcriptome data and whole proteome data. This list was then further filtered by the measurements of whole protein and phosphorylation data. We require that either the protein or phosphorylation is differentially expressed among sample groups. To construct a putative HGG network that links signal cascades to TF regulations. We applied similar rules as construction of kinase-to-kinase network introduced above. For the interplays between kinases and TFs, we incorporated relationship of kinase-substrate phosphorylation and TF-target information from PhosphoSitePlus and Encode database respectively.

**Pathway activity measurement using alterations of annotated functional phosphosites**

Most of pathway activity inference strategies were based on gene expression at transcripts level or proteins expression level, which are not accurate indexes of protein activity, especially for signal transduction proteins whose activity are mainly regulated by protein phosphorylation. Here we modified a transcripts expression based pathway activity inference strategy[159] to compute PI3K-AKT pathway activity using phosphorylation changes. Instead of defining a subset of genes in the pathway to optimize discriminative power through computational training, we directly selected proteins that have differential activity according to phosphorylation changes in reported functional sites compare HGG tumors to cortex to summarize pathway activity. Formula is as below:

$$a(P) = \sum_{i=1}^{k} Ci * Fi/\sqrt{k}$$

Where k is the number of proteins with different activity. Fi is calculated in two steps: average $\log_2$ fold changes of proteins with multiple activation and/or inhibition

sites that are differentially phosphorylated between protein $i$ in tumors over normal cortex was first calculated, and then this averaged value of each protein was applied for the formula above. $Ci$ is the functional annotation of phosphorylation. If the phosphorylation is reported to play a positive role in tumor biology, $Ci$ is +1; a negative role, $Ci$ is -1. Bootstrap was performed with replications of 10,000 times to determine the significance of pathway activity difference compare NTRK HGG to PDGFRA HGG.

### Combination of mouse and human HGG data to prioritize putative cancer genes

Since we demonstrated the distinct oncogenic potency of the two RTK cancer drivers in mice, oncogene-responsive changes can be restricted to genes with an expression pattern that correlates with the distinct oncogenic potency of two RTKs. Mouse transcripts expression that follow the expression of NTRK > PDGFRA > Cortex was first extracted. To be stringent, we require the fold difference between NTRK HGG and PDGFRA HGG larger than or equal to 2 and B.H. adjusted student T test P value <0.05. Moreover, only genes that have consistent expression pattern in either whole proteome or phosphoproteome were accepted as final oncogene-responsive changes in mice HGG. A list of genes that have higher expression in human cases with NTRK fusions than cases with PDGFRA mutations in transcriptome data were pulled out with a cutoff of P value 0.05 and fold change 2. Finally, the two lists of genes from mice and human were overlapped for convergent oncogene-responsive changes.

## Results

### Deep quantitative analysis of whole proteome and phosphoproteome of multiple HGG mouse models and immunoblotting validation

To provide a deep mouse HGG proteome and phosphoproteome landscape we used our newly developed MS pipeline with extensive peptide separation power and high mass resolution[148,150,151]. Mouse HGG samples were generated by intracranial implantation of p53-null primary astrocytes transduced with either PDGFRA D842V mutation or the TPM3-NTRK1 fusion, two oncogenic RTKs found in human HGGs[46,48] (**Figure 3-1a, b**, referred to as PDGFRA HGG and NTRK HGG, respectively). The HGG and normal mouse cortex (control) samples were submitted to proteome, phosphoproteome and transcriptome profiling. Multiplexed isobaric labeling was used to enable massively parallel proteome and phosphoproteome quantification of ten samples (**Figure 3-1c**). A total of 30 whole proteome peptide fractions and 20 phosphoproteome peptide fractions were acquired through a basic pH reverse phase liquid chromatography (LC) pre-fragmentation followed by an up to 9 hour acidic pH reverse phase LC to allow the maximum peptide separation[62,160]. As a result, 13,860 proteins (11,941 gene products, 200,454 peptides and 3,264,804 MS2 scans) and 30,431 phosphosites (5,959 phosphoproteins, 45,574 phosphopeptides, 1,829,889 MS2 scans) were identified (<1% false discovery rate, **Figure 3-1c**). Among them, 13,567 proteins (11,718 gene products)

and 28,527 phosphosites were quantified, representing the deepest HGG proteomic datasets available.

To evaluate the quality of the datasets, we examined the MS-based results of the two transduced human oncogenes and some well-known phosphorylation events, as well as the classification of all measurements. The protein expression levels of human PDGFRA D842V and TPM3-NTRK1 agreed with the HGG genotypes (**Figure 3-1d**). MS data of specific phosphosites were also consistent with immunoblot assays described previously in these HGG mouse models: AKT S473, PRAS40 T247, PDGFRA Y742, S6 S235 and S6 S236 (**Figure 3-2**)[46]. Principal component analysis and hierarchical clustering analysis revealed that the two RTK oncogenes drive distinct proteome, phosphoproteome and transcriptome profiles (**Figures 3-1e, f**, and **3-3a, b**). In the MS analysis, the intra-group replicated samples showed minimal variations with small standard deviation, whereas the inter-group comparisons exhibited differences with much larger standard deviation (**Figure 3-4a, b**). For transcriptome profiling, RNAseq replicates from a second cohort of HGGs displayed high reproducibility of these HGG mouse models ($R^2 > 0.95$, **Figure 3-5**). Together, these results indicate the high quality of our omics data sets and demonstrate that the two oncogenic RTKs drive HGGs with reproducibly distinct global proteome and phosphoproteome profiles.

We further analyzed the correlation and profiling depth of proteome and transcriptome. The transcript levels and protein abundances showed a moderate correlation (**Figure 3-3c**, $R^2 = 0.5$), consistent with previously reported datasets[35,161]. In 12,842 detected transcripts (FPKM > 1), 10,838 (84%) corresponding proteins were mapped by MS (**Figure 3-3d**). Deep proteomic profiling depth was also exemplified by high coverage of low abundance regulatory proteins, such as kinases (84%)[62,162]. In addition, we investigated peptide coverage of each protein in this shotgun proteomics analysis. More than 96% of proteins were identified by at least two peptides (**Figure 3-3e**), and the average coverage of theoretically observable protein sequences reached 42% (**Figure 3-3f**). Moreover, we estimated the phosphoproteomic profiling depth by comparing to all previously curated mouse phosphosites in the PhosphositePlus database, the most comprehensive protein modification database[153]. Our phosphoproteome covered approximately 68% of the mouse phosphosites collected from all cell types and tissues, and contained 12,354 novel phosphosites not in the database. In summary, these data present a paradigm of one of the deepest proteome and phosphoproteome analyzed in cancer studies.

## Globally differential regulation of the proteome and functional modules in the HGG tumors

We first identified differentially expressed (DE) proteins among mouse cortex, PDGFR and NTRK HGGs, and performed gene coexpression clustering, pathway analysis, and functional module classification by WGCNA[154] and ClueGO packages[155,156] (**Figure 3-6a**). A total of 4,703 DE proteins and 6,768 DE phosphosites (2,301 phosphoproteins) were identified and distributed into 5 whole proteome coexpression

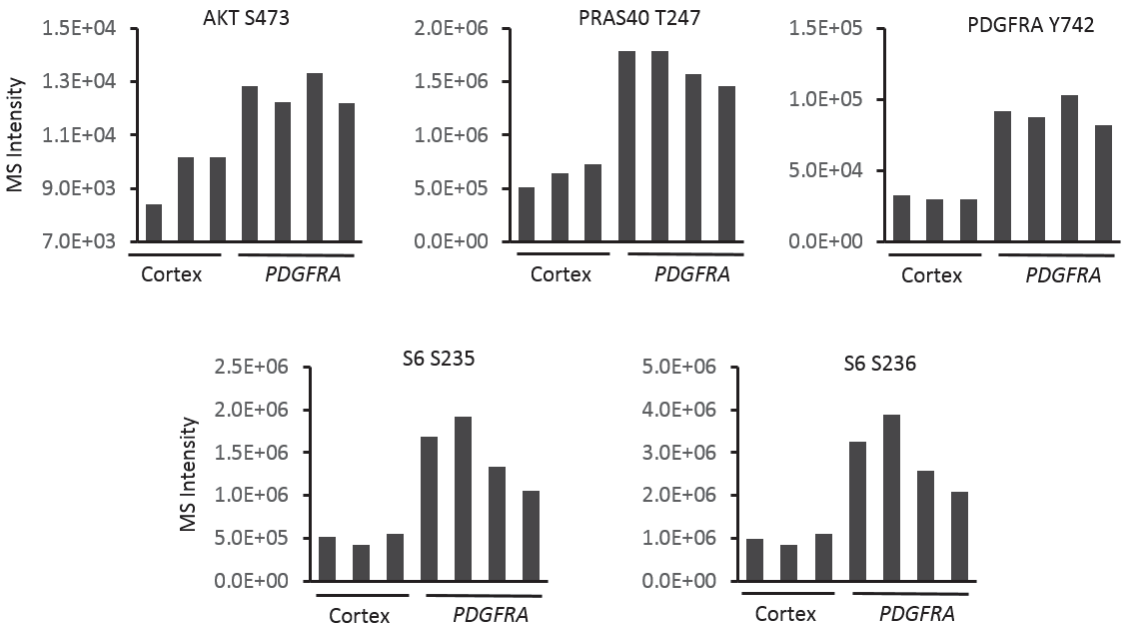



**Figure 3-2. MS-based quantification is accurate.**
MS measurements of phosphorylation events are highly consistent with previous immunoblot assays comparing PDGFRA-driven HGGs to normal cortex[46].

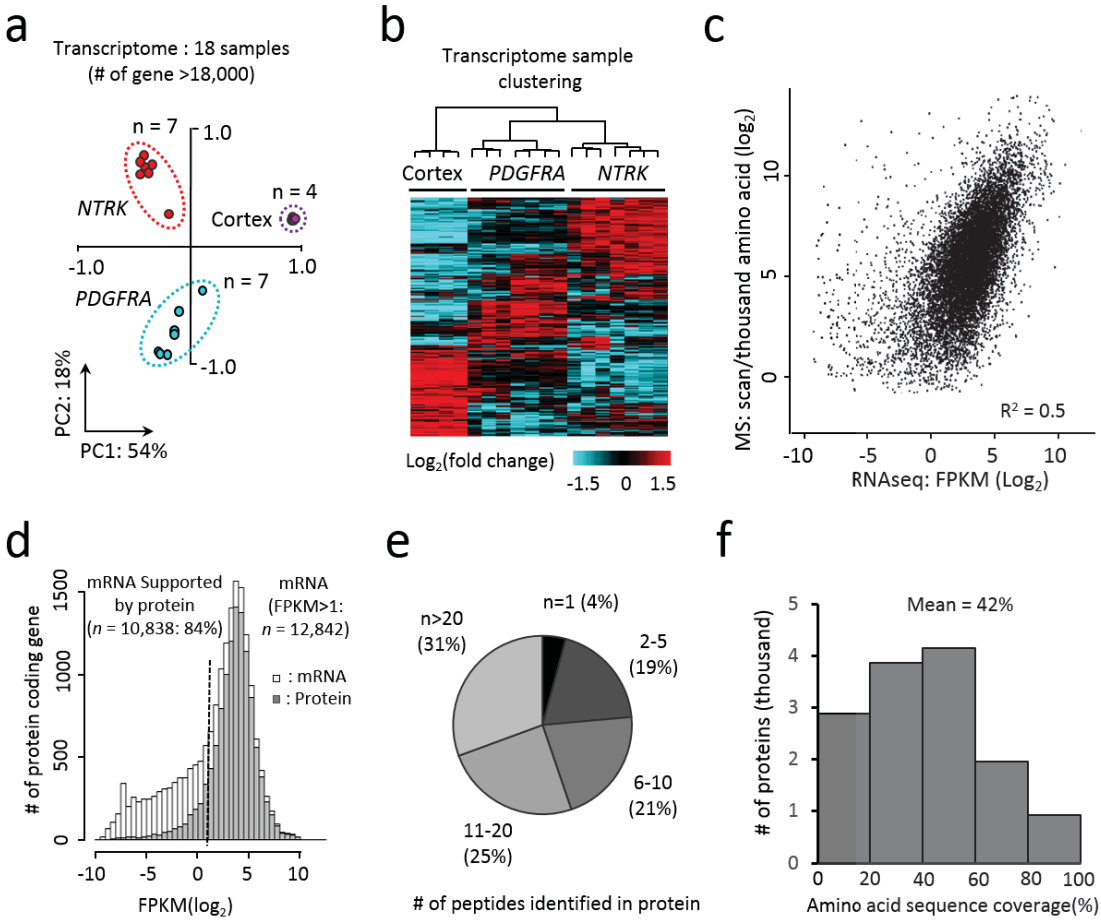



**Figure 3-3. Deep proteomic data achieved through extensive peptide fractionation.**

(a) Principal component analysis (PCA) of transcriptome separates samples by genotypes. Graph shows PCA analysis result (PC1 and PC2) for transcriptome. (b) Unsupervised hierarchical clustering of transcriptome clusters samples by genotypes. Heatmap shows hierarchical clustering of transcriptome using top 3,000 most variable transcripts. (c) Transcriptome abundance and proteome abundance displays moderate correlation. Scatter plot shows log2 level transcripts FPKM and their corresponding proteins scans per thousand amino acids. (d) High percentage (84%) of transcripts shows corresponding protein expression. Histogram distribution of the log2 level FPKM value of transcripts and the log2 level FPKM value of proteins. A threshold of FPKM equals 1 is marked. (e) Pie plot illustrates particularly high number of peptides identified for each protein using our MS-based analysis platform. (f) High percentage of theoretically observable amino acid sequence coverage achieved through our MS-based analysis platform. Histogram distribution shows amino acid sequence coverage of theoretically observable peptides for whole proteome.

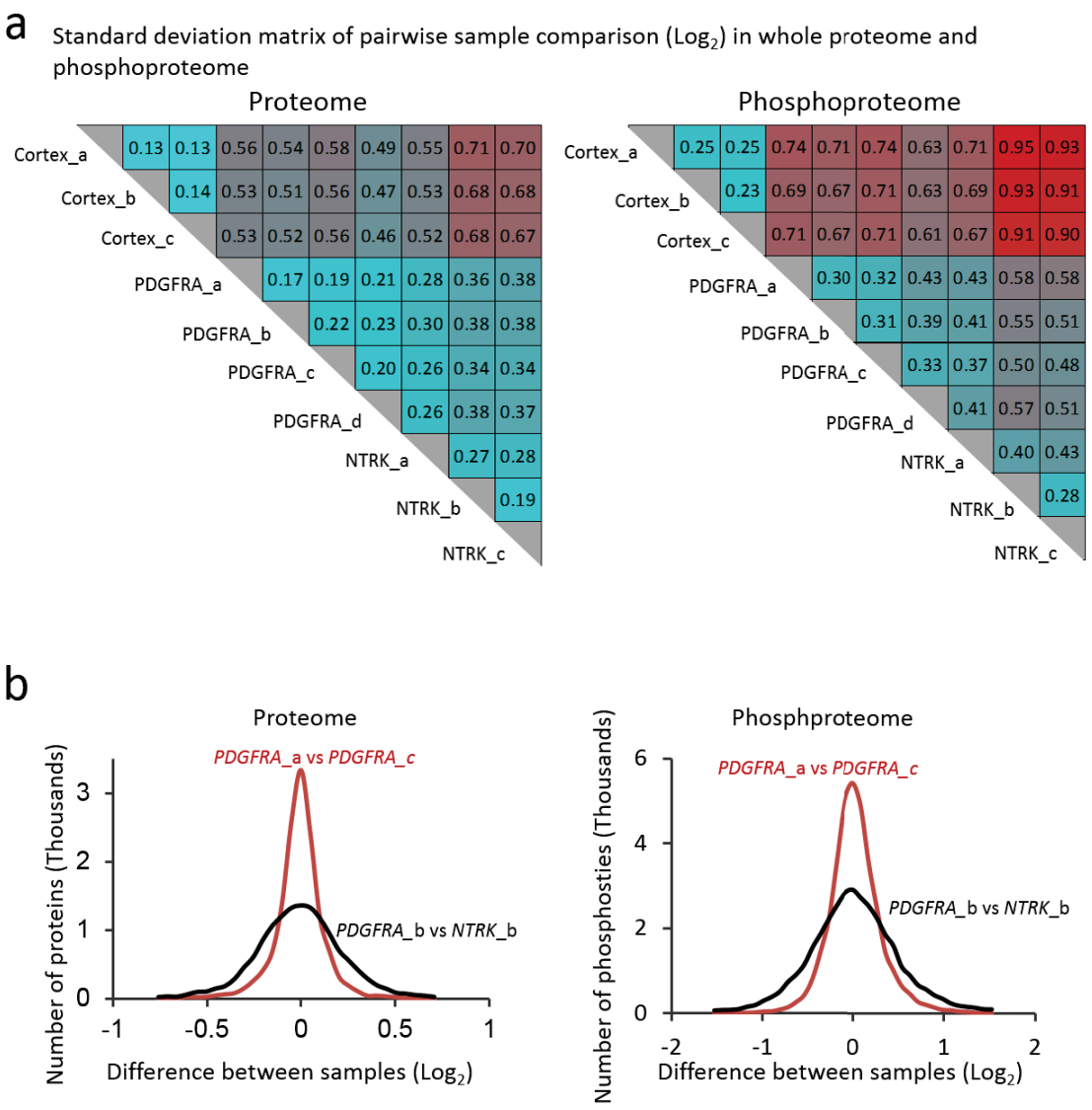



**Figure 3-4. MS-based proteomic analyses specify particularly small quantitative variations between replicates.**
(a) Pairwise standard deviation matrix of whole proteome and phosphoproteome displays particularly small variations between replicates. (b) Examples show distributions of representative Log2 level whole proteome and phosphoproteome variations between PDGFRA-driven HGG biological replicates and between two HGG tumors.

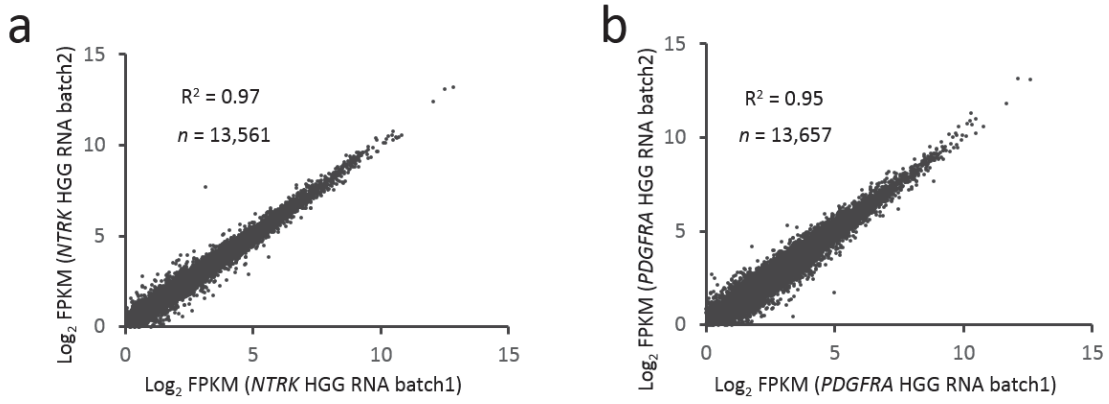
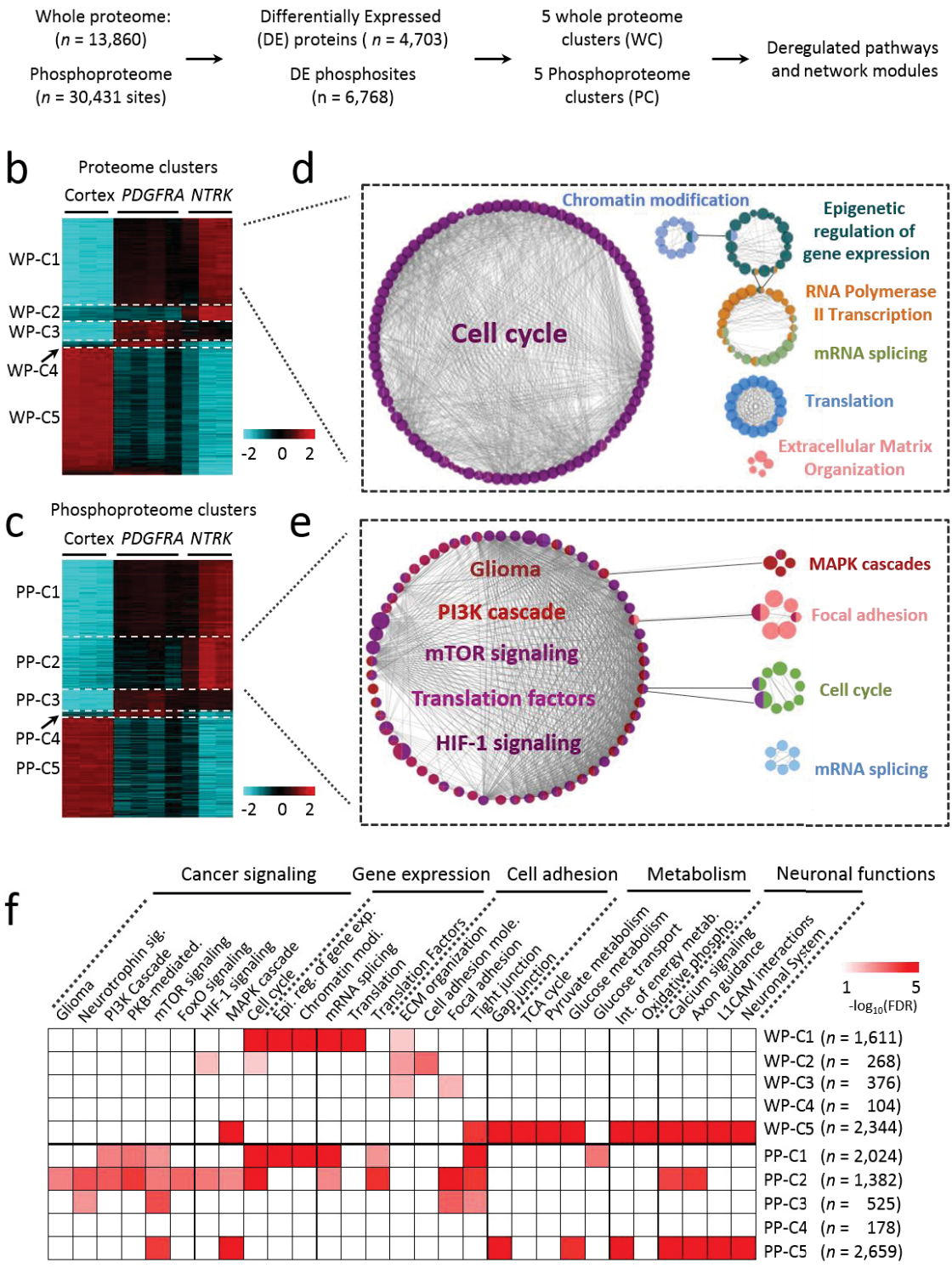



**Figure 3-5. HGG mouse models are reproducible.**
Comparisons of RNAseq results on batch2 and batch1 HGG mouse models show highly consistent transcript abundance from tumors generated by independent experiments. Scatterplots show FPKM values of genes in batch1 and batch2 HGG mouse models.

**Figure 3-6. Global network analyses using coexpression clustering and pathway functional grouping identify both canonical HGG network modules and multiple new pathways and network modules in HGGs.**
(a) Overview of pathway and network module analysis strategy. Differential expression (DE) analysis was carried out through ANOVA with a cutoff of BH adjusted p value 0.05. Fold change cutoff was set to 1.5 and 2 for proteome and phosphoproteome respectively. WGCNA package was applied for coexpression clustering analysis using DE genes. Each coexpression cluster was utilized for pathway and network module analysis using ClueGO. (b, c) Coexpression clustering analysis detects multiple DE protein or phosphoprotein coexpression clusters with distinct expression patterns. Heatmap shows expression patterns of DE proteins in whole proteome and phosphoproteome clusters identified through WGCNA. (d, e) Heavily interconnected network modules are identified in WP-C1 and PP-C2 respectively. Graphs show top network modules detected in WP-C1, PP-C2 respectively. Each node represents a pathway. Circular layout was applied to present network modules. Pathways that are functionally related are connected by edges and then grouped to network modules represented by distinct colors. Node size represents pathway enrichment significance. (f) Summary of pathways detected in each co-expression cluster. Representative pathways detected in each cluster were organized based on their general biological processes. Color scale represents B.H. adjusted p value derived from pathway analysis.

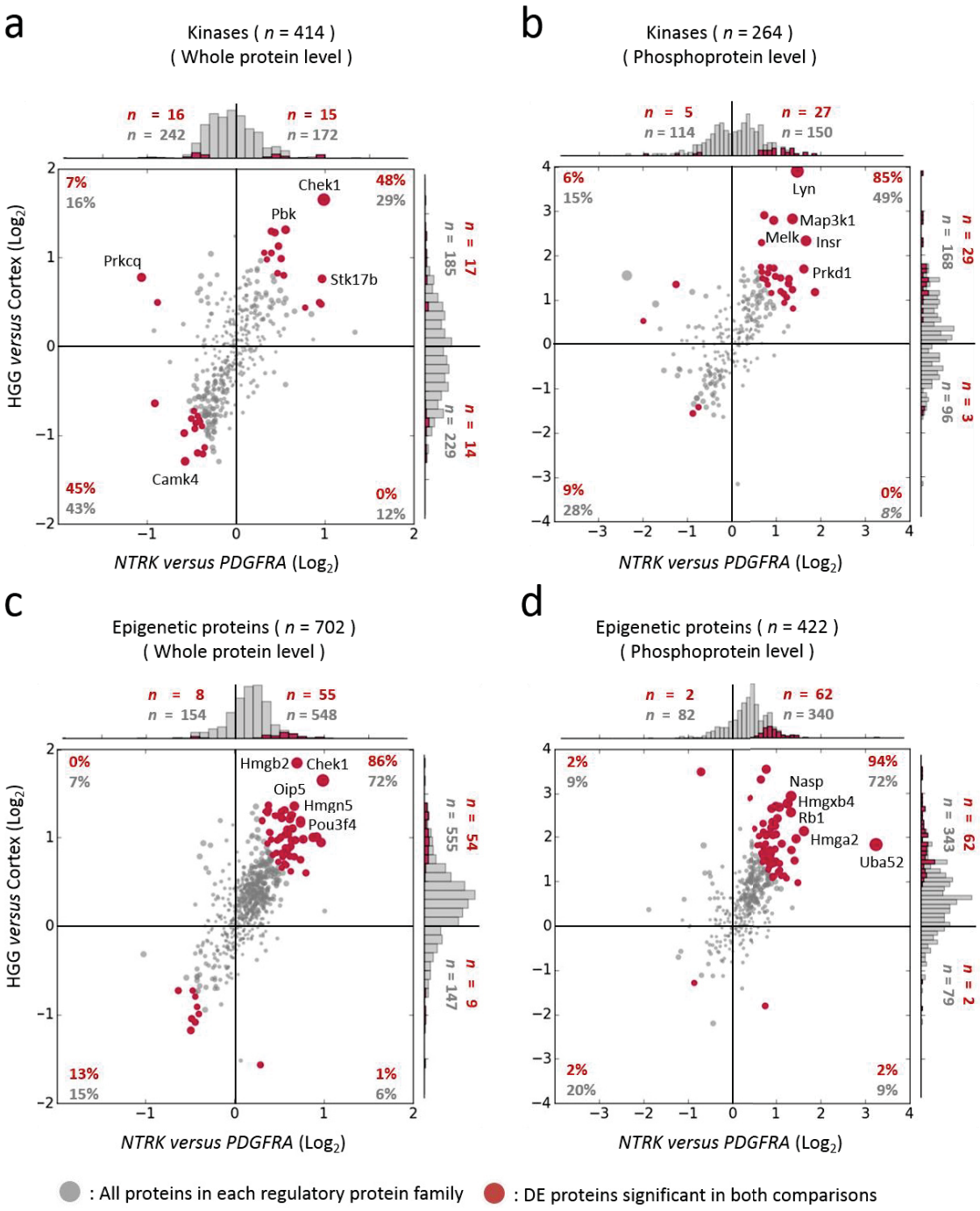

Whole proteome:
(n = 13,860)
Phosphoproteome
(n = 30,431 sites)
Differentially Expressed
(DE) proteins ( n = 4,703)
DE phosphosites
(n = 6,768)
5 whole proteome
clusters (WC)
5 Phosphoproteome
clusters (PC)
Deregulated pathways
and network modules
b
Proteome clusters
Cortex PDGFRA NTRK
WP-C1
WP-C2
WP-C3
WP-C4
WP-C5
-2 0 2
c
Phosphoproteome clusters
Cortex PDGFRA NTRK
PP-C1
PP-C2
PP-C3
PP-C4
PP-C5
-2 0 2
d
Chromatin modification
Cell cycle
Epigenetic
regulation of
gene expression
RNA Polymerase
II Transcription
mRNA splicing
Translation
Extracellular Matrix
Organization
e
Glioma
PI3K cascade
mTOR signaling
Translation factors
HIF-1 signaling
MAPK cascades
Focal adhesion
Cell cycle
mRNA splicing
f
Cancer signaling
Gene expression
Cell adhesion
Metabolism
Neuronal functions
Glioma
Neurotrophin sig.
PI3K Cascade
PKB-mediated.
mTOR signaling
FoxO signaling
HIF-1 signaling
MAPK cascade
Cell cycle
Epi. reg. of gene exp.
Chromatin modi.
mRNA splicing
Translation
Translation Factors
ECM organization
Cell adhesion mole.
Focal adhesion
Tight junction
Gap junction
TCA cycle
Pyruvate metabolism
Glucose metabolism
Glucose transport
Int. of energy metab.
Oxidative phospho.
Calcium signaling
Axon guidance
L1CAM interactions
Neuronal System
1 5
-log10(FDR)
WP-C1 (n = 1,611)
WP-C2 (n = 268)
WP-C3 (n = 376)
WP-C4 (n = 104)
WP-C5 (n = 2,344)
PP-C1 (n = 2,024)
PP-C2 (n = 1,382)
PP-C3 (n = 525)
PP-C4 (n = 178)
PP-C5 (n = 2,659)

clusters (WP-C) and 5 phosphoproteome coexpression clusters (PP-C) respectively (**Figure 3-6b, c**), leading to 67 functional modules. As expected, the two largest modules rewired in tumors compared with normal cortex are cell cycle (in WP-C1, **Figure 3-6d**) associated with tumor cell proliferation, and the PI3K signaling cascade (in PP-C2, **Figure 3-6e**), which transduces signals downstream of RTKs. Collectively, a series of module groups including cancer signaling, gene expression, cell adhesion, metabolism, and neuronal functions are rewired in HGG tumors (**Figure 3-6f**). Notably, three clusters (WP-C1, PP-C1, and PP-C2) display similar alteration patterns: cortex < PDGFRA HGG < NTRK HGG (**Figure 3-6b, c**), and the majority of known glioma pathways are enriched in these 3 clusters (**Figure 3-6f**), suggesting that NTRK HGG activates similar oncogenic pathways but with a greater magnitude of response at the global pathway level than PDGFRA HGG. Moreover, the majority of HGG cancer signaling pathways are only altered in the phosphoproteome but not in the whole proteome (**Figure 3-6f**), underlining the indispensable role of phosphoproteome profiling to decode oncogenic signaling. Thus, these results suggest RTK oncogenes drive massive rewiring of signaling networks at phosphorylation and/or protein expression level in the HGG mice.

We then investigated the global changes of regulatory family proteins including transcription factors (TFs), epigenetic genes, kinases and cancer genes in the HGG tumors. Regulatory proteins in general are present at low abundance[62], thus are difficult to analyze without highly sensitive methods. Nevertheless, our deep profiling systematically characterized both whole protein and phosphorylation levels of a large number of regulatory proteins (**Figure 3-7**). Strikingly, we observed a global increase of protein expression and phosphorylation of most of regulatory protein families in HGG tumors (P value <0.001). The majority of these proteins are expressed and phosphorylated even higher in NTRK HGG tumors when compared with PDGFRA HGGs. Indeed, most top DE genes show the expression pattern of NTRK > PDGFRA > cortex (**Figure 3-7b, c, d, e, f, g, and h**), including well-known master regulators (e.g. CHEK1, MAP3K1, PRKD1, INSR, and RB1) of HGG oncogenic pathways. Numerous other regulatory proteins (LYN, HMGB2, HMGA2, CD74, and CTNNB1) also fall into this pattern. Lyn (**Figure 3-7b**) is a SRC family tyrosine kinase that enhances Glut-4 translocation to the cell membrane to increase glucose uptake[163], a hallmark of cancer metabolism[164]. HMGB2 and HMGA2 (**Figure 3-7c, d**) are transcription and chromatin modulators that promote stemness and tumorigenicity in HGG[165]. CD74 (**Figure 3-7e**) is an attractive candidate target for immunotherapy that is present in limited amounts in normal tissues but high levels on a variety of hematological tumors[166,167]. CTNNB1 (**Figure 3-7f**) regulates cell adhesion and WNT signaling[168]. Thus, our results indicate an active role for TFs, epigenetic genes, kinases and cancer genes to reprogram signaling networks and maintain tumor homeostasis. Specifically, the NTRK genotype drives stronger global reprogramming than the PDGFRA genotype.

**Figure 3-7. Deep proteomic and phosphoproteomic profiling shows a global increase of protein expression and phosphorylation of most of regulatory protein families in HGG tumors.**

(a - h) Deep proteomic data analyses show a global increase of protein expression and phosphorylation of most of regulatory protein families (Kinase, epigenetic genes, transcription factors and cancer genes) in HGG tumors compare to cortex, with higher magnitude of increase in NTRK-driven HGG than PDGFRA-driven HGG. Scatter-histogram graphs of regulatory family proteins expression and phosphorylation comparing both HGG tumors to cortex and NTRK-driven HGGs to PDGFRA-driven HGGs. B.H. adjusted Student T test p values of 0.05 in both pairwise comparisons plus fold change (FC) cutoff of FC(NTRK/PDGFRA) * FC(HGG/cortex) > $1.5^2$ for proteome and >2 for phosphoproteome are applied for differential expression analyses. DE genes are shown in red. Magnitude of change is represented by dot size. Top five altered proteins and phosphoproteins with the largest magnitude of alterations are labeled. The distributions of pairwise differences are shown by histograms.


a
Kinases ( n = 414 )
( Whole protein level )
n = 16
n = 242
n = 15
n = 172
7%
16%
48%
29%
Chek1
Pbk
Prkcq
Stk17b
n = 17
n = 185
n = 14
n = 229
45%
43%
Camk4
0%
12%
HGG versus Cortex (Log2)
NTRK versus PDGFRA (Log2)
b
Kinases ( n = 264 )
( Phosphoprotein level )
n = 5
n = 114
n = 27
n = 150
6%
15%
85%
49%
Lyn
Map3k1
Melk
Insr
Prkd1
n = 29
n = 168
n = 3
n = 96
9%
28%
0%
8%
NTRK versus PDGFRA (Log2)
c
Epigenetic proteins ( n = 702 )
( Whole protein level )
n = 8
n = 154
n = 55
n = 548
0%
7%
86%
72%
Hmgb2
Chek1
Oip5
Hmgn5
Pou3f4
n = 54
n = 555
n = 9
n = 147
13%
15%
1%
6%
HGG versus Cortex (Log2)
NTRK versus PDGFRA (Log2)
d
Epigenetic proteins ( n = 422 )
( Phosphoprotein level )
n = 2
n = 82
n = 62
n = 340
2%
9%
94%
72%
Nasp
Hmgxb4
Rb1
Hmga2
Uba52
n = 62
n = 343
n = 2
n = 79
2%
20%
2%
9%
NTRK versus PDGFRA (Log2)
: All proteins in each regulatory protein family
: DE proteins significant in both comparisons

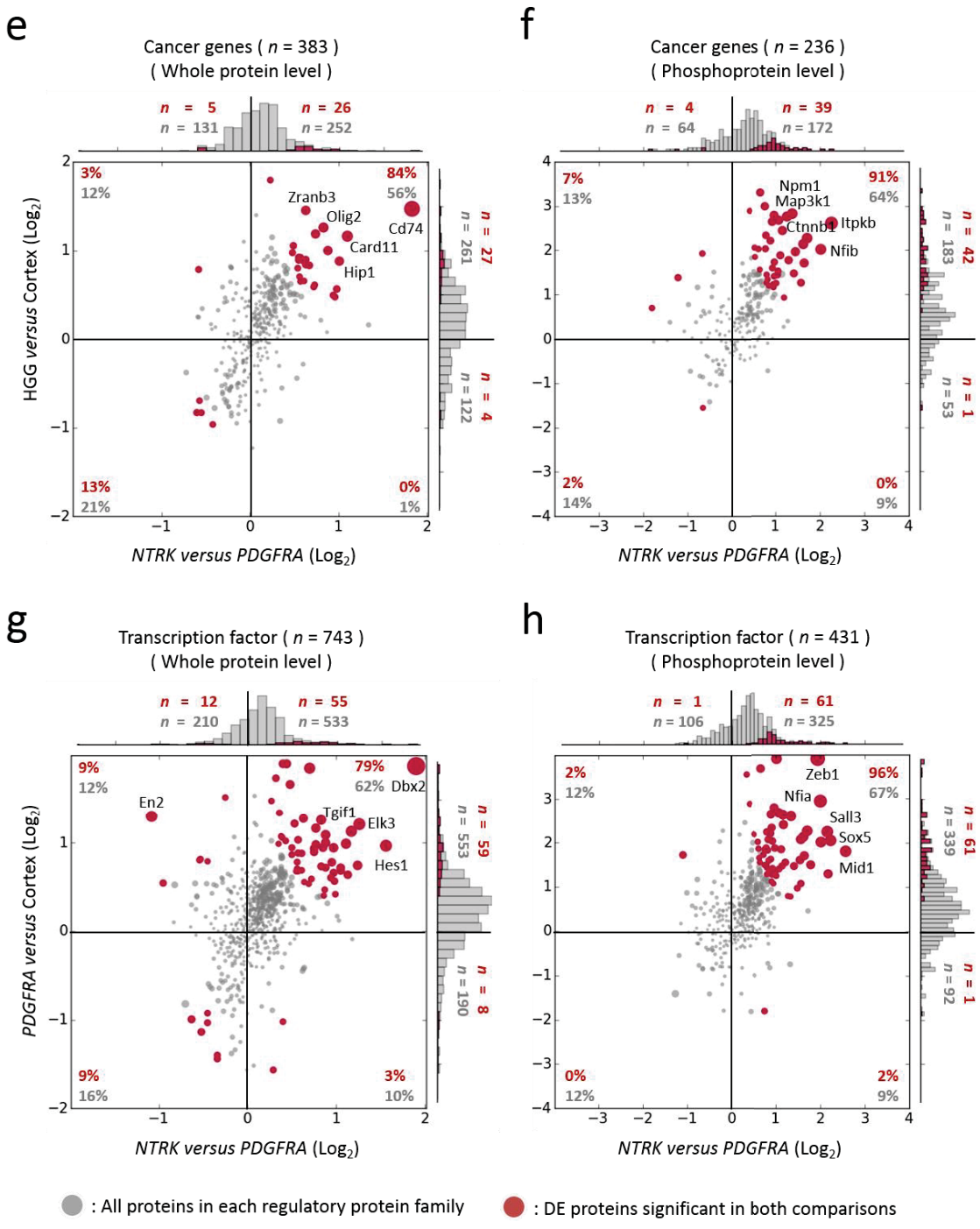


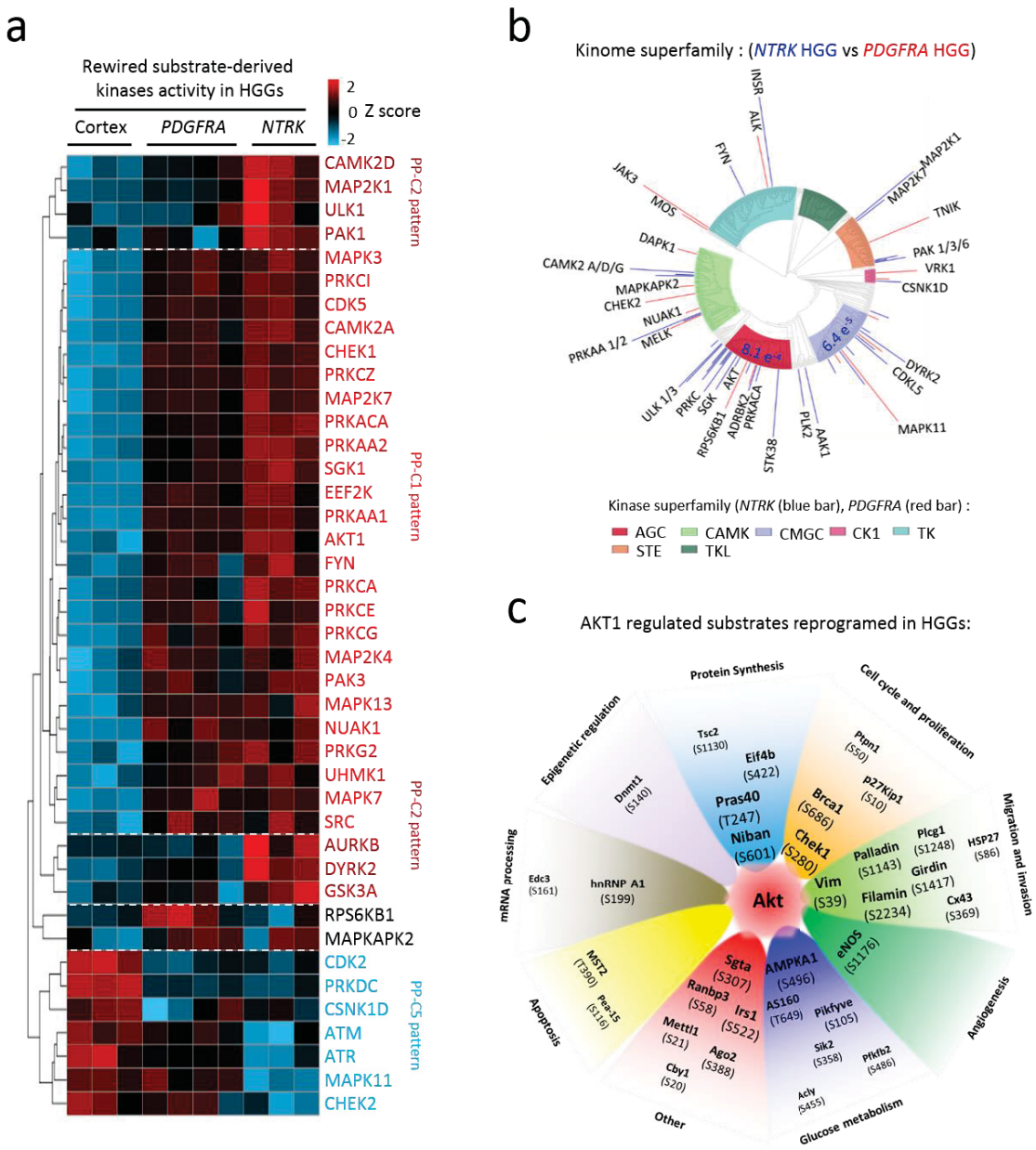

**Figure 3-7. Continued.**

**Multiple omics integration identifies master regulators (kinases and TFs) in the HGG models**

Since protein kinase activity can be inferred by substrate phosphorylation levels using computer programs, we used IKAP[157], a machine learning algorithm, to evaluate the activities of 187 kinases, 41 of which are reprogrammed in HGGs. Hierarchical clustering analysis classified these kinase activities into 3 major clusters (**Figure 3-8a**), all showing differential regulation among cortex, PDGFRA, and NTRK HGGs. Multiple known kinases in gliomagenesis are identified in HGGs, encompassing AKT, PKC, MAP Kinase cascade, and SRC family Kinases[39,169-171]. Other kinases regulating key intracellular systems are rewired as well, including AMPK (PRKAA1, PRKAA2) and p21-activated kinases (PAK1, PAK3). AMPK is a metabolic master sensor regulating glucose transporter GLUT4 production, fatty acid β-oxidation, and mitochondria biogenesis[172]. PAKs regulate cytoskeleton reorganization and cell motility[173]. HGGs also show higher levels of CDK5, CAMK2A, and CAMK2D, compared with normal cortex. Although these kinases are well-characterized regulators of neuronal function and synaptic plasticity, they are also expressed in glioblastoma, where they play roles in migration, invasion, mitochondrial regulation, and calcium signaling[174-176]. We further summarized the activities of these kinases at the level of kinase superfamilies. While AGC (cyclic nucleotide dependent family, protein kinase C family, ribosomal S6 family and related kinases), CMGC (cyclin-dependent kinases, mitogen-activated protein kinases, glycogen synthase kinases and cdk-like kinases) and CAMK (primarily kinases modulated by calcium/calmodulin) superfamily kinases are turned on significantly in HGG tumors (P value <0.001, **Figure 3-9**), NTRK HGGs display even higher activity in AGC and CMGC superfamilies than PDGFRA HGGs, supporting stronger cell proliferation signaling and cell cycle rewiring[177] (**Figure 3-8b**).

We constructed a kinase activation network by incorporating known kinase-to-kinase connections in the PhosphositePlus database, with consistent co-activation patterns in our datasets. This kinase activation network can be classified into 3 major functional groups (**Figure 3-8c**). Group 1 shows the co-activation of PKC, PKA, SRC, MAPK, and AKT, indicating strong and coordinated activation of the RTK-PI3K-AKT oncogenic pathway. Group 2 manifests co-activation of AMPKA1 and EEF2K, suggesting the rewiring of energy metabolism. In contrast, group 3 displays consistent attenuation of ATM, ATR, PRKDC, and CHEK2 activities, highlighting the inhibition of DNA repair, apoptosis, and cell cycle checkpoint functions in HGG tumors. Considering that AKT is the central node of this core HGG kinase network, we further analyzed the output of kinase activation on AKT substrates (**Figure 3-8d**). 34 AKT substrates (70% of DE substrates) show a phosphorylation pattern in agreement with AKT activity (**Figure 3-10**). The top activated AKT substrates are cell cycle and proliferation regulators (CHEK1 S280 and BRCA1 S686), central glucose metabolism regulators (e.g. AMPKA1 S496 and AS160 T649) and migration and angiogenesis regulators (e.g. eNOS S1176, VIM S39, and FLNC S2234, **Figure 3-8d**). Similar results were obtained for the co-regulation of other kinase-substrate connections (AMPKA1, CDK5, MAPK3, ATR, ATM, PAK1, and FYN, **Figure 3-11**). Collectively, our comprehensive kinase activity

**Figure 3-8. Deep phosphoproteome analysis reveals active kinases, kinase families and a central kinase-to-kinase network in HGG tumors.**
(a) Heatmap of hierarchical clustering of reprogrammed kinases with activity derived from substrate phosphorylation in HGG tumors. Kinase activity was inferred from substrate phosphorylation via a machine learning algorithm called IKAP. (b) Summarization of individual kinase activity into kinase superfamily shows stronger rewiring of AGC and CMGC superfamily in NTRK-driven HGG than PDGFRA-driven HGG. Kinome tree map shows pairwise kinase activity comparison inferred by phosphorylation of substrates. Chi-square P value of differentially phosphorylated kinase superfamilies are shown. Magnitudes of kinase activity difference are represented by length of bars located outside of the kinome tree circle. (c) Construction of core HGG kinase-to-kinase network. A kinase activation network was constructed by incorporating known kinase-to-kinase connections in the PhosphositePlus database, with consistent co-activation patterns in our datasets. The heat map keys show changes in protein expression (Whole), phosphorylation (Phospho.) and substrates-inferred activity (activity). Kinase activation network were constructed based on activity. This network can be further classified to three groups circled by colored ovals. Red and blue ovals indicate groups that are activated and inhibited in HGG respectively. (d) Substrates activated by AKT, the central hub of kinase-to-kinase network in HGGs. Figure shows AKT regulated substrates and their corresponding phosphosites that are organized according to their annotated functions. The magnitude of change is represented by substrate font size.


a
Rewired substrate-derived kinases activity in HGGs
2
0 Z score
-2
Cortex
PDGFRA
NTRK
CAMK2D
MAP2K1
ULK1
PAK1
MAPK3
PRKCI
CDK5
CAMK2A
CHEK1
PRKCZ
MAP2K7
PRKACA
PRKAA2
SGK1
EEF2K
PRKAA1
AKT1
FYN
PRKCA
PRKCE
PRKCG
MAP2K4
PAK3
MAPK13
NUAK1
PRKG2
UHMK1
MAPK7
SRC
AURKB
DYRK2
GSK3A
RPS6KB1
MAPKAPK2
CDK2
PRKDC
CSNK1D
ATM
ATR
MAPK11
CHEK2
PP-C2 pattern
PP-C1 pattern
PP-C2 pattern
PP-C5 pattern
b
Kinome superfamily : (NTRK HGG vs PDGFRA HGG)
INSR
ALK
FYN
JAK3
MOS
DAPK1
CAMK2 A/D/G
MAPKAPK2
CHEK2
NUAK1
PRKAA 1/2
MELK
ULK 1/3
PRKC
SGK
AKT
RPS6KB1
ADRBK2
PRKACA
STK38
PLK2
AAK1
MAPK11
CDKL5
DYRK2
CSNK1D
VRK1
PAK 1/3/6
TNIK
MAP2K
MAP2K1
8.1 e-4
6.4 e-5
Kinase superfamily (NTRK (blue bar), PDGFRA (red bar) :
AGC
CAMK
CMGC
CK1
TK
STE
TKL
c
AKT1 regulated substrates reprogramed in HGGs:
Protein Synthesis
Cell cycle and proliferation
Epigenetic regulation
Migration and invasion
mRNA processing
Apoptosis
Angiogenesis
Other
Glucose metabolism
Akt
Tsc2 (S1130)
Eif4b (S422)
Pras40 (T247)
Niban (S601)
Dnmt1 (S140)
Ptpn1 (S50)
p27Kip1 (S10)
Brca1 (S686)
Chek1 (S280)
Plcg1 (S1248)
HSP27 (S86)
Palladin (S1143)
Girdin (S1417)
Filamin (S2234)
Cx43 (S369)
Vim (S39)
eNOS (S1176)
Edc3 (S161)
hnRNP A1 (S199)
MST2 (T390)
Pea-15 (S116)
Sgta (S307)
Ranbp3 (S58)
Irs1 (S522)
Mettl1 (S21)
Ago2 (S388)
Cby1 (S20)
AMPKA1 (S496)
AS160 (T649)
Pikfyve (S105)
Sik2 (S358)
Pfkfb2 (S486)
Acly (S455)

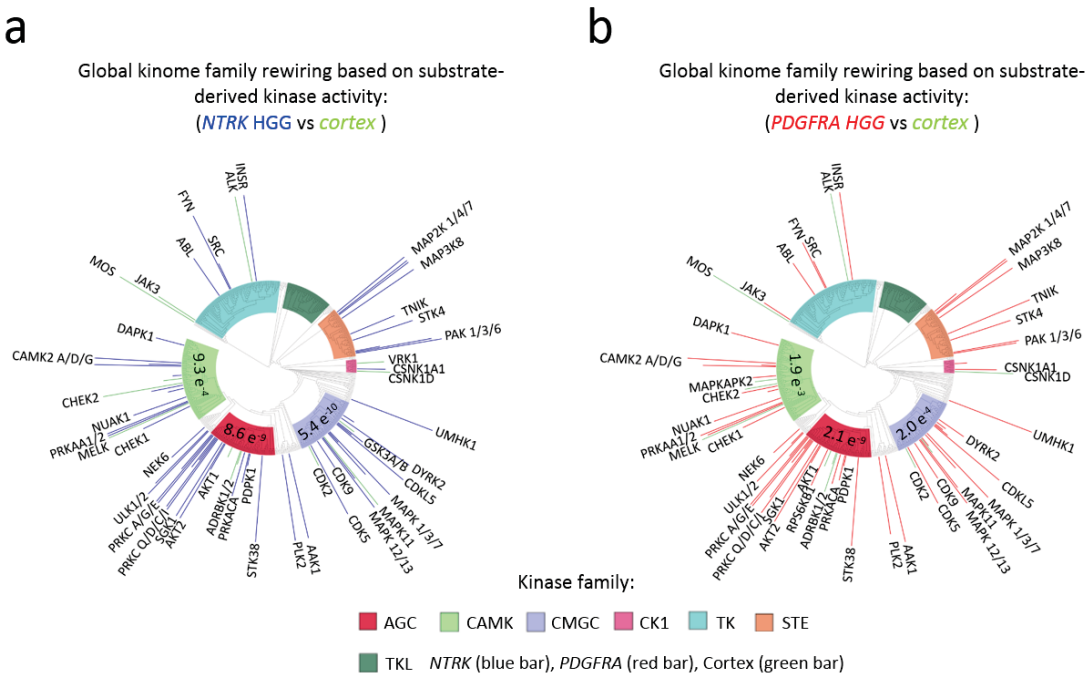



**Figure 3-9. AGC, CAMK and CMGC kinase superfamilies display higher activity in both HGG tumors compare to cortex.**

Kinome tree maps show kinase activity comparing NTRK-driven HGG to cortex and PDGFRA-driven HGG to cortex respectively. Pairwise comparisons of Kinase superfamilies with a Chi-square P values < 0.05 are labelled. Magnitude of kinase activity difference is represented by length of bars located outside of the kinome tree circles.

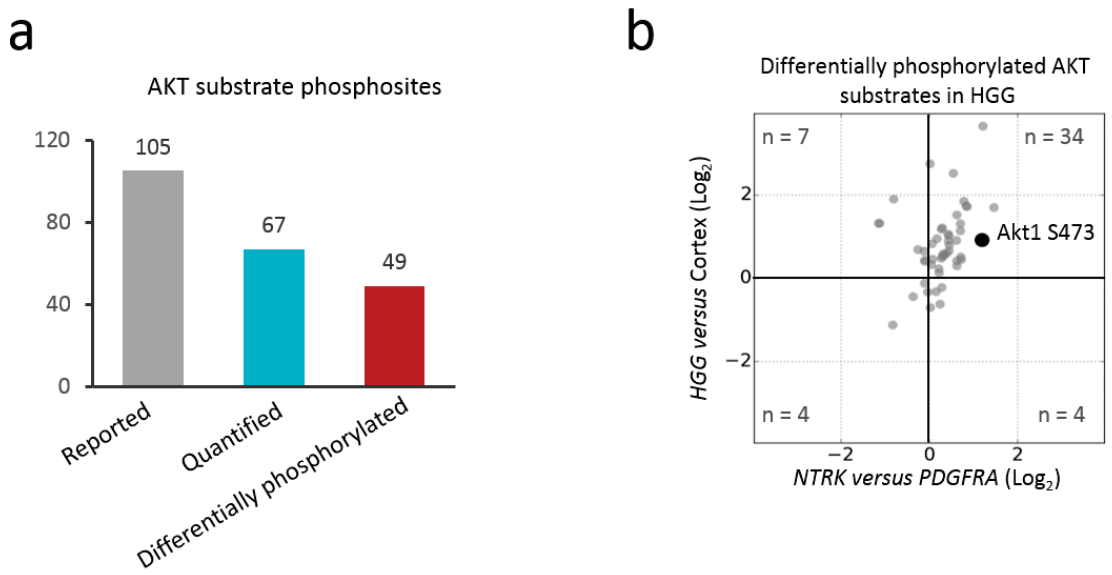



**Figure 3-10. Evaluation of AKT regulated substrates.**
(a) A high percentage of known AKT substrate sites were identified in the deep phosphoproteome. Bar graph shows reported substrate sites, quantified phosphosites and differentially phosphorylated sites in our data. (b) Differentially phosphorylated substrates that are regulated by AKT in HGGs. Scatter plot of log2 level changes compare HGG tumors to cortex and compare NTRK-driven HGG to PDGFRA-driven HGG. Substrates with the same phosphorylation pattern as the AKT active site (S473, located at upper right) were accepted as AKT regulated substrates.

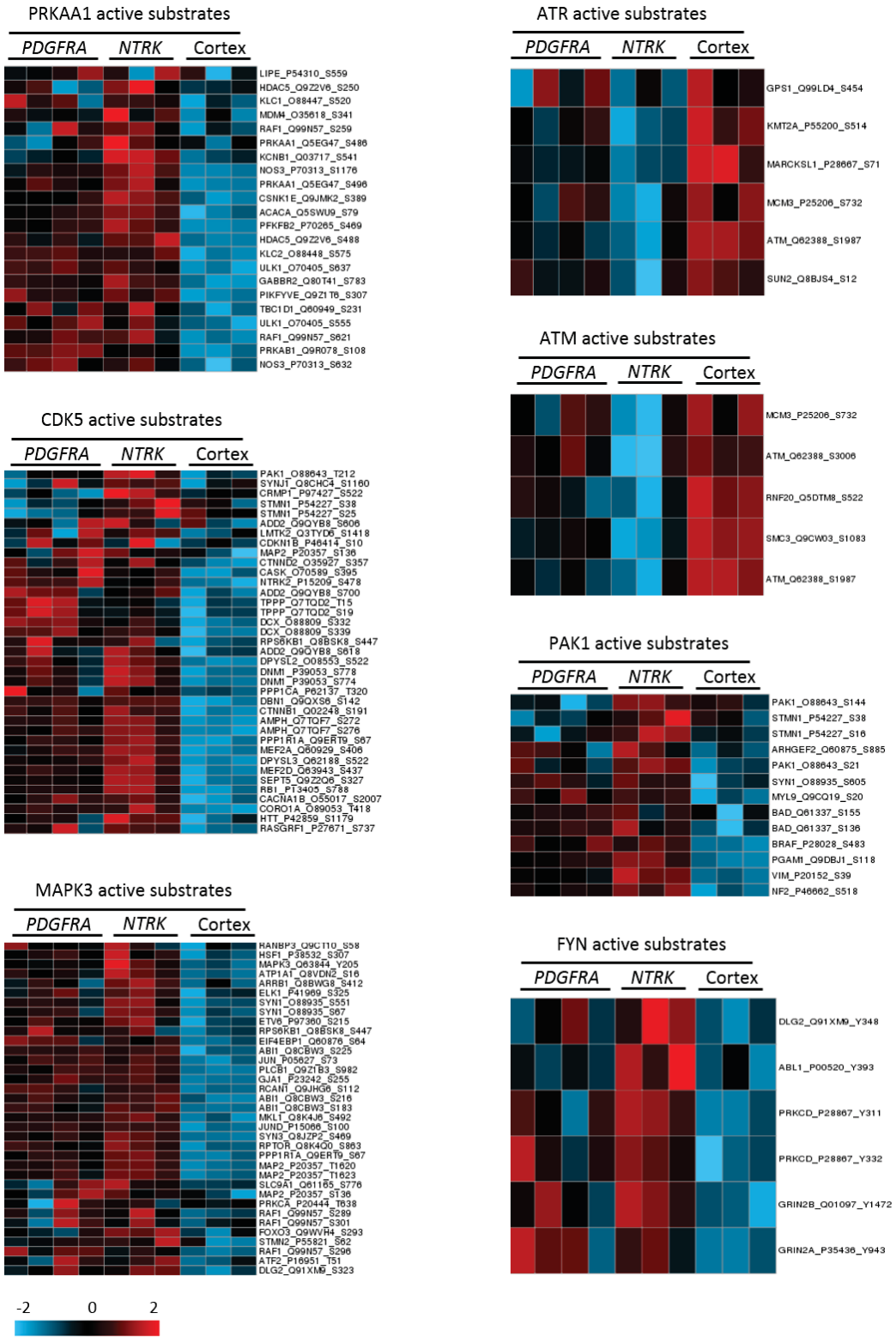



**Figure 3-11. Heatmaps display differentially phosphorylated substrates (with up-regulated phosphorylation in HGG tumors) of other active kinases derived from kinase-substrate analysis.**

analysis enables the identification of master kinases and central kinase activation networks in HGG tumors.

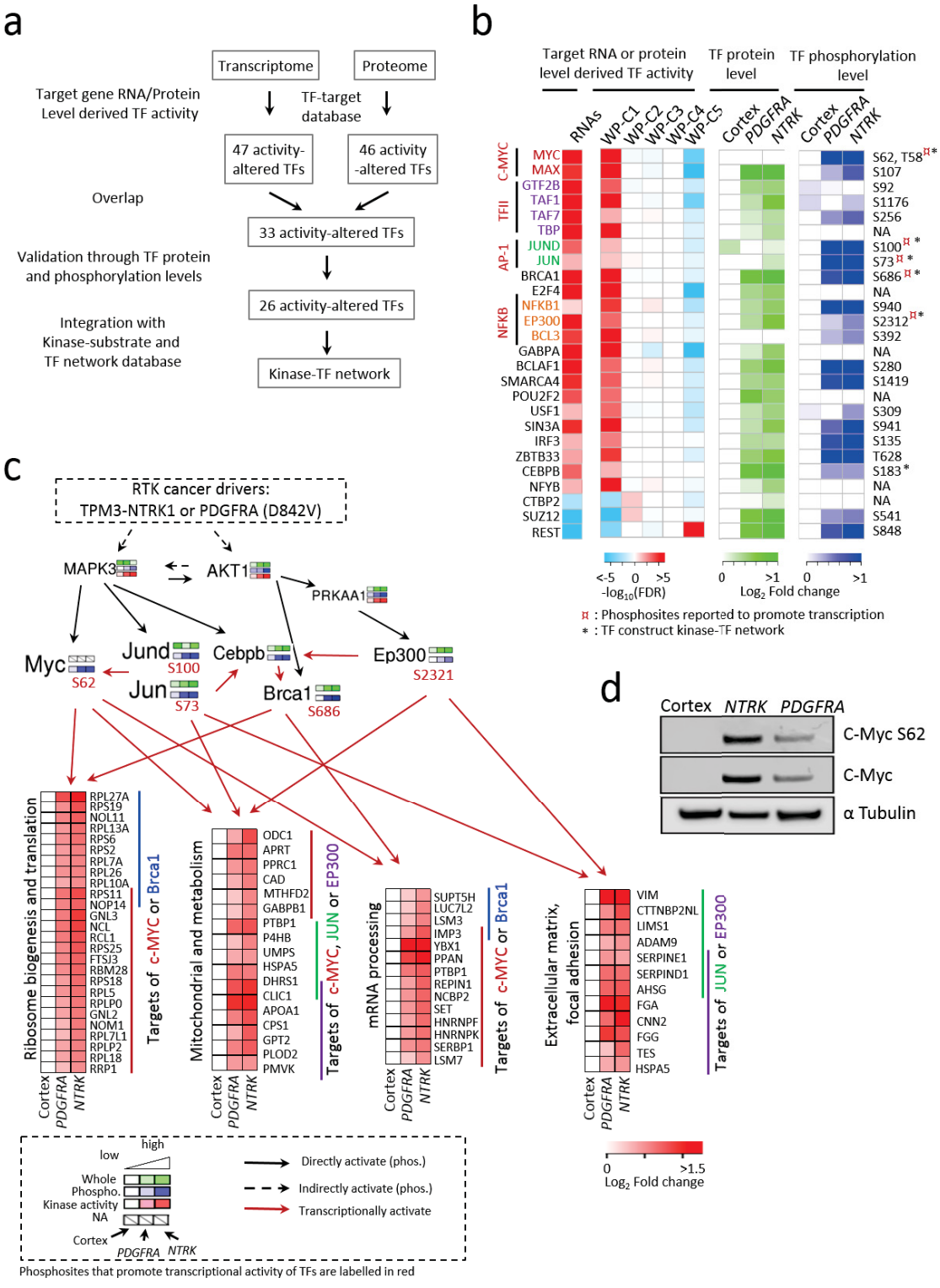
We also explored the activity of TFs through integrative analysis of transcriptome, proteome and phosphoproteome via a systems biology approach in multiple steps (**Figure 3-12a**). (i) TF activities were derived from target gene expression in either transcriptome or clustered proteome (WP-Cs), resulting in two lists of 47 and 46 TFs. (ii) Factors common to both lists showed consistent changes of 33 TFs. (iii) Additional data from whole proteome or phosphoproteome supported the activation of 26 out of 33 TFs (**Figure 3-12b**). For instance, 5 TFs show active status based upon the increase of phosphorylation at the activation sites (c-MYC S62, JUND S100, JUN S73, BRCA1 S686, and EP300 S2312). Among the most activated TFs, were the TFII family (GTF2B, TAF1, TAF7, and TBP) of general transcription factors, which assemble the RNA polymerase II pre-initiation complex and control general transcription rate; and the transcription suppressor REST, a chromatin modifier in brain[178]. Consistently, REST target gene expression was low in the tumors and high in normal cortex (**Figure 3-12b**, WP-C1, WP-C5), implicating a possible role of REST in HGG tumor transformation through suppression of target gene expression. Similar to REST, we also found the up-regulation of SUZ12, a polycomb repressive complex 2 (PRC2) component associated with silent chromatin[179] and CTBP2, a repressor that recruits histone deacetylases and methylases to target genes[180]. Thus, this integrative analysis reveals the activation of both TF activators and suppressors, which lead to distinct reprogramming of tumor cell transcriptome and proteome.

Finally, we developed a kinase-TF network by linking the 41 deregulated kinases with 26 TFs and gene targets (**Figure 3-12c**). This core network consists of TFs (c-MYC, JUND, JUN, EP300, BRCA1, and CEBPB) and kinases (e.g. AKT, MAPK, AMPK and CDK5). The c-MYC family (MYC, MAX) and AP-1 family (JUN, JUND) regulate a variety of central biological processes in tumorigenesis. Indeed, more than 100 c-MYC targets were transcriptionally active, strongly supporting the central role of c-MYC in HGG tumors. It is likely that c-MYC is activated by transcriptional up-regulation by JUN and/or JUND, as well as phosphorylation by MAPK3 and/or CDK5 (**Figure 3-12c**). We further validated c-MYC protein expression and phosphorylation in the HGG tumors by immunoblot assays (**Figure 3-12d**).

Close examination of 5 master TFs (c-MYC, JUND, JUN, EP300, and BRCA1) uncovered major target genes in HGGs (**Figure 3-12c**). The most transcriptionally activated biological processes are related to proliferation including ribosome biogenesis, translation, and mRNA processing (targets of c-MYC and BRCA1), energy metabolism including mitochondrial and metabolism (targets of c-MYC, JUN and EP300), and cell migration including extracellular matrix and focal adhesion (targets of JUN and EP300). In the kinase-TF network, c-MYC, JUN and EP300 are activated via phosphorylation by master energy and proliferation sensors kinases (AMPK and MAPK). Notably, multiple metabolic enzymes activated by these 3 TFs are rate-limiting during proliferation and energy stress, such as adenine phosphoribosyltransferase (APRT), regulating a nucleotide salvage pathway to synthesize purines *de novo*[181], and ornithine decarboxylase (ODC1)

**Figure 3-12. Integration of multiple deep omics data enables identification of active TFs and construction of a core kinase to transcriptional regulation network in HGGs.**

(a) Overview of the integrative analysis strategy for TF activity inference and construction of kinase to transcriptional regulation network by incorporating transcriptome, whole proteome, phosphoproteome, kinase-substrate database and TF-target database. (b) Active TFs in HGGs identified by integrative analysis. TF activities are indicated by the B.H. adjusted FDR values derived from either differential expressed target mRNAs or differential expressed target proteins in whole proteome clusters. MS-based quantification of protein expression and phosphorylation of TFs are also shown. (c) Integrative analysis reveals a putative core signaling network encompassing active kinases, active TFs, and transcriptionally activated target genes. Black arrows represent activation through phosphorylation. Red arrows represent TFs transcriptionally activate targets gene expression. Representative target genes that are transcriptionally activated are organized according to their functions. (d) Immunoblot assay on c-MYC validates the overexpression and activation of c-MYC. Western blotting was performed on c-Myc protein expression and c-Myc S62 phosphorylation, □-tubulin is shown as a loading control


a
Transcriptome
Proteome
Target gene RNA/Protein Level derived TF activity
TF-target database
47 activity-altered TFs
46 activity -altered TFs
Overlap
33 activity-altered TFs
Validation through TF protein and phosphorylation levels
26 activity-altered TFs
Integration with Kinase-substrate and TF network database
Kinase-TF network
b
Target RNA or protein level derived TF activity
TF protein level
TF phosphorylation level
RNAs
WP-C1
WP-C2
WP-C3
WP-C4
WP-C5
Cortex
PDGFRA
NTRK
C-MYC
TFII
AP-1
NFKB
MYC
MAX
GTF2B
TAF1
TAF7
TBP
JUND
JUN
BRCA1
E2F4
NFKB1
EP300
BCL3
GABPA
BCLAF1
SMARCA4
POU2F2
USF1
SIN3A
IRF3
ZBTB33
CEBPB
NFYB
CTBP2
SUZ12
REST
S62, T58¤*
S107
S92
S1176
S256
NA
S100¤*
S73¤*
S686¤*
NA
S940
S2312¤*
S392
NA
S280
S1419
NA
S309
S941
S135
T628
S183*
NA
NA
S541
S848
<-5 0 >5
-log10(FDR)
0 >1
Log2 Fold change
0 >1
¤ : Phosphosites reported to promote transcription
* : TF construct kinase-TF network
c
RTK cancer drivers:
TPM3-NTRK1 or PDGFRA (D842V)
MAPK3
AKT1
PRKAA1
Myc
S62
Jund
S100
Jun
S73
Cebpb
Brca1
S686
Ep300
S2321
Ribosome biogenesis and translation
RPL27A
RPS19
NOL11
RPL13A
RPS6
RPS2
RPL7A
RPL26
RPL10A
RPS11
NOP14
GNL3
NCL
RCL1
RPS25
FTSJ3
RBM28
RPS18
RPL5
RPLP0
GNL2
NOM1
RPL7L1
RPLP2
RPL18
RRP1
Targets of c-MYC or Brca1
Mitochondrial and metabolism
ODC1
APRT
PPRC1
CAD
MTHFD2
GABPB1
PTBP1
P4HB
UMPS
HSPA5
DHRS1
CLIC1
APOA1
CPS1
GPT2
PLOD2
PMVK
Targets of c-MYC, JUN or EP300
mRNA processing
SUPT5H
LUC7L2
LSM3
IMP3
YBX1
PPAN
PTBP1
REPIN1
NCBP2
SET
HNRNPF
HNRNPK
SERBP1
LSM7
Targets of c-MYC or Brca1
Extracellular matrix, focal adhesion
VIM
CTTNBP2NL
LIMS1
ADAM9
SERPINE1
SERPIND1
AHSG
FGA
CNN2
FGG
TES
HSPA5
Targets of JUN or EP300
low
high
Whole Phospho.
Kinase activity
NA
Directly activate (phos.)
Indirectly activate (phos.)
Transcriptionally activate
0 >1.5
Log2 Fold change
Phosphosites that promote transcriptional activity of TFs are labelled in red
d
Cortex
NTRK
PDGFRA
C-Myc S62
C-Myc
α Tubulin

for polyamine biosynthesis in response to growth stimulation. In summary, our systems biology approaches utilize multi-layer information to prioritize central HGG TFs, kinases and their interplay in HGG tumors.

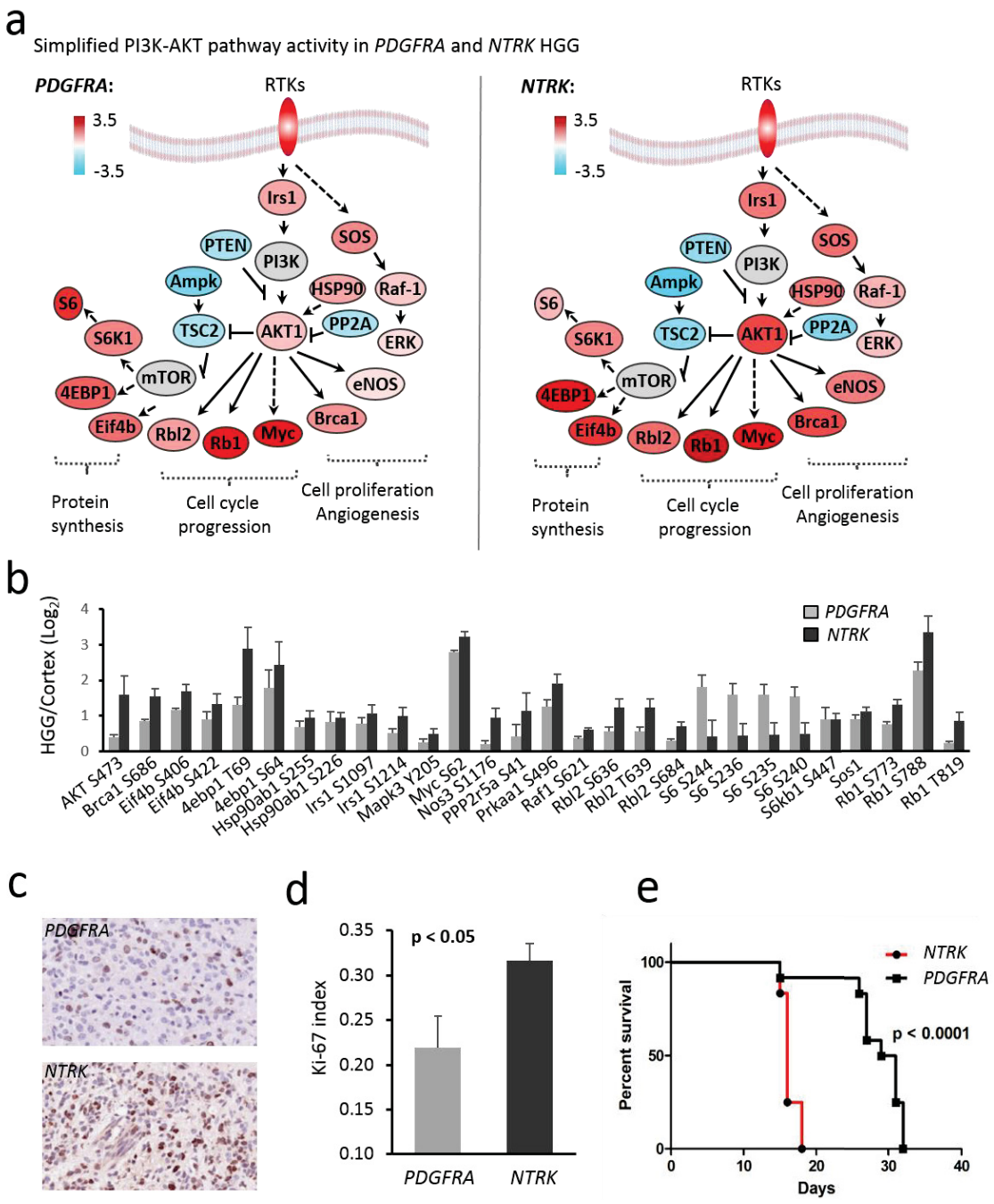
**NTRK HGG display stronger PI3K-AKT signaling activity, higher proliferation index and shorter latency than PDGFRA HGG**

The bioinformatics analysis suggests stronger global cancer network rewiring in NTRK HGG than PDGFRA HGG, indicating higher oncogenic potency of NTRK than PDGFRA mutations. To evaluate the oncogenic potency of the RTK cancer driver genes, we modified the PAC algorithm that was initially designed for gene expression analysis[159], to compute the summed PI3K-AKT signaling activity. The protein activity was derived from the levels of its phosphosites with known functions, which either promote or inhibit tumorigenesis (see Methods). In both HGGs, the PI3K-AKT pathway was clearly active and invoked similar downstream pathways, such as protein synthesis (S6, 4EBP1 and EIF4B), cell cycle progression (RB1, MYC and RBl2), cell proliferation and angiogenesis (BRCA1, eNOS, ERK) (**Figure 3-13a**). When comparing 27 regulatory phosphosites of these proteins that were statistically different between NTRK and PDGFRA HGGs, the majority ($n$ = 23) showed higher alteration in NTRK HGG than PDGFRA HGG (**Figure 3-13b**). Consistently, the NTRK HGG exhibited 1.45-fold greater PI3K-AKT signaling activity (P value < 0.05), suggesting that the TPM3-NTRK1 fusion gene harbors stronger oncogenic potency than the PDGFRA D842V.

To experimentally validate our predicted oncogenic potency of TPM3-NTRK1 and PDGFRA D842V, we analyzed cellular proliferative indexes and Kaplan-Meier survival curves of both HGG mice. The proliferative index was defined by the proportion of tumor cells that expressed the proliferation marker Ki67 (**Figure 3-13c**). Consistent with enhanced level of PI3K-AKT signaling, the proliferative index of NTRK HGG (0.32 ± 0.04) was 1.4 fold higher than that of PDGFRA HGG (0.22 ± 0.07, **Figure 3-13d**). NTRK HGG mice developed with much shorter latency than PDGFRA mice (median survival time of 16 days and 30 days, respectively, **Figure 3-13e**).

TPM3-NTRK1 and PDGFRA D842V both activated PI3K-AKT signaling, but with different potency. Strikingly, MS measurement showed higher PDGFRA protein expression in NTRK HGG than PDGFRA HGG. This was validated by Western blotting (**Figure 3-14a**). To distinguish human PDGFRA D842V oncogene product from mouse PDGFRA, we quantified endogenous mouse PDGFRA peptides and found that the TPM3-NTRK1 induced dramatic overexpression of mouse PDGFRA (**Figure 3-14b**). Many other RTKs (EphA2, Egfr, Flt4, Ptk7 and Ror2) also showed higher expression in NTRK HGG than PDGFRA HGG (**Figure 3-14c**). Transcriptomic measurement consistently indicated the up-regulation of these RTKs (**Figure 3-15**). Western blotting further confirmed EphA2 overexpression and activation reflected by concomitant phosphorylation (**Figure 3-14d**). To identify the mechanism driving increased RTK expression, we analyzed TF activities that promote RTK transcription according to the MSigDB database[182]. The TF activities were estimated by phosphorylation of active sites

**Figure 3-13. NTRK-driven HGG displays stronger PI3K-AKT signaling activity, higher cell proliferation index and shorter mice tumor onset latency than PDGFRA-driven HGG.**
(a) PI3K-AKT pathway is active in both PDGFRA-driven HGG and NTRK-driven HGG. Simplified PI3K-AKT pathway diagram shows activity change according to phosphorylation (Protein expression change of PTEN is included). Color scale represents difference between HGGs and cortex. (b) MS-based measurements of phosphorylation on reported activation or inhibition sites show that NTRK-driven HGGs in general have stronger PI3K-AKT pathway dysregulation than PDGFRA-driven HGGs. (c) Representative Ki-67 IHC-stained sections on PDGFRA- and NTRK-driven HGGs to examine the proliferative indexes of HGG tumors. (d) Bar plots of proliferation indexes according to Ki-67 IHC demonstrate more rapid tumor growth in NTRK-driven HGG ($n$ = 4) than PDGFRA-driven HGG ($n$ = 5, $p < 0.05$), consistent with higher oncogenic potency in NTRK-driven HGG derived from PI3K-AKT pathway activity computation. (e) K-M curve confirms more rapid mice tumor onset of NTRK-driven HGGs ($n$ = 12) compared to PDGFRA-driven HGGs ($n$ = 12).


a
Simplified PI3K-AKT pathway activity in PDGFRA and NTRK HGG
PDGFRA:
NTRK:
RTKs
3.5
-3.5
Irs1
SOS
PTEN
PI3K
Ampk
HSP90
Raf-1
S6
TSC2
AKT1
PP2A
ERK
S6K1
mTOR
eNOS
4EBP1
Eif4b
Rbl2
Rb1
Myc
Brca1
Protein synthesis
Cell cycle progression
Cell proliferation Angiogenesis
b
HGG/Cortex (Log2)
PDGFRA
NTRK
AKT S473
Brca1 S686
Eif4b S406
Eif4b S422
4ebp1 T69
4ebp1 S64
Hsp90ab1 S255
Hsp90ab1 S226
Irs1 S1097
Irs1 S1214
Mapk3 Y205
Myc S62
Nos3 S1176
PPP2r5a S41
Prkaa1 S496
Raf1 S621
Rbl2 S636
Rbl2 T639
Rbl2 S684
S6 S244
S6 S236
S6 S235
S6 S240
S6kb1 S447
Sos1
Rb1 S773
Rb1 S788
Rb1 T819
c
PDGFRA
NTRK
d
Ki-67 index
p < 0.05
PDGFRA
NTRK
e
Percent survival
NTRK
PDGFRA
p < 0.0001
Days

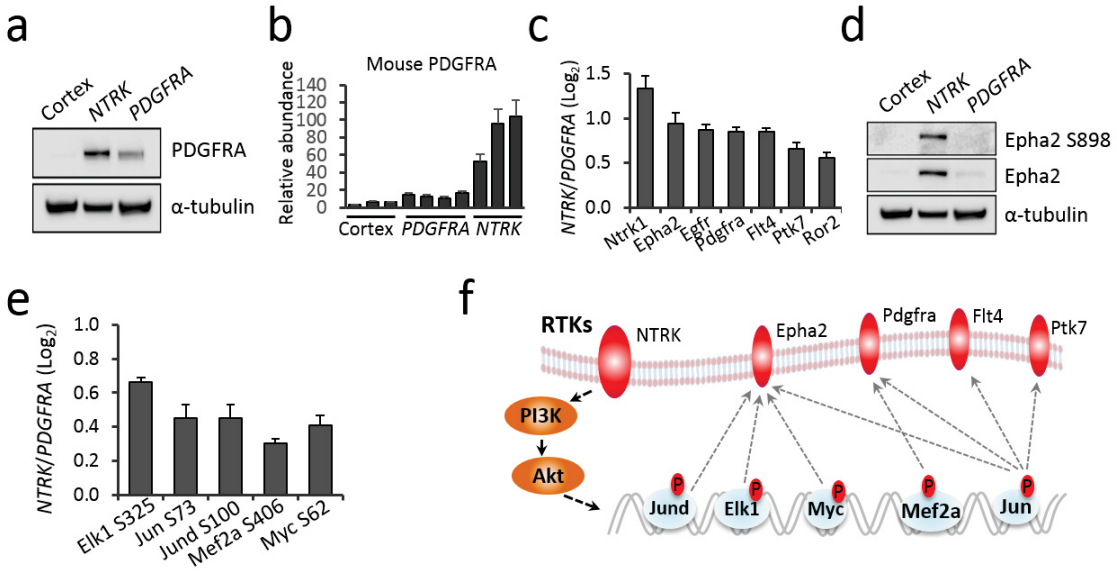



**Figure 3-14. NTRK fusion gene induces an enhanced overexpression and activation of other RTKs, suggesting a forward feedback loop within PI3K-AKT signaling.**

(a) Immunoblot assay on total PDGFRA expression is consistent with MS measurement. α-tubulin is included as a loading control. (b) NTRK fusion gene induces overexpression of endogenous mouse PDGFRA in NTRK-driven HGG. Mouse specific PDGFRA peptides were extracted and applied for quantification of endogenous PDGFRA expression. (c) NTRK fusion gene induces up regulation of multiple other RTKs in NTRK-driven HGG. Other RTKs that show differential expression between NTRK- and PDGFRA-driven HGGs are shown in bar plot. (d) Immunoblot assay on EPHA2 protein expression and phosphorylation (S898) demonstrates its overexpression and activation in NTRK-driven HGG. α-tubulin is included as a loading control. (e) TFs that promote the expression of the differentially expressed RTKs are also more active in NTRK- than PDGFRA-driven HGGs. Bar plot shows MS measurements of phosphorylation on activation sites of the TFs that promote the expression of differentially expressed RTKs. (f). Model shows NTRK fusion induces overexpression of other RTKs to form a positive feedback loop within PI3K-AKT pathway.

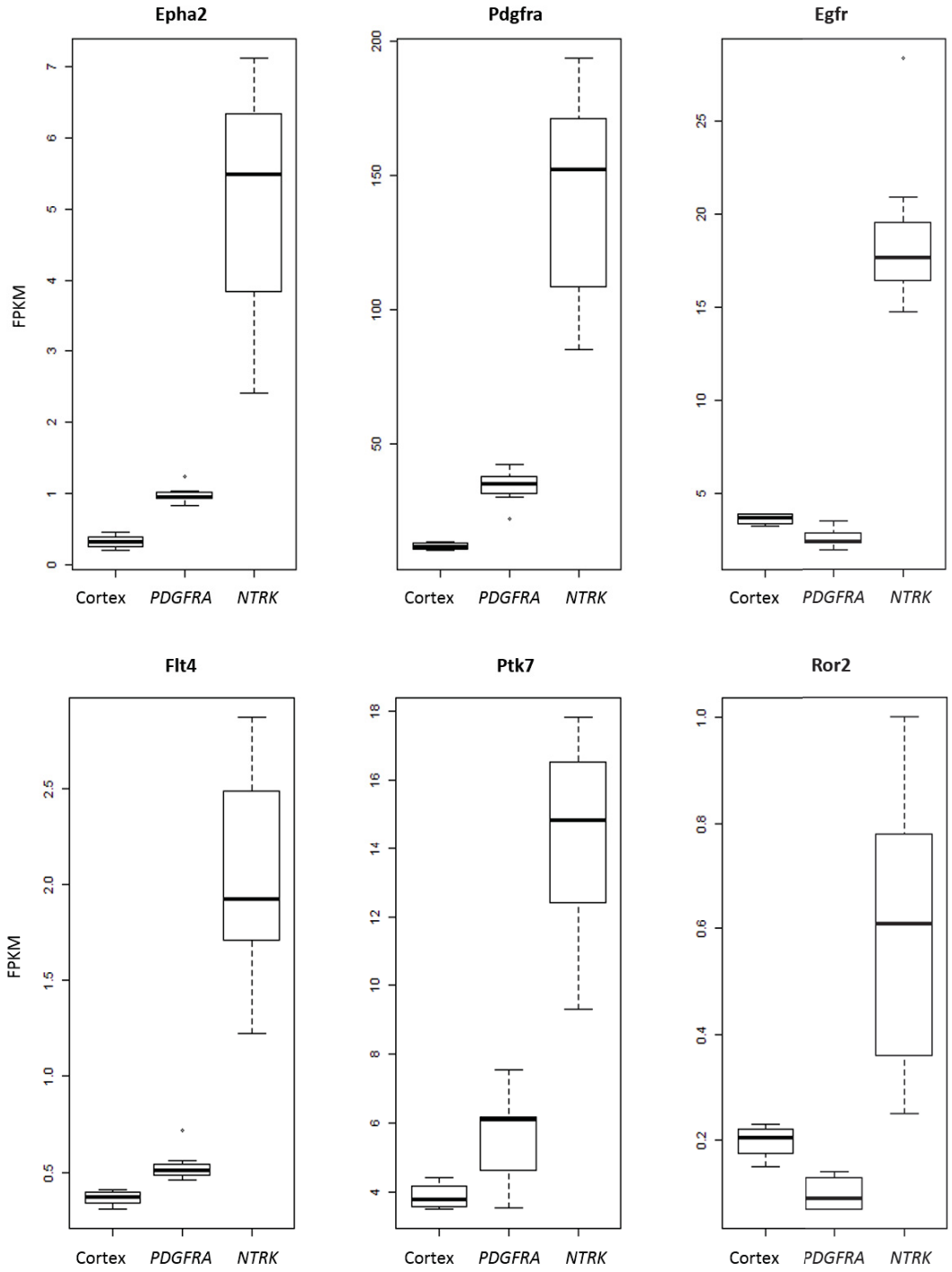



**Figure 3-15. NTRK fusion induces up-regulation of other RTKs at transcriptome level.**
Boxplots show corresponding transcripts expression of differentially expressed RTK proteins in cortex, PDGFRA- and NTRK-driven HGGs

(**Figure 3-14e**) and the target gene expression (**Figure 3-12**). Together, these bioinformatics and experimental findings demonstrate that the NTRK fusion gene induced an enhanced overexpression and activation of other RTKs, suggesting a forward feedback loop within PI3K-AKT signaling, resulting in a more aggressive tumor than PDGFRA-driven HGG (**Figure 3-14f**).

### Combination of mouse and human HGG data prioritizes putative cancer genes

As mouse modeling of human cancer is an effective avenue to define gene alterations for cancer initiation and progression[183], we used a cross species approach to integrate multi-omics mouse data with human cancer genomics data to explore cancer genes (**Figure 3-16a**). We first identified cancer driver responsive gene products ($n$ = 138) that are consistent across multi-omics data in mice HGG, and differentially expressed genes in cases with NTRK fusions compared to cases with PDGFRA mutations (transcriptome, $n$ = 375) in human pediatric HGG. The overlapping genes with consistent changes were accepted as putative HGG cancer driver responsive genes ($n$ = 20, **Figure 3-16b**). The majority of these genes were reported to function in cancer-related processes, including the regulation of cancer cell stemness, angiogenesis, tumor microenvironment, and invasion. For example, EPHA2 regulates cancer stem-like properties, drives self-renewal[184,185], mediates ligand-independent promotion of cell migration and invasion in human HGG[186] (**Figure 3-16c**). The expression of CD74 is associated with enhanced proliferation and invasion of multiple tumors, as well as patient survival and microglial response in glioblastoma[187] (**Figure 3-16d**). This analysis underlines the strength of inter-species analysis to prioritize a core subset of cancer-relevant candidates from massive multi-omics datasets.

## Discussion

As mRNA level is often moderately correlated with protein level[35], there is a need to profile both the transcriptome and proteome to obtain a full picture of gene expression in cancer biology. Here we demonstrate the power of deep proteomics coverage and integration of multi-omics datasets to probe molecular mechanisms underlying tumorigenicity. Recent developments of optimized long gradient LC-MS/MS system[148], refined phosphopeptide enrichment[151], and advanced bioinformatics tools[149,150] greatly improve the depth of proteomics profiling for cancer studies, detecting almost all of the expressed proteins. Such a high coverage allows the systematic analysis of proteins of low abundance, as exemplified by transcription factors and kinases. In parallel, a comprehensive phosphoproteome analysis offers complementary information about pathway/network activities, because many components in pathways are not changed at the protein level, but altered in phosphorylation states during signaling transduction. Although some known phosphosites are missed due to intrinsic limitations of the shotgun proteomics approach[162], we have detected nearly all NTRK and PDGFRA regulating pathways in the KEGG database in the deep phosphoproteome dataset.

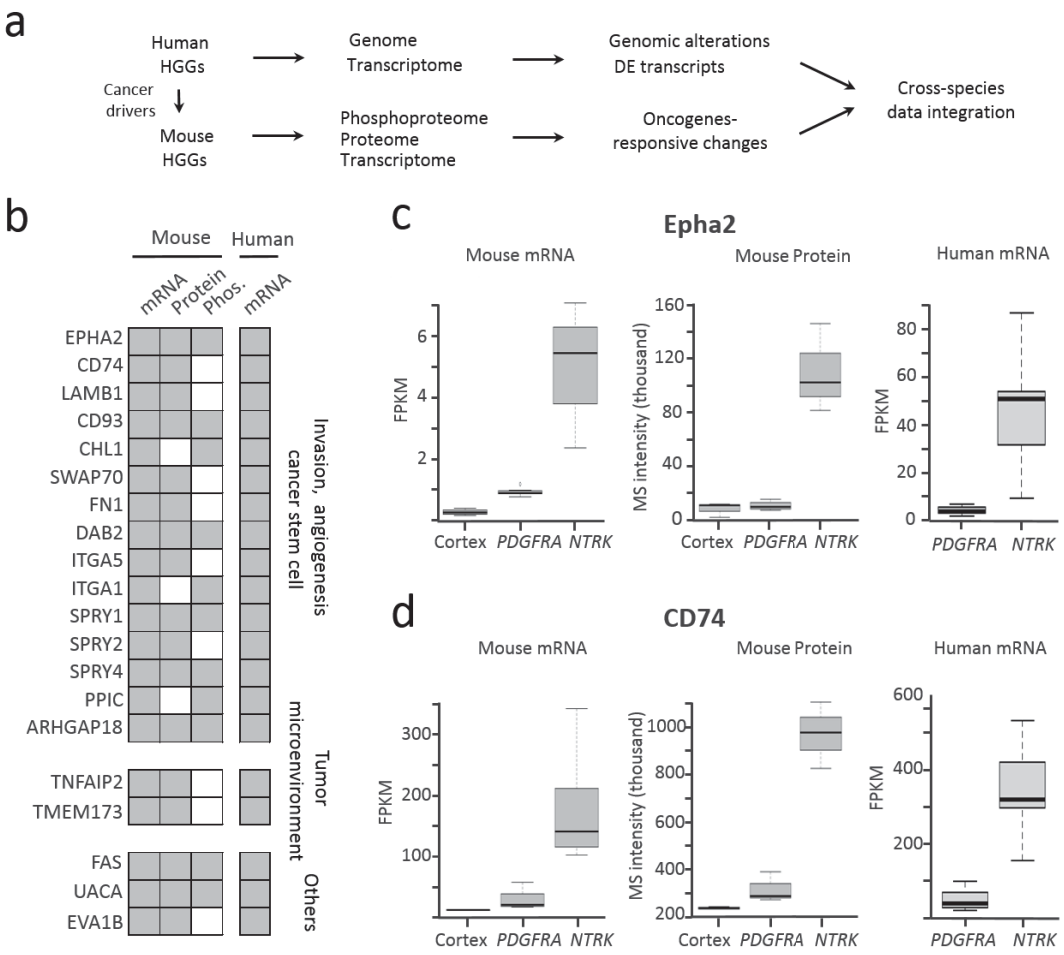



**Figure 3-16. Combination of mouse and human HGG data prioritizes putative cancer genes.**

(a) Overview of multi-omics analysis across species. We used RTK mutations identified in human HGGs to develop the mouse HGG models under study. Cancer driver responsive changes detected in multiple omics data in mouse were incorporated with human omics data to prioritize conserved cancer genes. (b) Cancer driver responsive changes identified through cross species integrative analysis are well-reported cancer genes in HGG. Genes that show consistent changes across mouse and human omics data were classified according to their cancer relevant functions. Mouse transcripts expression that follow the expression pattern of NTRK > PDGFRA > Cortex and the fold difference between NTRK HGG and PDGFRA HGG were larger than or equal to 2 and P value <0.05 with consistent whole protein level or phosphoprotein level changes were accepted as oncogene responsive changes in mouse. Human transcripts expression that show NTRK > PDGFRA with fold change larger than or equal to 2 and P value <0.05 were intersected with oncogene responsive changes identified in mouse to prioritize cross species conserved changes in human.(c, d) Boxplots show Epha2 and CD74 expression in multi-omics data in human and mouse. Human mRNA boxplots compare expression levels of EPHA2 or CD74 between pediatric HGGs with mutated PDGFRA and HGGs with NTRK fusion genes.

Isobaric labelling (e.g. TMT, and iTRAQ) is a powerful quantitative strategy for multiplexed deep proteomics profiling with high throughput and reproducibility[22,29] Although quantitative ratio compression often occurs with this method[22,188,189], it also reduces experimental variations, and therefore has almost no impact on differential expression analysis after scale normalization[22,29] (**Figure 3-17**). Moreover, our strategies of extensive peptide separation[148] with biological replicates facilitate statistical inference and largely reduce the effect of ratio compression.

Recurrent mutations in the RTK/RAS/PI3K signaling axis occur frequently in virtually all adult glioblastomas, more than half of pediatric glioblastomas, and diverse other tumor types[39,43,45]. While this implies that the PI3K pathway is an important therapeutic target, the response to small molecule inhibitors of the pathway is highly variable and often difficult to predict, likely due to varied consequences of specific mutations within the pathway, combinatorial effects with co-occurring mutations, complex feedback regulation within the pathway and cross-talk with other signaling pathways. In the present study, we investigated the sensitivity of integrated analysis of multiple omics datasets to identify differences in HGGs driven by two different glioma-associated RTK mutations in the same p53-null primary astrocyte population.

We presented a generic bioinformatics pipeline for prioritizing core signaling networks and master regulators in cancer proteomics studies. Massive reprogramming of molecular components occurs during the evolution from mortal to immortal status in cancer cells[38]. As improvement of profiling technologies allows the identification of thousands of these changes, prioritizing drivers and core regulators from the enormous amount of passenger changes becomes a rating-limiting step. Here, we first performed weighted co-expression clustering analysis to extract 10 proteome and phosphoproteome clusters from 4,703 differentially expressed proteins and 6,768 differentially phosphorylated phosphosites, which dramatically reduced the data complexity. This readily identified major pathways with well-established roles in glioma growth as well as clear connections with PI3K and mTOR signaling downstream of RTK activation[43,67]. Subsequently, co-regulated genes in each of the clusters were summarized to pathways and networks using the network analysis method, which further narrowed down these massive changes to 67 network modules. We also developed systematic protein activity inference strategies for kinases and transcription factors by integrating multi-omics data and a variety of databases to further prioritize 41 kinases, 26 TFs, and a core network consisting of 13 master regulators from these gene clusters and network modules. Importantly, this integrated approach extended beyond simple identification of pathways to illustrate differences between the two oncogenes. While oncogenic NTRK fusion genes are found in adult and childhood HGGs carrying multiple other mutations, they are also found in infant HGGs and childhood low-grade gliomas in which very few non-synonymous mutations can be detected[39,48,66,190,191]. NTRK-driven HGGs showed a higher amplitude of pathway activation compared with PDGFRA-driven HGGs including a feed-forward upregulation of other RTKs consistent with the higher proliferative index and shorter tumor latency of the mouse HGGs, and the ability of NTRK fusions to act as potent oncogenic drivers in primary human tumors with minimal co-occurring mutations. Finally, we overlapped the most significantly altered omics datasets in mouse HGGs back

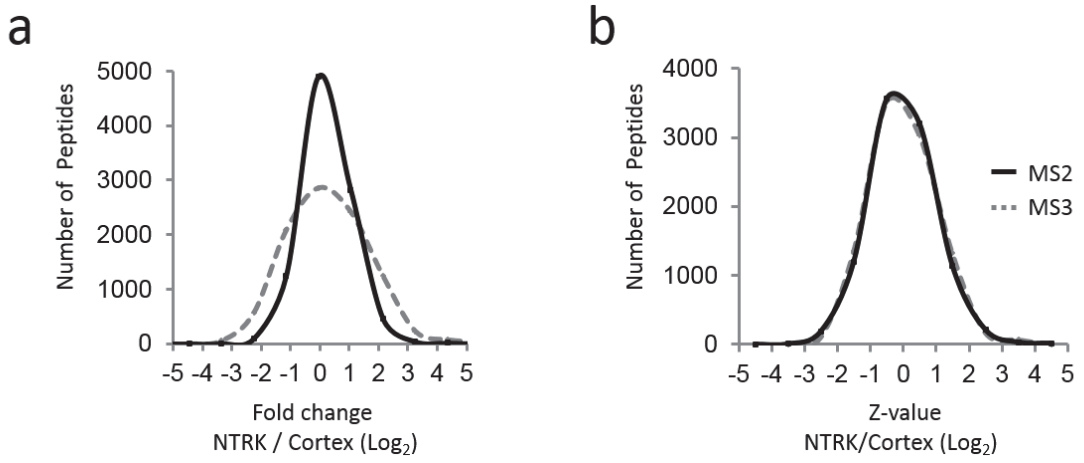



**Figure 3-17. TMT-based quantification using MS2 method has essentially no impact on protein differential expression analysis after Z scale normalization.** (a). Quantitative ratio compression occurred in TMT labeling strategy using MS2 compare to MS3 method. MS3 strategy can essentially eliminate ratio compression with the cost of more duty cycles and the use of low resolution MS2 data for identification, which often compromise peptide/protein identification. Comparison between MS2 and MS3 methods on the same sample using TMT labeling shows smaller difference/variance measured in MS2 method compare to MS3 method, suggesting quantitative ratio compression in MS2 analysis of TMT labeling. (b). Z scale normalization essentially eliminates the effect of ratio compression on protein differential expression analysis. Z scale transformation of the same data shows almost exact same Z value distribution in MS2 and MS3 method, suggesting similar amount of DE proteins with the same Z value cutoffs for differential expression analysis comparing MS2 and MS3 methods.

to human transcriptome data to search for consistent alterations driven by NTRK and PDGFRA mutations across species, resulting in a list of 20 convergent alterations in mouse and human. Indeed, many of the prioritized networks (e.g. AMPK-EEF2K) and proteins (e.g. EPHA2, CD74) are reported to be functional in HGGs[184,185,192].

With rapid improvement in omics technologies and accumulation of big datasets, this novel bioinformatics pipeline provides a general platform for prioritizing of master genes and core signaling networks in cancer omics study that will provide enhanced mechanistic understanding of the oncogenic process and illuminate potential therapeutic vulnerabilities.

# CHAPTER 4. INTEGRATED MULTI-OMICS ANALYSIS TO IDENTIFY A THERAPEUTIC VULNERABILITY IN RHABDOMYOSARCOMA

This study was a collaboration project between Dr. Michael Dyer's group and our group. Michael Dyer's group performed the genomic and epigenomic analyses, carried out drug screen assay and pre-clinical trials. Junmin and I designed the proteomics analysis experiments, and I performed the large-scale proteomics and phosphoproteome experiments and did the data analyses and integration.

## Introduction

To gain a better understanding of RMS disease recurrence and to provide additional preclinical models of pediatric solid tumors for the biomedical research community, Dr. Michael Dyer's group has established a unique collection of orthotopic patient derived xenografts (O-PDX) over the past 6 years[193]. RMS tumors had the most efficient engraftment rate and the fastest time to engraftment. They have established and characterized O-PDX tumors of translocation negative ERMS tumors and translocation positive ARMS tumors from diagnosis and recurrence[193]. They performed whole genome sequencing (WGS), whole exome sequencing (WES) and RNA-Seq of the patient tumor and the matched O-PDX. They also performed clonal analysis to profile the clonal distribution in the O-PDX tumors relative to the patient's tumor. This provides a unique resource to identify tumor vulnerabilities through integrative analysis with deep proteomic and phosphoproteomic data and test them in preclinical models in order to inform new clinical trails for RMS. This is particularly important because overall survival rates for RMS have not significantly improved in the past 20 years[194].

Dr. Michael Dyer's group attempted to target the RAS pathway in RMS by incorporating CDK4/6 inhibitors with MEK inhibitors as done for adult cancers with oncogenic RAS mutations[68-71]. While they achieved synergistic killing of a RAS mutant rhabdomyosarcoma cells in culture, there was no significant anti-tumor effect *in vivo* using O-PDX models of recurrent RMS (**Figure 4-1**). To determine how tumor cells escape the treatment and to identify a novel therapeutic vulnerability, we collaborated to perform epigenetic and proteomic/phosphoproteomic analysis of 17 RMS O-PDX tumors using advanced profiling strategies developed above[148-151,168] and integrated those data with genomic and gene expression data to identify tumor vulnerabilities

To define the chromatin states and more efficiently analyze the transitions thereof across the genome, Dr. Dyer's group performed chromatin Hidden Markov Modeling (chromHMM)[195] using all 756 ChIP-seq datasets. And they focused our analysis of the epigenomic landscape of rhabdomyosarcoma on 3 broad categories and integrated those with the DNA methylation analysis. The found that there were 98 genes upregulated in ERMS relative to ARMS that had a ERMS specific superenhancer and there were 174 genes that were downregulated in ERMS relative to ARMS that had an ARMS specific superenhancer (**Figure 4-2**). Many of these genes are implicated in myogenesis. Overall,

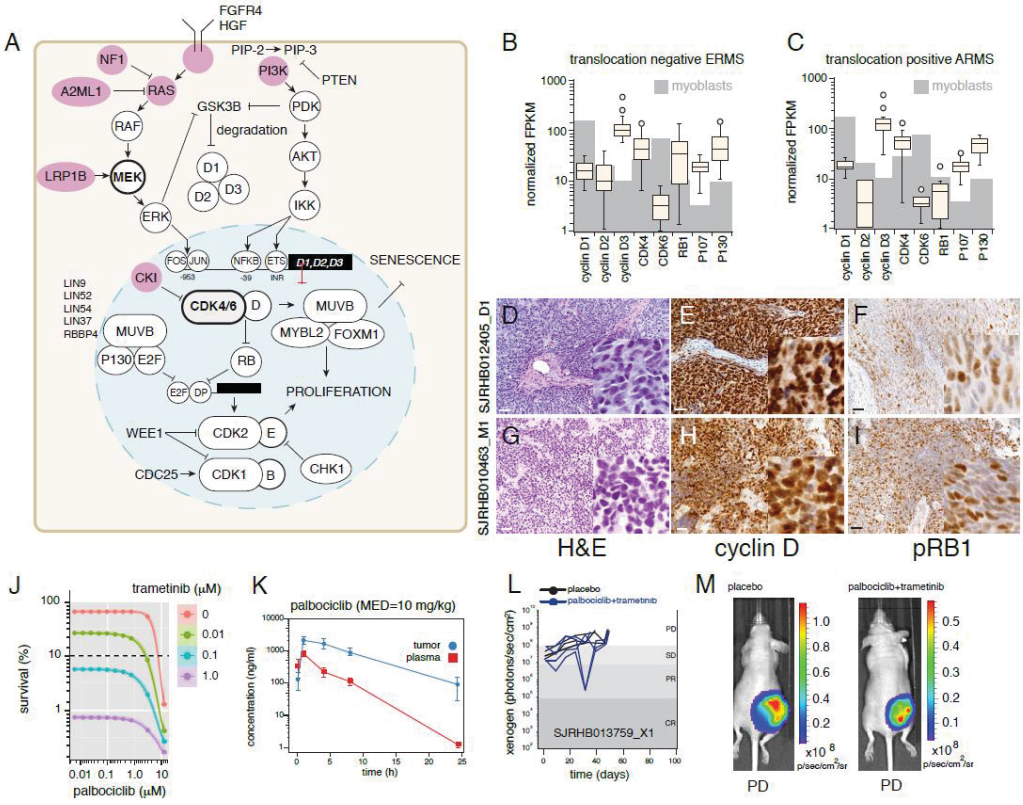



**Figure 4-1. Targeting CDK4/6 in Rhabdomyosarcoma.**
(A) Simplified pathway map for the RAS and CDK4/6 pathways. Recurrent mutations have been reported in the genes encoding the proteins highlighted in purple and those in bold (MEK and CDK4/6) were targeted with small molecule inhibitors in this study. (B,C) Boxplot of gene expression (FPKM) from RNA-seq for each of the indicated cell cycle genes across ERMS (B) and ARMS (C) tumors. The gray bars represent the expression of each gene in primary human myoblasts. (D-I) Representative micrographs of ERMS and ARMS patient tumor sections with H&E staining, immunohistochemical staining for cyclin D2 (E) and cyclin D3 (H) and phosphorylated RB1 (F,I). A magnified view is shown in each corner. (J) Representative combination drug sensitivity study for the RD RMS cell line showing 10 different concentrations of palbociclib and 4 different concentrations of trametinib. Survival is plotted relative to untreated cells (100% survival) and complete killing (0% survival) with a positive control cytotoxic compound. (K) Plot of concentration versus time for palbociclib in orthotopic RMS tumor xenografts (blue) and plasma (red) for 3 independent mice per treatment group. The mean and standard deviation are plotted for each time point. These data were used to calculate the MED by comparison to plasma levels for palbociclib from patients. (L) Representative plot of tumor burden as measured by bioluminescence over time during a preclinical phase II study with the SJRHB013759_X1 O-PDX tumor. M) Representative images of mice with progressive disease in the placebo treatment group and the palbociclib+trametinib treatment groups

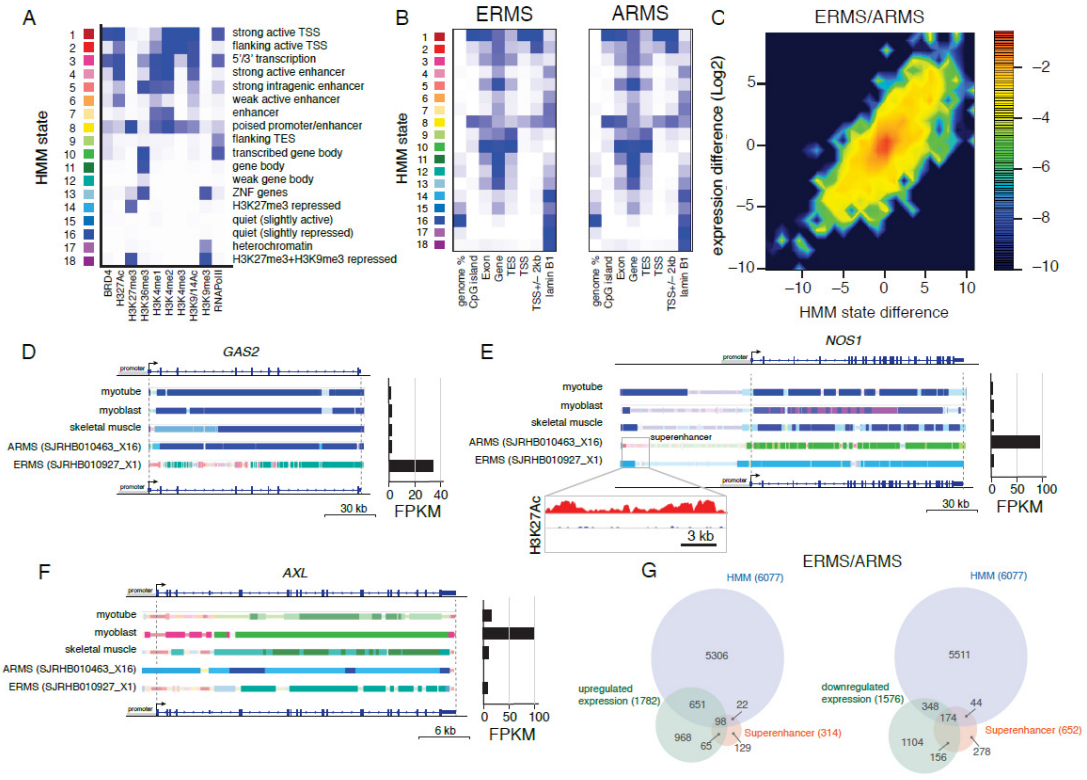



**Figure 4-2. Differences in epigenetic profiles correlate with promoter/enhancer activity.**

A) Heatmap of the 18 chromHMM states used in this analysis. B) Heatmap showing the proportion of the 18 chromHMM states in ERMS and ARMS for the annotated regions of the genome. For the heatmaps in A and B, the proportion of individual marks or regions is directly proportional to the intensity of the blue color. C) Heatmap of the correlation between the ratio of gene expression (ERMS/ARMS) on the y-axis and the ratio of HMM state difference for ERMS/ARMS. D) Representative chromHMM for the GAS2 gene that is selectively expressed in ERMS. The gene boundaries are marked by dashed lines and the colors for each state are indicated in panel (A). The 2 states that are the highest proportion are full-height bars, and the remaining states are half the height. The intensity of each bar is proportional to the percentage of each state across all stages for that gene. For the bars that are half the height, the intensity is scaled starting at 50% of maximum intensity. E) Representative chromHMM for the NOS1 gene that is selectively expressed in ARMS and has a tumor type specific superenhancer upstream of the gene (boxed H3K27me3 region). F) Representative chromHMM for the AXL gene that is selectively expressed in myoblasts. G) Venn diagram for the genes that are upregulated in ERMS relative to ARMS and those that are downregulated (green). The overlap with genes that have differences in chromHMM state (blue) or tumor type specific superenhancers (red) are shown. Abbreviations: chromHMM, chromatin hidden markov modeling; FPKM, fragments per kilobase million.

the ERMS tumors had epigenetic upregulation of genes enriched in pathways involved in extracellular matrix organization and morphogenesis during embryogenesis including limbs and the skeletal system. The ARMS tumors had epigenetic upregulation of genes enriched in pathways involved in muscle differentiation suggesting they have progressed further along the myogenic lineage than ERMS tumors. However, ARMS tumors also had upregulation of genes enriched in pathways involved in neurogenesis suggesting these tumors may have a mixed developmental phenotype. To further refine the myogenic lineage differences between ARMS and ERMS, these integrated analyses of the epigenome suggest that ARMS tumors are arrested at a later stage of muscle development (myogenic differentiation) than ERMS tumors (myogenic specification/determination).

Recently, mass spectrometry (MS)-based proteomics is emerging as the mainstream approach for unbiased analysis of the cancer proteome and phosphoproteome[38,196]. Together with advanced epigenetic profiling and DNA sequencing technologies, these methodologies provide an unprecedented opportunity for illuminating cancer therapeutic targets.

## Methods and Materials

Isobaric labeling, such as iTRAQ and TMT, is emerging as a powerful strategy for deep multiplexed proteomics analysis with high throughput and reproducibility[29,62,188,189,197]. One limitation associated with this method is that target peptide ions often co-isolated with other co-eluted peptide ions during LC-MS/MS analysis, which causes high noise level to compress quantitative ratios and decrease measurement accuracy. Fortunately, the ratio compression effect also alleviates experimental variations, and hence has only minor impact on protein differential expression analyses[29,198]. Moreover, the ratio compression can be diminished by extensive peptide fractionation, narrow isolation window, and post-MS correction[188]. Alternatively, the MS3 method can almost eliminate this ratio compression, but it requires longer duty cycles, specific MS settings, and the use of low resolution MS2 for identification, which often compromise the peptide/protein identification[189,199]. To balance the pros and cons associated with isobaric labeling, we implemented extensive fractionation through long gradient high resolution LC/LC-MS/MS to achieve deep proteome coverage and to reduce ion compression during quantification. In addition, we employed sample replicates to facilitate statistical inference of differentially expressed proteins, a widely adopted strategy used in proteomics analyses[200-202].

### RMS O-PDX tumor tissue, myoblast and myotube cells for proteome and phosphoproteome profiling

$10^8$ Myoblast and myotube cells per sample were collected, washed twice with 10 ml ice cold PBS, and followed by snap freezing. These processes were managed to be completed in 5 minutes. PDX-engrafted mice were anesthetized and perfused with 10 ml

PBS before sacrifice, center section of PDX tumor tissues were collected, homogenized, and followed by snap freezing. These steps were finished within 10 minutes.

**Protein extraction, digestion, labeling and pooling**

Protein extraction, digestion, labeling and pooling were performed similarly as previously described[201]. RMS PDX tissues, myoblast and myotube cells per sample were lysed in freshly prepared lysis buffer (50 mM HEPES, pH 8.5, 8 M urea, 0.5% sodium deoxycholate and phosphatase inhibitor cocktail (PhosphoSTOP, Roche)). Protein concentration of sample lysates were quantified by BCA protein assay (Thermo Fisher Scientific) with titrated BSA as a standard. ~1 mg proteins per sample were first digested with Lys-C (Wako, 1:100 w/w) at room temperature for 2 h, diluted 4 times with 50 mM HEPES, pH 8.5, and then further digested with trypsin (Promega, 1:100 w/w) for overnight at room temperature. 1% trifluoroacetic acid was added to quench the digestion reaction, followed by desalting with Sep-Pak C18 cartridge (Waters), and the desalted peptides were dried by speedvac. Samples were then resuspended in 50 mM HEPES, pH 8.5, and were labeled with 10-plex TMT reagents following the manufacturer's instruction. Lastly, 10 isobaric labeled samples were pooled together with equal amount, desalted again by Sep-Pak C18 cartridge and then speedvac dried.

**Offline basic pH reverse phase liquid chromatography**

The basic pH reverse phase liquid chromatography peptides pre-fractionation were performed on Agilent 1220 LC system as previously introduced[201]. The pooled TMT labeled sample was solubilized in buffer A (10 mM ammonium formate, pH 8) and separated on two XBridge C18 columns (3.5 μm particle size, 4.6 mm × 25 cm, Waters) into around 180 fractions with a 220 min long gradient started from 15% to 65% buffer B (95% acetonitrile, 10 mM ammonium formate, pH 8, flow rate: 0.4 ml/min). 5% of each combined fraction was dried for whole proteome analysis and the remaining 95% was dried by speedvac for phosphoproteome analysis.

**Refined phosphopeptide enrichment by $TiO_2$**

Phosphopeptide enrichment was performed following the refined protocol as previously introduced[198]. Briefly, peptides were added to clean $TiO_2$ beads (GL sciences) with a peptide-to-beads weight ratio of 1:4 in binding buffer (65% acetonitrile, 2% TFA, and 0.5 mM $KH_2PO_4$) and incubate for 20 min. Enriched phosphopeptides were washed, eluted, dried, and dissolved in 5% formic acid for LC-MS/MS analysis.

## Long gradient acidic pH reverse phase LC-MS/MS

The analysis was carried out based on our optimized platform as previously introduced[198,201,203]. The dried peptide fractions were reconstituted in loading buffer (5% formic acid), loaded on a reverse phase column (75 μm × 50 cm, 1.9 μm $C_{18}$ resin (Dr. Maisch GmbH, Germany)) interfaced with an FUSION or Q Exactive HF mass spectrometer (Thermo Fisher Scientific). Peptides were eluted by an up to 6h 15-65% gradient of buffer B. (buffer A: 0.2% formic acid, 5% DMSO; buffer B: buffer A plus 65% acetonitrile, flow rate: 0.25 μl/min). A butterfly portfolio heater (Phoenix S&T) was applied to heat the column at 65°C to reduce backpressure. The mass spectrometer was operated in data-dependent mode with MS1 settings of 60,000 resolution, $1 \times 10^6$ AGC target and 50 ms maximal ion time and top 20 MS/MS high resolution scans with MS2 settings of 1 m/z isolation window with offset 0.2, 60,000 resolution, 100 ms maximal ion time, $1 \times 10^5$ AGC target, HCD, 33 normalized collision energy, and 40 s dynamic exclusion (35 normalized collision energy and 20s dynamic exclusion for phosphoproteome).

## Peptide identification by JUMP, a tag-based hybrid search engine

Peptide identification was performed using our recently developed JUMP search engine with improved sensitivity and specificity[150]. Commercially available database search engines can be divided into two categories: tag-based *De novo* sequencing (e.g. PEAKS with limited sensitivity) and pattern-based database search (e.g. SEQUEST, MASCOT). The JUMP software integrates these two methods to score putative peptides, showing significant improvements compared with these commercially available tools[150]. The JUMP software has already been used in numerous publications[202,204-209]. Analysis was done similarly as previously described[201], MS/MS raw files were first converted into mzXML format and searched against a composite target/decoy database[108] for FDR estimation. The target protein database was compiled from the Uniprot mouse and human database (Human database: 88,965 protein entries; Mouse database: 52,738 protein entries, downloaded in February 2015), the decoy database was generated by reversing target protein sequences. Spectra were searched with ± 10 ppm mass tolerance for both precursor ions and product ions with fully tryptic restriction, static modification for TMT tag on N-terminus and lysine (+229.16293), dynamic modification for serine, threonine and tyrosine (+79.96633, for phosphoproteome analysis), three maximal modification sites, two maximal missed cleavages, and the assignments of a, b, and y ions. Peptide spectrum matches (PSM) were first filtered by MS mass accuracy (~2 ppm, ± 4 standard deviations). PSMs of doubly charged peptides with JUMP Jscore of > 30 were applied for global mass recalibration prior to the filtering. The qualified PSMs were first grouped by precursor ion charge state and then further filtered by Jscore and dJn values. Cutoffs were applied on these values and were adjusted until a protein FDR < 1% was achieved. If one peptide was shared by multiple proteins, the protein with the highest PSM will represent the peptide according to the rule of parsimony[202,210].

**Phosphosite assignment by the Lscore from the JUMP software suite**

To determine the reliability of phosphosites localization on peptides, we adopted the concept of the phosphoRS algorithm[211] to compute phosphosite localization scores (Lscore, 0-100%) in each PSM the same as previously described[201], Phosphosites were aligned to protein sequences to generate protein level Lscores in addition to the PSM Lscores. The protein Lscore was represented by the highest PSM Lscore if multiple PSMs were identified for one specific phosphosite. Since random assignments of PSMs containing ambiguous phosphosites often causes an excessively high number of unreliable phosphosites on proteins, we implemented series rules to alleviate the problem: (i) If the gap of PSM Lscores between the 1$^{st}$ and 2$^{nd}$ site > 10% for a singly phosphorylated peptide in one PSM, the top Lscore site was selected; (ii) Otherwise, we inspect the phosphosites in the corresponding proteins instead to select the sites with the highest protein Lscore. This allows low quality PSMs to borrow information inferred from the high quality PSMs; (iii) If both PSM and protein level Lscores were indistinguishable, a heuristic priority was assigned to phosphosites according to the order of occurrence: SP-motif, S, T and Y; (iv) If the PSMs did not satisfy any rule above, these PSMs were first sorted by JUMP Jscores, and then we selected protein phosphosites that had been determined by other PSMs of high Jscores.

**TMT-based protein and phosphosite quantification using the JUMP software suite**

This analysis was carried out in the following steps similarly as previously reported[201,212]: (i) TMT reporter ion intensities of each PSM were extracted; (ii) the raw intensities were corrected according to isotopic distribution of each labeling reagent; (iii) PSMs with very low reporter ion intensities were excluded (e.g. minimum intensity < 1,000 and median intensity < 5,000); (iv) sample loading bias was corrected by normalization with the trimmed median intensity of all PSMs; (v) the mean-centered intensities across samples were calculated; (vi) protein or phosphosite relative intensities were summarized by averaging related PSMs; (vii) protein or phosphosite absolute intensities were derived by multiplying the relative intensities by the grand-mean intensity of top three most highly abundant PSMs. To generate a combined quantification table from batches 1-3, a common sample (ARMS10468X) was included in each batch as internal standard. The batch effect was normalized for each protein or phosphosite by assuming that the abundance of the internal standard sample was equal among different batches. To generate a combined quantification table for batches 4 and 5, it was assumed that the mean of protein/peptide abundance across the ten sample of each batch was equal and batch-effect normalization was implemented by fitting a linear model of protein/peptide abundance (log-transformed intensity) on batch information using the removeBatchEffect function in the LIMMA R package[213].

### Differential expression analyses of proteome and phosphoproteome

Differential expression analyses of whole proteome and phosphoproteome were carried out using LIMMA R package, a software designed for the analysis of gene expression involving comparisons between many gene targets simultaneously[213,214]. LIMMA borrows information across genes by fitting linear models to overcome the problem of small sample size and complex experimental design[215], hence is ideal for differential expression analysis of TMT-based deep proteomic data. Briefly, (i) Linear models were fitted for expression data of each protein or phosphosite; (ii) Empirical Bayes method was used to borrow information across genes; (iii) P values were adjusted by the Benjamin Hochberg method; (iv) The adjusted P value cutoff of 0.05 was then applied; (v) Remaining proteins were further filtered by a fold change of 1.5 and 2.0 (equivalent to around 3 × SD of tumor biological replicates) in at least one group comparison for proteome and phosphoproteome respectively.

### Weighted gene co-expression network analysis (WGCNA) and pathway annotation

The analysis was done by the WGCNA R package[154,216] similarly as previously described (Tan, H. et al., 2017). Only DE proteins and phosphosites were applied to define the proteome and phosphoproteome co-expression clusters (i.e. WPCs and PPCs) respectively. (i) A Pearson correlation matrix was generated by calculating correlation between proteins (only positive correlations were considered), and was further raised to a power of 16 using the scale free topology principle to calculate an adjacency matrix[216]. (ii) Co-expression clusters were then determined by the hybrid dynamic tree-cutting method[217] with a height cutoff (e.g. 0.2) for merging modules; (iii) A consensus trend (eigengene) was calculated based on the first principal component for each co-expression cluster. Proteins were then assigned to the co-expression cluster with the highest correlation; (iv) Each co-expression cluster was then annotated using the Hallmark pathway database downloaded from MsigDB[218], Myogenic regulatory pathways that were not annotated in Hallmark database were manually extracted from KEGG database and added in) by Fisher's exact test. Pathways with a B.H. adjusted P value less than 0.05 were selected as deregulated pathways in each co-expression cluster.

## Results

### Quantitative analysis of whole proteome and phosphoproteome in rhabdomyosarcoma

To determine if the developmental arrest identified in epigenomic analysis is also reflected in the signal transduction cascades in RMS, we used the newly developed high-throughput MS pipeline with extensive peptide separation power and high mass resolution[38,148-151,201] to quantify the proteome and phosphoproteome of 12 O-PDX tumors (8 ERMS and 4 ARMS), normal human myoblasts and myotube using biological

duplicates. Batches of 10 samples were lysed, digested and labeled with 10 different TMT tags (**Figure 4-3**). To facilitate comparison across batches, we included a replicate across each batch (SJRHB010468_X1) to serve as an internal control. Together, we were able to quantify 13,403 proteins products from 10,332 genes and 12,653 phosphosites presented in all batches and validated several proteins and phosphoproteins by immunoblot (**Figure 4-3B, C**). The intra-batch reproducibility was 0.98 for both the whole proteome and phosphoproteome (**Figure 4-3D**). The inter-batch reproducibility was 0.95 for whole proteome and 0.94 for phosphoproteome (**Figure 4-3E**). Principle component analysis demonstrated clear separation of ARMS from ERMS and normal (myoblasts and myotube) samples for whole proteome and phosphoproteome (**Figure 4-3F, G**). Weighted gene co-expression network analysis (WGCNA)[154] of differentially expressed proteins and phosphoproteins identified 6 groups of proteins and phosphoproteins with differences across ARMS, ERMS and normal samples (**Figures 4-3H, I and 4-4**).

We performed pathway analysis on the differentially expressed proteins and phosphoproteins between the normal cells (myoblasts and myotube) and the RMS samples (ARMS and ERMS). As expected from our initial analysis of cyclin D–CDK4/6–RB signaling (**Figure 4-1**), the E2F target proteins and phosphoproteins were significantly upregulated in RMS relative to the normal cells (**Figure 4-5A**). Several other pathways associated with proliferation, cell cycle checkpoint regulation and DNA repair were also significantly altered in the tumor cells relative to normal human muscle (**Figure 4-5A**). Importantly, among the pathways that showed significant perturbation in RMS relative to normal muscle were those important for myogenesis including the WNT[219-221], HH[222,223], BMP[224-226], adenyl cyclase[227], p38/MAPK[228-231] and PI3K[232-234] pathways (**Figures 4-5A** and **4-6**). However, one of the limitations of pathway analysis using these types of consensus generic pathway maps is the lack of tissue or lineage specific signaling relationships. Therefore, we manually curated each of these 6 fundamental signal transduction pathways based on published literature for myogenesis and integrated our genomic, epigenomic, proteomic and phosphoproteomic data to advance our understanding of the developmental arrest in each RMS tumor. For example, BMP4 signaling through BMPR1A,B is implicated in patterning the early somite by blocking MYF5 and MYOD1 transcriptional programs required for myogenic specification in the early dermomyotome cells[235-238]. The BMP antagonists (NOG, CHRD, GREM1,2 and FST) are expressed in cells that will give rise to the muscle lineage at this early stage and are downregulated later during muscle maturation (**Figure 4-5B**). GREM1 and NOG mRNA are expressed at higher levels in myoblasts than myotubes and are variable across RMS samples (>5-fold). BMP4 mRNA levels are similar across normal muscle cells and RMS samples while BMPR1A,B are upregulated in RMS (**Figure 4-5C**). Importantly, the BMP antagonist (NOG, CHRD and GREM1) are significantly downregulated in RMS relative to myoblasts while BMPR1A is upregulated (**Figure 4-5D**).

In concert with BMP signaling, SHH is released from the notochord to specify muscle progenitors expressing MYF5 from the multipotent dermomyotome cells (**Figure 4-5B**). SHH is not expressed in the normal human muscle or RMS cells

**Figure 4-3. The proteome and phosphoproteome are distinct across ERMS, ARMS and myogenic precursors.**
A) Workflow for the proteomic and phosphoproteomic profiling of RMS and normal myotube and myoblasts. B) Representative immunoblots showing validation of differences in protein expression and phosphoprotein expression for NOS1, HMGA2 and phospho-S235/S236 of RPS6. C) Normalized relative fold of the proteins in (B) from the proteomic and phosphoproteomic analysis for each sample. The myotubes are indicated by (mt). D) Scatterplot of the quantitation of proteins across replicate samples in the same batch and replicate samples in different batches. E) Scatterplot of the quantitation of phosphoproteins across replicate samples in the same batch and replicate samples in different batches. F,G) Principle component analysis of the myoblasts (light green), myotubes (dark green), ERMS (red) and ARMS (purple) samples for the whole proteome (F) and the phosphoproteome (G). (H) Heatmap of the 6 groups of proteins that show significant differences across samples by sample grouping (normal, ERMS, ARMS). (I) Expression patterns of cluster 3 as a representative. Boxplots show log2 level abundance of RMS tumors samples relative to myoblasts. N indicates the normal myotubes (mt) and myoblasts (mb). Abbreviations: PC1 and PC2 are principle components 1 and 2, respectively

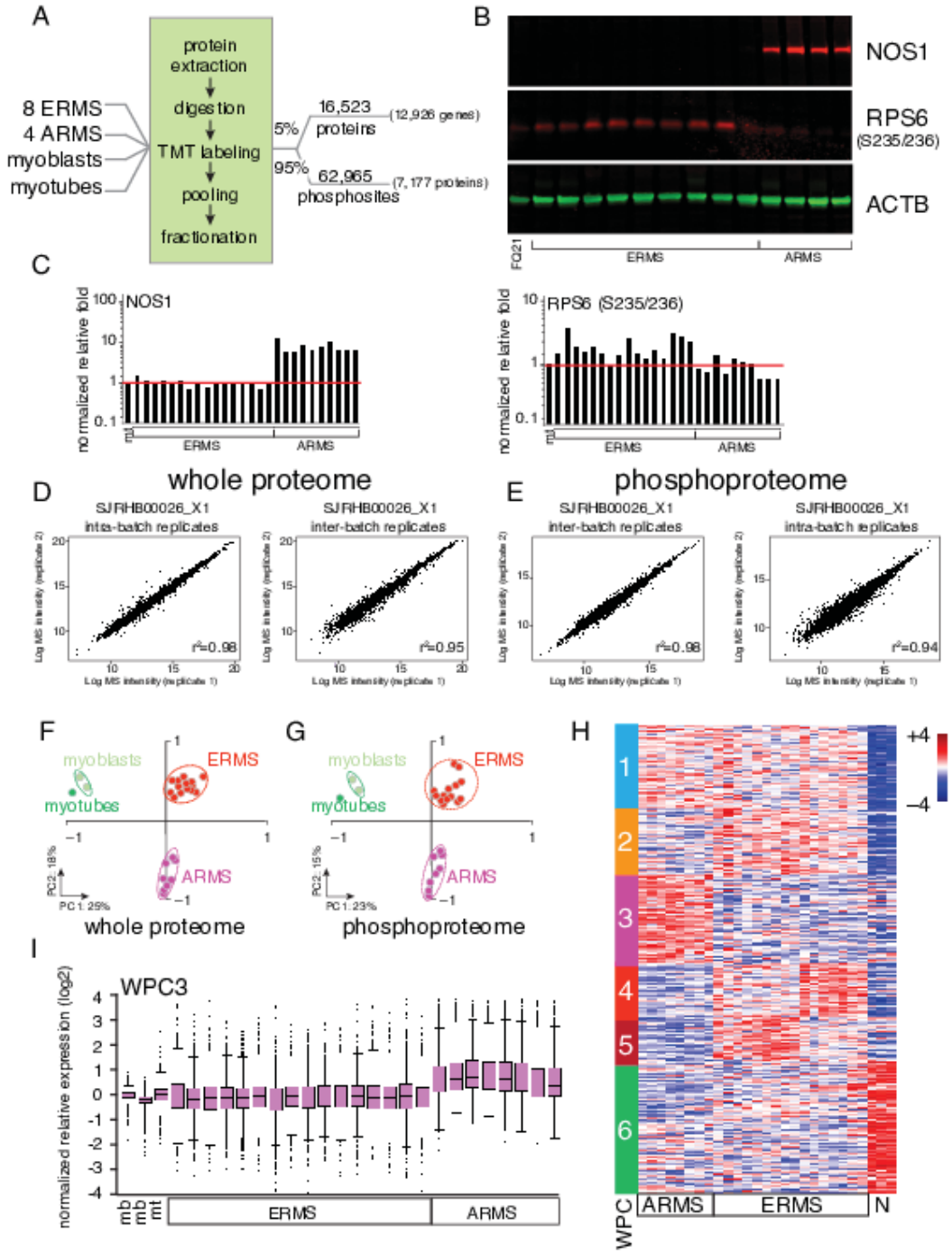

A
protein extraction
digestion
TMT labeling
pooling
fractionation
8 ERMS
4 ARMS
myoblasts
myotubes
5%
16,523 proteins
(12,926 genes)
95%
62,965 phosphosites
(7,177 proteins)
B
NOS1
RPS6 (S235/236)
ACTB
FC21
ERMS
ARMS
C
NOS1
RPS6 (S235/236)
normalized relative fold
mb
ERMS
ARMS
D
whole proteome
SJRHB00026_X1
intra-batch replicates
inter-batch replicates
r²=0.98
r²=0.95
E
phosphoproteome
SJRHB00026_X1
inter-batch replicates
intra-batch replicates
r²=0.98
r²=0.94
F
myoblasts
myotubes
ERMS
ARMS
PC2: 18%
PC1: 25%
whole proteome
G
myoblasts
myotubes
ERMS
ARMS
PC2: 15%
PC1: 23%
phosphoproteome
H
+4
−4
1
2
3
4
5
6
WPC
ARMS
ERMS
N
I
WPC3
normalized relative expression (log2)
mb
mb
mt
ERMS
ARMS

**Figure 4-4. Weighted gene co-expression network analysis of whole proteome and phosphoproteome.**
(A, C) Heatmap of the 6 groups of proteins and phosphoproteins that show significant differences across samples by sample grouping (normal, ERMS, ARMS). (B, D) Expression patterns of 6 whole proteome clusters (WPC) and 6 phosphoproteome clusters (PPC). Boxplots show log2 level abundance of RMS tumors samples relative to myoblasts.

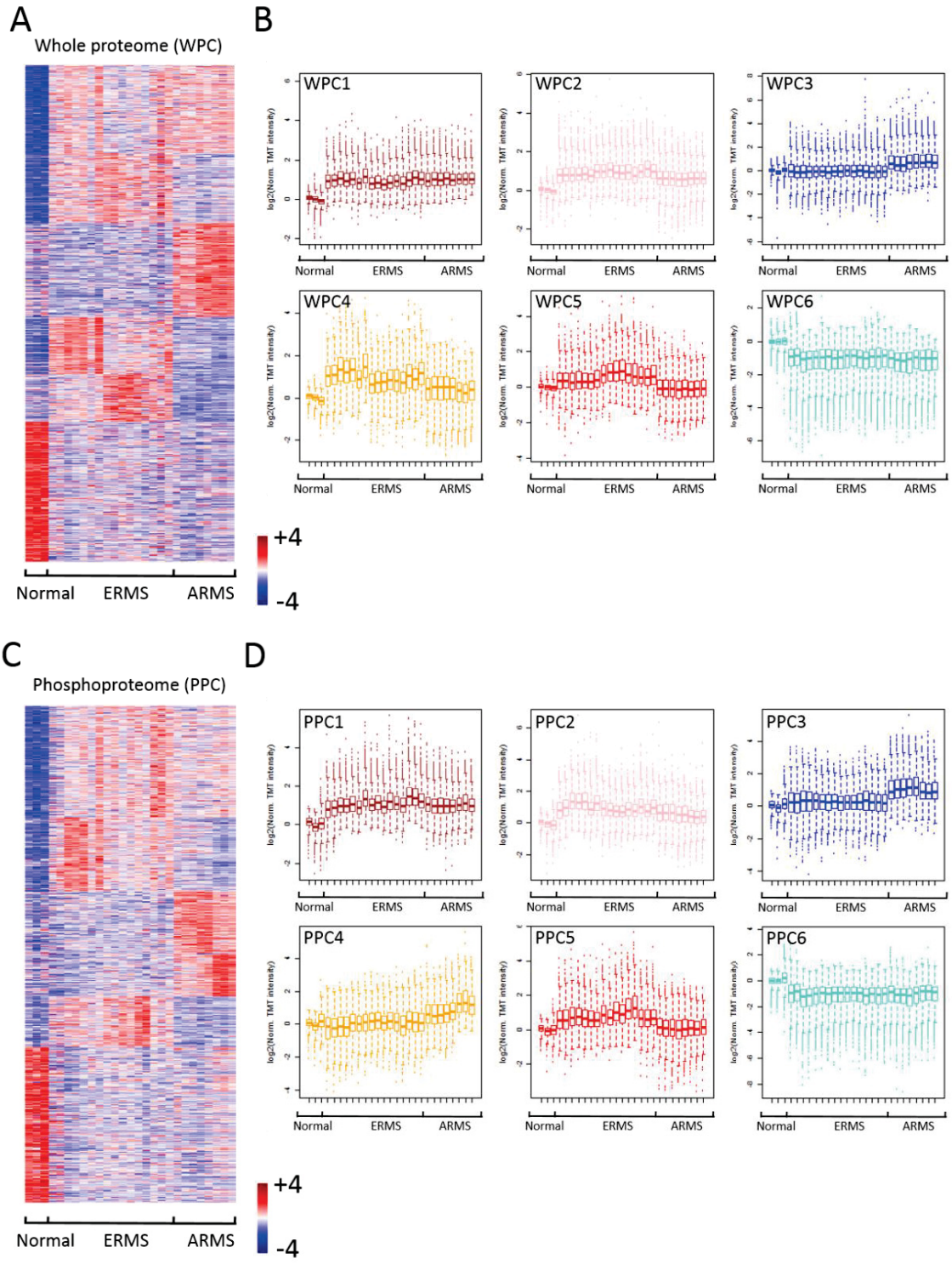
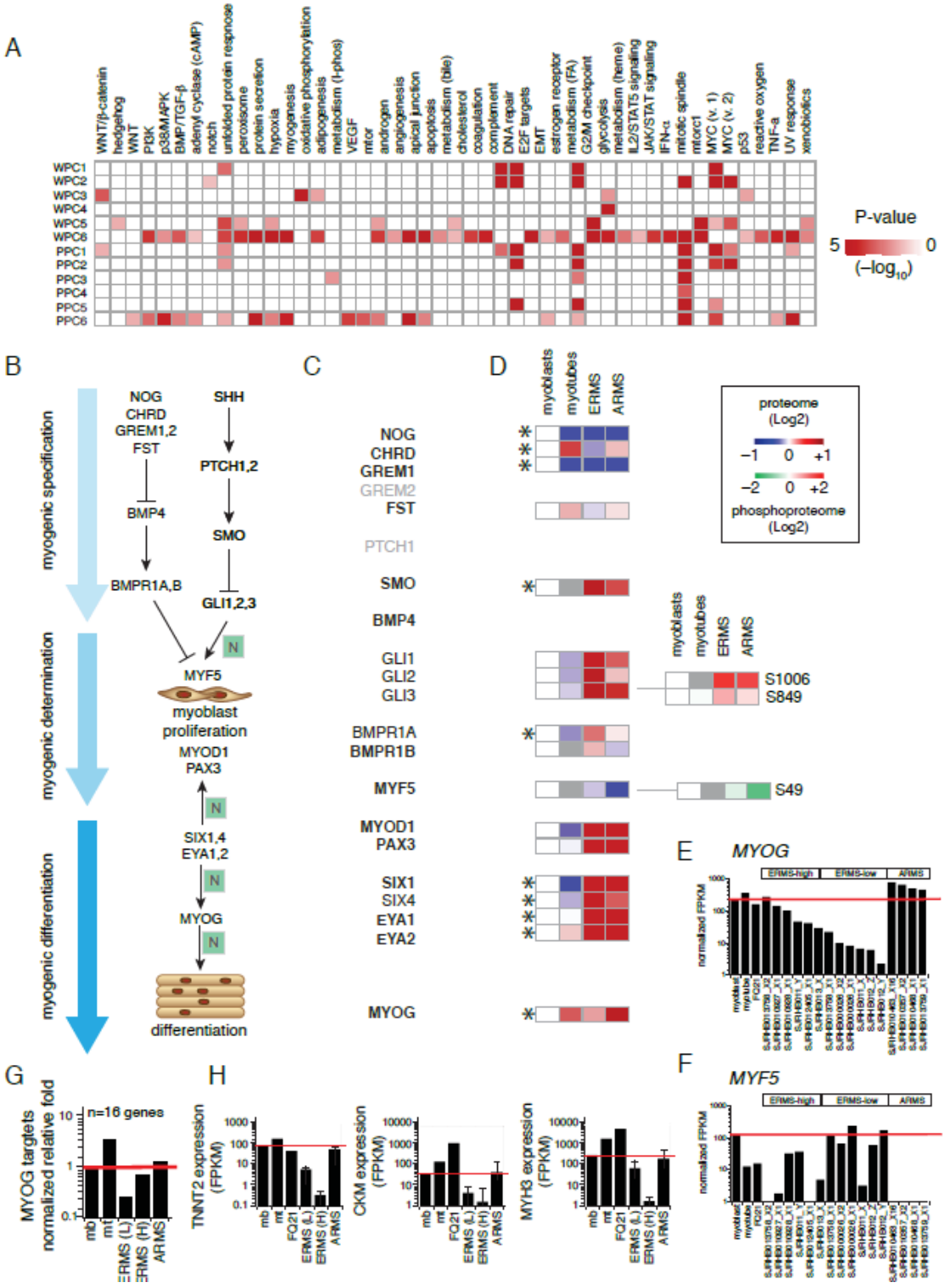

A
Whole proteome (WPC)
Normal
ERMS
ARMS
+4
-4
B
WPC1
WPC2
WPC3
WPC4
WPC5
WPC6
log2(Norm. TMT intensity)
Normal
ERMS
ARMS
C
Phosphoproteome (PPC)
Normal
ERMS
ARMS
+4
-4
D
PPC1
PPC2
PPC3
PPC4
PPC5
PPC6
log2(Norm. TMT intensity)
Normal
ERMS
ARMS

**Figure 4-5. Myogenic pathways are deregulated in RMS.**
(A) Heatmap of pathway analysis for the whole proteome groups (WP1-6) and the phosphoproteome (PPC1-6) groups. The intensity of each box if inversely proportional to the Log of the p-value. (B) Representative BMP and HH pathways for muscle development. Arrows indicate interactions that promote activity and bars represent interactions that block activity. The green boxed N represents perturbations in transcriptional targets. (C) The individual family members that are expressed in the cohort analyzed in this study are shown. The cutoff for inclusion in (C) is > FPKM of 1.0. Genes shown in gray are between FPKM of 1.0 and 10.0. Those in bold vary by more than 5-fold across samples in our analysis. (D) Heatmap of the protein and phosphoproteins normalized to myoblasts. The intensity is proportional to the Log2 of protein or phosphoprotein expression relative to myoblasts. The asterisk indicates those proteins that have statistically significant differences across samples. (E,F) Histogram of RNA expression (FPKM) for MYF5 and MYOG across normal samples, ARMS and ERMS. The red line indicates the expression in normal human myoblasts. FQ21 indicates fetal quadriceps at 21 weeks. Abbreviations: FPKM, fragments per kilobase million.


A
WNT/β-catenin
hedgehog
WNT
PI3K
p38/MAPK
BMP/TGF-β
adenyl cyclase (cAMP)
notch
unfolded protein response
peroxisome
protein secretion
hypoxia
myogenesis
oxidative phosphorylation
adipogenesis
metabolism (l-phos)
VEGF
mtor
androgen
angiogenesis
apical junction
apoptosis
metabolism (bile)
cholesterol
coagulation
complement
DNA repair
E2F targets
EMT
estrogen receptor
metabolism (FA)
G2/M checkpoint
glycolysis
metabolism (heme)
IL2/STAT5 signaling
JAK/STAT signaling
IFN-α
mitotic spindle
mtorc1
MYC (v. 1)
MYC (v. 2)
p53
reactive oxygen
TNF-a
UV response
xenobiotics
WPC1
WPC2
WPC3
WPC4
WPC5
WPC6
PPC1
PPC2
PPC3
PPC4
PPC5
PPC6
P-value
5
0
(–log10)
B
myogenic specification
myogenic determination
myogenic differentiation
NOG
CHRD
GREM1,2
FST
SHH
PTCH1,2
BMP4
SMO
BMPR1A,B
GLI1,2,3
N
MYF5
myoblast proliferation
MYOD1
PAX3
SIX1,4
EYA1,2
MYOG
differentiation
C
NOG
CHRD
GREM1
GREM2
FST
PTCH1
SMO
BMP4
GLI1
GLI2
GLI3
BMPR1A
BMPR1B
MYF5
MYOD1
PAX3
SIX1
SIX4
EYA1
EYA2
MYOG
D
myoblasts
myotubes
ERMS
ARMS
proteome (Log2)
−1 0 +1
−2 0 +2
phosphoproteome (Log2)
S1006
S849
S49
E
MYOG
ERMS-high
ERMS-low
ARMS
normalized FPKM
F
MYF5
G
MYOG targets normalized relative fold
n=16 genes
mb
mt
FQ21
ERMS (L)
ERMS (H)
ARMS
H
TNNT2 expression (FPKM)
CKM expression (FPKM)
MYH3 expression (FPKM)

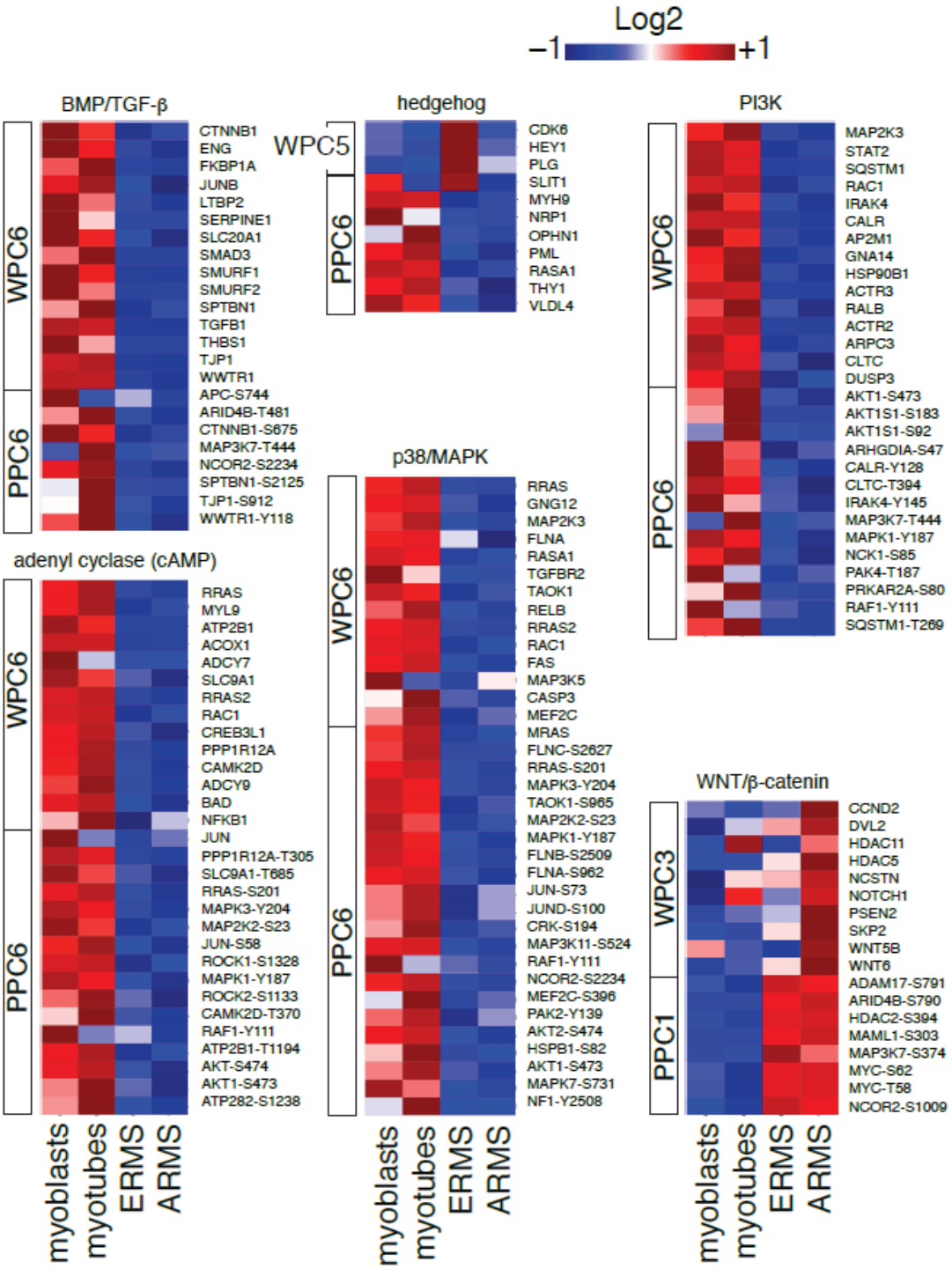



**Figure 4-6. Heatmaps of deregulated genes at protein expression and phosphorylation levels in pathways that are important for myogenesis.**

SHH binds the PTCH1/2 receptors, releases the SMO protein and this leads to activation of GLI transcription factors. Only PTCH1 mRNA is expressed in normal muscle and RMS but it expressed at low levels (<10 FPKM). There are 3 GLI family member and all 3 mRNAs are expressed in normal muscle and RMS (**Figure 4-5C**). GLI1,2 are activators of transcription and GLI3 is a repressor. Signaling through this pathway is implicated in transcriptional activation of MYF5 and formation of the MYOD+, MYF5+ committed myoblasts (**Figure 4-5B**). The GLI proteins are upregulated in RMS relative to myoblasts and levels in pathways that are important for myogenesis myotubes but the expression of MYF5 is lower in most of the RMS tumors relative to the normal human myotubes and myoblasts. This may be due to the GLI3 repressor and/or the downregulation of the BMP antagonist (**Figure 4-5D**). Indeed, both phosphorylation sites (S1006 and S849) on GLI3 required for transcriptional repressor activity are significantly enriched in the RMS phosphoproteome relative to myoblasts (**Figure 4-5D**). This example highlights the value of integrating proteome, phosphoproteome, transcriptome and epigenomic data.

To determine if there is variability across samples, we analyzed MYF5 expression and its transcriptional targets in each individual tumor because of the restricted window of expression of MYF5 in proliferating myoblasts during myogenic determination in development. All of the ARMS tumors had low levels of MYF5 expression relative to normal muscle and a subset of ERMS tumors had higher levels of MYF5 mRNA and protein (**Figure 4-5E**). The transcriptional target genes of MYF5 were also activated in the same pattern (**Figure 4-5F**). These data suggest that ARMS tumors have features of a later developmental stage than ERMS tumors after MYF5 is downregulated during myogenic differentiation. Consistent with this interpretation, the expression of MYOG protein and mRNA was significantly higher in ARMS tumors than in ERMS tumors (**Figure 5D, G**) as shown previously for patient tumors[239]. A similar analysis was carried out for the other deregulated myogenic pathways (**Figures 4-7, 4-8,** and **4-9**).

**Identification of an RMS vulnerability through integrated analysis**

It is not feasible to simultaneously target the 6 myogenic signal transduction pathways with molecular targeted therapeutics in a clinically relevant manner. Therefore, we focused on identifying a common regulatory mechanism such as a master transcription factor, kinase or phosphatase, protease or chaperone that also showed RMS selective drug sensitivity in our large high throughput screening drug database from the Childhood Solid Tumor Network[193]. Using this approach, we identified the HSP90 chaperone as an RMS specific vulnerability. Under normal conditions, an abundance of HSP90 protects cells from stress[240-242]. However, under disease conditions such as malignant transformation, the HSP90 reservoir is depleted and further stress such as chemotherapy can overload the protein homeostasis system, leading to cell death[240-242]. Many bona fide HSP90 clients are involved in signal transduction including key regulators of the WNT, HH, BMP, adenyl cyclase, p38/MAPK and PI3K pathways (**Figure 4-10A**)[241 243,244]. In addition, the transcription factor HSF1 has increased protein expression, increased phosphorylation and increased expression of its target genes in

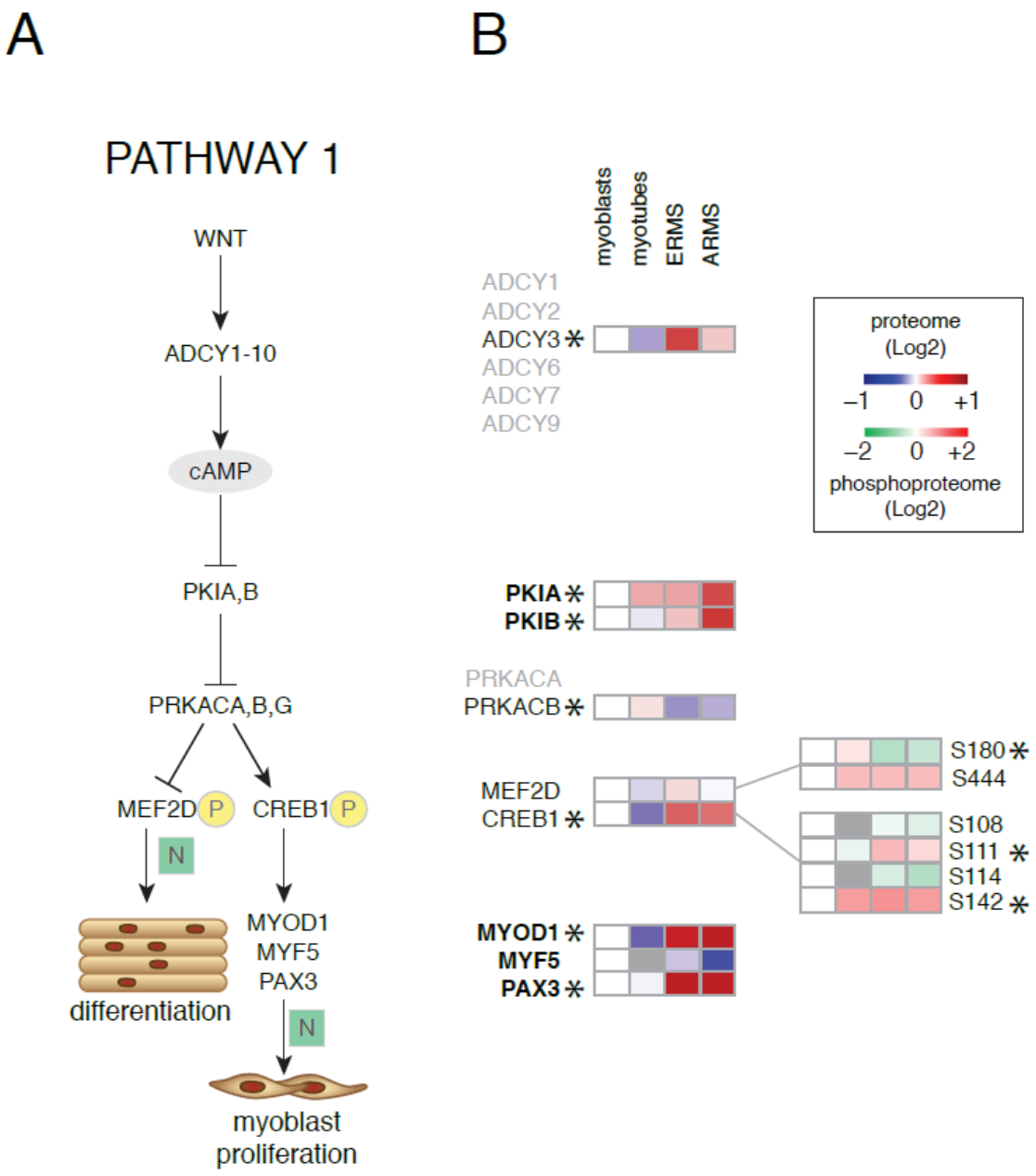



**Figure 4-7. Deregulation of WNT pathway for muscle development.**
A) WNT pathways for muscle development. Arrows indicate interactions that promote activity and bars represent interactions that block activity. The green boxed N represents perturbations in transcriptional targets. (B) The individual family members that are expressed in the cohort analyzed in this study are shown. The cutoff for inclusion in (B) is > FPKM of 1.0. Genes shown in gray are between FPKM of 1.0 and 10.0. Those in bold vary by more than 5-fold across samples in our analysis. Heatmap of the protein and phosphoproteins normalized to myoblasts. The intensity is proportional to the Log2 of protein or phosphoprotein expression relative to myoblasts. The asterisk indicates those proteins that have statistically significant differences across samples.

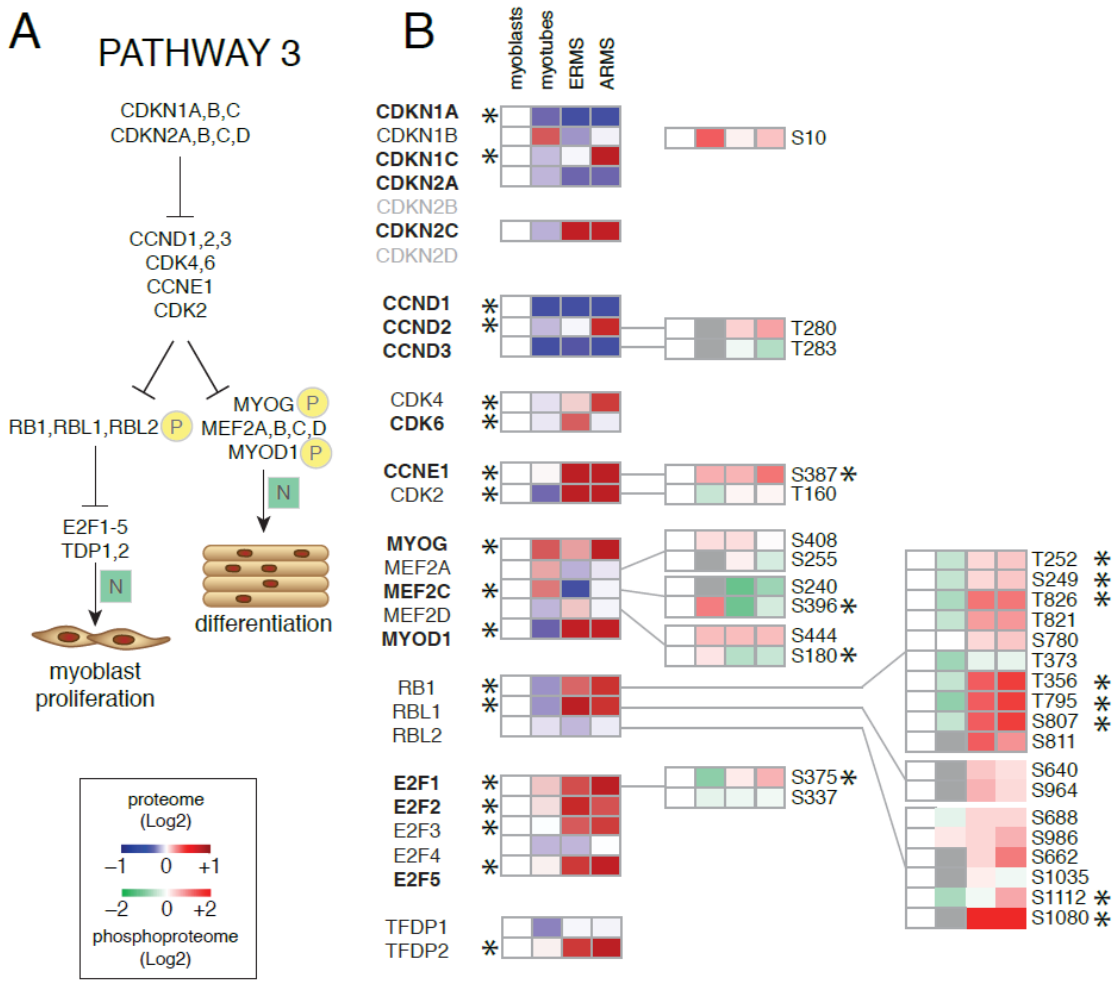



**Figure 4-8. Deregulation of adenyl cyclase pathway for muscle development.** A) Adenyl cyclase pathways for muscle development. Arrows indicate interactions that promote activity and bars represent interactions that block activity. The green boxed N represents perturbations in transcriptional targets. (B) The individual family members that are expressed in the cohort analyzed in this study are shown. The cutoff for inclusion in (B) is > FPKM of 1.0. Genes shown in gray are between FPKM of 1.0 and 10.0. Those in bold vary by more than 5-fold across samples in our analysis. Heatmap of the protein and phosphoproteins normalized to myoblasts. The intensity is proportional to the Log2 of protein or phosphoprotein expression relative to myoblasts. The asterisk indicates those proteins that have statistically significant differences across samples.

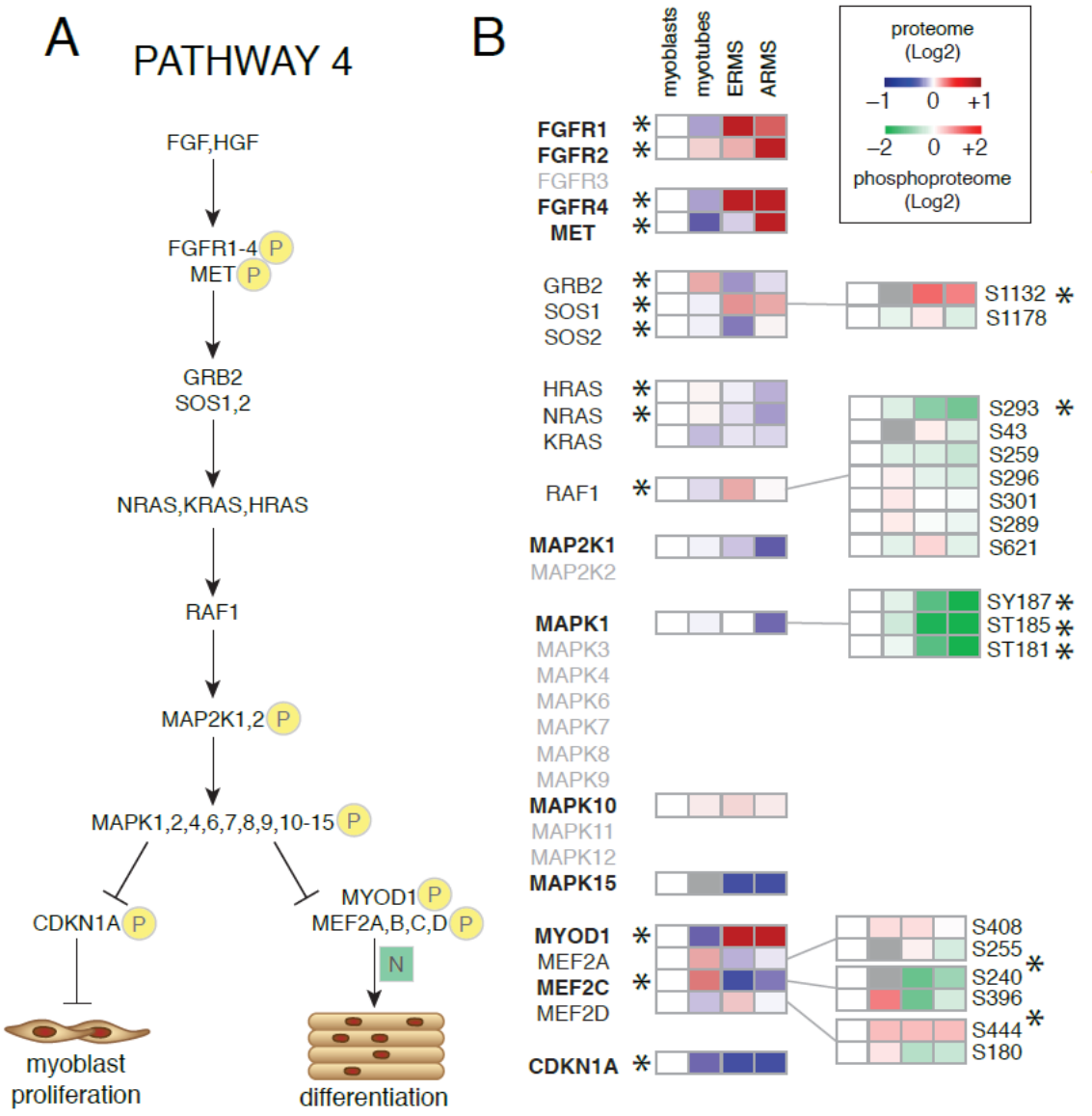



**Figure 4-9. Deregulation of MAPK pathway for muscle development.**
A) MAPK pathways for muscle development. Arrows indicate interactions that promote activity and bars represent interactions that block activity. The green boxed N represents perturbations in transcriptional targets. (B) The individual family members that are expressed in the cohort analyzed in this study are shown. The cutoff for inclusion in (B) is > FPKM of 1.0. Genes shown in gray are between FPKM of 1.0 and 10.0. Those in bold vary by more than 5-fold across samples in our analysis. Heatmap of the protein and phosphoproteins normalized to myoblasts. The intensity is proportional to the Log2 of protein or phosphoprotein expression relative to myoblasts. The asterisk indicates those proteins that have statistically significant differences across samples

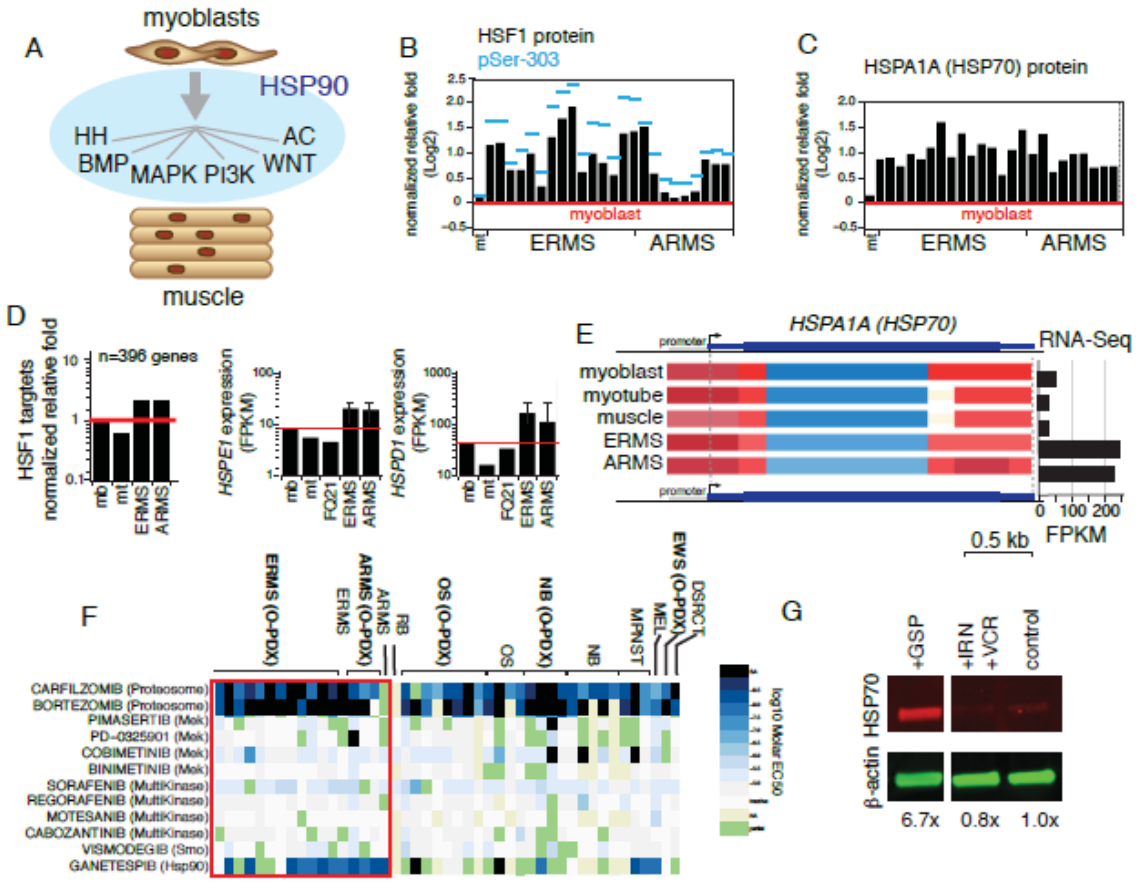



**Figure 4-10. HSP90 is a therapeutically relevant vulnerability in RMS.**
(A) Drawing of the 6 myogenic pathways that are deregulated in RMS that involve HSP90 regulation. (B,C) Histograms of HSF1 and HSP70 protein (blue lines) and phosphoprotein (black bars) expression across RMS samples relative to myoblasts (red line). (D) Quantitative expression of HSF1 target genes in RMS normalized relative to myoblasts as measure by RNA-Seq. (E) ChromHMM of HSP70 and corresponding gene expression by RNA-Seq (FPKM). (F) Heatmap of drug EC50 for a subset of the drugs tested in this study. The samples in bold are primary cultures of O-PDX tumors and those that are non-bold are established cell lines. (G) Immunoblot for HSP70 for the RMS RD cell line following treatment with ganetespib (GSP), irinotecan with vincristine (IRN+VCR) or untreated. The intensity of was quantitated and normalized to actin to calculate the fold difference for the treated samples relative to the control untreated sample. These data were used to calculate the MED by comparison to plasma levels for GSP from patients (see Supplemental Information). Abbreviations: GSP, ganetespib; IRN, irinotecan; VCR, vincristine; MED, murine equivalent dose.

RMS (**Figure 4-10B, C, D,** and **E**). Indeed, hyperphosphorylation of HSF1 is a hallmark of cellular stress (**Figure 4-10B**)[245]. Importantly, RMS cell lines and primary cultures of RMS O-PDXs are more sensitive to an HSP90 inhibitor (ganetespib) than other pediatric solid tumors (**Figure 4-10F**) and treatment of an RMS cell line (RD) led to rapid induction of HSP70 and other HSF1 target genes (**Figure 4-10G**). These data are consistent with recently published data showing that HSP70 is a vulnerability in RMS[246] and our unbiased proteome/phosphoproteome data indicating that the unfolded protein response which includes HSP90, was one of the most significantly deregulated pathways in RMS relative to normal developing muscle (**Figure 4-5A**).

To determine if the HSP90 inhibitor, ganetespib could potentiate the cytotoxic activity of conventional chemotherapeutic agents used to treat recurrent RMS or kinase inhibitors that are relevant to RMS, we performed combination drug screening experiments. Briefly, a set of 156 drugs were plated in 10-concentration dose response in triplicate and tested in combination with several different concentrations of ganetespib (100 nM, 32.4 nM, 10.8 nM and 3.5 nM). The drug combinations were tested on an RMS cell line (RD) and an RMS O-PDX (SJRHB00026_X1) in biological duplicates. Using the BRAID algorithm, we found that ganetespib could potentiate the killing of RMS cells lines and O-PDX tumors with conventional chemotherapeutic agents used to treat RMS.

To establish a clinically relevant murine equivalent dose for ganetespib (GSP), we performed pharmacokinetic studies of O-PDX tumor-bearing mice. We measured the plasma levels at 10 minutes, 1, 4, 8 and 16 hours after injection and then used the area under the concentration time curve (AUC) to calculate a murine equivalent dose matching the exposure in patients. We also measured the tumor penetration of GSP to relate exposure *in vivo* to sensitivity *in vitro*. The sustained levels of GSP in vivo were above those required in culture to potentiate the cytotoxic effect of conventional chemotherapeutic agents used to treat recurrent RMS (irinotecan (IRN) and vincristine (VCR)).

Preclinical phase I studies were then performed on CD1 non-tumor–bearing mice to establish the tolerability of GSP in combination with VCR and IRN at clinically relevant doses and schedules as described previously[247,248]. Briefly, we administered 4 courses (3 weeks per course) of therapy for 2 treatment groups. One treatment group had the low dose protracted schedule of IRN (daily x 5 x 2; equivalent to 200 mg/m2 total in patients) and the other group had a shorter schedule of IRN (daily x 5; equivalent to 250 mg/m2 total in patients). Both of these protracted schedules of IRN are commonly used to treat children with RMS and other solid tumors[249]. We measured body weight weekly and CBC-D with each course and we performed a necropsy at the end of the study. We found that at the MED of GSP was well tolerated in this triple drug combination in both treatment groups.

## Discussion

With the successful launch of the NCI-MATCH and other precision medicine trials an increasing number of cancer patients are receiving personalized treatment regimens based on the molecular characterization of their tumors. While several trials have demonstrated the feasibility of precision medicine for children with cancer[250,251], there are multiple challenges that are unique to pediatric cancer. First, the number of patients is small so molecular targeted therapy is prescribed based on the genetic lesion irrespective of the cancer type. If the cellular context of the targeted mutation impacts the effect of molecular targeted therapy, the responses will be difficult to interpret. Second, there are fewer actionable mutations in pediatric tumors that are currently druggable with available therapeutics making it difficult to accrue enough patients to provide statistically significant results. Taken together, these barriers will make it difficult to discern the benefit of precision medicine for children with cancer from clinical trials alone and complementary efforts in preclinical studies may help to identify the patients most likely to benefit from such interventions.

In this study, we sought to study one of the most common potentially druggable mutations in pediatric cancer—RAS mutant RMS. Based on previous genomic studies, approximately 50-75% of intermediate and high-risk RMS patients possess a mutation in the RAS pathway. In some adult cancers, oncogenic mutations in the RAS pathway promote tumorigenesis in part by increasing expression and activity of cyclin D/CDK4/6[252]. Therefore, in the pediatric NCI-MATCH trial that is under development, RAS mutant tumors may be treated with a CDK4/6 inhibitor (palbociclib) alone or in combination with an upstream inhibitor of the MEK kinase (trametinib). Consistent with this proposed mechanism, Dr. Dyer's group showed that RMS tumors have robust signaling through the RAS pathway leading to hyperphosphorylation of RB and deregulated proliferation (**Figure 4-1**). While the combination of CDK4/6 inhibitors (palbociclib and abemaciclib) with trametinib led to synergistic killing of RMS cells in culture, they failed to achieve efficacy *in vivo*. The pharmacokinetic and pharmacodynamics studies suggested that this was not due to a failure of the drugs to penetrate the tumor and alter target activity but rather compensatory changes in other signal transduction pathways that allow the tumors to continue to grow.

These disappointing results led us to consider the alternative hypothesis that signal transduction and transcriptional networks may be significantly different in pediatric solid tumors as compared to adult tumors because of their developmental origins. The dynamic and robust signaling that takes place during development may by co-opted in pediatric solid tumors making it much more difficult to target any single oncogenic pathway such as RAS.

In order to gain a deeper understanding of the transcriptional networks and the signal transduction pathways in RMS and to compare them to normal human muscle, we performed integrated epigenetic and proteomic/phosphoproteomic profiling of O-PDX RMS tumors and human myoblasts, myotubes and fetal skeletal muscle. The epigenetic analysis was instrumental in precisely mapping the developmental stage for each tumor

and the proteomic/phosphoproteomic profiling was used to identify deregulated signal transduction pathways. Taken together, using our newly developed bioinformatics pipelines, we showed that at least 6 pathways required for myogenesis were deregulated in RMS during the transition from myogenic determination to differentiation. It is not feasible to target each of these pathways simultaneously so we sought other more fundamental cellular pathways that are required to maintain this complex signaling milieu. We discovered that the unfolded protein response (UPR) is particularly important in RMS cells and HSP90 in particular is a key regulator of UPR. While HSP90 and UPR are deregulated in many cancers, our data suggest that it is a particular vulnerability in RMS due to the complex signaling pathways as a result of their developmental origins. We performed comprehensive pharmacokinetics and preclinical testing of an HSP90 inhibitor (ganetespib) and demonstrated efficacy in vivo.

Importantly, targeting HSP90 was efficacious across multiple aggressive and recurrent RMS tumors irrespective of their genetic lesions. These data suggest that for some tumor types, targeting fundamental pathways vulnerabilities may provide a greater anti-tumor effect than individualized molecular targeted therapy based on the somatic mutations found in the tumor.

**Elucidating the core signal transduction networks from proteomic and phosphoproteomic profiling of RMS**

Recent advances in proteomics has led to remarkable new insights into the complexity of the proteome and phosphoproteome in cancer and other cell types. In this study, we further advanced our understanding of the proteome by incorporating 10-plex isobaric labelling (TMT), extensive 2D LC peptide separation, high resolution mass spectrometer and comprehensive bioinformatics tools[148-151,201]. This powerful quantitative strategy enabled ultradeep proteome and phosphoproteome profiling with high throughput and reproducibility. Although quantitative ratio compression often occurs with MS/MS based isobaric labelling[21,253], experimental variations are also reduced, and therefore it has only a minor impact on differential expression analyses[120], moreover, the extensive peptide separation [150] coupled with biological replicates[130,201] facilitates statistical inference and largely diminishes the effect of ratio compression.

In our study, the proteomic and phosphoproteomic data were essential for interpreting the signal transduction networks in RMS and identifying an RMS specific tumor vulnerability. While we could infer the transcriptional outcome of several key myogenic signal transduction pathways in RMS from the RNA-Seq and epigenomic profiling, those data are very difficult to interpret because of complex interplay between multiple transcription factors at particular target genes. However, by integrating the proteomic and phosphoproteomic data, we were able to more accurately establish the status of the upstream signal transduction pathways and then validate those signaling networks by tracking key downstream transcriptional networks. Thus, by integrating the proteomic/phosphoproteomic data with the gene expression and epigenetic data, we increased our ability to advance our understanding of the developmental origins of RMS

and the activity of signal transduction pathways that are deregulated in these developmental tumors.

It is important to emphasize that while we used an unbiased approach to identify the deregulated pathways in RMS initially, it was essential to refine those pathway analyses in the context of the cellular lineage. That is, we refined the generic pathway gene and protein lists from the unbiased approach to incorporate the specific signaling molecules at each developmental stage in myogenesis. This type of manual analysis based on the extensive published literature on myogenesis was crucial for elucidating the 6 deregulated pathways in RMS and was made possible with the extensive quantitative proteomic and phosphoproteomic data in our study. Our data suggest that the complex signal transduction pathways that are altered in RMS reflect the key developmental stage when the tumors transition from myogenic determination to myogenic differentiation. For example, some tumor cells are arrested before the transition to myogenic differentiation and express higher levels of MYF5 and lower levels of MYOG. Other tumor cells exhibited further progression and expressed low levels of MYF5 and higher levels of MYOG. MYOD1, PAX3 and PAX7 were expressed more consistently across the individual tumors and showed no significant correlation with other developmental genes/proteins. An alternative interpretation is that both ERMS and ARMS tumors arise from muscle satellite cells and the transformation occurs at different stages along the pathway from satellite cell to myogenic differentiation. In either case, the ERMS tumors expressing high levels of MYF5 and low levels of MYOG are less mature than the ARMS tumors and the subset of ERMS tumors with low MYF5 and higher MYOG.

The proteomic/phosphoproteomic data also facilitated identification of deregulated pathways that may be targeted with novel therapeutic combinations. Beyond the developmental pathways, we identified DNA repair, G2/M checkpoint and mitotic spindle pathways among the most significantly altered pathways in RMS along with the UPR pathway. The DNA repair, G2/M and mitotic spindle pathways were exploited with molecular targeted therapy targeting the WEE1 kinase in a separate study (Stewart et al., in revision). The UPR pathway was exploited in this study with the HSP90 inhibitor. Under normal conditions, an abundance of HSP90 protects cells from stress. However, under extreme conditions, the HSP90 reservoir is rapidly depleted, leading to cell death. Many bona fide HSP90 clients are involved in signal transduction[241,243,244]. In RMS, HSP90 clients are significantly enriched, and HSP90 is upregulated. For example, when cells experience stress, the HSF-1 transcription factor is released from HSP90 and activates transcription of target genes such as HSP70[245]. Hyperphosphorylation of HSF1 is an indicator of cellular stress[220], and we found that HSF1 is hyperphosphorylated in RMS relative to myoblasts. Moreover, the direct target genes of HSF1, such as HSP70, are upregulated. Taken together, our data suggest that this protein-homeostasis pathway is an RMS vulnerability due in part to the burden on the HSP90 system of maintaining the 6 myogenic signal transduction pathways in RMS[242,254,255]. That is, the HSP90 levels are barely sufficient to manage the cellular stress in RMS and by increasing cellular stress with chemotherapy we may achieve synthetic lethality with low levels of HSP90 inhibitors such as GSP. Indeed, our drug sensitivity data demonstrated that RMS cell lines and O-PDX tumors are more sensitive to GSP than are other pediatric solid tumors.

In fact, the combination of GSP with chemotherapeutics at subtherapeutic doses that have no effect on their own led to synergistic killing of RMS cells in culture and in *vivo*. This is a particularly exciting result because another research group recently published a study confirming that the protein homeostasis pathway is a fundamental vulnerability in RMS[246].

**Implications for precision medicine**

Large-scale cancer genomics projects have revealed an unprecedented high-resolution map of the mutational landscape of diverse cancer types, ranging from common adult cancers to rare pediatric cancers. In parallel with these advances in cancer genomics, there is an expanding collection of molecularly targeted agents that have proven effective in various cancers. For some, an individualized treatment plan based on validated molecular diagnostic tests can be lifesaving such as the treatment of gastrointestinal stromal tumors with imatinib[256]. With the enormous demand for more effective and less toxic anticancer treatments, there are now several major efforts to aggressively expand genomic-based precision medicine in oncology. Indeed, the National Cancer Institute recently launched the Molecular Analysis for Therapy Choice (NCI-MATCH) trial for the treatment of recurrent or refractory adult solid tumors, and this initiative will soon be expanded to include pediatric solid tumors. However, there are many challenges in applying precision genomic based medicine for pediatric solid tumors. First, there are very few druggable mutations. Second, the NCI-MATCH trials are focused primarily on single drug treatment regimens with multiply recurrent or refractory disease. Thus, it is unlikely there will be improvement in outcomes for these heavily pretreated patients. Third, the match for each drug will be applied based on genomic data irrespective of tumor type. This is challenging because the oncogenic drivers and/or redundant/compensatory pathways may be dramatically different across diverse pediatric cancer types. Finally, there is very limited scientific justification or translational relevance for most of the molecular targeted agents used in the pediatric NCI-MATCH trial.

Our study provides an alternative to genomic-based precision medicine that is built on a foundation of comprehensive integration of diverse datasets and robust scientific justification and translational relevance for a new treatment regimen. We showed that the NCI-MATCH treatment regimen that may be used for recurrent RMS patients with oncogenic RAS mutations (CDK4/6 inhibitor+MEK inhibitor) is not effective in validated O-PDX models of recurrent RMS irrespective of RAS mutation status. Indeed, integrated analysis showed that at least 6 signal transduction pathways that are essential for myogenesis are active in RMS and this corresponds with the developmental stage when the myogenic program is halted in rhabdomyosarcomas. Drugs that target those individual pathways are not effective nor are combinations that target 2 pathways simultaneously. It is not feasible to attempt to target each of those pathways simultaneously so we turned to our integrated database to identify an upstream master regulatory mechanism that may be required across the diverse pathways. Our discovery that HSP90 is required for signal transduction through those pathways

provided an opportunity to take advantage of a unique vulnerability in rhabdomyosarcoma. More importantly, this vulnerability was present across RMS tumors in our analysis from intermediate and high risk ARMS and ERMS from diagnosis and recurrence. We provided scientific justification and translational relevance for targeting this pathway in RMS and the optimal treatment regimen is based on standard treatment for recurrent disease. This alternative of identifying fundamental vulnerabilities in a particular pediatric tumor type is a rational alternative to single agent, genomic based precision medicine across pediatric solid tumor histology that is proposed for pediatric NCI-MATCH. Importantly, this is an approach that can be readily applied to diverse pediatric solid tumors including very rare tumors because of the resources now available to the biomedical research community through the CSTN. By identify fundamental vulnerabilities that can be targeted with combinations of molecular targeted therapeutics and conventional chemotherapy that lead to synergistic killing as shown here or synthetic lethality as shown for Ewing sarcoma and providing scientific justification and translational relevance, we can advance new treatment regimens that have a much better chance of improving outcomes for children with solid tumors.

# CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

## Conclusions

### We established a robust proteomic analysis platform that enables near complete human proteome analysis

We have demonstrated a reverse phase-based, multidimensional long gradient LC-MS/MS platform suitable for deep proteomics analysis. We systematically examined and optimized various parameters of a 100 μm x 150 cm LC column packed with 5 μm reverse phase C18 beads. The column exhibits great robustness and reproducibility together with high peak capacity (~700) and loading capacity (optimal at 6 μg). Using this column in conjunction with basic pH LC and Q Exactive MS, the identification of a deep proteome of AD brain (>10,000 proteins) was achieved in about 4 days of MS instrument time.

### We developed a pipeline for accurate proteome quantification with high-throughput and genome-scale coverage

Multiplex isobaric labeling provides an efficient mass spectrometry technology for quantitative proteomics, but a common limitation of ratio compression leads to quantitative inaccuracy and often constrains its application. We demonstrate that the optimization of LC/LC-MS/MS settings, in combination of *y*1 ion-based post-MS correction, is capable of virtually removing the effect of interference and substantially enhancing the precision of measurements. The extensive LC/LC fractionation also allows deep profiling of proteome and protein modifications. Although we only analyzed TMT-labeled samples in this study, the principle of *y*1 ion-based correction is anticipated to be applicable to all other isobaric labeling approaches for analyzing a trypsinized proteome.

### We defined the cancer proteotypes, which successfully filled the gap between genotypes and phenotypes in different mouse HGGs

Using our novel proteomics pipelines, we analyzed 13,860 proteins (11,941 genes) and 30,431 phosphosites, representing the deepest proteomics study in a single mass spectrometry experiment. Two high-grade glioma (HGG) mouse models driven by mutated receptor tyrosine kinase (RTK) oncogenes, platelet-derived growth factor receptor alpha (*PDGFRA*) and neurotrophic receptor tyrosine kinase 1 (*NTRK1*) were analyzed and showed distinct global proteome and phosphoproteome landscapes. These proteome and phosphoproteome data showed that NTRK fusion genotype induces stronger deregulation of HGG pathways and drives a positive feedback loop compare to PDGFRA mutation genotype in mouse HGGs. Pathway activity computation, *in vitro* Ki-67 cell proliferation index, and *in vivo* mouse survival curve confirmed that HGG driven

by the *NTRK1* genotype has more severe HGG phenotypes than the *PDGFRA* genotype, and demonstrated that the *NTRK1* fusion has stronger oncogenic potency than the *PDGFRA* mutation. Together, these results present a new paradigm of the newly developed proteomics profiling techniques to successfully define cancer proteotypes, filling the gap between genotypes and phenotypes in cancers.

**We developed a bioinformatics pipeline to prioritize master regulators in cancer through integrating deep multi-omics data**

Massive reprogramming of molecular components occurs during the evolution from mortal to immortal status in cancer cells. As improvement of profiling technologies allows the identification of thousands of these changes, prioritizing drivers and core regulators from the enormous amount of passenger changes becomes a rate-limiting step. Here we presented a generic bioinformatics pipeline for prioritizing core signaling networks and master regulators in cancer proteomics studies. 10 proteome and phosphoproteome clusters were extracted from 4,703 differentially expressed proteins and 6,768 differentially phosphorylated phosphosites through weighted co-expression clustering analysis, which dramatically reduced the data complexity. This readily identified major pathways with well-established roles in glioma growth as well as clear connections with PI3K and mTOR signaling downstream of RTK activation. Subsequently, co-regulated genes in each of the clusters were summarized to pathways and networks using the network analysis method, which further narrowed down these massive changes to 67 network modules. We also developed systematic protein activity inference strategies for kinases and transcription factors by integrating multi-omics data and a variety of databases to further prioritize 41 kinases, 26 TFs, and a core network consisting of 13 master regulators from these gene clusters and network modules. Finally, we overlapped the most significantly altered omics datasets in mouse HGGs back to human transcriptome data to search for consistent alterations driven by NTRK and PDGFRA mutations across species, resulting in a list of 20 convergent alterations in mouse and human. Indeed, many of the prioritized networks (e.g. AMPK-EEF2K) and proteins (e.g. EPHA2, CD74) are reported to be functional in HGGs. Together, we demonstrated that our multi-omics integrative pipeline successfully extracted master regulators from thousands of passenger changes. Many of them being well-reported master regulators involved in HGG confirmed the robustness of our bioinformatics pipeline to handle big omics data in cancer.

**Integrative multi-omics analysis using our newly developed techniques and tools identified therapeutic vulnerabilities in rhabdomyosarcoma**

Encouraged by the success achieved on mouse HGGs analysis, we further advanced the application of these tools and technologies on large-scale, human specimen analysis on RMS O-PDX samples. 6 deregulated pathways in RMS that are essential for myogenesis were elucidated through comprehensive analysis of the proteome and phosphoproteome data. By integrating the proteomic/phosphoproteomic data with the

gene expression and epigenetic data, we advanced our understanding of the developmental origins of RMS and the activity of signal transduction pathways that are deregulated in these developmental tumors. These data aided the identification of a common tumor vulnerability (i.e. HSP90) in RMS. Using a combination of the HSP90 inhibitor GSP and chemotherapeutics reagents (i.e. VCR and IRN), we showed that the combination of GSP with chemotherapeutics at subtherapeutic doses that have no effect on their own led to synergistic killing of RMS cells in culture and *in vivo*.

## Future Directions

### The newly developed proteomic analysis pipeline is applicable to other complex biological systems

With the advances of genomic sequencing technologies, numerous comprehensive genomic landscapes have been generated for cancers. While we already know the oncogenic genome alterations for most cancers, how these cancer drivers lead to different proteotypes which end up with distinct cancers remains largely unclear. In this work, we presented a new paradigm of systems cancer proteome analysis. The strengths of this generic pipeline were exemplified by the applications on two cancer studies for defining cancer proteotype, revealing signaling networks and master regulators, and identifying cancer vulnerabilities. Although we only showed the cases on cancers, the principle of this approach is anticipated to be applicable to all other complex biological systems. Indeed, applications of our techniques on several other biological systems have been successful[110,205,207,208,257]. Since major genomic sequencing projects such as TCGA and PCGP have been largely accomplished, there are urgent demands of our techniques to explore cancer proteomic landscapes, we would like to continue applying this powerful technique on exploring the global proteomes of other cancers

### To provide a dynamic view of the molecular circuits, by which the targeted therapeutic strategy kills the rhabdomyosarcoma cells, through a time-resolved proteomic analysis

Although the large-scale proteomic analysis on rhabdomyosarcoma provided a comprehensive proteomic landscape of RMS. It only captured a snapshot of the global proteomic changes, there is no time-resolved information on the molecular circuits by which the oncogenomic alterations induce signaling transductions, activate downstream transcriptional reprograms, and result in systematic reprogramming of global cancer proteomes. To provide an enhanced mechanistic understanding of the oncogenic process and the molecular responses of cancer to drug treatments to overcome tumor recurrence , for the next step of the rhabdomyosarcoma project, we would like to study the time points resolved responses of rhabdomyosarcoma to the novel treatment strategies using GSP+VCR +IRN combination

**To identify cancer-derived proteins in the serum of xenograft-bearing rhabdomyosarcoma mice**

Most of established clinical cancer biomarkers are proteins produced by tumor cells[258]. Indeed, proteins that derived from cancer cells are considered as the best biomarkers to determine the status, size, and progression of the tumor[258,259]. Currently, the widely used cancer biomarker discovery protocols use comparisons of body fluid (e.g. serum) from cases with and without disease using mass spectrometry technologies[259,260]. While the strength of these approaches are limited by the fact that proteins detected are less likely to be produced by cancer cells, instead it is more likely to be introduced by secondary body defense mechanisms. The cancer xenograft model, in which human rhabdomyosarcoma are generated in an immune-deficient nude mice, can essentially eliminate these problems because proteins that presented in mouse serum that are identified to be human protein will, by definition, originate from human cancer cells. One inherent challenge of biomarker discovery study is the high dynamic range of proteins presented in serum. Fortunately, our novel deep proteomic profiling techniques with extraordinary sensitivity will largely diminish this problem. Moreover, the cancer proteome reference has already been produced through our large-scale proteomic landscape study on RMS tumors. Together, the RMS O-PDX tumor provides a perfect model and a unique opportunity to explore cancer biomarker discovery using our advanced MS-based proteomic pipeline.

## LIST OF REFERENCES


1 MacBeath, G. Protein microarrays and proteomics. *Nature genetics* **32 Suppl**, 526-532, doi:10.1038/ng1037 (2002).

2 Causier, B. Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass spectrometry reviews* **23**, 350-367, doi:10.1002/mas.10080 (2004).

3 Stevens, R. C., Yokoyama, S. & Wilson, I. A. Global efforts in structural genomics. *Science (New York, N.Y.)* **294**, 89-92, doi:10.1126/science.1066011 (2001).

4 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).

5 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207, doi:10.1038/nature01511 (2003).

6 Bensimon, A., Heck, A. J. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annual review of biochemistry* **81**, 379-405, doi:10.1146/annurev-biochem-072909-100424 (2012).

7 Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Molecular cell* **49**, 583-590, doi:10.1016/j.molcel.2013.01.029 (2013).

8 Altelaar, A. F., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews. Genetics* **14**, 35-48, doi:10.1038/nrg3356 (2013).

9 Mitchell, P. Proteomics retrenches. *Nature biotechnology* **28**, 665-670, doi:10.1038/nbt0710-665 (2010).

10 Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **80**, 273-299, doi:10.1146/annurev-biochem-061308-093216 (2011).

11 Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics* **3**, 1154-1169 (2004).

12 Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895-1904 (2003).

13 Frost, D. C., Greer, T. & Li, L. High-Resolution Enabled 12-Plex DiLeu Isobaric Tags for Quantitative Proteomics. *Anal. Chem.* **87**, 1646-1654, doi:10.1021/ac503276z (2015).

14 Rauniyar, N. & Yates, J. R., 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* **13**, 5293-5309, doi:10.1021/pr500880b (2014).

15 Chick, J. M. *et al.* Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**, 500-505, doi:10.1038/nature18270 (2016).

16 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

17 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).
18 Bantscheff, M. *et al.* Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol. Cell Proteomics* **7**, 1702-1713, doi:10.1074/mcp.M800029-MCP200 (2008).
19 Karp, N. A. *et al.* Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell Proteomics* **9**, 1885-1897, doi:10.1074/mcp.M900628-MCP200 (2010).
20 Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.* **8**, 5347-5355, doi:10.1021/pr900634c (2009).
21 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937-940, doi:10.1038/nmeth.1714 (2011).
22 Niu, M. *et al.* Extensive Peptide Fractionation and y1 Ion-based Interference Detection Enable Accurate Quantification by Isobaric Labeling and Mass Spectrometry. *Analytical chemistry*, doi:10.1021/acs.analchem.6b04415 (2017).
23 Olsen, J. V. *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648, doi:10.1016/j.cell.2006.09.026 (2006).
24 Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *European journal of biochemistry* **268**, 5001-5010 (2001).
25 Ptacek, J. *et al.* Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679-684, doi:http://www.nature.com/nature/journal/v438/n7068/suppinfo/nature04187_S1.html (2005).
26 Songyang, Z. *et al.* Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Current biology : CB* **4**, 973-982 (1994).
27 Hanash, S. & Taguchi, A. The grand challenge to decipher the cancer proteome. *Nat Rev Cancer* **10**, 652-660, doi:10.1038/nrc2918 (2010).
28 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).
29 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).
30 Romero, R. *et al.* The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *BJOG : an international journal of obstetrics and gynaecology* **113 Suppl 3**, 118-135, doi:10.1111/j.1471-0528.2006.01150.x (2006).
31 McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. *Molecular bioSystems* **11**, 2680-2689, doi:10.1039/c5mb00310e (2015).
32 Tang, Y. C. & Amon, A. Gene copy-number alterations: a cost-benefit analysis. *Cell* **152**, 394-405, doi:10.1016/j.cell.2012.11.043 (2013).

33 Wethmar, K., Smink, J. J. & Leutz, A. Upstream open reading frames: molecular switches in (patho)physiology. *BioEssays : news and reviews in molecular, cellular and developmental biology* **32**, 885-893, doi:10.1002/bies.201000037 (2010).
34 Barrett, L. W., Fletcher, S. & Wilton, S. D. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences : CMLS* **69**, 3613-3634, doi:10.1007/s00018-012-0990-9 (2012).
35 Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550, doi:10.1016/j.cell.2016.03.014 (2016).
36 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics* **13**, 227-232, doi:10.1038/nrg3185 (2012).
37 Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews. Molecular cell biology* **11**, 427-439, doi:10.1038/nrm2900 (2010).
38 Drake, J. M. *et al.* Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell* **166**, 1041-1054, doi:10.1016/j.cell.2016.07.007 (2016).
39 Jones, C. & Baker, S. J. Unique genetic and epigenetic mechanisms driving paediatric diffuse high-grade glioma. *Nat Rev Cancer* **14**, doi:10.1038/nrc3811 (2014).
40 Chen, J., McKay, R. M. & Parada, L. F. Malignant glioma: lessons from genomics, mouse models, and stem cells. *Cell* **149**, 36-47, doi:10.1016/j.cell.2012.03.009 (2012).
41 Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta neuropathologica* **131**, 803-820, doi:10.1007/s00401-016-1545-1 (2016).
42 Network, T. C. Corrigendum: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **494**, 506, doi:10.1038/nature11903 (2013).
43 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).
44 Huang, P. H., Xu, A. M. & White, F. M. Oncogenic EGFR signaling networks in glioma. *Science signaling* **2**, re6, doi:10.1126/scisignal.287re6 (2009).
45 Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)* **321**, 1807-1812, doi:10.1126/science.1164382 (2008).
46 Paugh, B. S. *et al.* Novel oncogenic PDGFRA mutations in pediatric high-grade gliomas. *Cancer research* **73**, 6219-6229, doi:10.1158/0008-5472.can-13-1491 (2013).
47 Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nature genetics* **44**, 251-253, doi:10.1038/ng.1102 (2012).

48 Wu, G. *et al.* The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nature genetics* **46**, 444-450, doi:10.1038/ng.2938 (2014).
49 Johnson, H. & White, F. M. Quantitative analysis of signaling networks across differentially embedded tumors highlights interpatient heterogeneity in human glioblastoma. *Journal of proteome research* **13**, 4581-4593, doi:10.1021/pr500418w (2014).
50 Pappo, A. S. *et al.* Survival after relapse in children and adolescents with rhabdomyosarcoma: A report from the Intergroup Rhabdomyosarcoma Study Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **17**, 3487-3493, doi:10.1200/jco.1999.17.11.3487 (1999).
51 Newton, W. A., Jr. *et al.* Histopathology of childhood sarcomas, Intergroup Rhabdomyosarcoma Studies I and II: clinicopathologic correlation. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **6**, 67-75, doi:10.1200/jco.1988.6.1.67 (1988).
52 Barr, F. G. Chromosomal translocations involving paired box transcription factors in human cancer. *The international journal of biochemistry & cell biology* **29**, 1449-1461 (1997).
53 Raney, R. B. *et al.* Rhabdomyosarcoma and undifferentiated sarcoma in the first two decades of life: a selective review of intergroup rhabdomyosarcoma study group experience and rationale for Intergroup Rhabdomyosarcoma Study V. *Journal of pediatric hematology/oncology* **23**, 215-220 (2001).
54 Williamson, D. *et al.* Fusion gene-negative alveolar rhabdomyosarcoma is clinically and molecularly indistinguishable from embryonal rhabdomyosarcoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 2151-2158, doi:10.1200/jco.2009.26.3814 (2010).
55 Scrable, H. *et al.* A model for embryonal rhabdomyosarcoma tumorigenesis that involves genome imprinting. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 7480-7484 (1989).
56 Ognjanovic, S., Linabery, A. M., Charbonneau, B. & Ross, J. A. Trends in childhood rhabdomyosarcoma incidence and survival in the United States, 1975-2005. *Cancer* **115**, 4218-4226, doi:10.1002/cncr.24465 (2009).
57 Shern, J. F. *et al.* Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors. *Cancer discovery* **4**, 216-231, doi:10.1158/2159-8290.cd-13-0639 (2014).
58 Chen, X. *et al.* Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer cell* **24**, 710-724, doi:10.1016/j.ccr.2013.11.002 (2013).
59 Alagesan, B. *et al.* Combined MEK and PI3K inhibition in a mouse model of pancreatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **21**, 396-404, doi:10.1158/1078-0432.ccr-14-1591 (2015).
60 Jokinen, E. & Koivunen, J. P. MEK and PI3K inhibition in solid tumors: rationale and evidence to date. *Therapeutic advances in medical oncology* **7**, 170-180, doi:10.1177/1758834015571111 (2015).

61 Haagensen, E. J., Kyle, S., Beale, G. S., Maxwell, R. J. & Newell, D. R. The synergistic interaction of MEK and PI3K inhibitors is modulated by mTOR inhibition. *British journal of cancer* **106**, 1386-1394, doi:10.1038/bjc.2012.70 (2012).
62 Zhou, F. *et al.* Genome-scale proteome quantification by DEEP SEQ mass spectrometry. *Nature communications* **4**, 2171, doi:10.1038/ncomms3171 (2013).
63 Xu, P., Duong, D. M. & Peng, J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. *Journal of proteome research* **8**, 3944-3950, doi:10.1021/pr900251d (2009).
64 Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Rev. Biomed. Eng.* **11**, 49-79, doi:10.1146/annurev-bioeng-061008-124934 (2009).
65 Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971-983 (2006).
66 Frattini, V. *et al.* The integrated landscape of driver genomic alterations in glioblastoma. *Nature genetics* **45**, 1141-1149, doi:10.1038/ng.2734 (2013).
67 Sturm, D. *et al.* Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer* **14**, 92-107, doi:10.1038/nrc3655 (2014).
68 Kwong, L. N. *et al.* Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma. *Nature medicine* **18**, 1503-1510, doi:10.1038/nm.2941 (2012).
69 Vora, S. R. *et al.* CDK 4/6 inhibitors sensitize PIK3CA mutant breast cancer to PI3K inhibitors. *Cancer cell* **26**, 136-149, doi:10.1016/j.ccr.2014.05.020 (2014).
70 Fleuren, E. D., Zhang, L., Wu, J. & Daly, R. J. The kinome 'at large' in cancer. *Nat Rev Cancer* **16**, 83-98, doi:10.1038/nrc.2015.18 (2016).
71 Puyol, M. *et al.* A synthetic lethal interaction between K-Ras oncogenes and Cdk4 unveils a therapeutic strategy for non-small cell lung carcinoma. *Cancer cell* **18**, 63-73, doi:10.1016/j.ccr.2010.05.025 (2010).
72 Low, T. Y. *et al.* Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell reports* **5**, 1469-1478, doi:10.1016/j.celrep.2013.10.041 (2013).
73 Yin, D. *et al.* Impact of NUDT15 polymorphisms on thiopurines-induced myelotoxicity and thiopurines tolerance dose. *Oncotarget*, doi:10.18632/oncotarget.14594 (2017).
74 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).
75 Giddings, J. C. Maximum number of components resolvable by gel filtration and other elution chromatographic methods. *Analytical chemistry* **39**, 1027-1028, doi:10.1021/ac60252a025 (1967).
76 Shen, Y. *et al.* Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. *Analytical chemistry* **77**, 3090-3100, doi:10.1021/ac0483062 (2005).
77 Wang, H. & Hanash, S. M. Increased throughput and reduced carryover of mass spectrometry-based proteomics using a high-efficiency nonsplit nanoflow parallel

dual-column capillary HPLC system. *Journal of proteome research* **7**, 2743-2755, doi:10.1021/pr700876g (2008).

78 Mellors, J. S. & Jorgenson, J. W. Use of 1.5-microm porous ethyl-bridged hybrid particles as a stationary-phase support for reversed-phase ultrahigh-pressure liquid chromatography. *Analytical chemistry* **76**, 5441-5450, doi:10.1021/ac049643d (2004).

79 Thakur, S. S. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & cellular proteomics : MCP* **10**, M110 003699, doi:10.1074/mcp.M110.003699 (2011).

80 Burgess, M. W., Keshishian, H., Mani, D. R., Gillette, M. A. & Carr, S. A. Simplified and efficient quantification of low-abundance proteins at very high multiplex via targeted mass spectrometry. *Molecular & cellular proteomics : MCP* **13**, 1137-1149, doi:10.1074/mcp.M113.034660 (2014).

81 Shi, T. *et al.* Long-gradient separations coupled with selected reaction monitoring for highly sensitive, large scale targeted protein quantification in a single analysis. *Analytical chemistry* **85**, 9196-9203, doi:10.1021/ac402105s (2013).

82 Senko, M. W. *et al.* Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. *Analytical chemistry* **85**, 11710-11714, doi:10.1021/ac403115c (2013).

83 Kocher, T., Swart, R. & Mechtler, K. Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Analytical chemistry* **83**, 2699-2704, doi:10.1021/ac103243t (2011).

84 Kocher, T., Pichler, P., Swart, R. & Mechtler, K. Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients. *Nature protocols* **7**, 882-890, doi:10.1038/nprot.2012.036 (2012).

85 Hsieh, E. J., Bereman, M. S., Durand, S., Valaskovic, G. A. & MacCoss, M. J. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *Journal of the American Society for Mass Spectrometry* **24**, 148-153, doi:10.1007/s13361-012-0508-6 (2013).

86 Pirmoradian, M. *et al.* Rapid and deep human proteome analysis by single-dimension shotgun proteomics. *Molecular & cellular proteomics : MCP* **12**, 3330-3338, doi:10.1074/mcp.O113.028787 (2013).

87 Wang, X., Stoll, D. R., Schellinger, A. P. & Carr, P. W. Peak capacity optimization of peptide separations in reversed-phase gradient elution chromatography: fixed column format. *Analytical chemistry* **78**, 3406-3416, doi:10.1021/ac0600149 (2006).

88 Zhou, F., Lu, Y., Ficarro, S. B., Webber, J. T. & Marto, J. A. Nanoflow low pressure high peak capacity single dimension LC-MS/MS platform for high-throughput, in-depth analysis of mammalian proteomes. *Analytical chemistry* **84**, 5133-5139, doi:10.1021/ac2031404 (2012).

89 Zhang, B., Pirmoradian, M., Chernobrovkin, A. & Zubarev, R. A. DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. *Molecular & cellular proteomics : MCP* **13**, 3211-3223, doi:10.1074/mcp.O114.038877 (2014).

90 Xu, P. *et al.* Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell* **137**, 133-145, doi:10.1016/j.cell.2009.01.041 (2009).
91 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods* **10**, 634-637, doi:10.1038/nmeth.2518 (2013).
92 Branca, R. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods* **11**, 59-62, doi:10.1038/nmeth.2732 (2014).
93 Scherl, A. *et al.* Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. *Analytical chemistry* **80**, 1182-1191, doi:10.1021/ac701680f (2008).
94 Vincent, C. E. *et al.* Segmentation of precursor mass range using "tiling" approach increases peptide identifications for MS1-based label-free quantification. *Analytical chemistry* **85**, 2825-2832, doi:10.1021/ac303352n (2013).
95 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548, doi:10.1038/msb.2011.81 (2011).
96 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP* **11**, M111 014050, doi:10.1074/mcp.M111.014050 (2012).
97 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-587, doi:10.1038/nature13319 (2014).
98 Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581, doi:10.1038/nature13302 (2014).
99 Hyman, B. T. & Trojanowski, J. Q. Consensus recommendations for the postmortem diagnosis of Alzheimer disease from the National Institute on Aging and the Reagan Institute Working Group on diagnostic criteria for the neuropathological assessment of Alzheimer disease. *Journal of neuropathology and experimental neurology* **56**, 1095-1097 (1997).
100 Na, C. H. *et al.* Synaptic protein ubiquitination in rat brain revealed by antibody-based ubiquitome analysis. *Journal of proteome research* **11**, 4722-4732, doi:10.1021/pr300536k (2012).
101 Gozal, Y. M. *et al.* Proteomics analysis reveals novel components in the detergent-insoluble subproteome in Alzheimer's disease. *Journal of proteome research* **8**, 5069-5079, doi:10.1021/pr900474t (2009).
102 Wang, Y. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11**, 2019-2026, doi:10.1002/pmic.201000722 (2011).
103 Dwivedi, R. C. *et al.* Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Analytical chemistry* **80**, 7036-7042, doi:10.1021/ac800984n (2008).
104 Song, C. *et al.* Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. *Analytical chemistry* **82**, 53-56, doi:10.1021/ac9023044 (2010).

105 Michalski, A. *et al.* Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular & cellular proteomics : MCP* **11**, O111 013698, doi:10.1074/mcp.O111.013698 (2012).
106 Dayon, L., Sonderegger, B. & Kussmann, M. Combination of gas-phase fractionation and MS(3) acquisition modes for relative protein quantification with isobaric tagging. *Journal of proteome research* **11**, 5081-5089, doi:10.1021/pr300519c (2012).
107 Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976-989, doi:10.1016/1044-0305(94)80016-2 (1994).
108 Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of proteome research* **2**, 43-50 (2003).
109 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).
110 Bai, B. *et al.* U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 16562-16567, doi:10.1073/pnas.1310249110 (2013).
111 Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Analytical chemistry* **68**, 1-8 (1996).
112 Shen, Y. *et al.* High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Analytical chemistry* **74**, 4235-4249 (2002).
113 Eriksson, J. & Fenyo, D. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nature biotechnology* **25**, 651-655, doi:10.1038/nbt1315 (2007).
114 Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13**, 723-726, doi:10.1002/pmic.201200451 (2013).
115 Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391-399, doi:10.1038/nature11405 (2012).
116 Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature biotechnology* **28**, 83-89, doi:10.1038/nbt.1592 (2010).
117 Hebert, A. S. *et al.* The one hour yeast proteome. *Molecular & cellular proteomics : MCP* **13**, 339-347, doi:10.1074/mcp.M113.034769 (2014).
118 Altelaar, A. F., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35-48, doi:10.1038/nrg3356 (2013).
119 Ow, S. Y., Salim, M., Noirel, J., Evans, C. & Wright, P. C. Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-

resolution HILIC fractionation. *Proteomics* **11**, 2341-2346, doi:10.1002/pmic.201000752 (2011).
120 Zhou, F. *et al.* Genome-scale proteome quantification by DEEP SEQ mass spectrometry. *Nat. Commun.* **4**, 2171, doi:10.1038/ncomms3171 (2013).
121 Savitski, M. M. *et al.* Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers. *Anal. Chem.* **83**, 8959-8967, doi:10.1021/ac201760x (2011).
122 Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8**, 933-935, doi:10.1038/nmeth.1716 (2011).
123 Wuhr, M. *et al.* Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **84**, 9214-9221, doi:10.1021/ac301962s (2012).
124 Ahrne, E. *et al.* Evaluation and Improvement of Quantification Accuracy in Isobaric Mass Tag-Based Protein Quantification Experiments. *J. Proteome Res.* **15**, 2537-2547, doi:10.1021/acs.jproteome.6b00066 (2016).
125 Sandberg, A., Branca, R. M., Lehtio, J. & Forshed, J. Quantitative accuracy in mass spectrometry based proteomics of complex samples: the impact of labeling and precursor interference. *J. proteomics* **96**, 133-144, doi:10.1016/j.jprot.2013.10.035 (2014).
126 Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* **12**, 3586-3598, doi:10.1021/pr400098r (2013).
127 Erickson, B. K. *et al.* Evaluating multiplexed quantitative phosphopeptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **87**, 1241-1249, doi:10.1021/ac503934f (2015).
128 Liu, J. M., Sweredoski, M. J. & Hess, S. Improved 6-Plex Tandem Mass Tags Quantification Throughput Using a Linear Ion Trap–High-Energy Collision Induced Dissociation MS3 Scan. *Anal. Chem.* **88**, 7471-7475, doi:10.1021/acs.analchem.6b01067 (2016).
129 Na, C. H. *et al.* Synaptic protein ubiquitination in rat brain revealed by antibody-based ubiquitome analysis. *J. Proteome Res.* **11**, 4722-4732, doi:10.1021/pr300536k (2012).
130 Tan, H. *et al.* Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome. *Proteomics* **15**, 500-507, doi:10.1002/pmic.201400171 (2014).
131 Wang, H. *et al.* Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. *J. Proteome Res.* **14**, 829-838, doi:10.1021/pr500882h (2015).
132 Pagala, V. R. *et al.* Quantitative protein analysis by mass spectrometry. *Methods Mol. Biol.* **1278**, 281-305, doi:10.1007/978-1-4939-2425-7_17 (2015).
133 Wang, X. *et al.* JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Mol. Cell Proteomics* **13**, 3663-3673, doi:10.1074/mcp.O114.039586 (2014).

134 Hanna, J. A. *et al.* PAX7 is a required target for microRNA-206-induced differentiation of fusion-negative rhabdomyosarcoma. *Cell Death Dis.* **7**, e2256, doi:10.1038/cddis.2016.159 (2016).
135 Li, Y. *et al.* JUMPg: an Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *J. Proteome Res.* **15**, 2309-2320, doi:10.1021/acs.jproteome.6b00344 (2016).
136 Mertz, J. *et al.* Sequential Elution Interactome Analysis of the Mind Bomb 1 Ubiquitin Ligase Reveals a Novel Role in Dendritic Spine Outgrowth. *Mol. Cell Proteomics* **14**, 1898-1910, doi:10.1074/mcp.M114.045898 (2015).
137 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).
138 Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43-50 (2003).
139 McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150-7158, doi:10.1021/ac502040v (2014).
140 Xu, P., Duong, D. M. & Peng, J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. *J. Proteome Res.* **8**, 3944-3950, doi:10.1021/pr900251d (2009).
141 Scheltema, R. A. *et al.* The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell Proteomics* **13**, 3698-3708, doi:10.1074/mcp.M114.043489 (2014).
142 Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477, doi:10.1016/j.cell.2013.09.034 (2013).
143 Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550-563, doi:10.1016/j.cell.2015.12.028 (2016).
144 Cohen, P. The origins of protein phosphorylation. *Nature cell biology* **4**, E127-E130 (2002).
145 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).
146 Rankin, S. L., Zhu, G. & Baker, S. J. Review: insights gained from modelling high-grade glioma in the mouse. *Neuropathology and applied neurobiology* **38**, 254-270, doi:10.1111/j.1365-2990.2011.01231.x (2012).
147 Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nature genetics* **48**, 768-776, doi:10.1038/ng.3590 (2016).
148 Wang, H. *et al.* Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. *Journal of proteome research* **14**, 829-838, doi:10.1021/pr500882h (2015).
149 Li, Y. *et al.* JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *Journal of proteome research* **15**, 2309-2320, doi:10.1021/acs.jproteome.6b00344 (2016).

150 Wang, X. *et al.* JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Molecular & cellular proteomics : MCP* **13**, 3663-3673, doi:10.1074/mcp.O114.039586 (2014).
151 Tan, H. *et al.* Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome. *Proteomics* **15**, 500-507, doi:10.1002/pmic.201400171 (2015).
152 Macek, B., Mann, M. & Olsen, J. V. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annual review of pharmacology and toxicology* **49**, 199-221, doi:10.1146/annurev.pharmtox.011008.145606 (2009).
153 Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520, doi:10.1093/nar/gku1267 (2015).
154 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).
155 Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)* **25**, 1091-1093, doi:10.1093/bioinformatics/btp101 (2009).
156 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).
157 Mischnik, M. *et al.* IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics (Oxford, England)* **32**, 424-431, doi:10.1093/bioinformatics/btv699 (2016).
158 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:http://www.nature.com/nature/journal/v489/n7414/abs/nature11247.html#supplementary-information (2012).
159 Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS computational biology* **4**, e1000217, doi:10.1371/journal.pcbi.1000217 (2008).
160 Mertins, P. *et al.* iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Molecular & cellular proteomics : MCP* **11**, M111 014423, doi:10.1074/mcp.M111.014423 (2012).
161 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).
162 Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat Biotech* **21**, 255-261 (2003).
163 Müller, G., Wied, S. & Frick, W. Cross Talk of pp125(FAK) and pp59(Lyn) Non-Receptor Tyrosine Kinases to Insulin-Mimetic Signaling in Adipocytes. *Molecular and Cellular Biology* **20**, 4708-4723 (2000).
164 Pavlova, N. N. & Thompson, C. B. The Emerging Hallmarks of Cancer Metabolism. *Cell metabolism* **23**, 27-47, doi:10.1016/j.cmet.2015.12.006 (2016).
165 Kaur, H. *et al.* The transcriptional modulator HMGA2 promotes stemness and tumorigenicity in glioblastoma. *Cancer letters* **377**, 55-64, doi:10.1016/j.canlet.2016.04.020 (2016).

166 Alinari, L. *et al.* Combination anti-CD74 (milatuzumab) and anti-CD20 (rituximab) monoclonal antibody therapy has in vitro and in vivo activity in mantle cell lymphoma. *Blood* **117**, 4530-4541, doi:10.1182/blood-2010-08-303354 (2011).
167 Stein, R. *et al.* CD74: a new candidate target for the immunotherapy of B-cell neoplasms. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 5556s-5563s, doi:10.1158/1078-0432.ccr-07-1167 (2007).
168 Brembeck, F. H., Rosario, M. & Birchmeier, W. Balancing cell adhesion and Wnt signaling, the key role of beta-catenin. *Current opinion in genetics & development* **16**, 51-59, doi:10.1016/j.gde.2005.12.007 (2006).
169 Chalhoub, N. & Baker, S. J. PTEN and the PI3-kinase pathway in cancer. *Annual review of pathology* **4**, 127-150, doi:10.1146/annurev.pathol.4.110807.092311 (2009).
170 do Carmo, A., Balca-Silva, J., Matias, D. & Lopes, M. C. PKC signaling in glioblastoma. *Cancer biology & therapy* **14**, 287-294, doi:10.4161/cbt.23615 (2013).
171 Engelman, J. A. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat Rev Cancer* **9**, 550-562, doi:10.1038/nrc2664 (2009).
172 Lopez, M., Nogueiras, R., Tena-Sempere, M. & Dieguez, C. Hypothalamic AMPK: a canonical regulator of whole-body energy balance. *Nature reviews. Endocrinology* **12**, 421-432, doi:10.1038/nrendo.2016.67 (2016).
173 Radu, M., Semenova, G., Kosoff, R. & Chernoff, J. PAK signalling during the development and progression of cancer. *Nat Rev Cancer* **14**, 13-25 (2014).
174 Liu, R. *et al.* Cdk5-mediated regulation of the PIKE-A-Akt pathway and glioblastoma cell invasion. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 7570-7575, doi:10.1073/pnas.0712306105 (2008).
175 Wee, S. *et al.* Selective calcium sensitivity in immature glioma cancer stem cells. *PloS one* **9**, e115698, doi:10.1371/journal.pone.0115698 (2014).
176 Xie, Q. *et al.* Mitochondrial control by DRP1 in brain tumor initiating cells. *Nature neuroscience* **18**, 501-510, doi:10.1038/nn.3960 (2015).
177 Fleuren, E. D. G., Zhang, L., Wu, J. & Daly, R. J. The kinome 'at large' in cancer. *Nat Rev Cancer* **16**, 83-98, doi:10.1038/nrc.2015.18
http://www.nature.com/nrc/journal/v16/n2/abs/nrc.2015.18.html#supplementary-information (2016).
178 Roopra, A., Qazi, R., Schoenike, B., Daley, T. J. & Morrison, J. F. Localized domains of G9a-mediated histone methylation are required for silencing of neuronal genes. *Molecular cell* **14**, 727-738, doi:10.1016/j.molcel.2004.05.026 (2004).
179 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).
180 Barroilhet, L. *et al.* C-terminal binding protein-2 regulates response of epithelial ovarian cancer cells to histone deacetylase inhibitors. *Oncogene* **32**, 3896-3903, doi:10.1038/onc.2012.380 (2013).

181 Silva, C. H., Silva, M., Iulek, J. & Thiemann, O. H. Structural complexes of human adenine phosphoribosyltransferase reveal novel features of the APRT catalytic mechanism. *Journal of biomolecular structure & dynamics* **25**, 589-597, doi:10.1080/07391102.2008.10507205 (2008).
182 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
183 Johnson, R. A. *et al.* Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature* **466**, 632-636, doi:10.1038/nature09173 (2010).
184 Binda, E. *et al.* The EphA2 receptor drives self-renewal and tumorigenicity in stem-like tumor-propagating cells from human glioblastomas. *Cancer cell* **22**, 765-780, doi:10.1016/j.ccr.2012.11.005 (2012).
185 Song, W., Ma, Y., Wang, J., Brantley-Sieders, D. & Chen, J. JNK signaling mediates EPHA2-dependent tumor cell proliferation, motility, and cancer stem cell-like properties in non-small cell lung cancer. *Cancer research* **74**, 2444-2454, doi:10.1158/0008-5472.can-13-2136 (2014).
186 Miao, H. *et al.* EphA2 mediates ligand-dependent inhibition and ligand-independent promotion of cell migration and invasion via a reciprocal regulatory loop with Akt. *Cancer cell* **16**, 9-20, doi:10.1016/j.ccr.2009.04.009 (2009).
187 Zeiner, P. S. *et al.* MIF Receptor CD74 is Restricted to Microglia/Macrophages, Associated with a M1-Polarized Immune Milieu and Prolonged Patient Survival in Gliomas. *Brain pathology (Zurich, Switzerland)* **25**, 491-504, doi:10.1111/bpa.12194 (2015).
188 Rauniyar, N. & Yates, J. R., 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of proteome research* **13**, 5293-5309, doi:10.1021/pr500880b (2014).
189 Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Meth* **8**, 937-940, doi:http://www.nature.com/nmeth/journal/v8/n11/abs/nmeth.1714.html#supplementary-information (2011).
190 Jones, D. T. *et al.* Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nature genetics* **45**, 927-932, doi:10.1038/ng.2682 (2013).
191 Zhang, J. *et al.* Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nature genetics* **45**, 602-612, doi:10.1038/ng.2611 (2013).
192 Leprivier, G. *et al.* The eEF2 kinase confers resistance to nutrient deprivation by blocking translation elongation. *Cell* **153**, 1064-1079, doi:10.1016/j.cell.2013.04.055 (2013).
193 Stewart, E. *et al.* The Childhood Solid Tumor Network: A new resource for the developmental biology and oncology research communities. *Developmental biology* **411**, 287-293, doi:10.1016/j.ydbio.2015.03.001 (2016).
194 Smith, M. A., Altekruse, S. F., Adamson, P. C., Reaman, G. H. & Seibel, N. L. Declining childhood and adolescent cancer mortality. *Cancer* **120**, 2497-2506, doi:10.1002/cncr.28748 (2014).

195 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).
196 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, doi:10.1016/j.cell.2016.05.069 (2016).
197 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).
198 Tan, H. *et al.* Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome. ..., doi:10.1002/pmic.201400171 (2015).
199 McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical chemistry* **86**, 7150-7158, doi:10.1021/ac502040v (2014).
200 Weekes, M. P. *et al.* Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell* **157**, 1460-1472, doi:10.1016/j.cell.2014.04.028 (2014).
201 Bai, B. *et al.* Deep Profiling of Proteome and Phosphoproteome by Isobaric Labeling, Extensive Liquid Chromatography, and Mass Spectrometry. *Methods in enzymology* **585**, 377-395, doi:10.1016/bs.mie.2016.10.007 (2017).
202 Bai, B. *et al.* U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proceedings of the National Academy of Sciences* **110**, 16562-16567, doi:10.1073/pnas.1310249110 (2013).
203 Wang, H., Yang, Y., Li, Y., Bai, B. & Wang, X. Systematic Optimization of Long Gradient Chromatography Mass Spectrometry for Deep Analysis of Brain Proteome. *Journal of proteome ...*, doi:10.1021/pr500882h (2014).
204 Martinez, J. *et al.* Molecular characterization of LC3-associated phagocytosis reveals distinct roles for Rubicon, NOX2 and autophagy proteins. *Nature cell biology* **17**, 893-906, doi:10.1038/ncb3192 (2015).
205 Lee, K. H. *et al.* C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. *Cell* **167**, 774-788 e717, doi:10.1016/j.cell.2016.10.002 (2016).
206 Joo, J. H. *et al.* The Noncanonical Role of ULK/ATG1 in ER-to-Golgi Trafficking Is Essential for Cellular Homeostasis. *Molecular cell* **62**, 491-506, doi:10.1016/j.molcel.2016.04.020 (2016).
207 Joo, J. H. *et al.* The Noncanonical Role of ULK/ATG1 in ER-to-Golgi Trafficking Is Essential for Cellular Homeostasis. *Molecular cell* **62**, 982, doi:10.1016/j.molcel.2016.05.030 (2016).
208 Gong, J. *et al.* The C. elegans Taste Receptor Homolog LITE-1 Is a Photoreceptor. *Cell* **168**, 325, doi:10.1016/j.cell.2016.12.040 (2017).
209 Churchman, M. L. *et al.* Efficacy of Retinoids in IKZF1-Mutated BCR-ABL1 Acute Lymphoblastic Leukemia. *Cancer cell* **28**, 343-356, doi:10.1016/j.ccell.2015.07.016 (2015).
210 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **4**, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).

211 Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research* **10**, 5354-5362, doi:10.1021/pr200611n (2011).
212 Mertz, J. *et al.* Sequential Elution Interactome Analysis of the Mind Bomb 1 Ubiquitin Ligase Reveals a Novel Role in Dendritic Spine Outgrowth. *Molecular & cellular proteomics : MCP* **14**, 1898-1910, doi:10.1074/mcp.M114.045898 (2015).
213 Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Robert Gentleman *et al.*) 397-420 (Springer New York, 2005).
214 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).
215 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
216 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**, Article17, doi:10.2202/1544-6115.1128 (2005).
217 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)* **24**, 719-720, doi:10.1093/bioinformatics/btm563 (2008).
218 Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)* **27**, 1739-1740, doi:10.1093/bioinformatics/btr260 (2011).
219 Cossu, G. & Borello, U. Wnt signaling and the activation of myogenesis in mammals. *The EMBO journal* **18**, 6867-6872, doi:10.1093/emboj/18.24.6867 (1999).
220 Brack, A. S., Conboy, I. M., Conboy, M. J., Shen, J. & Rando, T. A. A temporal switch from notch to Wnt signaling in muscle stem cells is necessary for normal adult myogenesis. *Cell stem cell* **2**, 50-59, doi:10.1016/j.stem.2007.10.006 (2008).
221 Stern, H. M., Brown, A. M. & Hauschka, S. D. Myogenesis in paraxial mesoderm: preferential induction by dorsal neural tube and by cells expressing Wnt-1. *Development (Cambridge, England)* **121**, 3675-3686 (1995).
222 Borycki, A. G. *et al.* Sonic hedgehog controls epaxial muscle determination through Myf5 activation. *Development (Cambridge, England)* **126**, 4053-4063 (1999).
223 Munsterberg, A. E., Kitajewski, J., Bumcrot, D. A., McMahon, A. P. & Lassar, A. B. Combinatorial signaling by Sonic hedgehog and Wnt family members induces myogenic bHLH gene expression in the somite. *Genes & development* **9**, 2911-2922 (1995).
224 Reshef, R., Maroto, M. & Lassar, A. B. Regulation of dorsal somitic cell fates: BMPs and Noggin control the timing and pattern of myogenic regulator expression. *Genes & development* **12**, 290-303 (1998).
225 Pourquie, O. *et al.* Lateral and axial signals involved in avian somite patterning: a role for BMP4. *Cell* **84**, 461-471 (1996).

226 Amthor, H., Christ, B., Weil, M. & Patel, K. The importance of timing differentiation during limb muscle development. *Current biology : CB* **8**, 642-652 (1998).
227 Chen, A. E., Ginty, D. D. & Fan, C. M. Protein kinase A signalling via CREB controls myogenesis induced by Wnt proteins. *Nature* **433**, 317-322, doi:10.1038/nature03126 (2005).
228 Wu, Z. *et al.* p38 and extracellular signal-regulated kinases regulate the myogenic program at multiple steps. *Mol Cell Biol* **20**, 3951-3964 (2000).
229 de Angelis, L. *et al.* Regulation of vertebrate myotome development by the p38 MAP kinase-MEF2 signaling pathway. *Developmental biology* **283**, 171-179, doi:10.1016/j.ydbio.2005.04.009 (2005).
230 Keren, A., Tamir, Y. & Bengal, E. The p38 MAPK signaling pathway: a major regulator of skeletal muscle development. *Molecular and cellular endocrinology* **252**, 224-230, doi:10.1016/j.mce.2006.03.017 (2006).
231 Bennett, A. M. & Tonks, N. K. Regulation of distinct stages of skeletal muscle differentiation by mitogen-activated protein kinases. *Science (New York, N.Y.)* **278**, 1288-1291 (1997).
232 Jiang, B. H., Zheng, J. Z. & Vogt, P. K. An essential role of phosphatidylinositol 3-kinase in myogenic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14179-14183 (1998).
233 Kaliman, P., Vinals, F., Testar, X., Palacin, M. & Zorzano, A. Phosphatidylinositol 3-kinase inhibitors block differentiation of skeletal muscle cells. *The Journal of biological chemistry* **271**, 19146-19151 (1996).
234 Jiang, B. H., Aoki, M., Zheng, J. Z., Li, J. & Vogt, P. K. Myogenic signaling of phosphatidylinositol 3-kinase requires the serine-threonine kinase Akt/protein kinase B. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2077-2081 (1999).
235 Bentzinger, C. F., Wang, Y. X. & Rudnicki, M. A. Building muscle: molecular regulation of myogenesis. *Cold Spring Harbor perspectives in biology* **4**, doi:10.1101/cshperspect.a008342 (2012).
236 Le Grand, F. & Rudnicki, M. A. Skeletal muscle satellite cells and adult myogenesis. *Current opinion in cell biology* **19**, 628-633, doi:10.1016/j.ceb.2007.09.012 (2007).
237 McKinnell, I. W., Parise, G. & Rudnicki, M. A. Muscle stem cells and regenerative myogenesis. *Current topics in developmental biology* **71**, 113-130, doi:10.1016/s0070-2153(05)71004-8 (2005).
238 Sabourin, L. A. & Rudnicki, M. A. The molecular regulation of myogenesis. *Clinical genetics* **57**, 16-25 (2000).
239 Kumar, S., Perlman, E., Harris, C. A., Raffeld, M. & Tsokos, M. Myogenin is a specific marker for rhabdomyosarcoma: an immunohistochemical study in paraffin-embedded tissues. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **13**, 988-993, doi:10.1038/modpathol.3880179 (2000).
240 Barrott, J. J. & Haystead, T. A. Hsp90, an unlikely ally in the war on cancer. *The FEBS journal* **280**, 1381-1396, doi:10.1111/febs.12147 (2013).

241 Karagoz, G. E. & Rudiger, S. G. Hsp90 interaction with clients. *Trends in biochemical sciences* **40**, 117-125, doi:10.1016/j.tibs.2014.12.002 (2015).
242 Prodromou, C. Mechanisms of Hsp90 regulation. *The Biochemical journal* **473**, 2439-2452, doi:10.1042/bcj20160005 (2016).
243 Taipale, M. *et al.* Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* **150**, 987-1001, doi:10.1016/j.cell.2012.06.047 (2012).
244 Taipale, M. *et al.* A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell* **158**, 434-448, doi:10.1016/j.cell.2014.05.039 (2014).
245 Akerfelt, M., Morimoto, R. I. & Sistonen, L. Heat shock factors: integrators of cell stress, development and lifespan. *Nature reviews. Molecular cell biology* **11**, 545-555, doi:10.1038/nrm2938 (2010).
246 Sabnis, A. J. *et al.* Combined chemical-genetic approach identifies cytosolic HSP70 dependence in rhabdomyosarcoma. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 9015-9020, doi:10.1073/pnas.1603883113 (2016).
247 Stewart, E. *et al.* Targeting the DNA repair pathway in Ewing sarcoma. *Cell reports* **9**, 829-841, doi:10.1016/j.celrep.2014.09.028 (2014).
248 Langenau, D. M., Sweet-Cordero, A., Wechsler-Reya, R. J. & Dyer, M. A. Preclinical Models Provide Scientific Justification and Translational Relevance for Moving Novel Therapeutics into Clinical Trials for Pediatric Cancer. *Cancer research* **75**, 5176-5186, doi:10.1158/0008-5472.can-15-1308 (2015).
249 Furman, W. L. *et al.* Direct translation of a protracted irinotecan schedule from a xenograft model to a phase I trial in children. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **17**, 1815-1824, doi:10.1200/jco.1999.17.6.1815 (1999).
250 Harris, M. H. *et al.* Multicenter Feasibility Study of Tumor Molecular Profiling to Inform Therapeutic Decisions in Advanced Pediatric Solid Tumors: The Individualized Cancer Therapy (iCat) Study. *JAMA oncology*, doi:10.1001/jamaoncol.2015.5689 (2016).
251 Parsons, D. W. *et al.* Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. *JAMA oncology*, doi:10.1001/jamaoncol.2015.5699 (2016).
252 Sherr, C. J. A New Cell-Cycle Target in Cancer - Inhibiting Cyclin D-Dependent Kinases 4 and 6. *The New England journal of medicine* **375**, 1920-1923, doi:10.1056/NEJMp1612343 (2016).
253 Rauniyar, N. & Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *Journal of proteome research* **13**, 5293-5309, doi:10.1021/pr500880b (2014).
254 Murphy, M. E. The HSP70 family and cancer. *Carcinogenesis* **34**, 1181-1188, doi:10.1093/carcin/bgt111 (2013).
255 Chen, Y. *et al.* Targeting HSF1 sensitizes cancer cells to HSP90 inhibition. *Oncotarget* **4**, 816-829, doi:10.18632/oncotarget.991 (2013).

256 Demetri, G. D. *et al.* Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *The New England journal of medicine* **347**, 472-480, doi:10.1056/NEJMoa020461 (2002).
257 Tan, H. *et al.* Integrative Proteomics and Phosphoproteomics Profiling Reveals Dynamic Signaling Networks and Bioenergetics Pathways Underlying T Cell Activation. *Immunity* **46**, 488-503, doi:10.1016/j.immuni.2017.02.010 (2017).
258 van den Bemd, G. J. *et al.* Mass spectrometric identification of human prostate cancer-derived proteins in serum of xenograft-bearing mice. *Molecular & cellular proteomics : MCP* **5**, 1830-1839, doi:10.1074/mcp.M500371-MCP200 (2006).
259 Sajic, T., Liu, Y. & Aebersold, R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics. Clinical applications* **9**, 307-321, doi:10.1002/prca.201400117 (2015).
260 Brawer, M. K. Prostate-specific antigen: current status. *CA: a cancer journal for clinicians* **49**, 264-281 (1999).

# VITA

Hong Wang was born in 1987. He attended Luzhou Medical College in Sichuan, China in 2006, majored in clinical laboratory medicine. He graduated with a medicine degree in June, 2011. At the same year, He joined the MIB track of integrated biomedical science program in the University of Tennessee Health Science Center. He joined the lab of Dr. Junmin Peng in the Department of Structural Biology and Developmental Neurobiology at St. Jude Children's Research Hospital and has been conducted research in mass spectrometry-based proteomic technique development and multi-omics integrative analysis of cancers.