INTEGRATION OF QUANTITATIVE AND FUNCTIONAL PROTEOMICS TO EXPLORE THE GLOBAL FUNCTIONAL LANDSCAPE OF POST-TRANSLATIONAL MODIFICATIONS

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

YAO GONG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DR. YUE CHEN, ADVISOR DR. RUI KUANG, CO-ADVISOR

JULY 2023

© YAO GONG 2023

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Yue Chen, and co-advisor, Dr. Rui Kuang, for their guidance, support, and mentorship throughout my doctoral journey. Their expertise, dedication, and encouragement have been invaluable in shaping my research and academic growth.

I am also immensely grateful to the members of my dissertation committee, Dr. Timothy Griffin and Dr. Jeongsik Yong, for their insightful feedback, constructive criticism, and valuable suggestions that have significantly contributed to the quality of this dissertation.

I would like to extend my appreciation to University of Minnesota, particularly Twin cities campus and Bioinformatics and Computational Biology (BICB) program for providing the necessary resources, facilities, and funding that enabled the successful completion of this research. I am grateful to all lab members at Chen lab and Kuang lab for their collaboration, assistance, and inspiring discussions. Their contributions and camaraderie have made the research environment stimulating and enjoyable.

I would like to thank my parents for their unwavering love, support, and belief in my abilities. Their constant encouragement and sacrifices have been instrumental in my academic pursuits.

I extend my heartfelt appreciation to my two roommates, Tianci Song and Jiacheng Liu, for their friendship and constant support during the ups and downs of my Ph.D. studies. Their presence made the journey more enjoyable and memorable.

In conclusion, this dissertation represents the culmination of years of hard work, dedication, and collaboration. I would like to acknowledge the inspiration and insights gained from the work of other researchers in my field. Their contributions have been a guiding light in my academic journey.

To all those whose names may not be mentioned but have played a role in my doctoral pursuit, your influence has not gone unnoticed, and I am sincerely grateful for your support.

Thank you all for being an integral part of this journey.

Dedication

This thesis is dedicated to my loving parents who unwaveringly supported me throughout my life. It is also dedicated to my two roommates at Eagan, Minnesota, U.S.A., we navigated through the ups and downs of my Ph.D. studies.

Abstract

Proteomics explores global-scale protein functions, with post-translational modifications (PTMs) expanding proteome complexity and functionality. Significant advancements in mass spectrometry (MS)-based quantitative proteomics have expanded the scope of PTM pathways. Meanwhile, functional proteomics bridges the annotation gap by linking uncharacterized proteins with biological functions. These two fields facilitate the exploration of the vast physiological function landscape of proteins and PTMs and aid in identifying disease-specific protein targets for the early detection and curation of diseases.

This research integrated the MS data analysis from quantitative proteomics and functional proteomics approaches and develop innovative bioinformatic strategies to interpret the functional landscape of PTM pathways in three aspects: up-stream enzymatic regulatory mechanisms that are responsible for adding and removing PTMs, comprehensive inventory and annotation analysis of PTM targets, and down-stream prediction of functional impacts of PTMs in protein structural stability and activities. First, to discover the up-stream enzymatic activities of a critical PTM pathway – ubiquitination, we constructed a quantitative ubiquitylome analysis algorithm and a stand-alone Python software package called UbE3-APA which is based on an interaction database for E3 ligase and ubiquitin sites to analyze the dynamics of ubiquitylome from MS-based quantitative proteomics data. The algorithm revealed the E3 ligase activity profiles of multiple global ubiquitylome studies effectively under various genetic and metabolic conditions. The algorithm developed in this project as well as a similar algorithm we developed in the Kinase Activity Profiling Analysis (KAPA) may be generally applicable to other PTM enzyme activity profiling analyses. Next, to comprehensively catalog an essential and yet under-studied oxygen-sensing posttranslational modification pathway – proline hydroxylation, we developed an online database and website platform, HypDB (https://hypdb.site), for hydroxyproline sites collection and functional annotation, while sharing the knowledge with the community. In the data collecting section, we evaluated the confidence of site-localization for identified sites and quantified their stoichiometry with corresponding MS spectra. And in the annotation section, we integrated multiple functional databases and developed bioinformatic strategies to study the enrichment of the hydroxyproline proteome in cellular pathways, structural domains, and tissue distributions at the levels of both the protein and modification site. All identified sites with annotation results were capsuled into HypDB websites, where its downloadable hydroxyproline spectra library allowed systematic data-independent DIA data analysis. Lastly, to understand the potential functional impacts of PTM in downstream pathways and protein activities, we developed bioinformatics strategies to identify the functional hydroxyproline proteome by integrated analysis of quantitative functional proteomics datasets. Our bioinformatic workflow explored the evolution conservation, protein turnover, and thermal stability profiles of hydroxyproline proteome in multiple species and different cell lines. Through individual and integrated analysis, we not only have a deeper understanding of the role of the global proline hydroxylation on protein structural stability under different conditions but also revealed significantly regulated hydroxyproline sites and substrate proteins that may serve as key targets in oxygen and metabolic sensing mechanisms in cellular activities. Collectively, this series of studies generate applicable models and novel knowledge in interpreting the functional network influenced by different PTMs.

Table of Contents

List of Figures

List of Abbreviations

PTM – post translational modification

MS – mass spectrometry

ESI – electrospray ionization

SILAC – stable isotope labeling by amino acids in cell culture

ICAT – isotope-coded affinity tags

cICAT – cleavable isotope-coded affinity tags

ICPL – isotope-coded protein label

iTRAQ – isobaric tags for relative and absolute quantitation

TMT – tandem mass tags

DiLeu – N, N-dimethyl leucine

LC-MS/MS – liquid chromatography coupled with tandem mass spectrometry

DDA – data-dependent acquisition

SIM – single-ion monitoring

RT – retention time

DIA – data-independent acquisition

SRM – selected reaction monitoring

PRM – parallel reaction monitoring

CID – collision-induced dissociation

XDIA – extended data-independent acquisition

HDMSE – high-definition MSE

PAcIFIC – precursor acquisition independent from ion count

AIF – all-ion fragmentation

SWATH-MS – sequential windowed acquisition of all theoretical fragment ion mass spectra

TOF – time-of-flight

WiSIM – wide selected-ion monitoring

diaPASEF – parallel accumulation-serial fragmentation combined with data-independent acquisition

TIMS – trapped ion mobility mass spectrometer

PSM – peptide-spectrum match

CHiMA – comprehensive histone mark analysis

TPP – trans-proteomic pipeline

TIC – total ion count

emPAI – exponentially modified protein abundance index

3D – three dimensional

PPI – protein-protein interactions

APMS – affinity purification mass spectrometry

UbE3-APA - ubiquitin E3 ligase activity profiling analysis

ESI – E3-substrate interactions

WT – wild type

H/L ratio – heavy to light ratio

siRNA - small interfering RNA

TNF – tumor necrosis factor

DMSO – Dimethyl sulfoxide

Hyp – hydroxyproline

HypDB – hydroxyproline database

P4HA – prolyl 4-hydroxylases

HIF – hypoxia-induced factor

PHD – prolyl hydroxylase domain

ccRCC – clear cell renal cell carcinoma

TK – tyrosine kinases

TKL – tyrosine kinase-like

CMGC – named after CDKs, MAPK, GSK, CLK families

RSA – relative solvent accessibility

SOD1 – superoxide dismutase

FBN1 – fibrillin-1

FDR – false discovery rate

CV – coefficient variance

BH – Benjamini-Hochberg

ETD - electron transfer dissociation

ET-HCD – electron transfer high energy collision dissociation

KNN – k-nearest neighbor

cALL – childhood acute lymphoblastic leukemia

ANOVA – analysis of variance

OG – orthologous groups

GO – gene ontology

KEGG – Kyoto encyclopedia of genes and genomes

Chapter I - Introduction: Post-Translational Modifications (PTMs), and its Studies with Quantitative Proteomics and Functional Proteomics Approaches

Yao Gong wrote this chapter

1.1 Introduction of PTMs and Early Development of PTM Identification and Functional Studies

1.1.1 Introduction to Proteomics and PTMs

Proteomics, a rapidly evolving field in biomedical research, focuses on the comprehensive study of proteins and their functions in biological systems. Recent discoveries have revealed that the complexity of the human proteome far surpasses that of the human genome. While the human genome was initially thought to contain only 20,000-25,000 protein-coding genes, estimates suggest that the total number of proteins in the human proteome is about 2 million[1]. This significant contrast highlights the fact that a single gene has the capacity to generate numerous distinct proteins.

At the transcriptome level, cells employ various mechanisms, such as genomic recombination, alternative transcription initiation, differential transcription termination, and alternative splicing, to generate diverse mRNA transcripts from a single gene.[2] These processes contribute to the increase in complexity and diversity from the transcriptome to the proteome. However, the full extent of protein diversity goes beyond transcript-level variations. It is further enhanced by a phenomenon known as protein post-translational modifications (PTMs).

PTMs refer to the covalent modifications that occur on proteins after translation. They play a critical role in expanding the functional repertoire of proteins and regulating their activities[3]. PTMs can occur through various mechanisms, such as the addition of chemical groups (e.g., phosphorylation, acetylation, methylation), enzymatic cleavage, or the attachment of small molecules (e.g., lipids, sugars)[3,4]. These modifications can influence protein localization, stability[5], interactions with other molecules[6], enzymatic activity, and signaling pathways, thereby impacting a broad range of cellular processes and contributing to the development of various disease states[7,8].

The exploration of PTMs and their functional consequences stands as a prominent focus within proteomics research. Leveraging the advancements in mass spectrometry, high-throughput technologies, and bioinformatics tools, PTM proteomics has become increasingly accessible and capable of globally identifying, quantifying, and characterizing PTMs in a high-throughput fashion. In biomedical studies, deciphering the proteome enables researchers to gain profound insights into the dynamic and functional aspects of cellular processes, while unraveling the intricate networks of protein interactions and signaling pathways that underlie normal physiological conditions and disease states. Consequently, PTM proteomic analyses facilitate the identification of disease-specific protein markers, thereby aiding in the early detection, diagnosis, and monitoring of various disorders, including cancer, cardiovascular diseases, and neurological conditions.

However, the journey of PTM proteomics commenced with fundamental explorations of the amino acid composition in proteins and the discovery of the initial protein modifications. These pioneering efforts laid the groundwork for our current understanding of PTMs and paved the way for remarkable advancements in the field of PTM proteomics.

1.1.2 History of PTM Discovery

At the turn of the 20th century, scientists embarked on unraveling the intricate chemical makeup of proteins. Through breaking down proteins with chemical methods, they sought to identify the individual amino acids and their distribution within proteins. This pursuit led to a groundbreaking discovery in 1906 by Levene and Alsberg[9], who identified phosphorylation, a post-translational modification, on the amino acid serine of vitellin. This finding marked the first documented protein modification and laid the foundation for the subsequent exploration of a wide range of post-translational modifications. The discovery of phosphorylation on serine occurred before the full characterization of all 20 canonical amino acids[10], making it particularly significant. There were still discussions and debates about the precise location of the phosphoryl group, until in 1932, Lipmann and Levene demonstrated that the phosphoryl group was attached to serine, establishing its status as a post-translational modification[11].

This discovery of the first post-translational modification paved the way for revealing various novel protein modifications and how the diverse array of these modifications regulated protein function. However, the initial discovery of protein modifications was followed by decades of relatively slow progress in the identification of new PTMs (Figure 1-1)[12], The characterization of hydroxylysine[13], for example, was not achieved until 1951, a remarkable 30 years after its initial observation in gelatin. The discovery and understanding of novel PTMs were hindered by technical challenges and the limited analytical tools available at the time. Methods in those days such as elemental analysis and nuclear magnetic resonance spectroscopy were inefficient for the analysis of PTMs due to their stoichiometric rarity, elemental similarity to proteins, and inherent lability. Furthermore, the concept of reversible PTMs had not yet been fully developed. For example, the notion that phospho-serine represented a distinct amino acid, rather than a reversible modification of canonical serine, persisted.

The turning point in the concept of PTM came in the mid-1950s when its reversibility was recognized in scientific literature. A significant breakthrough occurred in 1956 when it was revealed that covalent acyl-serine modifications were reversible[14], followed by the observation of similar reversibility in serine phosphorylation in 1960[15]. These findings supported the notion of a "high-energy bond" capable of releasing energy upon cleavage, which laid the foundation for explanations of various aspects of cellular metabolism. The updated concept triggered the boom of PTM discovery in 1970s, with approximately 40% of the currently known PTMs being identified by 1980 (Figure 1-1)[12].

During the late 1970s, the concept of PTMs encompassed a broader range of protein modifications compared to the current understanding[16]. For instance, N-terminal removal of signaling peptides and disulfide bond formation were considered PTMs. Transient cysteine modifications involved in the catalytic mechanism of proteins, however, were not recognized as PTMs yet. The field was in the process of discovering, documenting, and grappling with the definition of these newly identified modifications. In 1977, a significant review on PTMs mentioned the existence of over 140 amino acids, leading to discussions on whether these were modifications to canonical amino acids or distinct non-essential amino acids[17]. The cataloging of modified amino acids counted various combinations, with lysine exhibiting the highest number of modifications, while leucine, isoleucine, and

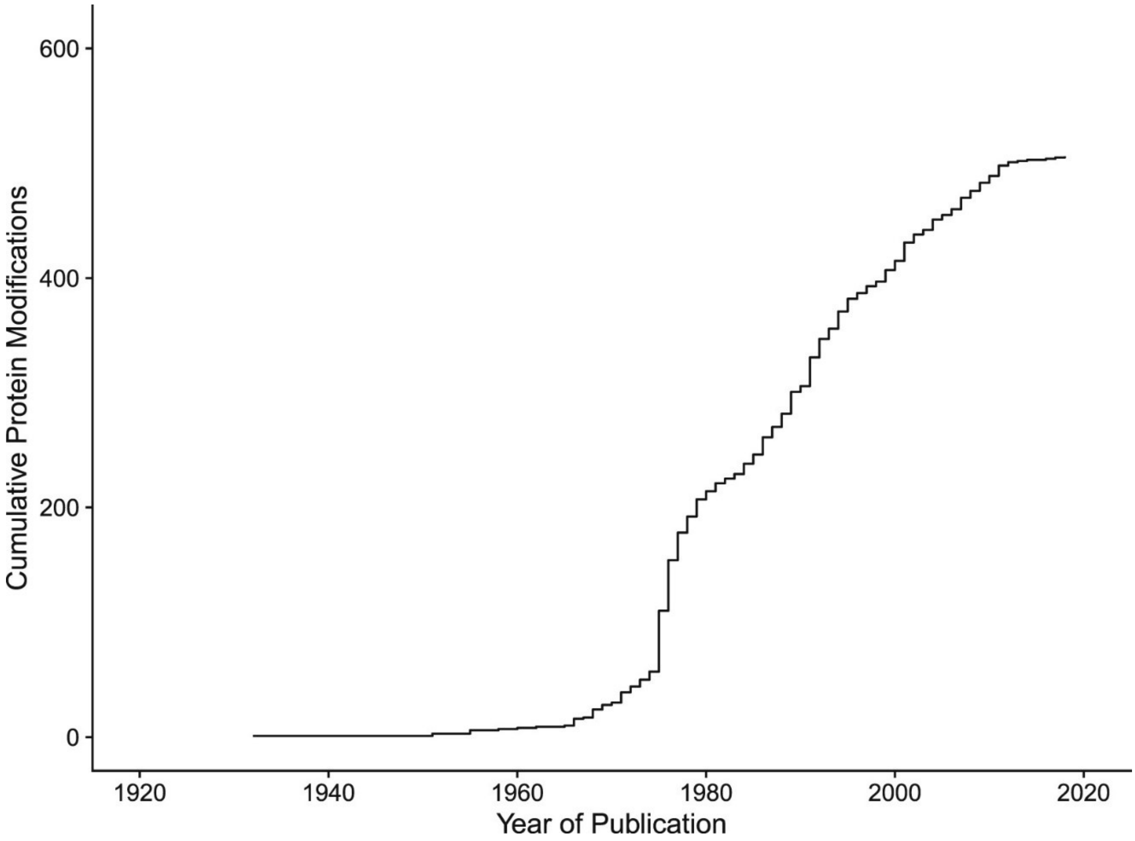


Figure 1-1 Cumulative PTMs Discovered in the Past Hundred Years

tryptophan were notably absent. Some modifications had plausible explanations, such as covalent modifications of coenzymes to enzyme active sites. Others, like N-alpha-acetylation or many methylation and halogenation modifications, had unknown functions at the time. This early catalog, although including species that are not considered PTMs today, played a crucial role in establishing a foundation for defining PTMs as a distinct class.

The rapid progress in PTM research awaited the development of protein mass spectrometry (MS) in the 1980s. MS, initially used in organic chemistry, became a powerful tool for the analysis of proteins and their modifications. The introduction of electrospray ionization (ESI) by Malcom Dole in 1968[18] revolutionized MS proteomics by gently ionizing macromolecules for analysis. Further advancements by John Fenn[19,20] and Koichi Tanaka,[21] Nobel laureates in Chemistry (2002), enabled sensitive detection of a wide range of PTMs and marked the early applications of MS technology in protein biology[22]. Despite the transformative potential of MS in PTM research, its initial impact on novel PTM discovery was limited. Most MS-based PTM identification protocols required prior enrichment of the modification, which depended on knowledge of the specific PTM being studied. Additionally, the selection of target masses for MS data analysis was essential, posing a challenge in detecting uncharacterized PTMs. Consequently, a significant portion of MS signals remained unattributed to known peptides or peptide-PTM adducts, what is now referred to as the "dark proteome." The field did not await long until MS technological advancements in sensitivity allowed for the quantification of relative PTM levels and absolute stoichiometries in MS-based studies. By the 1990s, MS had contributed to the identification of numerous new PTMs (Figure 1-1). As the catalog of known PTMs expanded, this enhanced resolution enabled the generation of large PTM profiles, initiated the development of bioinformatics approaches that analyze PTM data on a global scale, and provided insights into the regulatory role of many PTMs in protein activity.

1.1.3 Early Functional Studies of PTMs

The field aims to reveal the biological functions of PTMs simultaneously during the development of technologies to support PTM identification and quantification. This important direction converts data generated into useful knowledge but has posed significant challenges. Initially, early studies on PTMs focused on their detection and characterization, often without attempting to assign specific functions. The complexity of PTMs and their diverse effects on proteins, including activation, repression, translocation, degradation, and more, made it difficult to uncover their precise biological implications. However, notable exceptions in the mid-20th century emerged with the discovery of the functional roles of phosphorylation and acetylation.

Phosphorylation's biological significance was first revealed by Sutherland, Fischer, and Krebs, who demonstrated the interconversion of phosphorylase *a* and *b* and its regulation through reversible phosphorylation[23,24]. Their experiments involved enriching phosphorylase from liver tissue and measuring radioactive phosphate, establishing that phosphate could stimulate enzyme activity when “gained” by the enzyme. The process was also found to be reversible and under the influence of regulatory factors like epinephrine. Similarly, Allfrey *et al.* explored the acetylation of histones rich in lysine/arginine[25]. Working with calf thymus nuclei, they observed that acetylated histones inhibited RNA synthesis, indicating a regulatory role for acetylation. These pivotal studies provided initial insights into the functional consequences of PTMs and their involvement in multiple aspects of cellular processes and disease pathogenesis.

These earlier studies in biochemical proteomics were primarily centered around identifying and characterizing individual PTMs. However, with the advancements in MS, it is now possible to analyze thousands of proteins in a single sample. This high-throughput approach in proteomics aligns with the comprehensive analyses employed in other life science disciplines such as genomics, transcriptomics, and metabolomics, enabling a broader understanding of biological processes and their responses to various stimuli and disease conditions.

Enabled by the high throughput capability, there is an increasing need to quantitatively measure proteins due to their dynamic and interconnected nature. Quantitative proteomics plays a crucial role in elucidating global protein kinetics and the molecular mechanisms underlying biological processes. Meanwhile, as the catalog of PTMs expands and the number of quantified PTMs increases, the development of diverse functional approaches becomes even more essential in the field of PTM proteomics. These approaches aim to unravel the regulatory mechanisms of PTMs and their impact on protein function, translating PTM proteomics data into meaningful biological insights and therapeutic discoveries.

Section 1.2 of this chapter will delve into the latest advancements in MS-based PTM proteomics, highlighting the development of cutting-edge techniques and methodologies that facilitate PTM quantification. And Section 1.3 of this chapter will focus on functional approaches that elucidate the regulation of protein modifications in various biological contexts, shedding light on the role of PTMs in cellular processes and diseases. Together, these sections will provide a comprehensive overview of the current progress and future directions in PTM research, opening new avenues for therapeutic discoveries and advancing our understanding of protein regulation and cellular signaling networks.

1.2 Advances of Quantitative Proteomics Approaches and Their Application in PTM Studies

1.2.1 Brief Introduction of MS Procedure in Proteomics

Over the last two decades, MS has emerged as an indispensable technique for the discovery, identification, and quantification of proteins and PTMs, especially when comparing the relative expression level difference of various PTMs among protein samples[26]. Significant advancements in multiple steps of proteomics MS procedure, including sample labeling techniques, enrichment methods, liquid chromatography designs, MS instrumentations, acquisition modes, and analytical computational models have greatly expanded the scope and depth of PTM proteomics research. These advancements have enabled researchers to explore the vast landscape of PTMs in a comprehensive manner.

1.2.2 Advances in Sample Labeling

The first question that quantitative proteomics MS needs to solve is sample labeling. Mass spectrometry is not inherently quantitative, because proteolytic peptides show great variability in physiochemical properties. Neither do they ionize proportionally to the original copies of proteins, nor do data acquisition of MS capture intensity signals from all ionized peptides, both in turn results in mass spectrometric variability between runs. To distinguish and compare the same peptides from different samples in a single MS run, three main classes of labeling methods have been developed: *in vivo* labeling[27,28], *in vitro* labeling[29,30], and label-free quantification[31,32] (Figure 1-2). Each of these classes offers distinct advantages and limitations, allowing researchers to choose the most suitable approach based on their experimental requirements.

Methods labeling *in vivo* involve the incorporation of stable isotope labels into proteins within living systems. Techniques such as stable isotope labeling by amino acids in cell culture (SILAC)[28,33] and ^{15}N metabolic labeling[34–36] have significantly contributed to the field. These labeling methods rely on culturing cells or feeding animals with materials containing stable isotopes. After harvesting samples and comparing peptide pairs with different stable isotopes at the MS1 level, *in vivo* labeling enables accurate protein quantification while minimizing technical variability. However, these approaches are limited to studies involving cell culture or model organisms and are generally restricted to comparing no more than three samples.

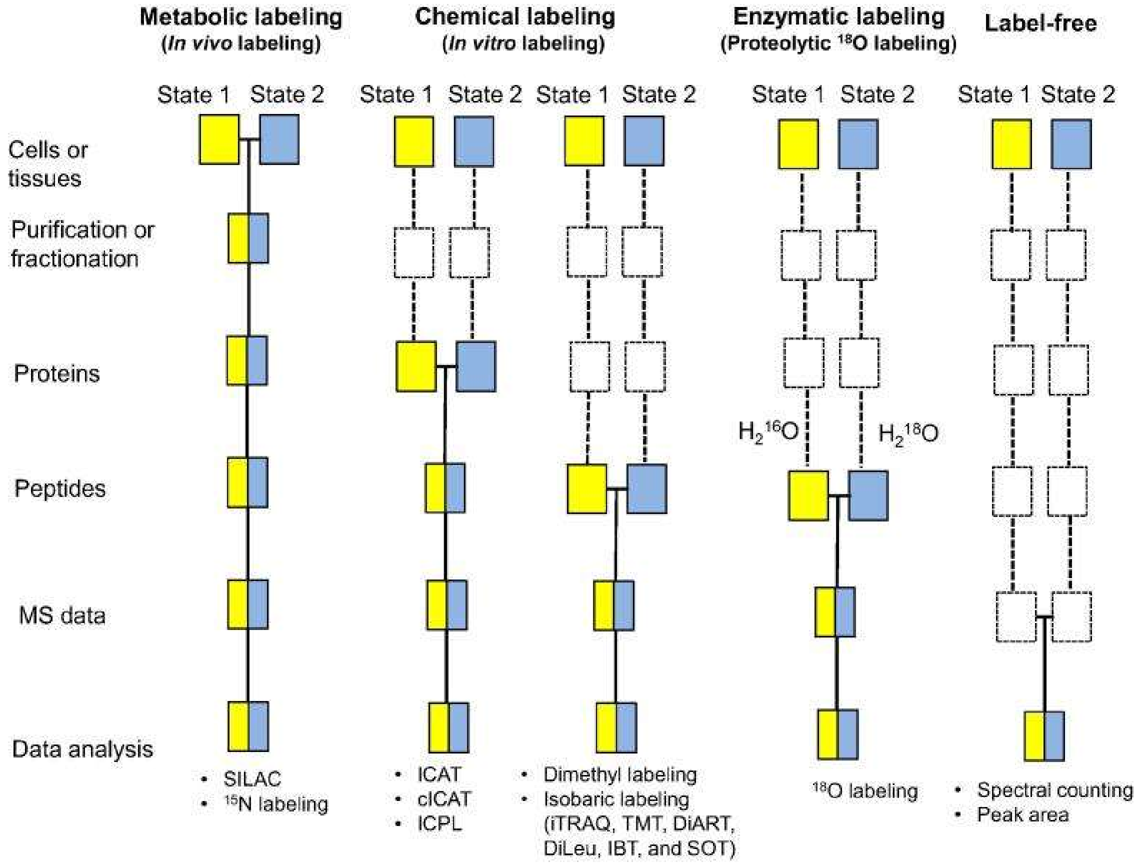


Figure 1-2 Procedure comparison of MS sample labeling methods

Techniques introduce mass tags into peptides or proteins using enzymatic or chemical processes that are *in vitro* labels. ^{18}O labeling[37,38], an example of enzymatic labeling, involves the exchange of ^{16}O for ^{18}O atoms at the C-terminal carboxyl group of digested peptides in the presence of $H_2{}^{18}O$ and proteases. Although simple in principle, ^{18}O labeling faces challenges such as isotopic peak overlap and variable labeling efficiency, limiting its widespread application[37]. Chemical labeling methods, on the other hand, offer greater versatility. Isotope-coded affinity tags (ICAT)[39], cleavable isotope-coded affinity tags (cICAT)[40], dimethyl labeling[41,42], isotope-coded protein label (ICPL)[43], and isobaric labeling[44] are among the in vitro chemical labeling techniques that have been developed.

Each of these methods presents unique advantages and drawbacks in terms of proteome coverage, quantification accuracy, and the ability to analyze specific protein groups.

Among these *in vitro* chemical labeling approaches, isobaric labeling methods have gained significant popularity. These techniques, including isobaric tags for relative and absolute quantitation (iTRAQ)[45,46], tandem mass tags (TMT)[47–50], N, N-dimethyl leucine (DiLeu)[51–53], and many other series, allowed the combination of different samples labeled with a series of mass tag variants to be analyzed by MS. Despite encountered challenges related to accuracy and precision[54,55], isobaric labeling methods have been widely adopted in quantitative proteomics. And recent studies have proved that through using a combination of several sets of isobaric tags, up to 45 samples could be compared in a single experiment[56].

In contrast to labeling techniques, label-free approaches rely on measuring chromatographic peak area, calculating ion intensity ratios, or counting MS2 spectra for comparisons between samples[29,57]. Label-free methods offer advantages such as unlimited sample comparisons in an experiment, more efficient protein identification and quantification, and a higher dynamic range of quantification compared to stable isotopic labeling approaches[58]. However, label-free approaches are susceptible to variations that affect MS data, leading to lower reproducibility and accuracy[30].

1.2.3 Data Dependent Acquisition and Targeted Proteomics

Secondly, data acquisition schemes draw attention as liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) become widely used in quantitative proteomics[59,60]. And the most used data collection method in this procedure involves quantifying peptides at the MS1-level and identifying them at the MS2-level. Traditionally, MS2-spectra were obtained by fragmenting precursors observed in MS1-spectra. This acquisition mode depends on MS1 spectra data, usually intensities and charge state of precursors, so-called data-dependent acquisition (DDA). Its MS2-spectra acquisition suffered from under-sampling due to precursor elution lengths and peptide co-elution[61]. However, advancements in mass spectrometry acquisition speed have alleviated this issue, allowing the fragmentation of even low-intensity MS1-peaks. As a result, attention has shifted towards addressing limitations arising from the restricted dynamic range of MS1-spectra[62].

Targeted proteomics focuses on a predefined set of peptides selected by the researcher. This approach typically involves longer accumulation times on single-ion monitoring (SIM) or MS2-scans to ensure sufficient signal and high-quality data for low-abundance targets. However, longer windows restrict the number of targeted peptides. To overcome varied expected chromatographic retention times (RTs) of target peptides between MS runs and reduce window length, target RTs are defined relative to RT standards mixed into the sample, and rescheduling is performed based on the observed RTs of these standards[63]. Dynamic control data acquisition was developed to streamline the scheduling process, where scheduling windows are adjusted based on real-time RT standards elution[64].

Deciding subsequent scans with information from scanned spectra emerged when the strategy of DDA occurs. Intelligent data acquisition in proteomics has evolved from traditional strategies such as dynamic exclusion lists to more advanced approaches like decision trees[65–67]. Early implementations of run-time decision-making processes were limited by instrument constraints, but the release of instrument application programming interfaces by Thermo Scientific has revolutionized the field[68]. This development enables efficient communication between the instrument's CPU and external data systems, expanding the possibilities for intelligent data acquisition.

1.2.4 Data Independent Acquisition

Meanwhile, data-independent acquisition (DIA)[69] has gained popularity in proteomics science due to its unbiased and reproducible nature[70,71]. DIA techniques, unlike data-dependent acquisition (DDA), acquire both MS1 and MS2 data without bias towards precursor ion selection. This overcomes the limitations of DDA, such as irreproducibility and under-sampling. DIA allows the acquisition of MS2 spectra for all precursor ions within isolation windows, enabling the complete coverage of the mass range without prior knowledge of specific precursors[72]. This approach retains accuracy, reproducibility, and consistency comparable to targeted data analysis methods like selected reaction monitoring (SRM) and parallel reaction monitoring (PRM)[70]. However, the computational deconvolution of the multiplexed fragmented spectra acquired in DIA poses a significant challenge.

Several DIA strategies have been developed since the early 2000s, with differing degrees of adoption and revisitation. Shotgun collision-induced dissociation (Shotgun CID) employs in-source fragmentations of all co-eluting ions from chromatographic separation, but its limited throughput and database search strategy hinder widespread usage[73]. The first formally named DIA technique utilized a scanning linear ion trap to sequentially isolate and fragment precursors[69]. It introduced the extended DIA (XDIA) computational pipeline for multiplexed spectra analysis[74]. MS^E and its high-definition variant ($HDMS^E$) alternate scans at different collision energies to obtain precursor and fragment ion spectra, originally designed for metabolite analysis but later extended to proteomics[75,76]. Precursor Acquisition Independent From Ion Count (PAcIFIC) employs gas-phase fractionation with narrow m/z ranges to enable effective peptide identification using conventional database search strategies, at the expense of increased sample injections[77]. All-ion fragmentation (AIF) employs a single orbitrap with no linear ion traps, allowing fragmentation of all ions without precursor isolation[78]. Notably, the introduction of targeted analysis of DIA data, based on statistical models from targeted proteomics, popularized DIA-based methods in proteomics[79,80]. And this concept was implemented by the sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH-MS) in fast-scanning Triple Time-of-flight (TOF) instruments[70].

Since then, several innovative DIA methods have been proposed to further enhance performance. MSX utilizes randomly selected narrow isolation windows, which can be demultiplexed into component spectra to improve precursor selectivity[81]. Scanning quadruple DIA (SONAR) employs wide continuously sliding precursor windows for fragmentation, enhancing the correlation of precursor and fragment ions[82]. Scanning SWATH combines continuous scanning with the first quadrupole and fast-duty cycles, enabling high-flow liquid chromatography for high-throughput proteomic analysis[83]. Wide selected-ion monitoring (WiSIM)-DIA combines conventional DIA with wide selected-ion monitoring windows to produce high-quality precursor ion chromatograms[84]. BoxCar acquisition increases the MS1 dynamic range[62], and PulseDIA utilizes iterative gas phase fractionation to decrease window width, improving specificity and sensitivity[85]. The diaPASEF method combines parallel accumulation-serial fragmentation with DIA in a trapped ion mobility mass spectrometer (TIMS), incorporating ion mobility as the fourth dimension of spectral library searching[86].

1.2.5 Dissolving the MS Spectra through Database Searching Engine

The last key question is the development of computational models that dissolve the complex chromatogram-spectra space generated by MS and convert them into proteins and PTMs. Relatively short sequences, different fragmentation methods, incomplete MS2 spectra, various kinds of PTMs, and labeling groups all add difficulty to figure out the exact sequences of peptides and proteins. Despite these challenges, researchers developed various algorithms to fit all kinds of MS data generated with different labeling groups through different acquisition schemes. A general database search engine procedure starts with filtering out low-intensity peaks and retaining only the most intense ones in MS2 spectra. Then the given sequence library is also processed by simulating enzymatic cleavage and fragmentation, generating theoretical MS2 spectra for comparison with the experimental spectra. Search parameters specified by the user customize the comparison process, and algorithm-specific scores assess the quality of each comparison. This yields a list of peptide candidates known as peptide-spectrum matches (PSMs), along with their corresponding scores. And mzIdentML format has been developed as a standard for exchanging peptide and protein identification results, with many search engines supporting direct export to this format[87].

Database searching originated as an efficient alternative to spectrum sequencing, pioneered by the SEQUEST algorithm[88]. Followed by Mascot[89], a commercial alternative with server-based infrastructure and easily interpretable scores, gained popularity in the scientific community. Numerous search engines, including free and open-source options, have since been developed. The increasing number of search algorithms over time reflects the demand for improved and faster algorithms capable of handling larger datasets. Till 2016, SEQUEST and Mascot were searching engines with the most citations, followed by early open-source alternatives like OMSSA[90] and X! Tandem[91,92]. Also, there was a significant increase in the usage of Andromeda[93] as part of the MaxQuant software[94] highlighting the importance of incorporating the search engine into a global data interpretation pipeline.

The scoring function used for PSMs is an important aspect of the database searching approach. A well-calibrated scoring system is necessary to differentiate between different spectra. For instance, Mascot employs a probability-based scoring method where the total score relates to the probability of the observed match being a random event. SEQUEST utilizes two scoring functions: one to select a limited range of peptide candidates for each spectrum (Sp), and another that measures the cross-correlation between the observed and theoretical spectra (Xcorr)[95]. The Andromeda search engine employed a scoring method based on the probability of chance matches between the number of matched ions and the total number of theoretical ions[93]. And there have been new approaches developed to further improve the performance of peptide identification[96,97].

The database search approach is often followed by a second round of searching against a decoy database to reduce the false discovery rate[98,99]. This procedure ensures that the remaining identified peptides have an FDR higher than the predefined cutoff for peptides identified from the decoy database. Decoy database searching helps estimate a threshold for removing identifications with low statistical confidence, leading to a higher percentage of true positive hits. Analysis platforms like Mascot and MaxQuant commonly integrate this strategy into their standard pipelines. However, as the size of protein sequence databases increases, the target-decoy search strategy becomes computationally inefficient due to the doubled searching space compared to the original search. To address this issue, novel searching strategies including decoy-free[100] and small decoy database[101] have been developed. Recently, a novel approach called Comprehensive Histone Mark Analysis (CHiMA) has been introduced, replacing the target-decoy–based methodology with matched fragment ions criteria to address the small mark size of histone[102]. By employing this new criterion on analyzed datasets, CHiMA nearly doubled the number of previously reported marks in histone.

1.2.6 Methodologies that Facilitate PTM Identification

Identification of PTMs in MS spectra can be computationally challenging due to the large number of possible combinations of positions and modifications. Usually, search engines reduce the search space by only searching PTMs required by users in variable modifications. Meanwhile, methods like ModifiComb[103], PTMselect[104], and G-PTM-D[105] have been developed to decrease computational requirements or increase result accuracy. PEAKS PTM[106], integrated into PEAKS Studio software, searches unassigned spectra with high de novo scores against identified proteins to identify PTMs. Additionally, sequence data from genomic or transcriptomic databases can aid peptide identification by providing possible protein sequences, a strategy commonly used in proteogenomics[107].

MetaMorpheus is a novel tool that incorporates multi-notch searches in global PTM discovery, achieving a higher number of identified PTMs with increased search speed compared to the G-PTM-D approach[108].

1.2.7 Quantification Approaches that Fit Data with Various Labels

Protein and PTM abundance quantification in quantitative proteomics can be achieved using labeled methods or label-free methods. Labeled methods include MS1-based labeling and MS2-based labeling. In MS1-based labeling, different samples are labeled with distinct isotope patterns in the MS1 spectra. And software tools like MaxQuant[109], PVIEW[110], and XPRESS[111] have been developed to handle MS1 labeling experiments, allowing analysis of samples from multiple labeling methods and calculating relative abundance based on labeled peptide pairs.

In MS2-based labeling, quantification signals are detected in the low mass range of the MS2 spectra. Isobaric labeling is commonly used in MS2-based labeling, where samples are labeled with isobaric tags for relative and absolute quantification. Instrument vendors provide commercial software such as ProteinPilot and Proteome Discoverer for analyzing raw data generated by their spectrometers. Free software tools like iTracker[112], IsobariQ[113], and Libra[114] are also available for processing and analyzing data from isobarically labeled samples. Trans-Proteomic Pipeline (TPP)[115] provides a full workflow from raw MS data to quantitative results analysis with the integration of a collection of modules. Skyline[116] and OpenMS[117] are also integrated platforms that provide similar analysis for MS data.

Label-free quantification involves acquiring spectra for different samples separately and addressing variations introduced by LC-MS/MS experiments between runs. Intensity normalization methods like Total Ion Count (TIC) and MaxLFQ[32] are implemented to mitigate variations. Software tools like Mascot Distiller[118], MaxQuant[109], VIPER[119], and Skyline[116] offer label-free quantification based on the signal intensity of peptide precursor ions or spectral counting. The exponentially modified protein abundance index (emPAI)[120] integrated into Mascot and machine learning models can also be used for quantification[121].

1.2.8 Establishment of Whole Proteome PTM Databases

Over the years, significant advancements in PTM analysis strategies have greatly expanded our understanding of the breadth and depth of modifications. These methodologies have enabled the quantification of a wide range of PTMs, contributing to the rapid accumulation of identified and quantified PTM sites under different conditions. To facilitate data collection and communication within the scientific community, numerous large-scale PTM databases and platforms have emerged.

For instance, phosphorylation, one of the most extensively studied PTMs, serves as an illustrative example. In 2010, the Mann lab conducted a comprehensive study in which they quantified 20,443 unique phosphorylation sites on 6,027 proteins, analyzing their dynamics within the context of the cell cycle[122]. In a separate project, the Gygi lab measured the stoichiometry of 5,033 phosphorylation sites from Saccharomyces cerevisiae in triplicates[123]. These groundbreaking studies, focusing on the entire proteome, spurred a rapid increase in the study of quantitative phosphoproteomics. By 2019, the PhosphoSitePlus database alone had recorded approximately 300,000 groups of phosphorylation sites[124]. Furthermore, advancements in quantification methods have facilitated studies involving multiple conditions, various PTMs, and different species[125–

[132]. Overall, these developments in PTM analysis have significantly enhanced our understanding of the diverse landscape of modifications, and this global view of PTMs has paved the way for the involvement of more functional approaches in elucidating their biological significance.

1.3 Functional Proteomics Approaches and Their Application in PTM Studies

1.3.1 Unbalanced Annotation, Dark Proteome and Functional Proteomics

Understanding the intricate interplay between protein and PTM function is crucial for unraveling the mechanisms underlying physiological processes and developing targeted interventions for various diseases. While the biological function of more and more individual protein and PTM sites have been revealed with biochemistry methods, a great majority of them discovered through MS remained in the dark proteome. It has been observed that the attention level differs significantly towards different proteins, with one of the most extensively studied proteins in the human proteome being p53. It receives an average of two publications per day, highlighting its prominence in research[133]. However, the functional understanding of numerous human proteins remains largely unexplored[134–137]. This creates a significant bias, as approximately 95% of life science publications concentrate on a select group of 5,000 well-studied human proteins.[138] Although the sequencing of the human genome was expected to address this bias by providing opportunities to investigate previously unknown genes, 75% of publications continued to focus on genes that have been studied before the release of the genome in 2011[138]. The consequence of this annotation inequality is an impediment to biomedical progress[139]. Mechanistic studies exploring gene-disease associations tend to prioritize proteins that are already well-known, contributing to what is commonly referred to as the street-light effect[140]. Consequently, many uncharacterized proteins receive little attention despite substantial evidence from omics studies linking them to human diseases[134]. Formulating hypotheses regarding the mechanistic molecular function of an uncharacterized protein presents inherent challenges and complexities. However, functional proteomics has the potential to play a crucial role in closing the annotation gap by systematically linking uncharacterized proteins with proteins of known function, enabling their assignment to specific cellular processes. By broadening the scope of the investigation beyond traditional laboratory conditions and incorporating innovative functional experimental designs, it becomes possible to expand the exploration of uncharacterized proteins and reconstruct the intricate network that connects them.

1.3.2 Evolution Conservation of Protein Sequence and PTM Sites

Proteins and PTMs are pivotal in a wide range of biological functions, providing researchers with diverse avenues to explore their functional landscapes. The protein sequence serves as an indispensable feature that influences both protein structure and function. Sequence conservation analysis, which has matured through the analysis of diverse genomes, has become a well-established tool in this regard. With the availability of genome sequences, protein sequences can be generated in silico, facilitating sequence-based conservation analysis[141,142]. Conservation analysis goes beyond the examination of amino acid residues and extends to encompass PTM sites. It has been observed that PTM sites with known functions often exhibit significant conservation[143–145]. Building on this observation, the concept of an evolutionary tree can be integrated with identified PTM sites, enabling the prioritization of experimental validation based on the conservation of modification sites[146]. An illustrative example of such an approach is the work conducted by the Korgan lab[143]. They performed an analysis involving 11 eukaryotic species and compiled a comprehensive dataset of approximately 200,000 PTM sites. Their study explored evolutionary conservation at both the sequence and modification levels, leading to the identification of functionally important PTMs and regulatory regions. Through this integrative approach, they gained insights into the functional significance of PTMs by examining their evolutionary conservation patterns.

1.3.3 Relate Protein 3D Structure with Protein Functions

The three-dimensional (3D) structure of proteins is a crucial feature that significantly influences protein function. Determining the protein's 3D structure not only reveals the secondary structure and biochemical environment of each amino acid residue but also provides information about protein regions, domains, and potential interaction partners. AlphaFold, developed through deep learning algorithms and based on extensive protein sequence and corresponding 3D structure data, has achieved remarkable accuracy in predicting protein structures[147,148]. This powerful tool covers almost the complete human proteome and proteomes of over 20 model organisms, providing a structural prediction database (AlphaFold DB; https://alphafold.ebi.ac.uk). This resource enables proteome-wide structural investigations, promising to enhance our understanding of the relationship between protein structure and function.

The application of AlphaFold has also been proposed to have immense potential for incorporating and analyzing PTMs[149]. A recent study conducted by the Mann lab took advantage of AlphaFold's structural predictions and integrated PTM data on a proteome-wide scale using deep learning-based models[150]. They combined essential features of protein domains with functionally relevant predictions, leading to the discovery of numerous sites with potential regulatory roles. Furthermore, they employed 3D proximity analysis to illustrate spatial coregulation examples of potential PTM crosstalk, involving multiple PTMs. To facilitate sharing of their findings with the research community, the lab also developed programs that assist researchers in processing and visualizing protein and PTM 3D structures.

1.3.4 Interaction between Proteins Implied Their Locations and Functions

The field of protein-protein interactions (PPIs) is closely linked to protein function. PPIs not only indicate that interacting proteins coexist in the same cellular location at certain points but also imply their functional relevance in specific biological processes. The advancement of affinity purification mass spectrometry (APMS) has greatly assisted in studying PPIs. The BioPlex project, leveraging APMS technology, has generated a vast human interactome database with high proteome coverage[151,152]. The latest version, BioPlex 3.0, encompasses results from affinity purification of 10,128 APMS experiments in 293T cells, capturing 118,162 interactions involving 14,586 proteins. Additionally, results from 5,522 immunoprecipitations in HCT116 cells are also included in the database.

These interaction networks provide rich information about protein localization and function. The OpenCell project took a different approach by employing an endogenous fluorescent tagging technique, which enabled the visualization of protein locations within living cells[148]. By conducting immunopurification-mass spectrometry experiments using these tagged cells, they obtained a physical interaction network consisting of 1,310 proteins, along with their corresponding cellular localization map. Both the BioPlex and OpenCell databases, along with other databases that collect biochemical results of protein interactions such as BioGriD[153], offer valuable insights into the spatial organization and functional relationships of proteins. These databases provide a wealth of information that can be leveraged to understand the interplay between protein interactions and PTMs.

1.3.5 Emerging Approaches: Turnover Analysis and Thermal Profiling

Turnover analysis and thermal profiling are two emerging approaches that provide insights into the dynamics and stability of proteins. Protein turnover is intricately linked to its synthesis and degradation processes, making it a powerful tool for bridging the gap between protein discovery and functional understanding. The advent of dynamic SILAC or pulse SILAC workflows[154,155], which involve feeding cells with stable isotopic amino acids for a certain period of time, has enabled turnover analysis in protein and PTM studies under various environments[156–158]. Thermal profiling, on the other hand, investigates the denaturation process of proteins as temperature increases and proves valuable for exploring protein biological functions and protein-protein interactions[159]. In recent years, there has been a surge in thermal profiling studies focusing on characterizing the thermal profiles of proteomes under different conditions. These investigations have explored the influence of factors such as solution concentration[160], PTMs[161], cell line differences, and proteoform groups[162]. Together, turnover analysis and thermal profiling contribute to our understanding of dynamics, stability, and functional implications and proteins and PTMs.

1.3.6 Diversity of Functional Approaches in Proteomics and PTM Field

In summary, the construction of a functional annotation network for proteins and PTMs can be achieved through multiple approaches, extending beyond the categories mentioned above. Changes in protein expression levels and PTM stoichiometry in response to diseases, drug treatments, and different stimuli can provide valuable insights into their functional roles. Furthermore, integrating data from genomics, transcriptomics, metabolomics, and functional annotations verified through biological experiments allows for a more comprehensive understanding of the functional proteome.

By combining information from various omics disciplines and experimental validations, we can gradually unravel the intricate connections and interactions within the functional proteome. This holistic approach enables us to paint a more complete picture of protein function, taking into account their expression patterns, post-translational modifications, and their impact on cellular processes. As our knowledge and technological capabilities continue to advance, the integration of diverse datasets and experimental evidence will facilitate a deeper understanding of the functional landscape of proteins and PTMs.

Chapter II – Development of UbE3-APA for Annotating Quantitative Proteomics Results Functionally

Yao Gong[1,2], Yue Chen[1,2]

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

[2]Bioinformatics and Computational Biology Program, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

This chapter contains an original research article previously published.

Gong Y, Chen Y. UbE3-APA: a bioinformatic strategy to elucidate ubiquitin E3 ligase activities in quantitative proteomics study. *Bioinformatics*. 2022;38(8):2211-2218. doi:10.1093/bioinformatics/btac069

Reproduced with permission from Oxford University Press, Copyright © 2022

2.1 Summary

Ubiquitination is widely involved in protein homeostasis and cell signaling. Ubiquitin E3 ligases are critical regulators of ubiquitination that recognize and recruit specific ubiquitination targets for the final rate-limiting step of ubiquitin transfer reactions. Understanding the ubiquitin E3 ligase activities will provide knowledge in the upstream regulator of the ubiquitination pathway and reveal potential mechanisms in biological processes and disease progression. Recent advances in mass spectrometry-based proteomics have enabled deep profiling of ubiquitylome in a quantitative manner. Yet, functional analysis of ubiquitylome dynamics and pathway activity remains challenging.

Here, we developed a UbE3-APA, a computational algorithm and stand-alone python-based software for Ubiquitin E3 ligase Activity Profiling Analysis. Combining an integrated annotation database with statistical analysis, UbE3-APA identifies significantly activated or suppressed E3 ligases based on quantitative ubiquitylome proteomics datasets. Benchmarking the software with published quantitative ubiquitylome analysis confirms the genetic manipulation of SPOP enzyme activity through overexpression and mutation. Application of the algorithm in the re-analysis of a large cohort of ubiquitination proteomics study revealed the activation of PARKIN and the co-activation of other E3 ligases in mitochondria depolarization-induced mitophagy process. We further demonstrated the application of the algorithm in the DIA (data-independent acquisition)-based quantitative ubiquitylome analysis.

Source code and binaries are freely available for download at URL: https://github.com/Chenlab-UMN/Ub-E3-ligase-Activity-Profiling-Analysis, implemented in python and supported on Linux and MS Windows.

2.2 Introduction

Ubiquitylation is a key protein post-translational modification (PTM) involved in diverse cellular processes including protein homeostasis, cell signaling and epigenetic regulations. Its E1–E2–E3 cascades linking an isopeptide bond between the c-terminus of ubiquitin and a lysine residue of the target protein[163] to form a mono- or a polymer chain of ubiquitin with eight distinct linkage types. Ubiquitylation not only acts as the essential modification in protein degradation through proteasome that accounted for the breakdown of over 80% of the proteins[164], but it also plays a crucial role in non-degradative functions, including regulation of protein trans-location, protein–protein interactions and enzymatic activity[165]. Within the process of ubiquitylation, E3, also known as the ubiquitin ligase, mediates the ubiquitination substrates specificity[163,166]. Changes in the E3 ligase activities will lead to changes in ubiquitination of its target proteins, and further regulate various downstream cellular processes including cell-cycle, apoptosis and transcription regulation.[167] Studies have found that dysfunction of E3 ligases and these cellular functions may lead to neurodegenerative diseases[168,169], cardiovascular diseases[170] and development of cancer[171,172], while therapies and drugs have been developed to target specific E3 ligases for potential clinical applications[173,174]. Therefore, it is important to develop strategies that evaluate the activities of different E3 ligases in a system-wide manner.

Recent advances in mass spectrometry have enabled deep profiling of PTM pathways. Combining with quantitative proteomics strategies such as SILAC and isobaric labeling, proteomics analysis allows system-wide profiling of PTM dynamics at the site-specific level. Such quantitative information provides a rich resource to develop computational tools evaluating PTM pathway activities[175]. Recent efforts studying kinase activities based on quantitative phosphorylation datasets have led to the development of several tools, including PTMsigDB[176], IKAP[177], KinasePA[178], KSEA[179] and KEA3[180]. Among these models, KEA3 collected 24 kinase substrate libraries from different sources as their database and test significance of kinase integrating sum rank tests results of all libraries, while the rank-sum test in PTMsigDB was supported by a collection of site-specific PTM signature of perturbations, kinase states and pathway activities from published studies[176]. IKAP used a non-linear optimization routine to find enriched kinase, KinasePA used the direction pathway analysis to study insulin pathways[178] and KSEA applies z-score test to find differentially activated kinases. Despite these advances in kinase analysis, there is lack of bioinformatic strategies for evaluating ubiquitin E3 ligase activities.

Improvements in biochemical enrichment and chemical labeling strategies have allowed global quantification of ubiquitination dynamics[181–184] and measurement of site-specific ubiquitination stoichiometries[185]. Recent bioinformatic efforts have led to the development of multiple enzyme–substrate databases in the ubiquitination pathway[186–190]. In this study, based on an integrated resource of ubiquitin E3 ligase and substrate network, we proposed a computational strategy UbE3-APA, Ubiquitin E3 ligase Activity Profiling Analysis (Figure 2-1), for systematic evaluating ubiquitin E3 ligase activity based on quantitative ubiquitylome analysis. The model was validated with two published large-scale proteomics studies with different biological context[191,192] and confirmed known regulatory mechanisms in the pathway.

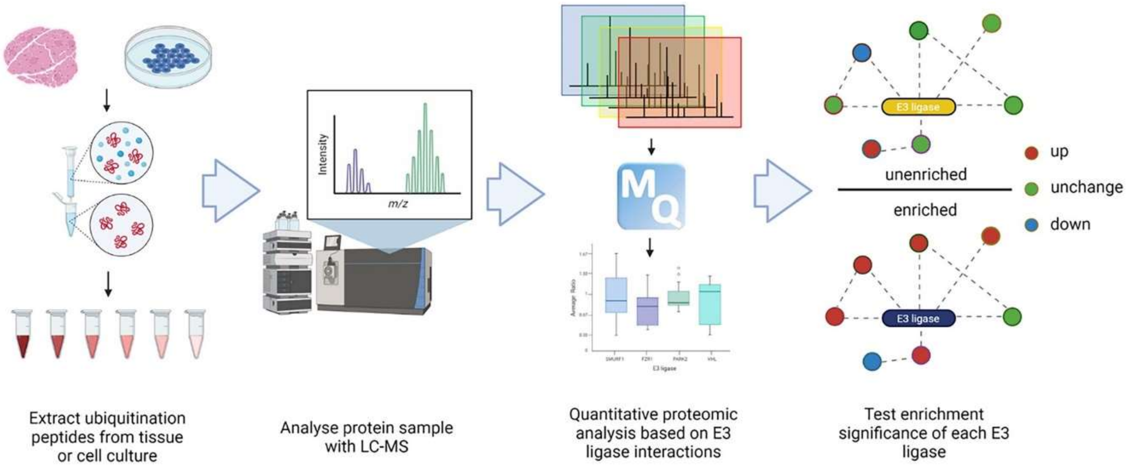

Intensity
m/z
MQ
E3 ligase
unenriched
enriched
E3 ligase
up
unchange
down
Extract ubiquitination peptides from tissue or cell culture
Analyse protein sample with LC-MS
Quantitative proteomic analysis based on E3 ligase interactions
Test enrichment significance of each E3 ligase

Figure 2-1 A workflow of analyzing quantitative ubiquitylation proteome with E3 ligases activity profiling analysis

2.3 Results

2.3.1 Establish a Comprehensive ESI Network

We collected datasets with rich information of ESIs supported by biological experiments from multiple sources, including the UbiBrowser, the Ubinet and another published ESI database[186,189,190]. Comparing the information from the three data sources showed a largely overlapping information with some differences (Figure 2-2). After removing redundancy and discrepancies, we established an integrated database for human ubiquitin E3 ligase–substrate interaction that includes 354 E3 ligases and 2501 interactions. All E3 ligases interactions collected in the database were built in the package to allow comprehensive analysis.

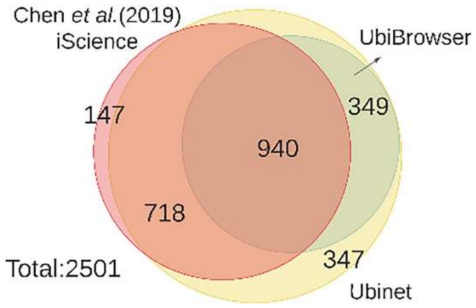


Figure 2-2 Overlap and integration of the E3 ligase–substrate data resources

2.3.2 Profile E3 Ligase Activity in Quantitative Ubiquitylome

With the establishment of a comprehensive E3 ligase–substrate database, the program collects the site and protein-specific data from quantitative ubiquitylome studies. To profile E3 ligase activity based on this data, we reasoned that a physiologically meaningful changes in ubiquitin E3 ligase activities should be reflected on the overall changes of the ubiquitination abundance of their corresponding targets (Figure 2-3a). As the ubiquitylome proteomics analysis mainly provide site-specific quantification of ubiquitination, the program offers the option to calculate the averaged site ratios of a target protein to represent the ubiquitination changes of each protein for the protein-level E3 ligase activity profile

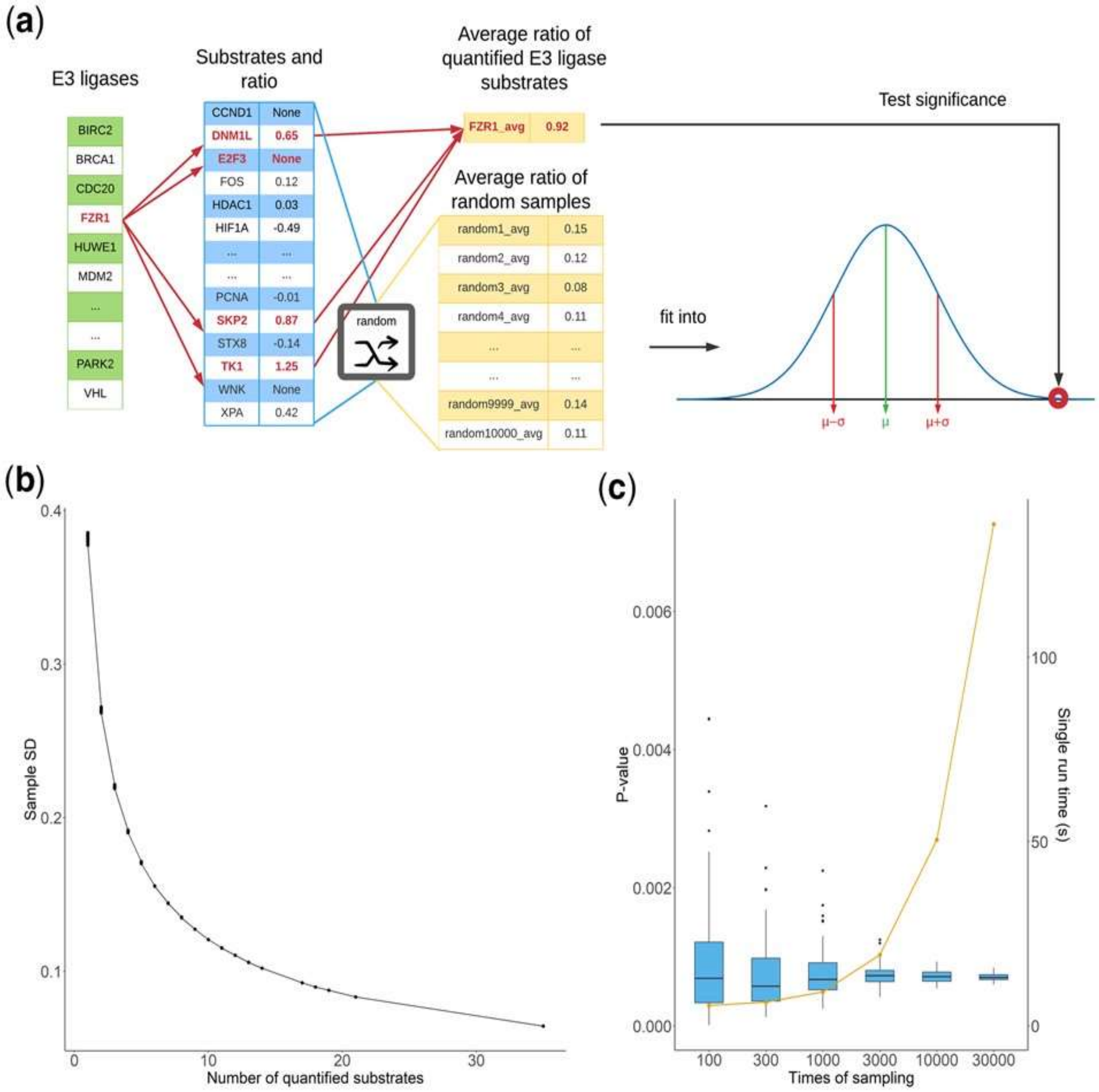


Figure 2-3 Establishing the statistical model for UbE3-APA Statistical modeling to evaluate an E3 ligase activity profile in UbE3-APA analysis. (b) Variations of standard deviations of the sampling ratios toward E3 ligases with different number of quantified substrates (testing with a dataset in the SPOP study). (c) Variations of SPOP P-value and analysis time for single run in seconds (s) with various number of sampling times for E3 ligase activity profiling testing with a dataset in the SPOP study

analysis. Then, all quantified ubiquitination proteins for a specific E3 ligase in the dataset will be collected as a sample group. The same number of quantified ubiquitination proteins as the number of quantified substrates for any given E3 ligase will be randomly selected from the ubiquitylome dataset in a bootstrapping procedure. Based on the Central Limit Theorem, the average quantification ratios of each group of randomly selected ubiquitinated proteins should form a normal distribution. Based on this reference distribution, we can estimate the statistical significance of the average ubiquitination ratio of the substrate group for an E3 ligase (Figure 2-3a). A significant change in the activity profile can indicate a significant increase or decrease of E3 ligase activity in the context of experimental conditions compared to the overall changes in ubiquitination dynamics in the background. The protein-level E3 ligase activity profile analysis was used for the downstream applications.

To further explore how the number of quantified substrates and the times of sampling may affect the analysis processes and results, we performed tests with different parameters (Figure 2-3b-c). Our data showed that if more substrates of any given E3 ligases were quantified, the standard deviation of randomly selected sample ratios for the activity profile analysis decreased, which therefore led to more statistically significant estimation of E3 ligase substrate ratios based on the distribution (Figure 2-3b). This built-in mechanism of our model certainly supports the notion that E3 ligase activity profile could be better assessed if more E3 ligase substrates were quantified. Our test of the sampling process showed that increasing the number of random samplings in the bootstrapping process could reduce the variation of P-values calculated based on the sampling distribution and therefore led to more reliable and precise estimation of the statistical significance (Figure 2-3c). On the other hand, increasing the number of random samplings would also cost more time per run and reduce the efficiency of the analysis (Figure 2-3c). Considering both efficiency and reliability of the results, we have selected 10 000 times of repeats for the random selection process in E3 ligase activity profiling analysis.

2.3.3 Validate UbE3-APA Workflow with the Quantitative Ubiquitylome Analysis of SPOP E3 Ligase

To validate our algorithm, we chose a quantitative proteomics study that aimed to characterize ubiquitination dynamics that was mediated by SPOP, an E3 ligase that is frequently mutated in prostate cancer and affects the regulation of downstream pathways in cancer progression[192]. This study included two sets of quantitative proteomics experiments and each set of experiment aimed to quantify the ubiquitination dynamics upon the overexpression of vector control, SPOP-wild-type (WT) and one of the two SPOP-mutants F133L and Y87N. Both mutations are naturally occurring mutations in prostate cancer and known to suppress the SPOP-WT induced ubiquitylation. The quantitative analysis was performed using SILAC workflow with the expression of each form of SPOP-WT or vector control pairing to one of the SPOP-mutant.

Using UbE3-APA workflow, we analyzed the normalized quantitative ubiquitination ratios included in their Supplementary Material across all six pairs of SILAC experiments. The activity profiling analysis showed that the SPOP activity was significantly enriched in cells overexpressing SPOP-WT when comparing to cells overexpressing vector control or either one of the SPOP mutants (Figure 2-4a). When comparing cells expressing SPOP mutant and vector control, our model found no significant changes in SPOP activity. These profiling analysis results matched well with SPOP ubiquitylation activity differences expected in the original study. The activity profiling analysis also allowed us to generate volcano plots with the statistical significance test and quantification ratios. Two examples with the Y87N (L)—WT (H) group and the Y87N (L)—mutant (H) group were shown (Figure 2-4b-c). As clearly shown, when SPOP-WT was overexpressed, the SPOP activity increased significantly comparing to the overexpression of SPOP-Y87N mutant with significantly increased SILAC H/L ratios of SPOP target proteins, while the SPOP activity did not change when comparing the cells overexpressing SPOP-Y87N mutant and vector control.

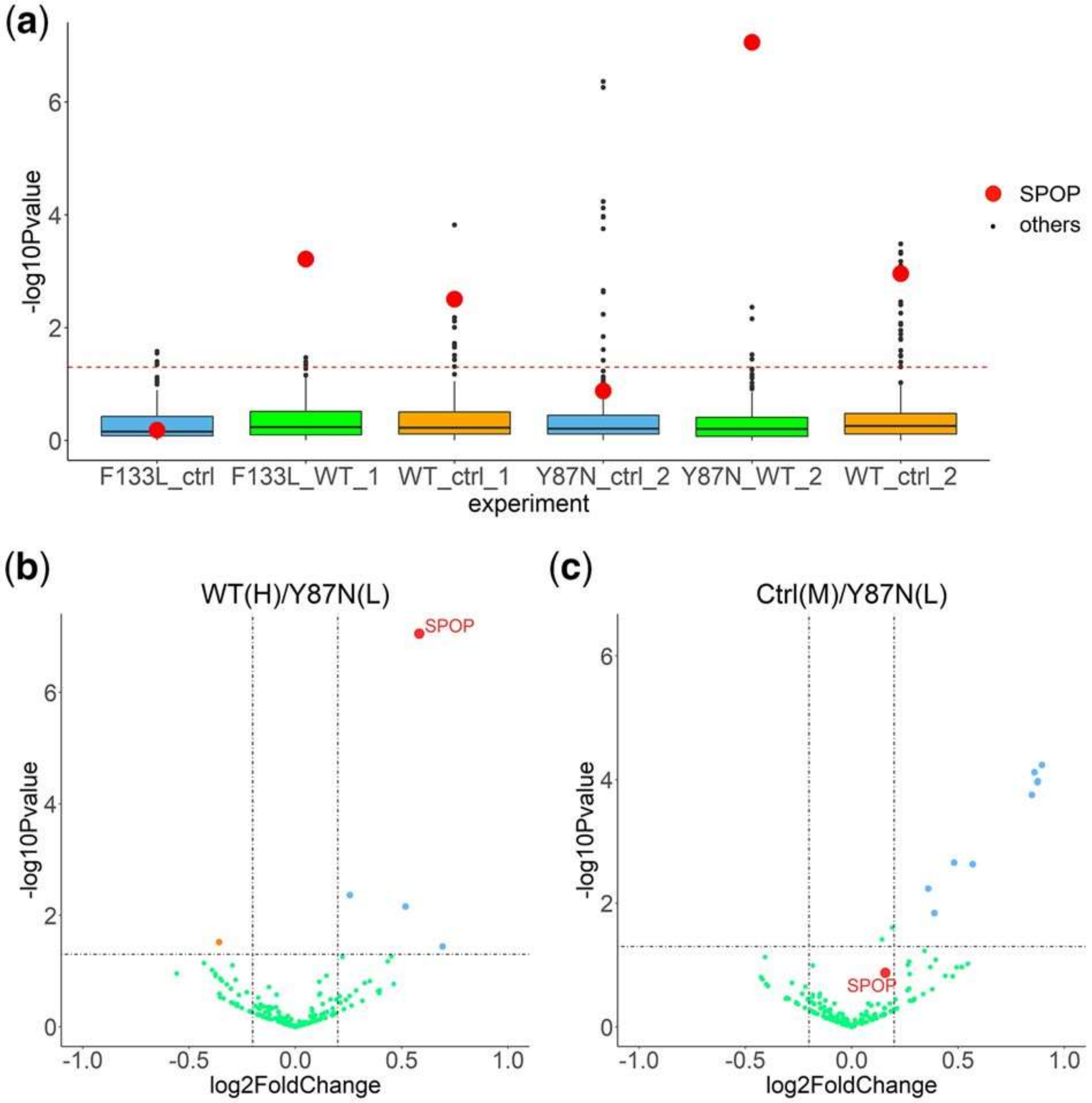


Figure 2-4 Protein level E3 ligase activity profiling results of the SPOP study. Ctrl, vector control; WT, wild-type-SPOP; 133 L, SPOP-F113L; 87 N, SPOP-Y87N; 1 and 2, experiment set one and two. (a) All experiments of the SPOP study in box plot, each box contained enrichment P-values of all E3 ligases in one SILAC experiment. (b) SPOP-Y87N to SPOP-WT group in volcano plot. (c) SPOP-Y87N to vector control group in volcano plot

2.3.4 Apply UbE3-APA model to profile E3 ligase activities in response to mitochondrial depolarization

We applied UbE3-APA workflow to analyze a quantitative ubiquitylome study that focused on PARKIN and global ubiquitylation network in response to mitochondrial depolarization[191]. We applied UbE3-APA workflow to analyze a quantitative ubiquitylome study that focused on PARKIN and global ubiquitylation network in response to mitochondrial depolarization[191]. This large-scale study included 73 quantitative ubiquitylome proteomics analysis to explore the dynamics of the ubiquitination pathways under various mitochondrial depolarization treatment as well as in cells with different genetic background, and detailed information of experiment condition of each group we collected from Supplementary Data in original paper[191] were listed. Analysis of all the datasets with UbE3-APA workflow showed that when mitochondria was not damaged, there was not an apparent PARKIN activity even when PARKIN was overexpressed. Once the mitochondria were polarized, there was a significant increase of PARKIN activity (Figure 2-5a). Inhibition of Pink1, the upstream kinase activating PARKIN, abolished the activation of PARKIN as expected when mitochondria was depolarized. When mitochondria were depolarized, we could not see an apparent activation of PARKIN based on ubiquitylome analysis data (Figure 2-5b). But when cells were treated with bafilomycin, an autophagy inhibitor, there was a strong indication of activation of PARKIN upon mitochondria depolarization (Figure 2-5c), suggesting that the ubiquitinated substrates of PARKIN could not be efficiently degraded. Therefore, this data agrees well with the knowledge that PARKIN activation led to mitochondria degradation through autophagy process and it also suggested that PARKIN substrates are mainly degraded through autophagy pathways. In addition, the dataset also included the proteasome inhibition experiment upon mitochondria depolarization. Interestingly, our analysis showed that the inhibition of proteasome activity alone showed no strong enrichment of PARKIN substrate ubiquitination, which could suggest that either PARKIN substrates were mainly degraded through processes other than proteasome degradation (such as autophagy), or PARKIN was not activated upon mitochondria depolarization when proteasome was inhibited (Figure 2-5c). Analysis of the dataset with the cotreatment of cells with both proteasome inhibitor and autophagy inhibitor showed that PARKIN was indeed not activated upon proteasome inhibition because even autophagy inhibitor treatment failed to enrich PARKIN ubiquitination substrates (Figure 2-5c). This finding agreed well with previously published observation that upon proteasome inhibition, mitochondria depolarization failed to induce mitochondria fragmentation despite PARKIN translocation to mitochondria[193].

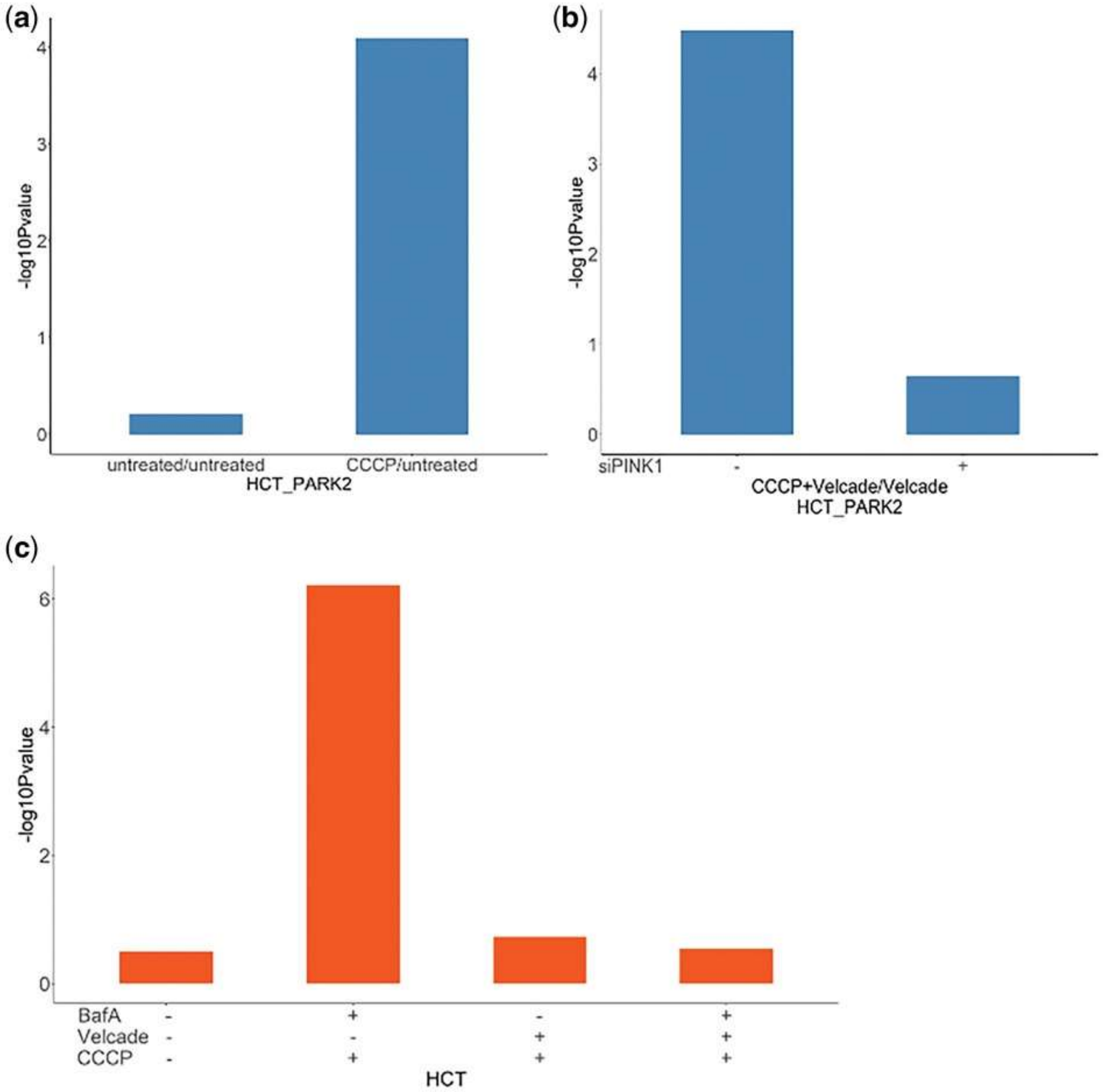


Figure 2-5 Ubiquitin E3 ligase activity profiling analysis of PARKIN under various genetic and chemical treatment. (a) PARKIN activity difference in HCT116 cells with PARK2 overexpression and with or without carbonyl cyanide m-chlorophenyl hydrazone (CCCP, mitochondrial depolarization inducer) treatment. (b) PARKIN activity difference in HCT116 cells with PARK2 overexpression and with or without Pink1 inhibition. (c) PARKIN activity difference in HCT116 cells with the treatment of bafilomycin (BafA, an autophagy inhibitor) and/or Velcade (a proteosome inhibitor)

We then integrated the analysis of the ubiquitylome dynamics in all experimental conditions in the mitochondria depolarization study and plot the E3 ligase activity profiles in heatmap (Figure 2-6). The E3 ligases were clustered with hierarchical clustering based on how similar their activity profiles change under different treatment conditions. Out of 203 E3 ligases profiled by our model across all experimental conditions, three E3 ligases (FZR1, AMFR, MARCHF5) showed a very similar activity pattern to PARKIN across most of the experimental conditions. Since the activity profiles of E3 ligases were analyzed based on their corresponding substrates, it was likely that E3 ligases showed similar activity profiles when they shared common substrates. For better clarification, we mapped the E3-ligase—substrate interaction networks for the four E3 ligases (Figure 2-7a). The network indicated the shared and unique connection between each E3 ligase and corresponding substrates, the number of times the substrates were quantified under all conditions and the significance of ubiquitination level changes for the substrate proteins. We can clearly see that the unique substrates of AMFR and FZR1 did not change significantly to contribute to activity profiles and their activity changes were mainly caused by the changes of ubiquitination levels in substrates shared with PARKIN. Only MARCHF5 had unique substrates whose ratios were changed significantly and similarly along with those unique substrates of PARKIN. For better clarification of the data, we included 'Group' mode to the result output. In this mode, the software will group E3 ligases that share substrates together if the E3 ligases do not have unique substrate and only the E3 ligases that contain all the substrates in the group were labeled as leading E3 ligases of the group. We re-analyzed all the data using the Group mode and identified 127 E3 ligase groups. Then, we performed correlation analysis of all the E3 ligase groups in distance matrix. The data clearly showed that MARCHF5 and PARKIN shared the most similar activation profiles (Figure 2-7b). This finding confirmed that MARCHF5 was also activated during mitochondrial depolarization, which agrees well with the previous finding that PARKIN-dependent ubiquitination targets MARCHF5 for translocation and activation[194].

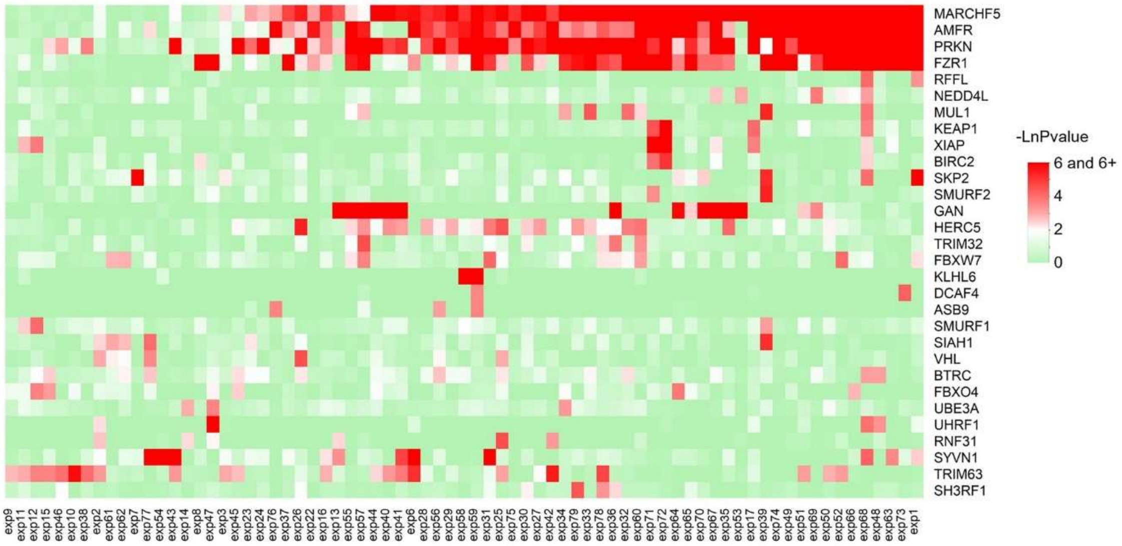


Figure 2-6 E3 ligase co-activation profiles across all experiments involved in the mitochondria depolarization study. The P value enrichment of E3 ligases across 73 groups of experiments were -ln transformed. Only E3 ligases that were enriched ($P < 0.05$) in at least one experiment were included in this heatmap.

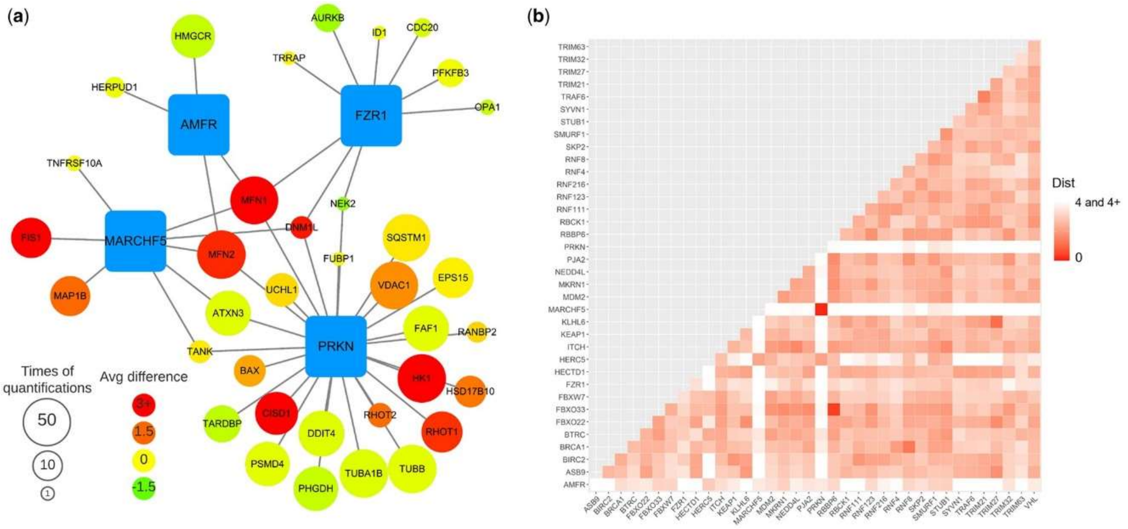


Figure 2-7 Regulation networks between four E3 ligase, PARKIN, MARCHF5, AMFR and FZR1 in the PARKIN study.(a) Interaction network of four E3 ligases and their substrates found in the mitochondria depolarization experiments. Blue square represents E3 ligases, circle represents substrates, edge represents Enzyme–substrate Interaction (ESI), size of circles indicates the number of groups (out of 73 experimental groups) this substrate being quantified and used for the corresponding E3 ligase activity analysis, and the color of the circle indicates the difference in average log2 quantification ratio between the substrate ratio and the group average. (b) Correlation matrix and heatmap of leading E3 ligase activity profiles from all experiments. Color gradient represents 2-D Euclidean distance (Dist) between a pair of leading E3 ligase activity profiles. Only E3 ligases with P < 0.2 in at least one experiment were included.

2.3.5 Apply UbE3-APA Model to DIA Dataset

To further explore the usage of our UbE3-APA model, we applied it to two recently published quantitative ubiquitylome studies based on DIA analysis. The first study explored how the ubiquiyltome was associated with the TNF signaling pathways[195] by treating cells with or without TNF. We collected site-specific intensities of all replicates in the treated group and mock group and calculated the intensity ratios between TNF treated and mock treated cells. Then we performed UbE3-APA analysis on the protein level in the grouped mode. The activity profiling analysis revealed several distinct up- and down-regulated E3 ligases under TNF treatment (Figure 2-8). Among these up-regulated E3 ligases, TRAF2 and TRAF6 were members of tumor necrosis factor receptor-associated factors whose ubiquitination activity was crucial in the TNF signaling pathways[196]. In

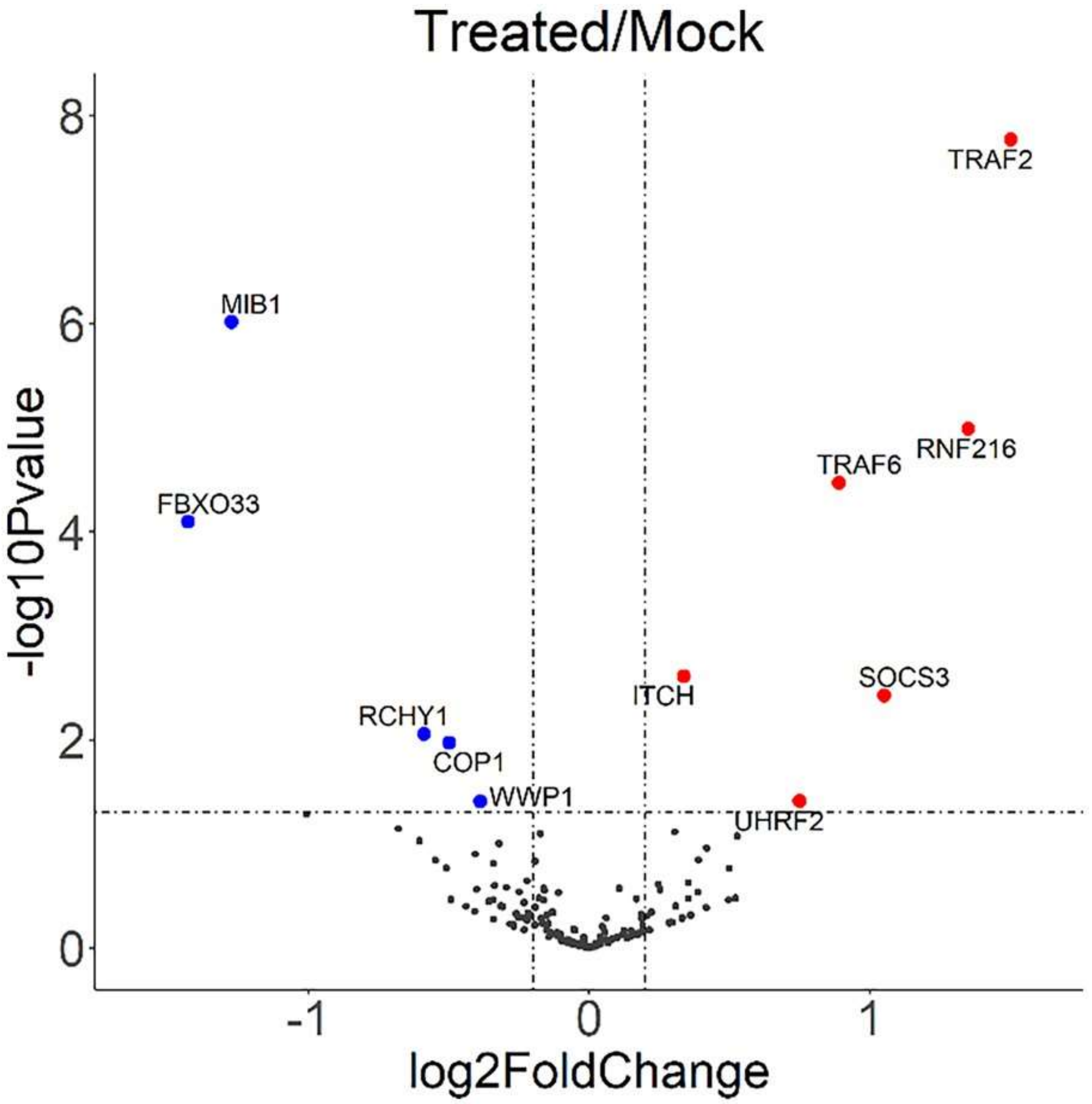


Figure 2-8 E3 ligase activation analysis based on the ubiquitination dynamics between TNF treated and mock treated cells in the TNF study

addition, our analysis also identified up-regulation of activity for RNF216, SOCS3 and down-regulation of activity for MIB1 and FBXO33.

The second study investigated the ubiquitylome changes in response to the inhibition of deubiquitinase USP7 by siRNA knockdown or chemical inhibitor such as FT671[196]. We extracted their DIA-based ubiquitylome data under these four conditions: siCTRL+DMSO, siCTRL+FT671, siUSP7+DMSO and siUSP7+FT671 and then calculated the ratio difference between FT671 and DMSO treatments under either siCtrl background or siUSP7 background. Analyzing the two pairs of ubiquitylome datasets with UbE3-APA revealed differential activation profiles of E3 ligases upon the FT671 treatment with siUSP7 background or siCTRL background (Supp Figure 2-1). In agreement with the findings in the published study, more E3 ligases showed altered activity profiles upon FT671 treatment but these activity changes were attenuated under siUSP7 background, suggesting that FT671 treatment was specific in targeting USP7 activity in cells.

2.4 Discussion

Advances in quantitative proteomics have enabled large-scale profiling of ubiquitination substrates, also known as the ubiquitylome. Application of quantitative proteomics in ubiquitylome analysis revealed the key ubiquitination targets in the biological processes and determined the downstream signaling pathways that were most significantly affected by the ubiquitination process. Yet, few studies systematically examine the upstream regulatory pathways of ubiquitination. Analysis of regulatory enzyme activities has been largely limited to a few well-selected targets of each enzyme. Recent advances in the collection and biological validation of E3 ligase–substrate database provide a great opportunity to use ubiquitylome quantitative analysis as an activity-readout to profile the ubiquitin E3 ligases.

In this study, we developed a statistical framework and workflow to identify the ubiquitin E3 ligase activity in a high-throughput and unbiased manner. This open-source python package enabled effective profiling of E3 ligase activities through robust statistical analysis based on quantitative ubiquitylation results. In the case study of SPOP E3 ligase ubquitylome analysis, our model correctly validated the SPOP activity upon the overexpression of SPOP WT and mutant forms with various activity. Application of our workflow to analyze the ubiquitylome dynamics upon mitochondria depolarization confirmed that activation of PARKIN E3 ligases under various conditions and unexpectedly discovered the role of proteasome inhibition on PARKIN activation. Our statistical framework allowed us to collect the E3 ligase activity profiles across multiple conditions. Application of our workflow to profile 73 quantitative ubiquitylome analysis enabled clustering analysis of E3 ligases across the experimental conditions and revealed the co-activation of PARKIN and MARCHF5 upon mitochondrial depolarization. Our methods were further applied to two studies with DIA ubiquitomes, and in both case studies, E3 ligases related with the treatment proved by previous papers were revealed by our model through activation profile changes.

The statistical framework described in this study can be generally applied to other PTM pathway analysis. We have recently applied the workflow and developed Kinase Activity Profiling Analysis (KAPA) to identify iron deficiency induced activation of AMPK pathway in neuronal cells[197]. Our study demonstrated that it is possible to apply statistical analysis workflow to systematically profile E3 ligase activity. However, we also recognize that efficient analysis of E3 ligase activity in a system-wide manner is limited by several factors. First, the number of E3 ligase–substrate interactions in our knowledgebase is still limited compared to other PTMs such as phosphorylation and acetylation. Our integrated database from various sources contained 2354 gene-level interactions of humans in total, which is quite small comparing to 13 855 gene-level interactions between phosphorylation sites and kinases collected by PhosphoSitePlus in human[198]. Continued effort in the high throughput discovery of E3 ligase and substrate interaction is needed to expand the knowledgebase for more reliable and confident analysis of upstream regulatory enzyme activities.

Secondly, although current high throughput proteomics have allowed in-depth quantification of ubiquitylome in single experiment, the data-dependent analysis (DDA) often suffers from limited reproducibility and reduced quantification precision. For example, in our analysis of mitochondria depolarization ubiquitylome study, for E3 ligase FZR1 and AMFR, they have 56 and 14 substrate proteins respectively based on our ESI database, but only 10 and 4 substrates were found at least once in all 73 groups of experiments. Therefore, their activity profiles were affected by the shared substrates with PARKIN. If more substrates were reproducibly quantified, the analysis profiles of the two E3 ligases could be more accurate and informative. Application of DIA for ubiquitylome analysis as we demonstrated in our study will certainly help address this challenge[195].

Lastly, currently E3 ligase and substrate interaction database has been largely based on the protein-level and there is limited knowledge on the site-specificity of ubiquitination regulatory pathways comparing to the knowledge on the kinase-phosphorylation regulatory network. Lack of site-specific regulation information presents a challenge to reveal potential regulatory enzyme activities on overlapping protein substrates but on distinct target sites., making it less precise compared to other well-studied PTMs based on the regulatory network between enzymes and sites, for example, phosphorylation and acetylation. It requires the continued development technologies to identify major enzyme target sites in the ubiquitination pathway. Future updates of the program will include an updated E3 ligase–substrate interactions database with the potential for site-specific enzyme activity analysis.

2.5 Materials and Methods

2.5.1 Collecting E3–substrate Interactions

We integrated ubiquitin E3 ligases and substrates relationship data from the following three sources: UbiBrowser[189], Ubinet[190,199] and a multidimensional database collection[186]. UbiBrowser is an extensive database that collects interactions between E3 ligases and substrates. They incorporate E3–substrate interactions (ESIs) from both literature manual curations and predictions, which were based on a set of biological features and Bayesian models, in their database. Another online platform that updated recently, Ubinet, focuses on ESI collection, prediction, and visualization across different species. They predicted ESIs based on the substrate specificity of E3 ligases extracted from experiment verified interactions. To characterize the interaction network between E3 ligases and their substrates, the Chen group collected ESIs from a variety of sources: E3net[188], hUbiquitome[187], Uniprot[200], and BioGRID[201]. They collected ESIs directly from the first two sources, and they gathered the interaction between E3 ligase and proteins in the last two databases through data mining and selected those physical interactions supported by low throughput methods. By integrating the Ub E3–substrate interaction network from these three resources, we established the database for Ub E3 ligase activity profiling analysis

(Supplementary Table S2). Only the interactions that were supported by literature from all sources above were integrated into our database. Those interactions solely supported by prediction models were not included.

2.5.2 Algorithm Development for Ubiquitin E3 Ligase Activity Profiling Analysis

The E3 ligase activity analysis model profiles E3 ligase activities based on a bootstrapping procedure by evaluating the difference between the quantitative ratios of E3 ubiquitination targets and the overall background. Firstly, the program collects the quantitative ratios of identified ubiquitination sites and proteins from proteomics analysis. To normalize the site-specific ubiquitination ratios for statistical analysis, the program offers two options—computationally normalized values or protein-normalized values. Computationally normalized ubiquitination ratios are often provided by the quantification software. For example, MaxQuant provides normalized site-specific ubiquitination ratios based on the median ratios of all quantified sites. To obtain protein-normalized ubiquitination ratios, the program will fetch the original ratios of both the ubiquitination sites and their corresponding ubiquitination proteins. The protein quantification ratios should be calculated excluding ubiquitinated peptides. The protein-normalized ubiquitination site ratios are calculated by dividing the original site ratios by the original ratios of the corresponding proteins. The normalized site-ratios are then log2 transformed for downstream analysis and averaged to generate the quantification ratios of ubiquitination protein substrates.

Secondly, based on the integrated ubiquitin E3 ligase–substrate database, the program iteratively analyzes each E3 ligase and extracts all substrates quantified for each E3 ligase in the quantitative datasets. Then, for each E3 ligase, the program collects the total number of quantified targets and the average quantification ratios of its targets.

Thirdly, the program performs randomized selection from the quantification datasets. At this point, the program offers two options for enrichment testing—protein-level profile analysis and site-level profile analysis. For protein-level profile analysis, the program randomly selects the same number of ubiquitylation proteins as the number of targets for a specific E3 ligase and then computes the average quantification ratios of selected ubiquitination proteins. For site-level profile analysis, the program randomly selects the same number of ubiquitylation sites as the number of sites quantified for known targets of a specific E3 ligase and then computes the average quantification ratios of selected ubiquitination sites. The random selection process is repeated various times for every E3 ligase for parameter optimization and the data analysis in this study was performed with 10 000 repeats.

Lastly, the average ratios of randomly selected groups of ubiquitination proteins or sites were fit into a normal distribution based on the central limit theorem. Based on this distribution, the program calculates the statistical significance for the averaged ratio of the E3 ligase protein targets or sites quantified in the dataset using the formula below.

$$z = \frac{\bar{s} - \frac{\Sigma \bar{r}}{10000}}{\sigma_{\bar{r}}}$$

$$p = 2 \cdot \left| \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \right|$$

Here, $\bar{s}$ stands for the average quantification ratios of ubiquitination substrate proteins or sites quantified for a given E3 ligase in the dataset, $\bar{r}$ stands for the average ratio of one group of randomly selected proteins or sites, $\sigma_{\bar{r}}$ stands for the standard deviation for the distribution of all the average ratios of randomly selected groups of ubiquitination proteins or sites, z stands for z-score and p stands for the P-value.

The program outputs tab-delimited results in text format. To generate a more concise output file, the program offers the option to group E3 ligases. The grouping is helpful as some E3 ligases share ubiquitination targets and depending on the analysis depth, not all targets are quantifiable in the datasets. E3 ligases are grouped together with the leading E3 ligases have all the ubiquitination substrates in this group while the remaining E3 ligases in the group only accounts for a subset of the ubiquitination substrates in the group with no unique substrates. In this way, users can filter out E3 ligases that have no unique substrates but get enriched because of a few common substrates as shown in our analysis result, and therefore, put more emphasis on the E3 ligases whose activity profiles dominate the ubiquitination dynamics. The workflow was written into a python package, and can be accessed from either PyPI, the standard way of installing python package or our GitHub webpage.

2.5.3 Analysis of Large-scale Quantitative Ubiquitylome Proteomics Datasets

We benchmarked our algorithm using two published global ubiquitylome analysis. First, we collected original ubiquitylation ratios generated by two studies. One study focused on how SPOP-mutant affects ubiquitylome and prostate cancer[192]. From this study, we collected the protein-normalized median log2 ratio of quantified ubiquitination site under each experimental condition. The other one focused on the relation between mitochondrial depolarization and PARKIN-dependent ubiquitylome[191]. In this study, we collected the log2 site ratios of quantified ubiquitination site under each experimental condition. Second, we reorganized original data into different tables according to experimental groups described in literature. In this way, the SPOP related data was reorganized into six groups, containing two of mutant-control, mutant-wild-type, and mutant-wild-type each. Meanwhile, the PARKIN-related data was reorganized into 73 groups of experiments treated with different chemicals or with various genetic backgrounds. Thirdly, the protein-level UbE3-APA analysis was applied to profile E3 ligase activities in both studies. Ubiquitin site ratios from both studies were log2 ratios, so we analyzed them directly without further log transformation. For the PARKIN study, the grouped protein-level analysis was performed. In the group mode, the E3 ligases in the results were clustered when they are sharing the same set of quantifiable targets and the E3 ligase that has the greatest number of quantifiable substrates in the group was defined as the leading E3 ligase. Correlation of E3 ligase profiles between each pair of experiments was calculated with the two-dimensional Euclidean distance between E3 ligase activity profiles in experiments.

We further applied our model to two recently published ubiquitylome studies with data-independent acquisition (DIA) analysis. First, we gathered original ubiquitylation intensities generated by two studies. One study explored how tumor necrosis factor (TNF) treatment affects ubiquitylome[195]. And we collected the average log2 intensities of quantified ubiquitination site of treated group and mock group from this study respectively. The other study investigated ubiquitylome changes under USP7 inhibition with chemical inhibition and knockdown methods[202]. In this study, we collected the average log2 site intensities of quantified ubiquitination site under each experimental condition respectively. Second, we reorganized original intensity data into ratios by comparing different treatment groups described in literature. In this way, we calculated the Treated/Mock ratios in TNF treatment study. And we calculated siCTRL+FT671/siCTRL+DMSO and siUSP7+FT671/siUSP7+DMSO in the USP7 study. Lastly, we applied protein-level UbE3-APA analysis for both studies in the grouped mode.

2.5.4 Model Accessibility and Utility

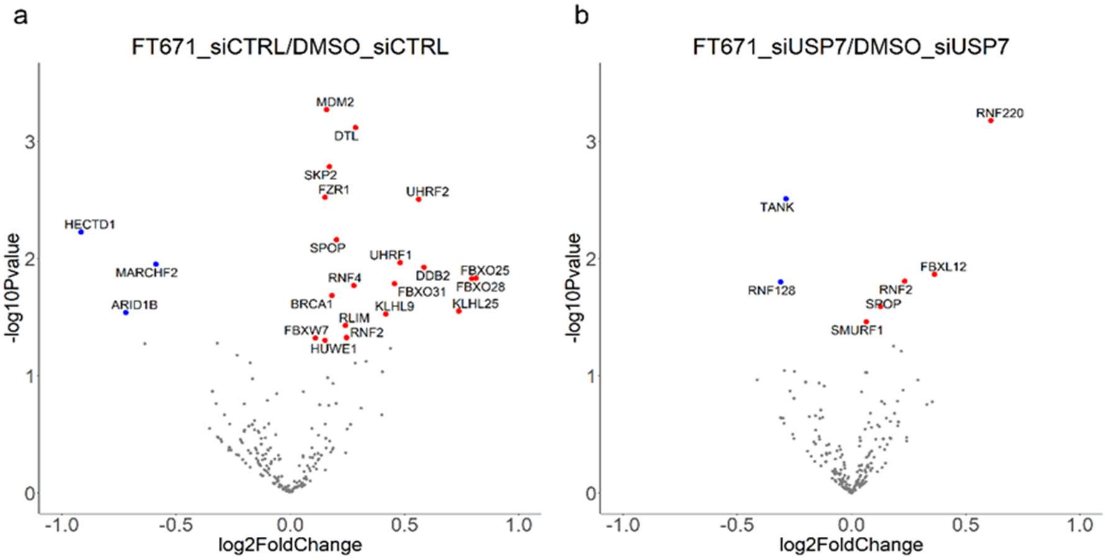
We packed the whole UbE3-APA model into a python3 library on PyPI to make it accessible. And the most direct way of installing the library is executing the pip command from a python console. For Unix/macOS/Windows users, use 'python -m pip install ube3_apa' for installation.

The main function that performs that analysis is e3enrich. It takes two tables and a set of parameters as standard input. The first of two tables is the site ratio table which records the protein ID, the site position and site ratio of every ubiquitylation site in this experiment. The second one, the protein ratio table, contains information about protein ID and the ratio of different proteins instead of sites. The input of this table is optional and triggers normalization of site ratio by corresponding protein ratio. Other parameters can be modified to fit different types of protein ID inputs, change the directory that results are generated and select various output formats based on research focus.

All related files including the code, the E3 ligase substrate dictionary, example input files were also uploaded to GitHub, which can be downloaded from the following link: https://github.com/Chenlab-UMN/Ub-E3-ligase-Activity-Profiling-Analysis.

2.6 Acknowledgements

The authors thank Gaurav Behera and other members of the Chen lab for helpful suggestions and discussions.



Supp Figure 2-1 E3 ligase activity profiling results of the USP7 study. a) Volcano plot of E3 ligase activity profiles comparing FT671+siCTRL and DMSO+siCTRL treatment conditions and b) volcano plot of E3 ligase activity profiles comparing FT671+siUSP7 and DMSO+siUSP7.

Chapter III – HypDB: Platform and Database for Identified Hydroxylated Proline Sites and Their Function Annotation in Diverse Perspectives

Yao Gong[1,2], Gaurav Behera[1], Luke Erber[1], Ang Luo[1], Yue Chen[1,2]

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

[2]Bioinformatics and Computational Biology Program, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

This chapter contains an original research article previously published.

Gong Y, Behera G, Erber L, Luo A, Chen Y. HypDB: A functionally annotated web-based database of the proline hydroxylation proteome. *PLOS Biol.* 2022;20(8):e3001757. https://doi.org/10.1371/journal.pbio.3001757

Reproduced with permission from Gong *et al*, Copyright © 2022

3.1 Summary

Proline hydroxylation (Hyp) regulates protein structure, stability, and protein–protein interaction. It is widely involved in diverse metabolic and physiological pathways in cells and diseases. To reveal functional features of the Hyp proteome, we integrated various data sources for deep proteome profiling of the Hyp proteome in humans and developed HypDB (https://www.HypDB.site), an annotated database and web server for Hyp proteome. HypDB provides site-specific evidence of modification based on extensive LC-MS analysis and literature mining with 14,413 nonredundant Hyp sites on 5,165 human proteins including 3,383 Class I and 4,335 Class II sites. Annotation analysis revealed significant enrichment of Hyp on key functional domains and tissue-specific distribution of Hyp abundance across 26 types of human organs and fluids and 6 cell lines. The network connectivity analysis further revealed a critical role of Hyp in mediating protein–protein interactions. Moreover, the spectral library generated by HypDB enabled data-independent analysis (DIA) of clinical tissues and the identification of novel Hyp biomarkers in lung cancer and kidney cancer. Taken together, our integrated analysis of human proteome with publicly accessible HypDB revealed functional diversity of Hyp substrates and provides a quantitative data source to characterize Hyp in pathways and diseases.

3.2 Introduction

Proline hydroxylation (Hyp), first discovered in 1902, is an important protein posttranslational modification (PTM) pathway in cellular physiology and metabolism[203–206]. As a simple addition of a hydroxyl group to the imino side chain of proline residue, the modification is found to be evolutionarily conserved from bacteria to humans. In mammalian cells, Hyp is largely mediated through the enzymatic activities of 2 major families of prolyl hydroxylases—collagen prolyl 4-hydroxylases (P4HAs)[207–209] and hypoxia-induced factor (HIF) prolyl hydroxylase domain (PHD) proteins[210–214], while there are no known enzymes capable of removing protein-bound Hyp yet. Since the activity of prolyl hydroxylases depends on the cellular collaboration of multiple co-factors, including oxygen and iron, as well as several metabolites, such as alpha-ketoglutarate, succinate, and ascorbate, the Hyp pathway is an important metabolic-sensing mechanism in the cells and tissues.

The most well-characterized Hyp targets are collagen proteins and HIFα family of transcription factors. Hyp on collagens mediated by P4Hs is critical to maintaining the triple-helical structure of the collagen polymer and enabling the proper protein folding after translation. Indeed, adding an electronegative oxygen on the proline 4R position promotes the trans-conformation and stabilizes the secondary structure of collagen[203]. Inhibition of collagen Hyp destabilizes the collagen and prevents its export from the ER, therefore inducing cell stress and death[215–217]. HIFα transcription factors are essential to mediate hypoxia-response in mammalian cells[218–220]. Hyp of HIFα proteins mediated by PHD proteins under normoxia condition is recognized by pVHL in the Cullin 2 E3 ligase complex, which leads to rapid ubiquitination and degradation of HIFα proteins[221,222]. Hypoxia condition inhibits HIFα Hyp and degradation, enabling the transcriptional activation of over 100 hypoxia-responding genes[223–225].

In the past two decades, numerous studies driven by advances in mass spectrometry-based proteomics technology have reported the identification and characterization of diverse new

Hyp targets and the important roles of the modification in physiological functions[226–231]. Hyp has been well known to affect protein homeostasis and the classic example is the PHD-HIF-pVHL regulatory axis. A similar mechanism also regulates the turnover of diverse key transcriptional, metabolic, and signaling proteins, including β2AR, NDRG3, ACC2, EPOR, G9a, and SFMBT1, etc.[232–236]. In addition to pVHL-mediated protein degradation, Hyp also regulates substrate degradation by affecting its interaction with deubiquitinases. For example, the hydroxylation of Foxo3a promotes substrate degradation by inhibiting the interaction with deubiquinase Usp9x, and hydroxylation of p53 enhances its interaction with deubiquitinases Usp7/Usp10 to prevent its rapid degradation[237,238]. P4H-mediated Hyp has also been known to regulate the stability of diverse substrates including AGO2 and Carabin[239,240]. In addition to protein degradation, Hyp can also affect protein–protein interaction to regulate signaling and transcriptional activities. For example, PKM2 hydroxylation promotes its binding with HIF1A for transcriptional activation, Hyp of AKT enhances the interaction with pVHL to inhibit the kinase activity of AKT, and PHD1-mediated hydroxylation of Rpb1 is necessary for its translocation and phosphorylation[241–244]. More recently, TBK1 hydroxylation was identified and found to induce pVHL and phosphatase binding, which decreases its phosphorylation and enzyme activity, while the loss of pVHL hyperactivates TBK1 and promotes tumor development in clear cell renal cell carcinoma (ccRCC)[229,245].

Despite these advances, there is a lack of an integrated and annotated knowledgebase dedicated to Hyp, which underappreciates the functional diversity and physiological significance of this evolutionarily conserved metabolic-sensing PTM pathway. To fill the knowledge gap, we developed a publicly accessible Hyp database, HypDB (http://www.HypDB.site) (Supp Figure 3-1). The development of the HypDB provides 3 main features—first, a classification-based algorithm for confident identification of Hyp substrates; second, integrated resources based on exhaustive manual literature mining, large-scale LC-MS analysis, and curated public database; and third, a collection of a large spectral library for LC-MS-based site-specific identification from a variety of cell lines and tissues. Furthermore, stoichiometry-based quantification of Hyp sites allows quantitative comparison of site abundance across various proteins and tissues, and the extensively annotated Hyp proteome enables deep bioinformatic analysis, including network connectivity, structural domain enrichment, and tissue-specific distribution study. The online database system allows the community-driven submission of LC-MS datasets to be included in HypDB annotation and the direct export of precursor and fragmentation with spectral library that enables the development of targeted quantitative proteomics and data-independent analysis workflow. We hope that the HypDB will provide critical insights into the functional diversity and network of the Hyp proteome and aid in further mechanistic studies on the physiological roles of the metabolic-sensing PTM pathway in cells and diseases.

3.3 Results

3.3.1 Database Construction and Analysis Workflow

To construct a bioinformatic resource for metabolic-sensing Hyp targets, we developed HypDB, a MySQL-based relational database on a public-accessible web server (Figure 3-1 and Supp Figure 3-2). It was constructed based on 3 main resources to comprehensively annotate human Hyp proteome (Figure 3-1). First, manual curation of literature through PubMed (searching term: “proline hydroxylation” and time limit between 2000 and 2021) was performed by 2 independent curators, which yielded 1,287 research journal articles. Site identification was extracted from each journal article, and its corresponding protein was mapped to UniProt protein ID if possible. Manual curation of the research articles focused on the sites that were biochemically investigated with multiple evidence including mass spectrometry, mutagenesis, western blotting as well as in vitro or in vivo enzymatic assays. Analyzed Hyp site identifications were then matched against the existing data in the database to reduce redundancy. Second, the database included extensive LC-MS-based direct evidence of Hyp site identifications based on the integrated analysis of over 100 LC-MS datasets of various human cell lines and tissues (see Experimental methods). The datasets were either downloaded from publicly accessible server or produced in-house. Each dataset was analyzed through a standardized workflow using MaxQuant search engine, and the Hyp site identifications were filtered and imported into the HypDB with a streamlined bioinformatic analysis pipeline specified in detail below. Our collection of MS-based evidence of Hyp identifications from cell lines and tissues likely revealed a significant portion of Hyp sites that can be potentially identified by deep proteomic analysis as evidenced by our observation that the rate of unique Hyp site addition from each dataset decreased significantly despite the increased collection of datasets in the database (Supp Figure 3-2B). Third, the HypDB also integrated Hyp identification annotated in the public UniProt database. For better clarification, the database records indicate whether the site was uniquely reported by the UniProt database or by both UniProt annotation and evidence from large-scale LC-MS analysis.

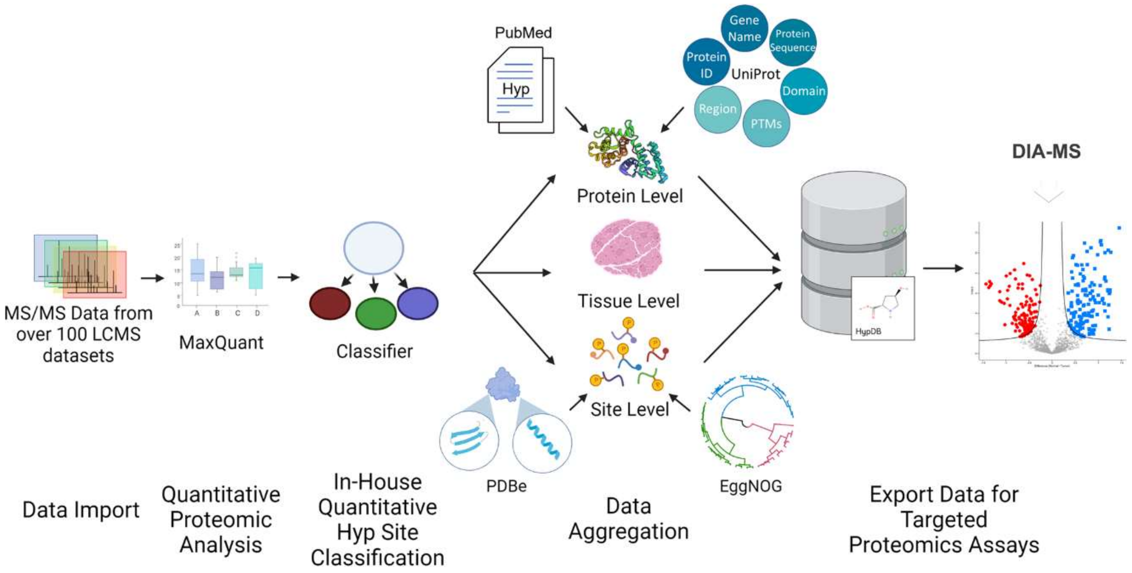
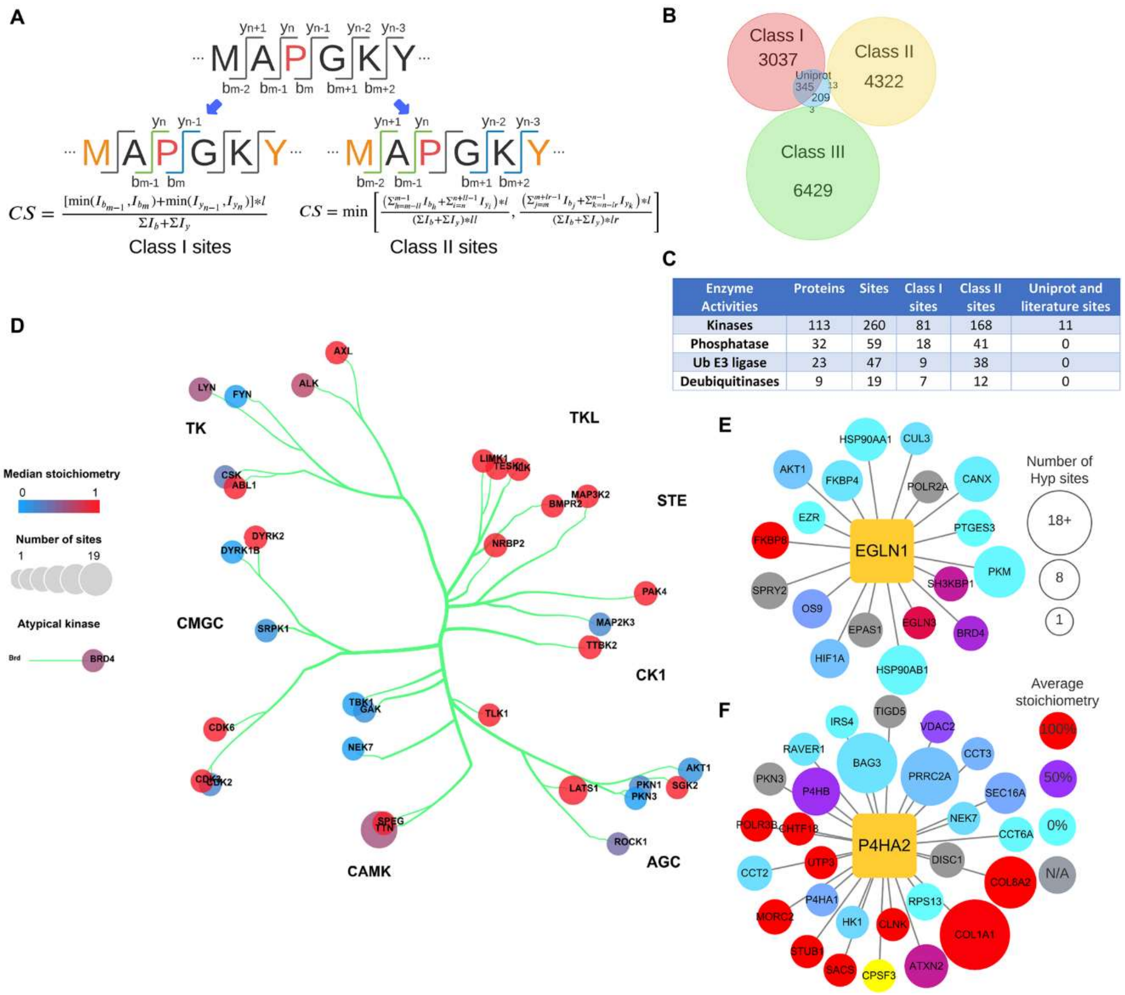


Figure 3-1 Workflow of establishing HypDB database and webserver. HypDB was constructed through deep proteome profiling analysis of human tissues and cell lines, manual literature mining, and integration with UniProt data source. Classification-based algorithm was applied to extract confident identifications, and site-specific bioinformatic analysis with stoichiometry-based quantification revealed the biochemical pathways involved with human Hyp proteome. MS-based Hyp library further enabled DIA-MS quantification of Hyp proteome in cells and tissues. DIA, data-independent acquisition; Hyp, proline hydroxylation.

We implemented stringent criteria for data importing and classification from LC-MS-based identifications. To import data into the HypDB, LC-MS-based identification of Hyp site from database search analysis was first analyzed by a classification-based algorithm to determine the confidence of Hyp site identification and localization (Figure 3-2A). The classification was performed using the best scored MS/MS spectrum of a Hyp site in each dataset analysis. The algorithm classified Hyp identifications that can be exclusively localized to proline residue based on consecutive b- or y-ions as Class I sites. The algorithm classified the Hyp identifications that cannot be exclusively localized based on MS/MS spectrum analysis but can be distinguished from 5 common types of oxidation artifacts (methionine, tryptophan, tyrosine, histidine, phenylalanine) mainly induced during sample preparation as Class II sites. Other Hyp identifications that were reported by the MaxQuant database search software (with 1% false-discovery rate at the site-level and a minimum Andromeda score of 40) were grouped as Class III sites. We further developed a site-localization score using the relative intensities of key fragment ions to index the level of confidence in site localization with MS/MS spectrum analysis for Class I and Class II sites (Experimental methods). Each dataset was analyzed by the classification algorithm separately, and the best classification evidence for each Hyp site was selected and reported on the HypDB website to indicate the confidence of site localization. The classification-based algorithm provides the specificity and reliability required for an accurately annotated database while maintaining all possible identifications as searchable records. And the localization credit score distribution of Class I and Class II sites were shown in Supp Figure 3-2C-D.



$$CS = \frac{[\min(I_{b_{m-1}}, I_{b_m}) + \min(I_{y_{n-1}}, I_{y_n})] * l}{\Sigma I_b + \Sigma I_y}$$

Class I sites

$$CS = \min\left[\frac{(\Sigma_{h=m-ll}^{m-1} I_{b_h} + \Sigma_{i=n}^{n+ll-1} I_{y_i}) * l}{(\Sigma I_b + \Sigma I_y) * ll}, \frac{(\Sigma_{j=m}^{m+lr-1} I_{b_j} + \Sigma_{k=n-lr}^{n-1} I_{y_k}) * l}{(\Sigma I_b + \Sigma I_y) * lr}\right]$$

Class II sites

Enzyme Activities	Proteins	Sites	Class I sites	Class II sites	Uniprot and literature sites
Kinases	113	260	81	168	11
Phosphatase	32	59	18	41	0
Ub E3 ligase	23	47	9	38	0
Deubiquitinases	9	19	7	12	0

Figure 3-2 Substrate diversity of the human Hyp proteome. (A) Illustration of classification-based algorithm to identify confident Hyp sites. (B)Venn diagram of Class I, II, III Hyp sites identified from MS analysis and manually curated UniProt sites. (C) PTM regulatory enzymes identified as Hyp substrates. (D) Kinase tree classification showing the distributions of kinases as Hyp substrates in different kinase families, including AGC (named after PKA, PKG, PKC families), CAMK (leaded by calcium/calmodulin-dependent protein kinases), CK1 (cell kinase 1), CMGC (named after CDKs, MAPK, GSK, CLK families), STE (homologs of the yeast STE counterparts), TK (tyrosine kinases), and TKL (tyrosine kinase-like). (E) Hydroxyproline proteins that interact with EGLN1. (F) Hydroxyproline proteins that interact with P4HA2. Hyp, proline hydroxylation; PTM, posttranslational modification.

To evaluate the site-specific prevalence of Hyp, a stoichiometry-based quantification strategy was integrated into the analysis workflow using the previously established principles[229,246]. Briefly, the Hyp stoichiometry was calculated by dividing the summed intensities of the peptides containing the Hyp site identification with the total intensities of the peptides containing the same proline site in the dataset. HypDB recorded all available site-specific Hyp stoichiometry analysis from various cell lines and tissues, which allowed site-specific quantitative analysis of modification abundance across cell and tissue types. And the median stoichiometry of all stoichiometry measurements for any specific site was calculated and reported on the HypDB website.

To further explore the functional association of Hyp proteome, several bioinformatic annotation strategies were integrated into the analysis workflow as a part of the data importing process. These stand-alone workflows include evolutionary conservation analysis, solvent accessibility analysis, and protein–protein interface analysis. Evolutionary conservation analysis compared the conservation of Hyp sites with other proline sites on the same protein and performed a statistical test to determine if the Hyp site is more evolutionarily conserved than non-Hyp sites. Solvent accessibility analysis analyzed the sequence of the substrate protein with DSSP package and calculated the likelihood of solvent accessibility for each Hyp site. Protein–protein interaction interface analysis extracted the domain interaction residues from the 3DID database based on PDB structure analysis and matched them against the Hyp site in the database to identify the Hyp site that is localized in the interface and more likely to interfere with protein–protein interaction.

All information above was integrated into several tables and linked through foreign keys as the schema in Supp Figure 3-2A. Complete information on all Hyp sites was organized in 2 major tables including a redundant site table , which stored all Hyp sites identified in different tissues and cell lines including annotated MS/MS spectra, site-specific abundance and sample source information, and a nonredundant site table which merged the LC-MS-based evidence from different sources at the site-specific level and also integrated with the sites collected from UniProt and manual curation of literatures.

3.3.2 Validation of the Hyp Site Classification Strategy

To validate our classification-based strategy for confidence Hyp site identification, we performed comparative analysis of Hyp site identifications from each class with manually curated UniProt Hyp identifications. Our analysis showed that the Class I sites alone covered over 60% sites annotated in the UniProt, and a combination of Class I and II sites covered about 63% of the UniProt sites, while very few UniProt annotated sites overlapped with the Class III sites (Figure 3-2B), suggesting that our Hyp site localization and classification algorithm allowed the collection of highly confident Hyp identification and significantly improved the reliability of LC-MS-based Hyp site analysis. To further probe the current state of the Hyp proteome, we performed extensive bioinformatic analysis for functional annotation of the Hyp proteome based on more confident Hyp site identifications in HypDB, which excluded Class III only Hyp sites whose LC-MS evidence cannot distinguish them from potential oxidation artifacts.

3.3.3 Mapping Human Proline Hydroxylation Proteome

HypDB currently collected 14,413 nonredundant Hyp sites out of 59,436 Hyp site records through large-scale deep proteomics analysis of different tissue, cell lines, manual curation of literatures, and integration with UniProt database. Among 14,413 nonredundant Hyp sites, 3,382 sites were categorized as Class I sites, 4,335 sites were categorized as Class II sites, and 6,432 were categorized as Class III sites (Figure 3-2B). In addition, the database contained 55 sites from literature mining and 209 sites that were integrated from the UniProt database. We applied enrichment analysis with Gene Ontology molecular function annotation and found that Hyp substrates are widely involved in diverse cellular activities, from nucleotide binding and cell adhesion to enzymatic activities such as oxidoreductase and ligases (Supp Figure 3-3). Excluding Class III Hyp sites, we identified a total of 113 kinases (260 sites), 32 phosphatases (59 sites), 23 E3 ligases (47 sites), and 9 deubiquitinases (19 sites) as Hyp substrates (Figure 3-2C). Statistical analysis showed a specific enrichment of kinases in Hyp proteome ($p = 0.037$), suggesting a potentially broad crosstalk between Hyp and kinase signaling pathways (Figure 3-2D). Comparing Hyp substrates with the interactome of prolyl hydroxylases in BioGRID[247], we identified 22 Hyp proteins with 68 sites that were known to interact with EGLN1/PHD2, 17 Hyp proteins with 34 sites that were known to interact with EGLN2/PHD1, 416 Hyp proteins with 861 sites that were known to interact with EGLN3/PHD3, 58 Hyp proteins with 156 sites that were known to interact with P4HA1, 31 Hyp proteins with 296 sites that were known to interact with P4HA2, and 26 Hyp proteins with 66 sites that were known to interact with P4HA3 (Figure 3-2E-F). The numbers of Class I, Class II Hyp sites, and Hyp proteins that interact with each prolyl hydroxylase were collected in Supp Figure 3-4.

To determine if Hyp site is more accessible to solvent, we collected 3D structures of proteins from PDBe and UniProt and calculated the relative solvent accessibility (RSA) of each proline residual on proteins with hydroxyproline sites with the DSSP package[248,249]. To examine if there is an RSA difference between Hyp sites and non-Hyp sites on protein with Hyp sites, we performed a 2-tail t test and found no significant difference in the distribution of solvent accessibility, suggesting that Hyp does not necessarily target solvent accessible proline residues (Supp Figure 3-5A). To determine if Hyp targets proline sites that are more evolutionarily conserved, we performed evolutionary conservation analysis through extensive sequence alignment of protein orthologs across species based on EggNOG database[250] and statistically compared the conservation of Hyp sites with the conservation of all proline on the same protein. Our data showed that about 49% sites were evolutionarily conserved with statistical significance ($p < 0.05$) (Supp Figure 3-5B). To determine if Hyp could play a potential role in domain–domain interactions, we analyzed data of known domain-based interactions of 3D protein structures from HypDB nonredundant site database. We identified 168 unique Hyp sites that were located at the interface of the interaction. These data suggested the potential involvement of Hyp in directly regulating protein–protein interaction. For example, Hyp at position 14 on Superoxide dismutase (SOD1) will form a hydrogen bonding with a neighboring chain Gln16 in a dimeric structure and potentially promote the stabilization of the dimer (Supp Figure 3-5C).

3.3.4 Functional Features of Proline Hydroxylation Proteins

We performed GO enrichment tests and other functional annotations on proteins that contain Class I, II, literature, or UniProt sites (Figure 3-3A). Our analysis revealed that Hyp substrates are highly enriched in metabolic processes such as response to toxic substances ($p < 10^{-26}$) and organic cyclic compound catabolic process ($p < 10^{-14}$), mRNA splicing ($p < 10^{-26}$) and structural functions such as NABA collagens ($p < 10^{-35}$), supramolecular fiber organization ($p < 10^{-41}$), and cell morphogenesis involved in differentiation ($p < 10^{-18}$). To determine if the Hyp proteome prefers to be involved in protein–protein interactions, we extracted a human protein interaction database from STRING with a cutoff score of 0.7, and then, extracted all the interactions containing two Hyp proteins based on the STRING database. Based on these data, we performed network connectivity analysis by comparing the number of interactions of Hyp proteins with the distribution of the number of interactions from randomly selected human proteins with 10,000 times of repeats. Our data showed that Hyp substrates are significantly involved in the protein–protein interaction network ($p < 0.0001$) (Figure 3-3B). We further performed protein complex enrichment analysis using manually curated CORUM database, and our analysis showed that Hyp proteome is significantly enriched with many known protein complexes (Supp Figure 3-6), such as TNF-alpha/NF-kappa B signaling complex 6 (Supp Figure 3-7A), TLE1 corepressor complex (Supp Figure 3-7B), DGCR8 multiprotein complex (Supp Figure 3-7C), Nop56p-associated pre-rRNA complex (Supp Figure 3-7D), and PA700-20S-PA28 complex (Supp Figure 3-7E), suggesting that Hyp targets proteins in multiple pathways that affects signaling and gene expression. Using MCODE clustering analysis, we extracted significantly enriched clusters from Hyp proteome interaction network, and these highly connected clusters of Hyp substrates suggested that Hyp targets important cellular activities including regulation of mRNA splicing, hypoxia response, and focal adhesion (Figure 3-3C-E).

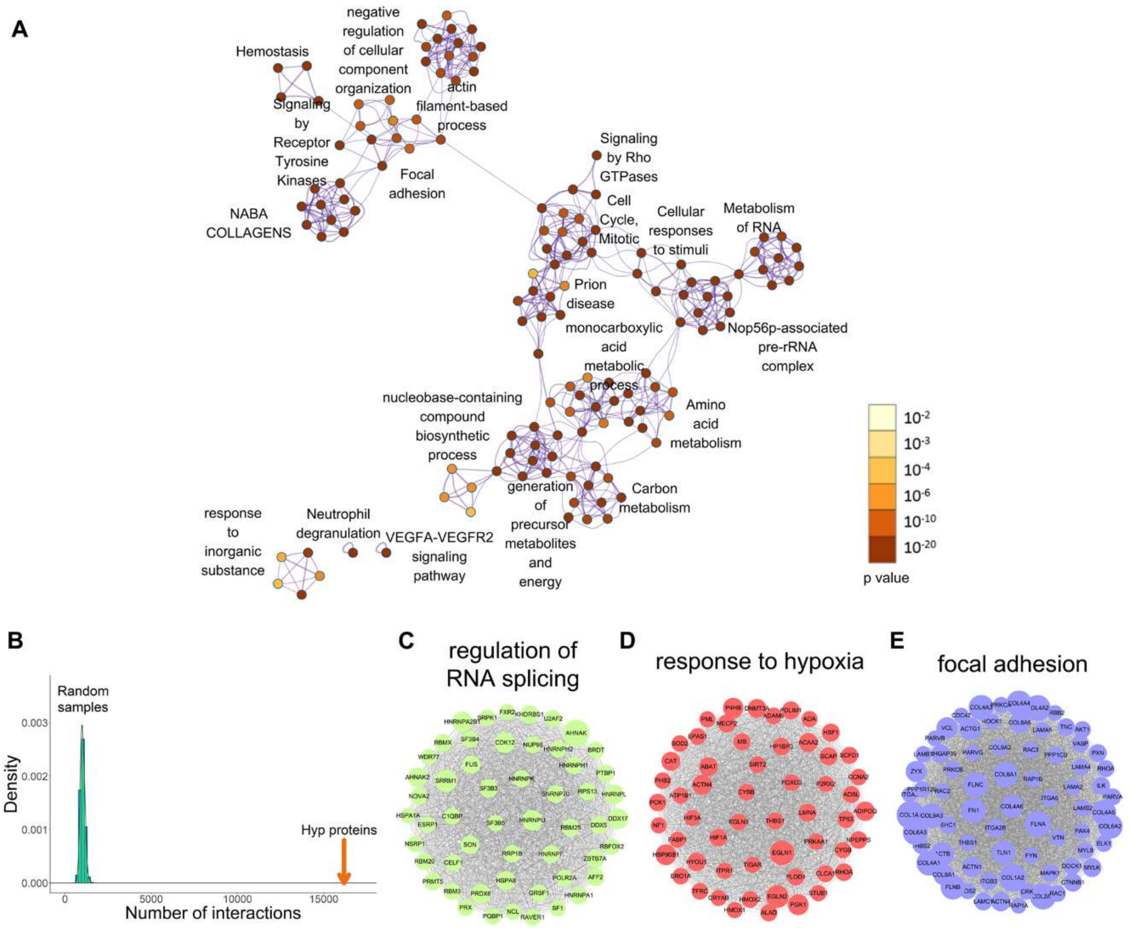


Figure 3-3 Gene enrichment and connectivity analysis of HypDB. (A) Interaction network of top 20 enriched functional annotation clusters of HypDB proteins. (B) Bootstrapping-based analysis of hydroxyproline protein interactions comparing to a distribution of protein interactions from random samples with the same number of human proteins. (C) Hydroxyproline proteins enriched in the regulation of RNA splicing. (D) Hydroxyproline proteins enriched in the response to hypoxia. (E) Hydroxyproline proteins enriched in focal adhesion.

3.3.5 Structural and Motif Features of Proline Hydroxylation Sites

We analyzed the local sequence context around Hyp sites (excluding Class III sites) using the MoMo software tool[251]. As we expected, Hyp sites with PG motif and GPPG motif were highly enriched ($p < 10^{-10}$) which is characteristic for collagen protein families (Figure 3-4A and Supp Figure 3-8A). In addition to collagen, we identified 33 proteins with similar motif to collagen, and these proteins may be potential substrates of prolyl-4-hydroxylases. Other than the collagen-like motif, we also identified CP motif ($p < 10^{-6}$) (Figure 3-4A), and proteins containing CP motifs are highly enriched in focal adhesion ($FDR < 0.05$). To remove the high background of sites with collagen-like Hyp motifs, we filtered out sites with local sequence contexts in PG motif. Our re-analysis identified that acidic amino acids were enriched at the +1 position to form PD motif (Figure 3-4A). PD motif containing proteins were highly enriched in metabolic pathways ($FDR < 0.05$). The number and proportion of Hyp sites represented in the HypDB proteome that appeared in the motifs above are shown in Supp Figure 3-8B. As Hyp sites may have crosstalk with other protein, our analysis revealed 2,386 phosphorylation sites and 535 ubiquitination sites that have been identified very close to the Hyp sites (Supp Figure 3-8C).

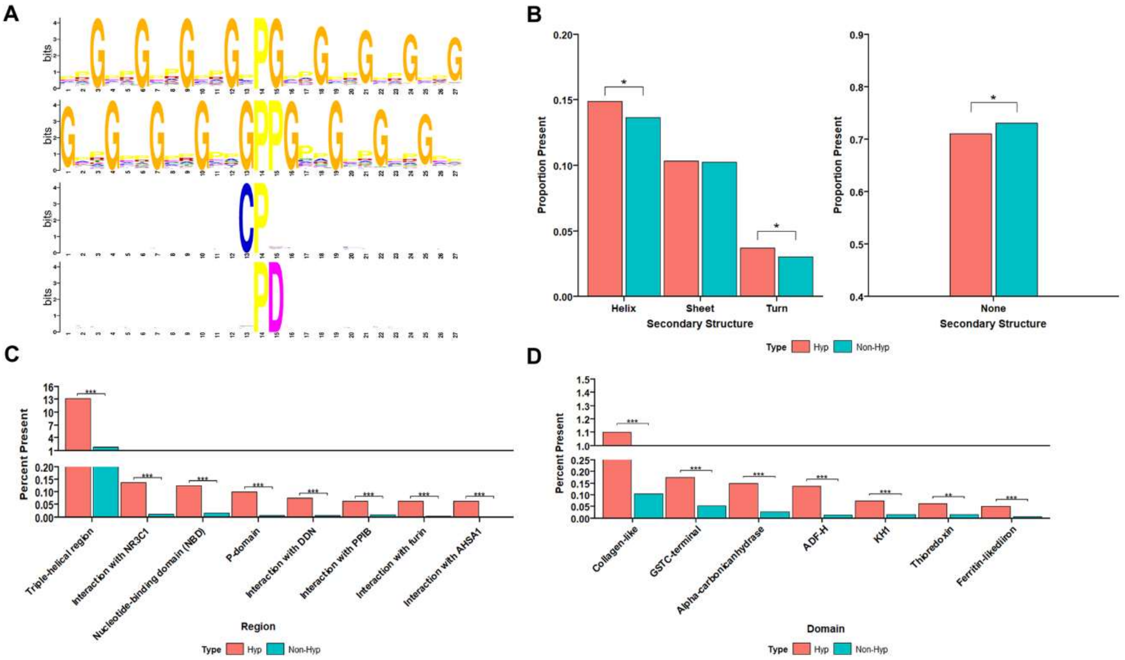


Figure 3-4 Motif and protein feature analysis of HypDB. (A) Motif enrichment analysis with the flanking sequences of Hyp sites identified PG, GPPG and CP motifs (adj $p < 10^{-6}$) and repeated analysis with the flanking sequences of Hyp sites after filtering out PG motif sequences identified PD motif (adj $p < 10^{-6}$). (B) Secondary structure enrichment of Hyp sites based on PDB protein structures (*$p < 0.05$). (C) Functional region enrichment analysis of Hyp sites based on region localizations on proteins in UniProt (***$p < 0.001$). (D) Functional domain enrichment analysis of Hyp sites based on domain localizations on proteins in UniProt (***$p < 0.001$, **$p < 0.01$).

To determine the structural features of Hyp sites, we extracted all Hyp proteins with known secondary structures. These proteins contain 2,279 Hyp sites and 27,159 non-Hyp sites on sequences that have experimentally determined PDB structure. We then classified structure features into helix, sheet, turn, and non-structure regions and performed statistical analysis to compare the secondary structure features of Hyp sites and non-Hyp sites. We found that Hyp indeed preferentially targets proline residues that are localized in the helix ($p < 0.05$) and turn secondary structures ($p < 0.05$) (Figure 3-4B left panel). Accordingly, we observed a depletion of Hyp sites outside of a secondary structure feature (Figure 3-4B right panel).

As secondary structures may not fully represent functional structural features, we developed a similar statistical analysis strategy to determine the site-specific enrichment of Hyp sites on functional domains or structural regions. In contrast to the traditional domain enrichment analysis using Pfam or Interpro for protein-level analysis, our strategy enabled site-specific enrichment analysis of domains or regions based on UniProt annotation. Application of this strategy revealed diverse known and novel structural features that were highly enriched with Hyp, such as the triple-helical region, which is characteristic for collagen protein family (Figure 3-4C). In addition to the triple-helical region, our analysis revealed more than 10 functional regions and domains that were highly enriched with Hyp, including p-domain ($p < 10^{-6}$), NBD domain ($p < 10^{-6}$), thioredoxin domain ($p < 10^{-2}$), and ferritin-like domain ($p < 10^{-6}$) (Figure 3-4C-D). These data revealed previously unexpected role of Hyp targeting functional domains in diverse cellular pathways.

3.3.6 Site-specific Stoichiometric Quantification of Hyp Proteome

Compared to relative quantification, stoichiometry analysis measures the prevalence and dynamics of the modification in a physiologically meaningful manner[185,229,252]. Our mass spectrometry-based deep proteome profiling enables site-specific quantification of Hyp stoichiometries across multiple tissues and cell lines. Our data showed that site-specific abundance of Hyp varies widely from below 1% to nearly 100% with an overall median stoichiometry of 7.89% (Figure 3-5A). Indeed, a bulk portion of the Hyp sites have either very low or very high stoichiometries. To investigate the functional differences between sites with different stoichiometry, we divided proteins into 5 quantiles based on average stoichiometry measurement for the same site across all cells and tissues (Figure 3-5B). The 4 cutoffs 5%, 20%, 80%, and 95% were selected so that each quantile contained a similar number of Hyp sites. We then performed GO enrichment and functional annotation on the 5 quantiles respectively and performed hierarchical clustering with correlation coefficient. Our data showed that proteins in immune response and neutrophil activation pathways are enriched with low to medium stoichiometry, and proteins in cell adhesion and system development are enriched with medium to high stoichiometry (Figure 3-5B). We also saw a significant enrichment of proteins involved in chromatin assembly and RNA processing but the stoichiometry of hydroxylation on those proteins seemed to be very low (Figure 3-5B). Combining site-specific functional feature annotation and stoichiometry analysis, we performed stoichiometry-based clustering of Hyp-targeted functional domains. Our data showed that ODD region that is known to regulate hydroxylation-mediated protein degradation of HIFα was enriched with medium stoichiometry, and triple-helical region on collagen, whose hydroxylation is required for its maturation, was enriched with high stoichiometry (Figure 3-5C). Furthermore, our analysis revealed stoichiometry-based enrichment of kinase domains at medium stoichiometry, GATA1 interaction domains at high stoichiometry, nucleotide-binding domains at low to medium stoichiometry, and histone-binding domains at low stoichiometry (Figure 3-5C).

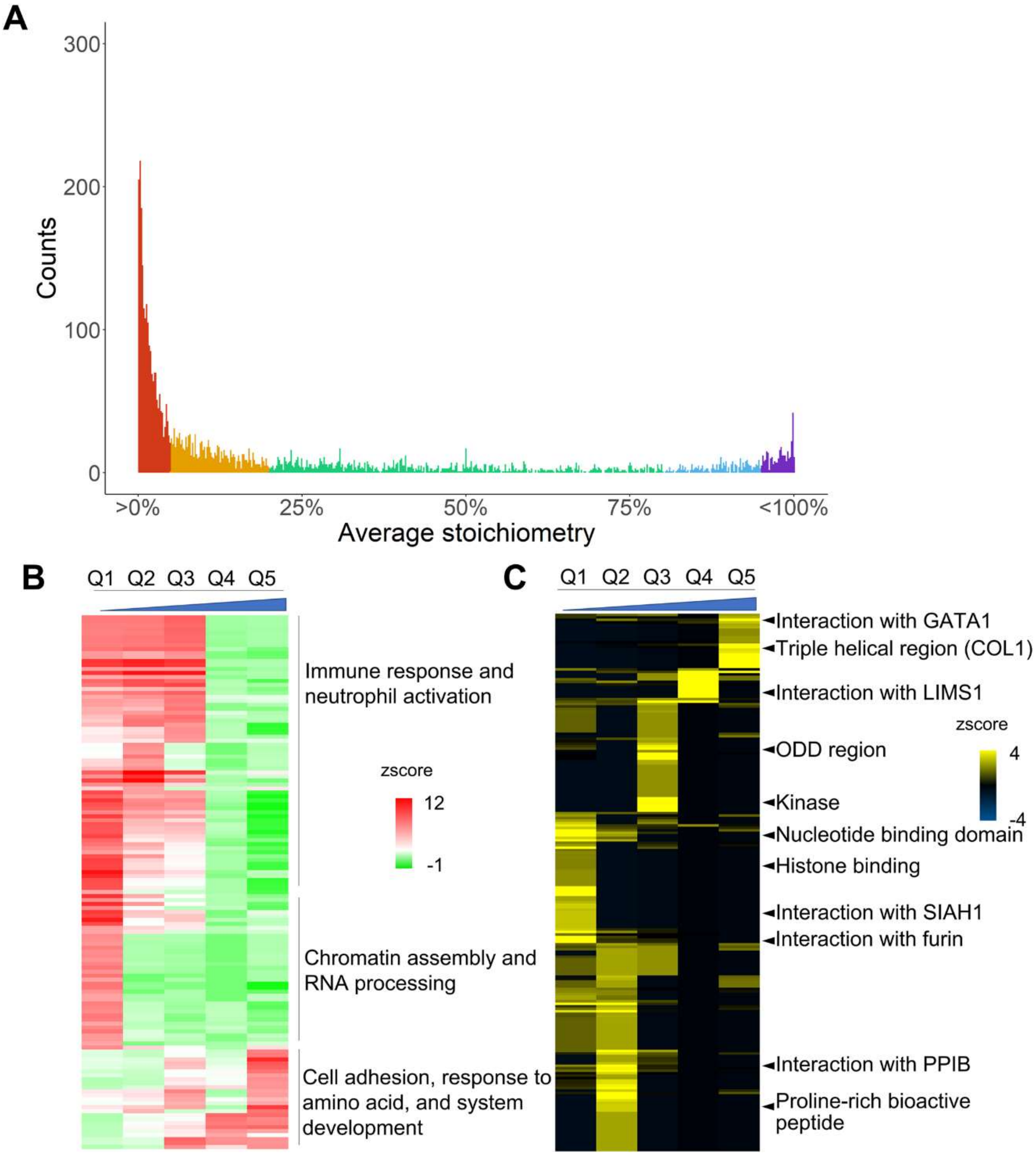


Figure 3-5 Stoichiometry-based functional enrichment analysis of the Hyp proteome. (A) Stoichiometry distribution of the Hyp sites divided into 5 quantiles—Q1, Q2, Q3, Q4, and Q5, from low to high stoichiometry with 4 cutoffs of 5%, 20%, 80%, and 95% respectively. (B, C) Hierarchical clustering of GO biological processes enrichment of Hyp proteins (B) and functional region enrichment of Hyp sites on proteins in UniProt (C) across the 5 quantiles.

3.3.7 Tissue-specific Distribution of Hyp Proteome

The collection of mass spectrometry-based identification of Hyp proteome enabled cross-tissue comparative analysis. Indeed, at individual protein level, we observed a wide distribution of Hyp abundance for the same site and between different sites across different tissue (Figure 3-6A and Supp Figure 3-9). For example, Fibrillin-1 (FBN1) was identified with 22 Hyp sites of which 17 were Class I or II sites. Hyp1090 on EGF_CA repeat showed consistent high Hyp stoichiometry (71% to 96%) across 4 different tissues (testis, colon, heart, and rectum), while Hyp1453 on another EGF_CA repeat showed varied Hyp stoichiometry (3% to 50.5%) across the same 4 tissues (testis, colon, heart, and rectum) (Figure 3-6A). In another example, 6-phosphogluconate dehydrogenase (PGD) was identified with 8 Hyp sites with half of them belonging to Class I or II sites. Hyp169 on the NAD-binding domain showed relatively low stoichiometries in heart, liver, and ovary (7.6% to 11.6%) but much higher stoichiometries in gut and B cell (21.9% and 75.6%) (Supp Figure 3-9B). We performed pathway enrichment analysis of Hyp and clustering of the enrichment across the tissues. Our data showed that Hyp proteome varied dramatically in terms of pathway and abundance among tissues (Figure 3-6B-C). For example, in lung, the Hyp proteome is mainly involved in collagen synthesis and tissue development, and it has relatively low portion of unique Hyp sites, but in liver, the Hyp proteome is heavily involved in diverse metabolic and translational processes with many liver specific Hyp targets (Figure 3-6B-C). Interestingly, clustering analysis showed that tissues sharing similar physiological functions tend to share similar Hyp profiles and are therefore clustered together. Testis and ovary, for example, have similar enrichment of Hyp proteins related to chromosome organization, DNA repair, and other DNA-related metabolic processes (Figure 3-6D). Hyp proteomes in urinary bladder and prostate are co-enriched in regulation of proteolysis and morphogenesis of different tissues. CD4 T cells and CD8 T cells are enriched with Hyp proteins related to chromatin remodeling and immune system development. Liver showed a distinctive enrichment pattern compared to other tissues, and its Hyp proteome is strongly enriched in various metabolic and catabolic processes. Meanwhile, 4 of these tissues: ovary, testis, liver, and prostate, co-enriched in neutrophil activation involved in immune response (Figure 3-6D).

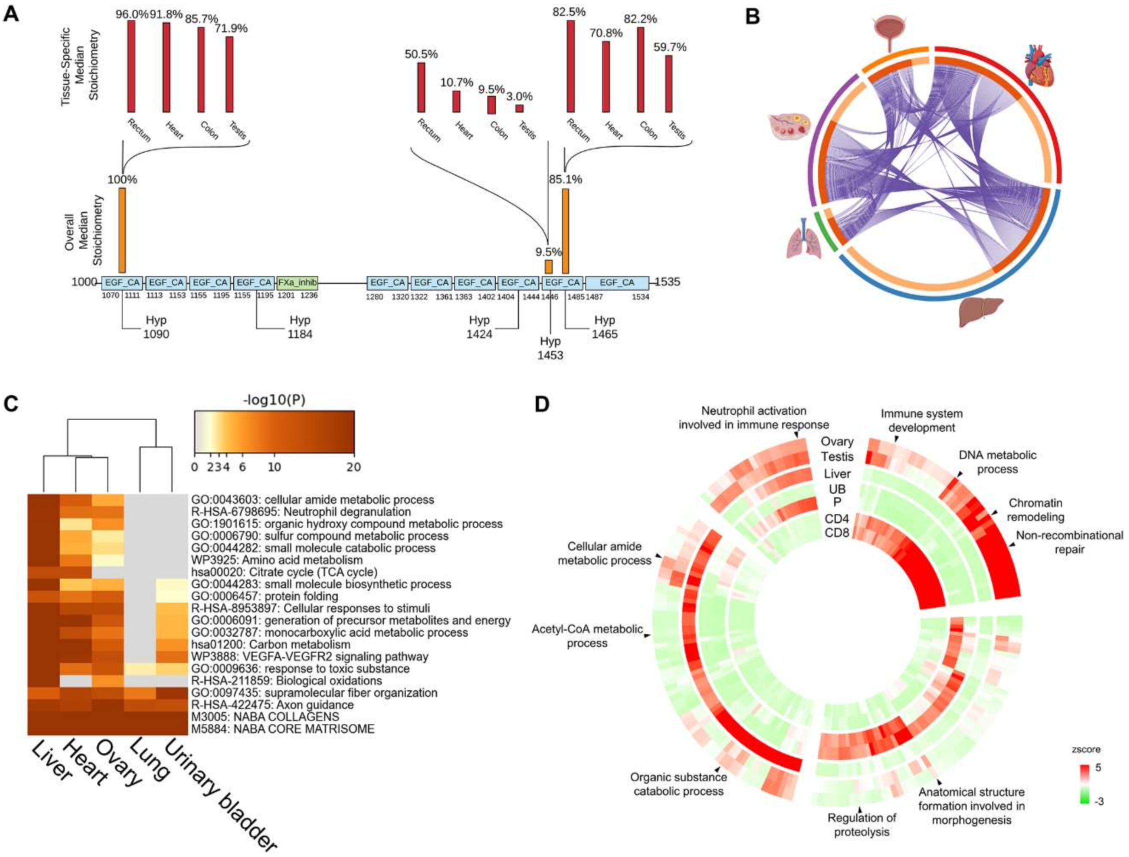


Figure 3-6 Hyp proteome distributions in different tissues. (A) An example showing varying stoichiometries of Hyp sites across different types of tissue for FBN1 with protein domains labeled in colored boxes. (B) Correlation plot of Hyp proteins in 5 different tissues: heart, liver, lung, ovary, and urinary bladder with the size of arc shows relative number and the purple curved lines showing overlap proteins (C) Heat map of the top 20 enriched functional annotations of the Hyp proteins in 5 tissues. (D) GO biological process enrichment heat map of the Hyp proteins across 7 tissues. Refer to Sheet A in S2 Table, Sheet D-E in S8 Table for the underlying data of Fig 6B–6D. CD4, CD4 T cells; CD8, CD8 T cells; FBN1, Fibrillin-1; Hyp, proline hydroxylation; P, prostate; UB, urinary bladder.

3.3.8 Data-independent Acquisition (DIA) Analysis of Hyp Targets with HypDB-Generated Spectral Library

DIA has been developed in the past 10 years as a powerful strategy for reliable and efficient quantification of proteins and PTM sites[69,70,79,156,157,195,202,253,254]. Our extensive collection of the MS-based evidence for human Hyp sites provided an ideal resource to establish a DIA workflow for global, site-specific quantification of Hyp targets in cells and tissues. To this end, our web server has integrated functions for the direct export of annotated MS/MS identification of Hyp sites for selected proteins, cell line, tissue, or at a proteome scale. The Export function provided 2 options—exporting the peptide precursor m/z only or exporting formatted MS/MS spectra. The former option can generate a target m/z list that can be used as an inclusion list for targeted quantification of Hyp sites on selected proteins or sites. The latter option can directly generate spectral library used for DIA analysis. Using the Export function, the current HypDB allowed the generation of a comprehensive Hyp spectral library in the NIST Mass Search format (msp) consisting of 6,000 precursor ions, 5,307 peptides, representing 7,717 Class 1 and 2 sites from 3,022 proteins. The webserver was also integrated with the various options for selective exporting. To demonstrate the applicability of our resource in DIA analysis workflow, we analyzed 2 recently published large-scale DIA analysis datasets[79,253]. Both datasets applied DIA analysis to quantify protein dynamics in the multiple replicates of paired normal and tumor samples.

The study by Kitata and colleagues analyzed global protein profiles of lung cancer with 5 pairs of tumor and normal tissues in triplicate analysis for a total of 30 DIA-based LC-MS runs[253]. As a routine procedure in DIA analysis, we first performed database searching of data-dependent acquisition (DDA) data in the dataset. Then, using the spectral library generated from the DDA data in the same study, we performed DIA analysis of all tumor and normal tissues with replicates. The analysis quantified 1,339 Class 1 and 2 Hyp sites from Kitata and colleagues' study (1% FDR). Next, we applied the HypDB-generated spectral library and repeated the DIA analysis. Our result showed that using the HypDB-generated spectral library led to more than double the total number of Hyp sites using a DDA-based spectral library with 3,015 Hyp sites identified while covering more than 83% of the nonredundant Hyp sites identified using the 2 spectral libraries, suggesting that the application of the HypDB-generated spectral library was sufficient to cover majority of the Hyp identifications and significantly increased the sensitivity of Hyp proteome coverage (Figure 3-7A). DIA analysis with a combined library generated by both HypDB and DDA identified 3,651 Hyp sites and 1,249 Hyp proteins (1% FDR). To determine the reproducibility of the quantification, we calculated the distribution of the percentage of coefficient variance (%CV) for DIA analysis of Hyp sites. Our data showed that %CV varied between 2% and 15% with a median value around 5% (Figure 3-7B), similar to the %CV distribution observed in the DIA analysis of proteins and phosphoproteins[253]. Given the high reproducibility of the quantification, we filtered the Hyp sites with a global 1% q-value cutoff (2,283 sites) and performed hierarchical clustering analysis of Hyp sites quantified with normalized intensity in tumor and normal lung tissues (Figure 3-7C). Our data clearly showed that site-specific Hyp quantification was sufficient to cluster and distinguish tumor versus normal tissue. To identify significantly up- or down-regulated Hyp sites in tumor tissues, we performed a 2-sample t test and analyzed the data in the volcano plot (Figure 3-7D). The analysis allowed us to identify 142 Hyp sites that were significantly up-regulated and 178 Hyp sites that were significantly down-regulated in tumor tissue (5% permutation-based FDR). The dynamically regulated Hyp sites showed strong characteristics that were distinct between tumor and normal tissue. Interestingly, we observed subtype dependent Hyp dynamics on collagen proteins. Collagen subtypes IV and VI showed significantly down-regulated Hyp level across multiple sites in tumor samples, while collagen subtype X showed significantly increased Hyp (Figure 3-7D). Since Hyp promotes the structural stability of collagens, such changes likely indicated a significantly increase in stability for collagen X and decrease in stability for collagen IV and VI in lung cancer tissue compared to the normal tissue. Our finding agreed well with a very recent publication indicating a pro-metastatic role of up-regulated collagen X in lung cancer progression[255]. In addition, we also identified significant up-regulation of Hyp on glycolysis enzymes pyruvate kinase (PKM), enolase (ENO1), and autophagy protein Parkin (PARK7) in tumor tissue (Figure 3-7D). P4HB, a member of the collagen prolyl 4-hydroxylase enzyme, also showed significant increase in Hyp (Figure 3-7D), likely due to increased prolyl 4-hydroxylase activity in lung cancer[256].

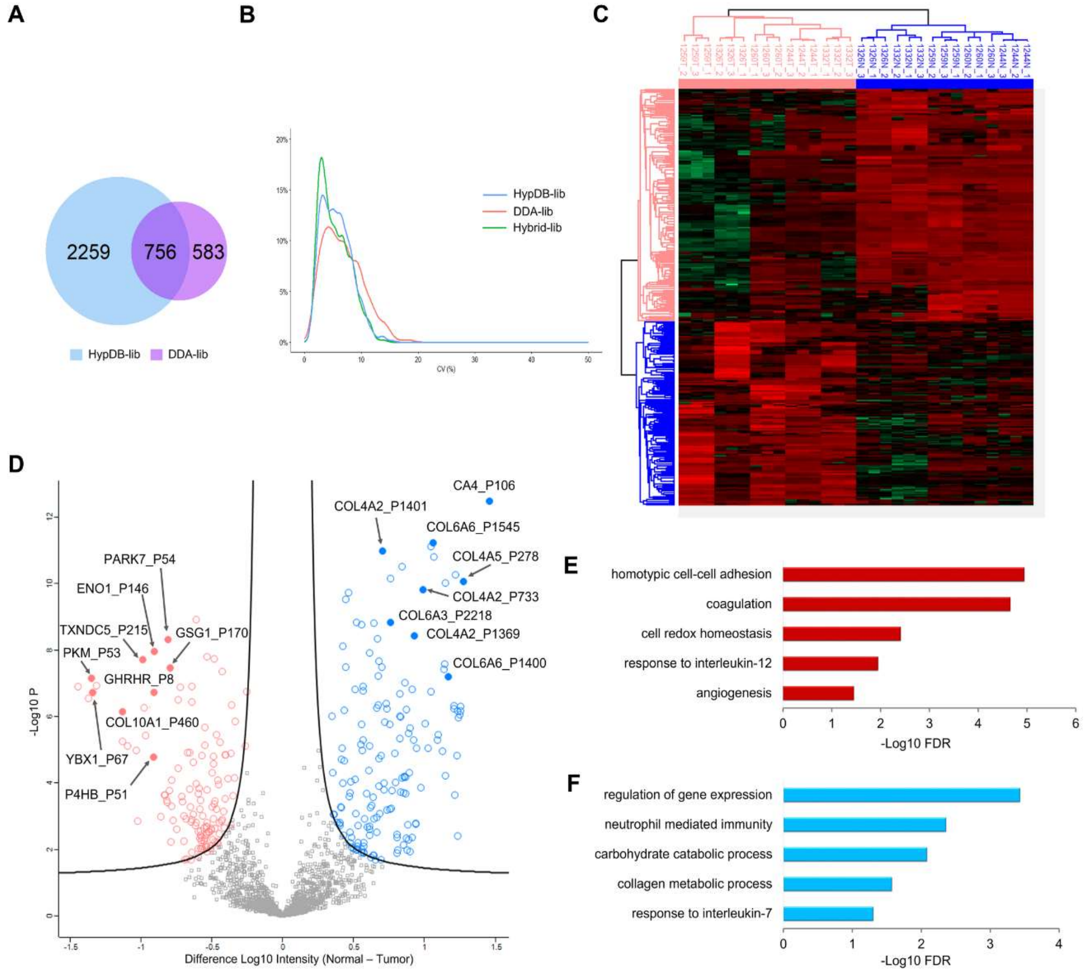


Figure 3-7 Label-free quantification of the Hyp proteome in lung cancer with DIA analysis. (A) Venn diagram of DIA-based Hyp site identifications using HypDB-generated library and the library generated by the DDA in Kitata and colleagues study. (B) Distribution of %CV for Hyp sites quantified with HypDB-generated library, DDA-generated library, or the hybrid library that combined both sources. (C, D) Hierarchical clustering (C) and volcano plot (D) of significantly up- or down-regulated Hyp sites in normal (blue) and tumor (red) tissues in the DIA analysis. (E, F) Significantly enriched GO biological processes among up-regulated (E) and down-regulated (F) Hyp proteins in tumor with at least 1-fold change after normalizing with protein abundance changes.

In another study, Guo and colleagues applied DIA analysis to quantitatively profile kidney cancer proteome and the dataset consisted of an analysis of 18 normal tissues and 18 tumor tissues[79]. Following the same workflow, we first performed DDA analysis and then applied DDA-generated Hyp library to quantify Hyp substrates in tissues. The DDA library-based analysis only quantified 387 Hyp sites from all replicate analysis. Application of the HypDB-generated spectral library increased the number of Hyp site quantifications by more than 5 times, identifying 2,510 sites (Supp Figure 3-10A). Our result confirmed that HypDB-generated library greatly increased the Hyp sequence coverage and analysis sensitivity. DIA analysis with a combined library generated by both HypDB and DDA analysis identified 2,556 Hyp sites and 981 Hyp proteins (1% FDR). To test the reproducibility among replicate tissues, we performed a correlation matrix analysis using the corrplot package in R. Our data showed that quantitative analysis of Hyp substrates allowed efficient clustering and segregation of tumor versus normal tissues (Supp Figure 3-10B). After global q-value filtering and intensity normalization, we analyzed 1,160 Hyp sites across all samples with pair-wise t test, and our analysis identified 12 up-regulated sites and 24 down-regulated Hyp sites in tumor (5% permutation-based FDR) (Supp Figure 3-10C).

To understand whether the differential abundance of Hyp sites between the normal and tumor tissues was due to changes in the abundance of corresponding proteins, we compared the log2 transformed average site ratios to the log2 transformed average protein ratios for both Kitata and colleagues and Guo and colleagues datasets (Supp Figure 3-11 A-B). We found that more than 82% of the Hyp sites in Kitata and colleagues dataset and at least 37% of the Hyp sites in Guo and colleagues dataset could be quantified with the corresponding protein abundance. From the correlative analysis between site ratios and protein ratios, we noticed a certain degree of linearity, suggesting the changes in the abundance of some Hyp sites were indeed driven by the changes in the abundance of corresponding proteins (Supp Figure 3-11A-B). We also noticed that a significant portion of Hyp site dynamics did not correlate with protein abundance changes. To this end, we calculated 95% confidence interval along the bisector correlation lines that represent equal ratios of Hyp site and protein abundance changes for all Hyp sites with corresponding protein quantification ratios. Our analysis showed that 78% of the Hyp sites in Kitata and colleagues dataset and 35% of the Hyp sites in Guo and colleagues' dataset showed significant deviation in site abundance changes from the corresponding protein abundance changes (Supp Figure 3-11 A-B). The correlation analysis therefore identified Hyp substrates that showed differential changes in abundances compared to the corresponding protein abundance changes. We further extracted only the significantly up- or down-regulated Hyp sites based on DIA analysis and compared their dynamics with corresponding protein abundance changes (Supp Figure 3-11C-F). Notably, in Kitata and colleagues dataset, the protein abundance of COL1A2 and COL14A1 was similar between tumor and normal tissues, while the abundance of the Hyp sites on each of those proteins were well above or below the 95% confidence interval (Supp Figure 3-11C,E). The correlation analysis also confirmed the down-regulation of Hyp abundance on collagen subtypes IV and VI in Kitata and colleagues lung cancer dataset with the protein-level normalization, while showing that the up-regulation of Hyp abundance on collagen subtype X in tumor was due to the up-regulation of the protein abundance (Supp Figure 3-11C,E). In Guo and colleagues' dataset, significantly changed Hyp sites showed good correlation with corresponding protein dynamics, while the Hyp sites of CRK and TPI1 showed much larger increase or decrease in abundance compared to those of their total proteins, suggesting differential activities of the Hyp pathways for each substrate (Supp Figure 3-11D,F).

To reveal the functional significance of up-regulated or down-regulated Hyp substrates in both datasets, we performed functional annotation enrichment analysis with Hyp substrates whose site ratios showed at least 1-fold increase or decrease with protein abundance normalization. Analysis of Kitata and colleagues dataset showed that the biological processes related to homotypic cell–cell adhesion, coagulation, cell redox homeostasis, response to interleukin-12, and angiogenesis were significantly enriched among up-regulated Hyp substrates (Figure 3-7E), while processes related with regulation of gene expression, neutrophil-mediated immunity, carbohydrate catabolism, collagen metabolic process, and response to interleukin-7 were significantly enriched among down-regulated Hyp substrates (Figure 3-7F) (BH corrected FDR < 0.05). The analysis of Guo and colleagues dataset showed that Hyp proteins up-regulated in kidney cancer were strongly enriched in KEGG pathways including ECM-receptor interaction, focal adhesion, glyoxylate/dicarboxylate metabolism, and tryptophan metabolism (Supp Figure 3-10D), while pathways including biosynthesis of amino acids, fructose/mannose metabolism, pathgenic E. coli infection and PI3K-Akt signaling were significantly enriched among down-regulated Hyp proteins in tumor tissue (Supp Figure 3-10E) (BH corrected FDR < 0.05).

3.4 Discussion

A grand challenge in functional analysis of PTM pathways is the lack of annotation resources to profile modification substrates and annotate enzyme-target relationships. Hyp is a key oxygen and metabolic-sensing PTM that governs the cellular programs in response to the hypoxia microenvironment and micronutrient stress. Earlier studies of Hyp mainly focused on its role in structural stability and maturation of cytoskeletal proteins such as collagens. In the past several decades, extensive biochemical studies on HIF pathways as well as other new Hyp substrates suggests that Hyp is widely involved in regulating protein–protein interaction, protein stability, signal transduction, metabolism, and gene expression. Growing evidence has also suggested that specific Hyp pathways play critical roles in cancer development, metastasis, heart disease, and diabetes. Systematic categorization and functional annotation of Hyp proteome will provide comprehensive understanding and important physiological insights into Hyp-regulated cellular pathways as well as potential therapeutic strategies targeting metabolic-sensing Hyp pathways in diseases.

To address this need, we developed HypDB, an integrated online portal and publicly accessible server for functional analysis of Hyp substrates and their interaction networks. HypDB collected various data sources for comprehensive coverage of Hyp proteome, including manual curation of published literature, deep proteomics analysis of tissues, and cell lines, as well as integration with annotated UniProt database. The site-localization and classification algorithm enabled efficient extraction of confident Hyp substrate identification from LC-MS analysis. Our identification of highly confident Hyp substrates expanded the current annotation of human Hyp targets in UniProt by over 40-fold. Streamlined data processing and stoichiometry-based Hyp quantification allowed site-specific comparative analysis of Hyp abundance across 26 human organs and fluids as well as 6 human cell lines. We collected 14,413 Hyp sites from various origins, and 86% of the top 500 Hyp sites with the most repeat identifications in various MS datasets were structural proteins, which matched well with one of its most important molecular functions.

Bioinformatic analysis of the first draft of human Hyp proteome offers critical insights into the functional and structural diversity of the modification substrates. The analysis not only revealed diverse cellular pathways enriched with Hyp proteins including mRNA processing, metabolism, cell cycle, and signaling, but also demonstrated for the first time that Hyp preferentially targets protein complexes and protein–interaction networks, indicating important roles of Hyp in fine-tuning protein structural features and mediating protein–protein interactions. Indeed, analysis of the expanded Hyp proteome with site-level secondary structure enrichment analysis indicated a significant enrichment of Hyp sites on the alpha-helix, while site-level enrichment analysis of functional domains and regions revealed novel protein domain features that are preferentially targeted by Hyp, such as P-domain, NBD domain, ferritin-like domain, and thioredoxin. These findings suggested potentially important roles for Hyp-mediated regulation of domain stability or activity that are worthy of further biochemical investigation.

MS-based analysis of Hyp proteome allows the stoichiometry-based quantification of Hyp abundance at the site-specific level. By classifying Hyp substrates based on stoichiometry dynamics, we revealed the enrichment of functional domains and activity with very high stoichiometry, indicating that Hyp on those domains may be required for the protein function, which is similar to collagen. In comparison, the oxygen-sensing ODD domain was enriched with median stoichiometry and nucleotide, or histone-binding domains were enriched with low stoichiometry. Such differences may suggest differential activities of prolyl hydroxylases targeting various functional domains. Comparative analysis of Hyp stoichiometry across tissues also indicated variations in modification abundance at the site-specific level. Such variation may be attributed to the differential metabolic and gene expression profiles in various tissues.

The collection of MS-based identification of Hyp proteome in HypDB established an annotated spectral library for Hyp-containing peptides that were identified, and site localized with high confidence. Such extensive spectral library enabled reliable and sensitive analysis of deep proteomic analysis of human cells and tissues with DIA. Application of the HypDB-generated spectral library in DIA analysis demonstrated excellent data reproducibility, significantly improved the coverage of Hyp proteome in cancer proteome analysis and revealed novel enrichment of Hyp sites that were significantly up-regulated or down-regulated in cancer tissues.

Although the current edition of HypDB (v1.0) is limited to the human proteome, future development of HypDB will include Hyp proteome in other species. Comparative analysis of Hyp targets from diverse species will allow evolutionary conservation analysis of Hyp sites and identify functionally important Hyp targets in protein structure and activity. Further application of the HypDB-generated spectral library in tissue analysis will enable the discovery of novel Hyp targets in disease animal models or patient samples and potentially lead to the development of clinically relevant therapeutic strategies.

3.5 Materials and Methods

3.5.1 MS Raw Data Analysis

We collected MS data from the human proteome draft[257] , deep proteome analysis of human cell lines[258] , PHD interactome analysis[228,246], and Hyp proteome analysis[229] as well as IP-MS analysis of Flag-tagged HIF1A. All MS raw data collected above were searched with MaxQuant (version 1.5.3.12) against the UniProt human database while having carbamidomethyl cystine as fixed modification and protein N-terminal acetylation, methionine oxidation, and Hyp as variable modification. Most of the raw data had trypsin as the digestion enzyme, while a few samples used other digestion enzymes, for example, LysC and GluC, based on the experimental procedure of original projects. Maximum missing cleavage number was set to two and the identification threshold was set at 1% false discovery rate for concatenated reversed decoy database search at protein, peptide, and site levels.

3.5.2 Site Localization Classification and Scoring

To filter out low confidence sites, we developed the site localization classification algorithm. Based on the experience that sites are localized more accurately when more ion fragments are found in corresponding MS2 spectra helping to localize the modification mass shift, our algorithm divided sites into 3 classes according to their modification localization confidence: exclusive localized sites in Class I, sites nonexclusive but distinguishable from similar modifications in Class II, and the rest in Class III (Figure 3-2A).

For a site to be classified as Class I site, a pair of b-ions or y-ions separating the proline from other amino acids must be found to localize it exclusively. In this way, a mass shift caused by hydroxylation can only occur on that specific proline. And we gave credits to that ion pair in the scoring function for Class I sites as follows:

$$CS = \frac{\left[\min\left(I_{b_{m-1}}, I_{b_m}\right) + \min\left(I_{y_{n-1}}, I_{y_n}\right)\right] * \mathrm{l}}{\sum I_b + \sum I_y}$$

where *CS* stands for credit score, *I* stand for intensity of different ion fragments, for example, I_{b_m} stands for the intensity of bm-ion, and *l* stands for peptide length. We gave credit to the pair of b-ions and y-ions that localize hydroxylation exclusively. The one with lower intensity within the pair will be selected, and we calculate the credit score based on the ratio of their intensities to average ion intensity on the same peptide.

Hydroxylation that cannot be exclusively localized but distinguishable from occurring on other prion-to-oxidize amino acid residuals are classified as Class II because we can infer that hydroxylation occurs on proline in this case. As all ions that separate proline from nearest amino acid may get oxidized easily, we gave credits to all ions that help to separate them in the scoring function for Class II sites as follows:

$$CS = \min\left[\frac{\left(\sum_{h=m-l}^{m-1} I_{b_h} + \sum_{i=n}^{n+ll-1} I_{y_i}\right) * \mathrm{l}}{\left(\sum I_b + \sum I_y\right) * ll}, \frac{\left(\sum_{j=m}^{m+lr-1} I_{b_j} + \sum_{k=n-l}^{n-1} I_{y_k}\right) * \mathrm{l}}{\left(\sum I_b + \sum I_y\right) * ll}\right]$$

where *ll* and *lr* for distance between hydroxylated proline and nearest prion-to-oxidation amino acid residual on the left side and right side. Instead of only giving credit to the pair next to the side, for Class II sites, we gave credits to all ions that contributed to separate Hyp with other prion-to-oxidation amino acid residues. We require that Hyp site contains at least 1 fragment ion on both left and right flanking sequences excluding terminal fragment ions. After that, we also calculate the ratio between the average intensity of selected ions and all ions on both sides, and the credit score is determined by the weaker side.

Sites that belong to neither Class I nor Class II are classified as Class III sites. There are chances that Class III sites are Hyp on other positions or other modifications that are identified falsely. Due to their low credibility, we do not score them and only use more confident Hyp identifications, which include Class I, Class II, UniProt, and literature sites for bioinformatic analyses.

3.5.3 Stoichiometry Calculation

We calculate the stoichiometry of each hydroxyproline site according to the total peptide intensity and modified peptide intensity. For a specific site, we collect all modified and unmodified peptides that contain this site from MS data. Then, we get stoichiometries by dividing total modified peptide intensity by total peptide intensity. Site stoichiometries in different samples are calculated separately, so there might be multiple original stoichiometries for one site in the same tissue or cell line. We take the average stoichiometry for analysis in the following steps in this case.

3.5.4 Statistical Enrichment of Pathways, Functional Annotations, Domains, and Complexes

We use R packages including “GO.db,” “GOstats,” and “org.Hs.eg.db” to perform enrichment analysis including Pfam, Kegg, and Gene Ontology—biological processes, molecular function, and cellular compartment. We collected proteins of Class I, Class II, UniProt, and literature sites from HypDB and performed a hypergeometric test for each term in the annotations above. Enrichment significance is log transformed, and we used Benjamini–Hochberg correction to check the enrichment significance with a cutoff of 0.05.

Meanwhile, we performed enrichment tests by sample and stoichiometry quantiles, respectively. For sample-specific enrichment tests, proteins with hyp sites discovered in different tissues and cell lines are analyzed, respectively. While in the other group, we divided proteins into 5 quantiles according to the average stoichiometry of corresponding sites across all samples. The stoichiometry ranges for 5 quantiles are [0%, 5%), [5%, 20%), [20%, 80%), [80%, 95%), and [95%, 100%). We also perform the log transformation and cluster the samples or quantiles according to the enrichment difference in different terms.

We also used Metascape for functional annotations and visualizations.

3.5.5 Motif Enrichment Analysis

The protein sequences of the proteins represented in HypDB were downloaded from the UniProt database. In-house Python scripts were written to extract peptides that contained Hyp sites that passed our stringent filtering criteria. These peptides were extended to the length of 27 amino acids and centered around the hydroxylated proline residue. The prealigned peptides were uploaded to the MoMo (version 5.4.1) web application[251]. All protein sequences that were obtained from the UniProt database were set as the background for the analysis. Within the MoMo web application, the motif-x algorithm was selected. The minimum number of occurrences for a motif was set to 20. The sequence logos were generated by the MoMo web application.

3.5.6 Secondary Structure Analysis

The positions for the secondary structures of the proteins represented in HypDB were downloaded from the UniProt database. In-house Python scripts were developed to determine the number of Hyp sites and non-Hyp sites found in secondary structure features for regions of proteins that have a known PDB structure.

3.5.7 Network Connectivity Analysis

All Class I, Class II, UniProt, and literature sites in HypDB are collected and transformed into 7,321 ENSP IDs with UniProt. Then, we look for interactions in the String database having both nodes in the ENSP list, and there are 16,176 interactions in total. To test the connectivity significance, we randomly picked 7,321 ENSP IDs from UniProt proteins and counted interactions whose both nodes were included by the randomly selected sample in the String database. The pick and count process are repeated 10,000 times, and these interaction counts from random samples are compared with the corresponding number of sites from HypDB.

We also built a protein–protein interaction network with these hyp proteins. From which we then selected some highly interconnected subnetworks that carry different biological functions with the help of Cytoscape software and the Mcode module.

3.5.8 Solvent Accessibility Analysis

With information from PDBe and UniProt, we matched hydroxyproline proteins with corresponding pdb ID and protein structures in pdb files. Then, we use R package bio.PDB.DSSP to interpret pdb files that contain structural information and calculate the solvent accessibility of each proline residual in the protein structure using the Sander and Rost accessible surface area (ASA) values. Then, all accessibilities are divided by maximum accessibility of proline to get the relative accessibility number between 0 and 1.

3.5.9 Protein–protein Interface Analysis

The interacting domain pairs and instances of domain–domain interactions of 3D protein structures were downloaded from 3DID (https://3did.irbbarcelona.org/index.php). In-house Python scripts were developed to analyze the number of Hyp sites interacting with another residue and the number of Hyp sites within 3 residues of an interacting residue.

3.5.10 Evolutionary Conservation Analysis

Evolutionary conservation analysis of Hyp sites was performed using EggNOG ortholog database (v5.0) and EggNOG-mapper online portal[250]. Briefly, first, using EggNOG-mapper, Hyp proteins were mapped to the corresponding ortholog groups. Next, Hyp sites and non-Hyp proline sites on Hyp proteins were aligned to ortholog sequences using MAFFT algorithm[259]. The number of matches a Hyp site or non-Hyp proline site to a proline for the same positions in ortholog sequences and the total number of sequences in the ortholog group were recorded. Lastly, HyperG test was performed for each Hyp site based on normalized number of matches to proline residues in ortholog sequences for Hyp sites and non-Hyp sites, as well as the total number of any amino acid residues in ortholog sequences for the same position as the Hyp sites or non-Hyp sites.

3.5.11 Development of Website and MySQL Database

The website serves as a front-end interactive interface of the database. It was developed using HTML, CSS, Javascript, and PHP and works on a Linux-Apache-MySQL-PHP (LAMP) server architecture. The front-end was designed using the Bootstrap framework. Associated protein data are fetched using APIs from several sources. Protein sequences, identifiers, and descriptions are fetched from entries in the UniProtKB/Swiss-Prot knowledgebase[260], protein secondary structure data are fetched from PDBe[261], and domains are fetched from Pfam[262] . The protein sequences are displayed on the website using neXtProt Sequence Viewer (https://github.com/calipho-sib/sequence-viewer). The spectral graphs on the website are visualized using d3.js (https://d3js.org/). The backend of the website utilizes PHP to interface with a MySQL database that contains the data as shown in Supp Figure 3-2A.

3.5.12 Transfection and Immunoprecipitation of HIF1A

Transfection and overexpression of Flag-tagged HIF1A was performed following a procedure as previously described[263]. Flag-tagged HIF1A plasmid (Sino Biological) was transfected into 293T cells with polyethylenimine. Cells were treated with 10 µm proteasome inhibitor MG-132 (Apexbio) for 4 hours prior to harvesting. Approximately 24 hours after transfection, cells were washed with cold PBS buffer and lysed in lysis buffer (150 mM NaCl, 50 mM Tris-HCL, 0.5% NP-40, 10% glycerol (pH 7.5), protease inhibitor cocktail (Roche)) on ice for 15 to 20 minutes. Then, the cell lysates were clarified by centrifugation prior to the incubation with anti-FLAG M2 affinity gel (Sigma) for 6 hours at 4°C. After incubation, the M2 gel was washed with wash buffer (cell lysis buffer with 300 mM NaCl) for 3 times and then eluted with 3× Flag peptide (ApexBio). The eluate was mixed with 4× SDS loading buffer and boiled, and then, loaded onto homemade SDS-PAGE gel and stained with Coomassie blue (Thermo Fisher).

3.5.13 In-gel Digestion and LC-MS Analysis of HIF1A

A large gel piece covering a wide MW range above 100 kDa was cut out and subjected to reduction/alkylation and in-gel digestion with trypsin (Promega) as previously described[185]. Tryptic peptides were desalted with homemade C18 StageTip and resuspended in HPLC Buffer A (0.1% formic acid) before being loaded onto a capillary column (75 μm ID and 20 cm in length) in-house packed with Luna C18 resin (5 μm, 100 Å, Phenomenex). The peptides were separated with a linear gradient of 7% to 35% HPLC Buffer B (0.1% formic acid in 90% acetonitrile) at a flow rate of 200 nl/min on Dionex Ultimate 3000 UPLC and electrosprayed into a high-resolution Orbitrap Lumos mass spectrometer (Thermo Fisher). Peptide precursor ions were acquired in Orbitrap with a resolution of 120,000 at 200 m/z, and peptides were fragmented with Electron Transfer/High Energy Collision Dissociation (EThcD) with calibrated charge-dependent ETD parameters and ETD Supplemental Activation and acquired in Top12 data-dependent mode sort by highest charge state and lowest m/z as priority settings. Raw data were analyzed by Maxquant software following the same procedure and parameter setting as previously published dataset as described above.

3.5.14 Usage of HypDB Website

A dedicated website with integrated MySQL database was established to host the HypDB service. The database schema includes 4 tables representing redundant Hyp site identifications, nonredundant Hyp site identifications, interaction interface analysis, evolutionary conservation analysis, and solvent accessibility analysis. Each record in the site identification table is assigned a unique HypDB site ID. The website was designed with the Bootstrap framework (v4.1.3) and features several key functions including a Search bar, Protein information page, Site information page, Database summary, Upload/contribute page, and Download/export page.

The Search bar allows the user to input a UniProt accession number or Gene name of the protein of interest, and the server will use the information to extract and display a ranked list of most similar entries in real time. Clicking on an entry will bring the user to the protein information page where protein identifiers, description, and protein sequence are displayed. All Hyp sites are identified on the protein sequence as well as known acetylation and phosphorylation sites from PhosphoSitePlus database[198] are highlighted by different colors. The list of Hyp sites is further displayed below the sequence in the table that includes the site properties including localization class, localization score, stoichiometry, solvent accessibility, and evolutionary conservation information. Hyp site table is followed by properties of Hyp proteins including protein–protein interaction, secondary structure, functional domains, and domain–domain interactions. Hyp sites identified with MS/MS evidence in the HypDB have a “Details” button displayed for each site in the site table on the protein information page. Clicking on the Details button will bring the user to the peptide information page where the best identified MS/MS spectrum for the site is displayed with annotations of fragment ions.

The Contribute/Upload page allows the community to contribute raw MS/MS identifications to the HypDB through an embedded Google Form. Information regarding the raw data type, location, sample type, database searching parameters as well as user information will be entered into the database. Raw data will be downloaded and processed using the same streamlined workflow. The data will pass through the classification and site-localization analysis process and annotated with the bioinformatic workflows as described above. The final data will be deposited into the HypDB to share with the research community.

The Export/Download page allows the community to download the complete dataset deposited in the HypDB including both redundant and nonredundant modification site tables. In addition, the Export function enables users to select a list of proteins, tissues of interests, filter sites based on localization credit class, MS fragmentation type, proteolytic enzyme used in proteomics analysis, as well as specify the precursor ion m/z of Hyp proteins for export to set up targeted quantification method when acquiring data or export the collected spectral libraries of Hyp sites from the selected Hyp proteins to perform database searching with DIA.

3.5.15 Construction of DDA-based Spectral Libraries

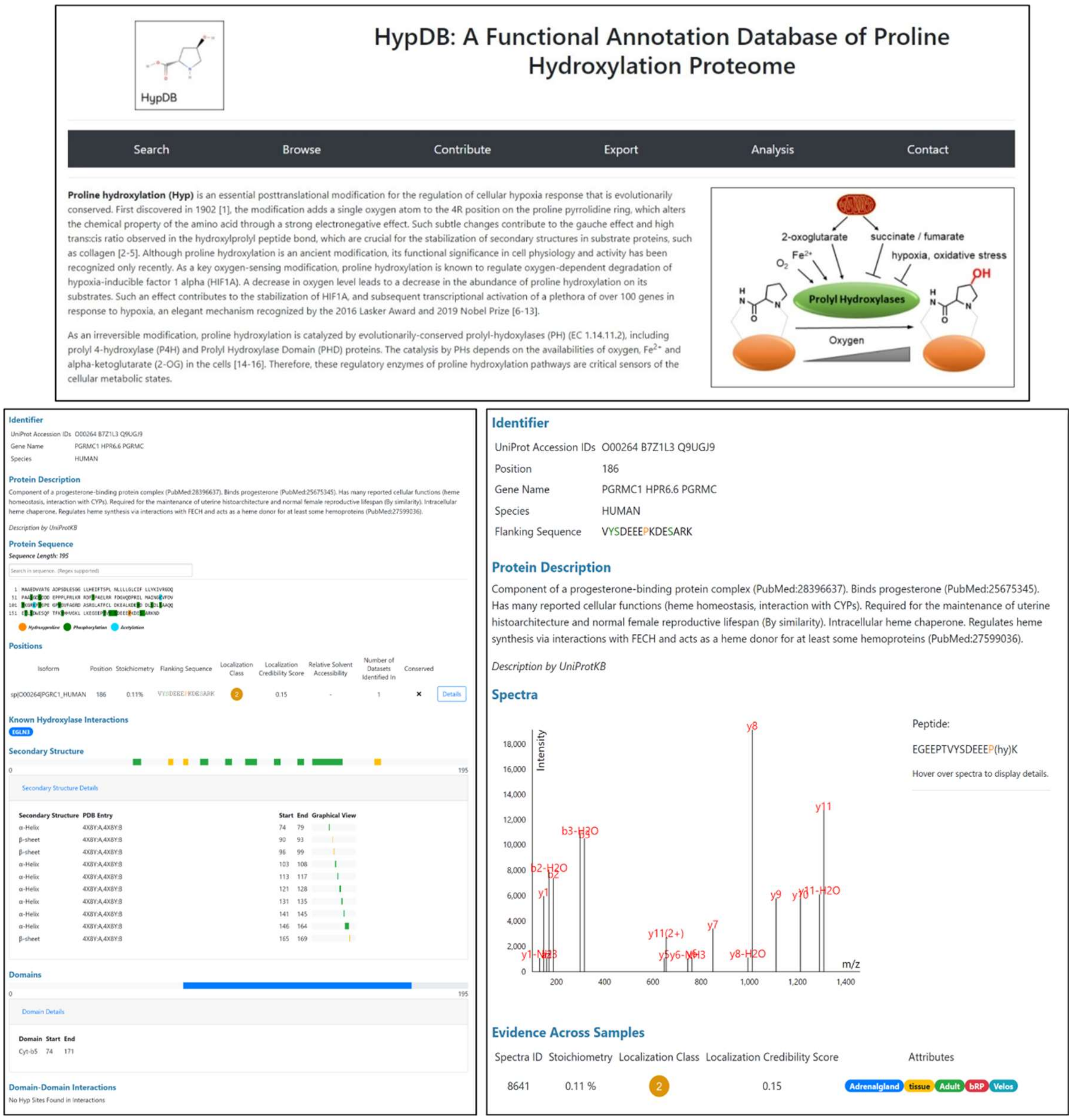
To construct the study-specific DDA-based spectral libraries from Kitata and colleagues and Guo and colleagues studies, a database search of the DDA data from each study was performed by MaxQuant (version 1.5.3.12). The parameters for the search engine were slightly modified from the parameters reported by the authors of each study. The maximum number of cleavages was set to two and the threshold for identification was set at 1% FDR. The variable modification of Hyp was included in addition to the variable modifications that the authors of each study reported. The spectral data for Hyp sites were compiled into an msp-formatted spectral library.

3.5.16 DIA Data Analysis

DIA data were analyzed using DIA-NN (v1.8)[264]. The default workflow for analysis using a spectral library was followed (https://github.com/vdemichev/diann). The DIA data from Kitata and colleagues and Guo and colleagues studies were analyzed separately with DIA-NN. FDR (q-value) for protein groups and Hyp site identification was set at 1.0%. The analysis of the DIA data from each study was performed with spectral library from various sources: HypDB Library, Study-Specific DDA-based Library, and Combined Library generated by both HypDB and Study-Specific DDA Analysis for both Hyp peptide identifications and non-Hyp peptide identifications. DIA-NN further applied global q-value filtering and intensity normalization to generate Hyp site matrix output for Hyp sites that were confidently quantified across all samples. Python scripts developed in-house to process the output from DIA-NN to be Hyp site nonredundant. The matrix output from each study with nonredundant Hyp site quantification was used for clustering, annotation enrichment analysis, and visualization using the Perseus software platform[265]. Missing values were imputed using a normal distribution, and the data were hierarchically clustered. The processed site-nonredundant Hyp intensity data from DIA-NN was also analyzed and visualized using R. Missing values were imputed using the k-nearest neighbor (KNN) method in the NAguideR tool[266].

3.6 Acknowledgements

We would like to thank the members of the Chen lab for helpful discussion and suggestion. We appreciate Christopher Lee of the College of Biological Sciences at the University of Minnesota for the help with the initial testing of the website. We would also like to thank Changzhi Wang for the initial development of evolutionary conservation analysis workflow in this study and Yi-Cheng Sin for helping with LC-MS analysis. We are grateful for the Masonic Cancer Center and the Center for Mass Spectrometry and Proteomics for providing LC-MS access and analysis. Some figures were created with BioRender.com.



HypDB: A Functional Annotation Database of Proline Hydroxylation Proteome

Proline hydroxylation (Hyp) is an essential posttranslational modification for the regulation of cellular hypoxia response that is evolutionarily conserved. First discovered in 1902 [1], the modification adds a single oxygen atom to the 4R position on the proline pyrrolidine ring, which alters the chemical property of the amino acid through a strong electronegative effect. Such subtle changes contribute to the gauche effect and high trans:cis ratio observed in the hydroxylprolyl peptide bond, which are crucial for the stabilization of secondary structures in substrate proteins, such as collagen [2-5]. Although proline hydroxylation is an ancient modification, its functional significance in cell physiology and activity has been recognized only recently. As a key oxygen-sensing modification, proline hydroxylation is known to regulate oxygen-dependent degradation of hypoxia-inducible factor 1 alpha (HIF1A). A decrease in oxygen level leads to a decrease in the abundance of proline hydroxylation on its substrates. Such an effect contributes to the stabilization of HIF1A, and subsequent transcriptional activation of a plethora of over 100 genes in response to hypoxia, an elegant mechanism recognized by the 2016 Lasker Award and 2019 Nobel Prize [6-13].

As an irreversible modification, proline hydroxylation is catalyzed by evolutionarily-conserved prolyl-hydoxylases (PH) (EC 1.14.11.2), including prolyl 4-hydroxylase (P4H) and Prolyl Hydroxylase Domain (PHD) proteins. The catalysis by PHs depends on the availabilities of oxygen, Fe^{2+} and alpha-ketoglutarate (2-OG) in the cells [14-16]. Therefore, these regulatory enzymes of proline hydroxylation pathways are critical sensors of the cellular metabolic states.

Identifier

UniProt Accession IDs	O00264 B7Z1L3 Q9UGJ9
Gene Name	PGRMC1 HPR6.6 PGRMC
Species	HUMAN

Protein Description

Component of a progesterone-binding protein complex (PubMed:28396637). Binds progesterone (PubMed:25675345). Has many reported cellular functions (heme homeostasis, interaction with CYPs). Required for the maintenance of uterine histoarchitecture and normal female reproductive lifespan (By similarity). Intracellular heme chaperone. Regulates heme synthesis via interactions with FECH and acts as a heme donor for at least some hemoproteins (PubMed:27599036).

Description by UniProtKB

Protein Sequence

Sequence Length: 195

Search in sequence.. (Regex supported)

```
  1 MAAEDVVATG ADPSDLESGG LLHEIFTSPL NLLLLGLCIF LLYKIVRGDQ
 51 PAASGDSDDD EPPPLPRLKR RDFTPAELRR FDGVQDPRIL MAINGKVFDV
101 TKGRKFYGPE GPYGVFAGRD ASRGLATFCL DKEALKDEYD DLSDLTAAQQ
151 ETLSDWESQF TFKYHHVGKL LKEGEEPTVY SDEEEPKDES ARKND
```

● Hydroxyproline ● Phosphorylation ● Acetylation

Positions

Isoform	Position	Stoichiometry	Flanking Sequence	Localization Class	Localization Credibility Score	Relative Solvent Accessibility	Number of Datasets Identified In	Conserved	
sp|O00264|PGRC1_HUMAN	186	0.11%	VYSDEEEPKDESARK	2	0.15	-	1	×	Details

Known Hydroxylase Interactions

EGLN3

Secondary Structure

Secondary Structure Details

Secondary Structure	PDB Entry	Start	End	Graphical View
α-Helix	4X8Y:A,4X8Y:B	74	79	
β-sheet	4X8Y:A,4X8Y:B	90	93	
β-sheet	4X8Y:A,4X8Y:B	96	99	
α-Helix	4X8Y:A,4X8Y:B	103	108	
α-Helix	4X8Y:A,4X8Y:B	113	117	
α-Helix	4X8Y:A,4X8Y:B	121	128	
α-Helix	4X8Y:A,4X8Y:B	131	135	
α-Helix	4X8Y:A,4X8Y:B	141	145	
α-Helix	4X8Y:A,4X8Y:B	146	164	
β-sheet	4X8Y:A,4X8Y:B	165	169	

Domains

Domain Details

Domain	Start	End
Cyt-b5	74	171

Domain-Domain Interactions

No Hyp Sites Found in Interactions

Identifier

UniProt Accession IDs	O00264 B7Z1L3 Q9UGJ9
Position	186
Gene Name	PGRMC1 HPR6.6 PGRMC
Species	HUMAN
Flanking Sequence	VYSDEEEPKDESARK

Protein Description

Component of a progesterone-binding protein complex (PubMed:28396637). Binds progesterone (PubMed:25675345). Has many reported cellular functions (heme homeostasis, interaction with CYPs). Required for the maintenance of uterine histoarchitecture and normal female reproductive lifespan (By similarity). Intracellular heme chaperone. Regulates heme synthesis via interactions with FECH and acts as a heme donor for at least some hemoproteins (PubMed:27599036).

Description by UniProtKB

Spectra

Peptide:

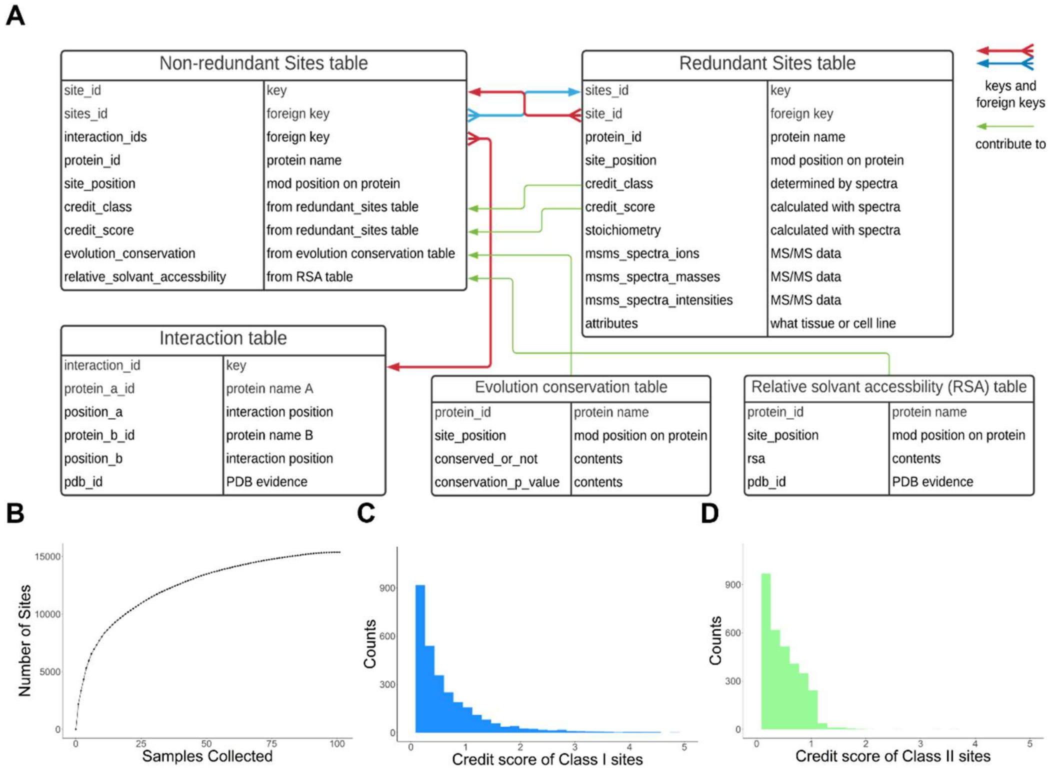
EGEEPTVYSDEEEP(hy)K

Hover over spectra to display details.

Evidence Across Samples

Spectra ID	Stoichiometry	Localization Class	Localization Credibility Score	Attributes
8641	0.11 %	2	0.15	Adrenalgland tissue Adult bRP Velos

Supp Figure 3-1 Screen shots of the HypDB web portal with front page (top), protein-level view (bottom left), and peptide-level view (bottom right).



Non-redundant Sites table

site_id	key
sites_id	foreign key
interaction_ids	foreign key
protein_id	protein name
site_position	mod position on protein
credit_class	from redundant_sites table
credit_score	from redundant_sites table
evolution_conservation	from evolution conservation table
relative_solvant_accessbility	from RSA table

Redundant Sites table

sites_id	key
site_id	foreign key
protein_id	protein name
site_position	mod position on protein
credit_class	determined by spectra
credit_score	calculated with spectra
stoichiometry	calculated with spectra
msms_spectra_ions	MS/MS data
msms_spectra_masses	MS/MS data
msms_spectra_intensities	MS/MS data
attributes	what tissue or cell line

Interaction table

interaction_id	key
protein_a_id	protein name A
position_a	interaction position
protein_b_id	protein name B
position_b	interaction position
pdb_id	PDB evidence

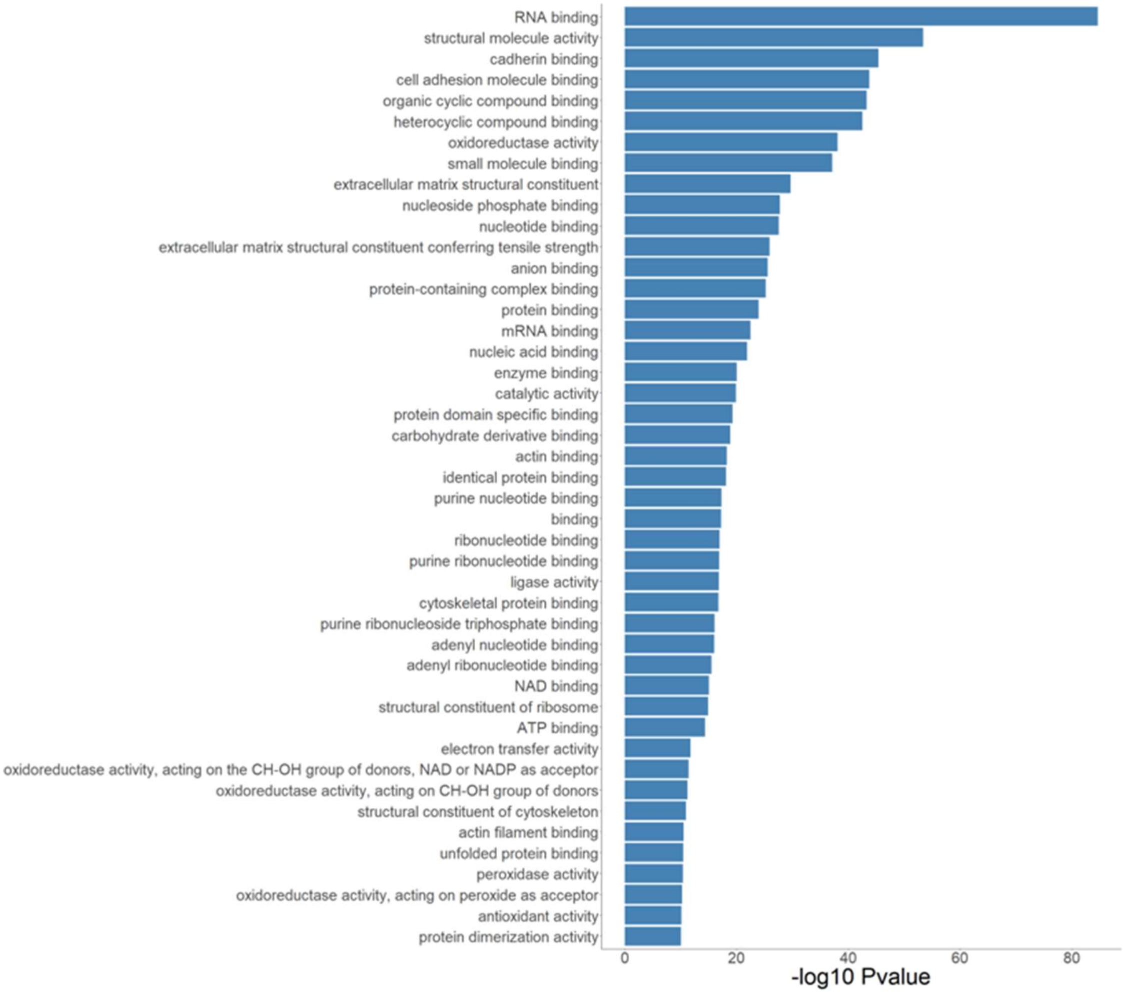
Evolution conservation table

protein_id	protein name
site_position	mod position on protein
conserved_or_not	contents
conservation_p_value	contents

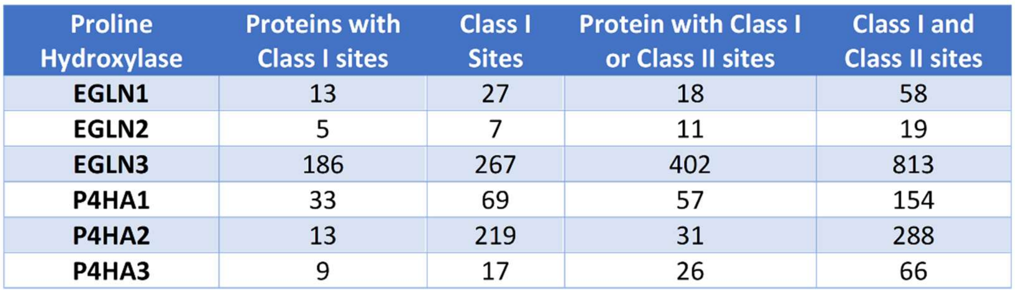
Relative solvant accessbility (RSA) table

protein_id	protein name
site_position	mod position on protein
rsa	contents
pdb_id	PDB evidence

Supp Figure 3-2 The schema of the MySQL database of HypDB and distribution of site properties.

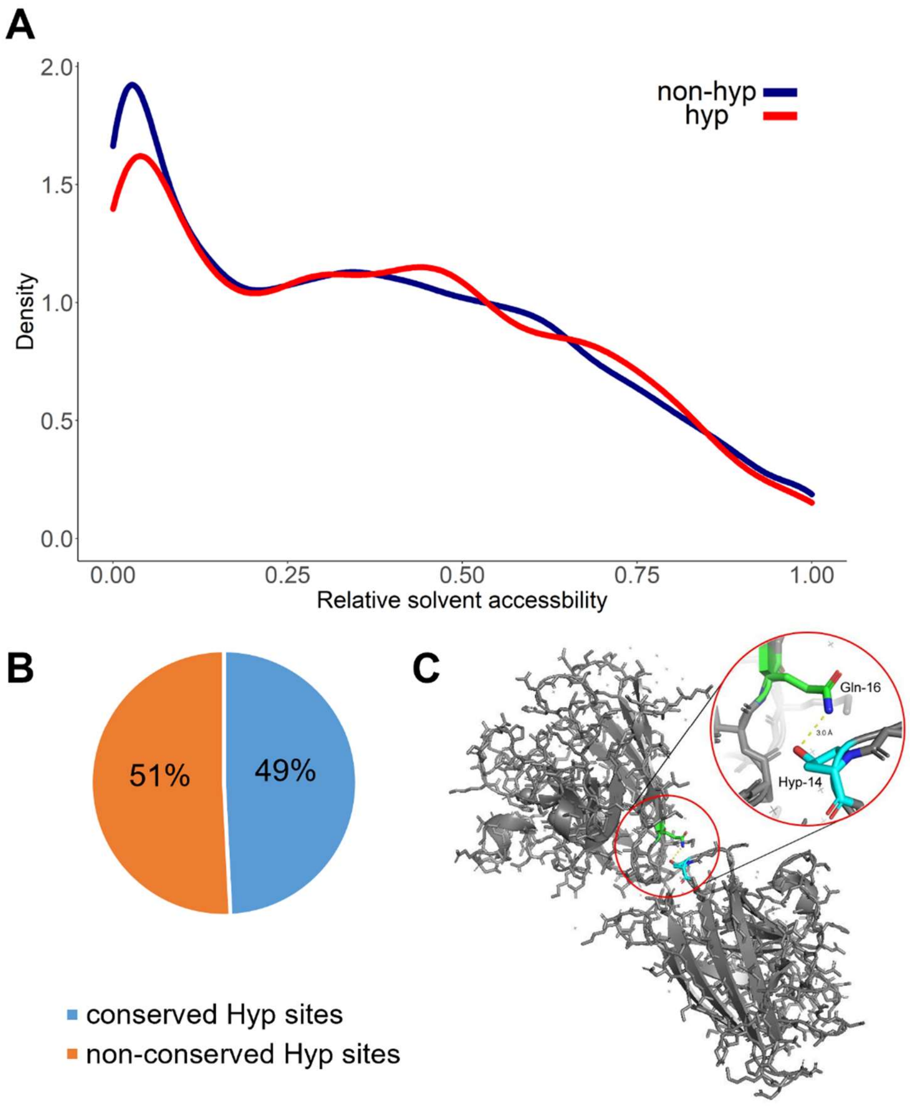


Supp Figure 3-3 Statistical enrichment analysis of Gene Ontology Molecular Function annotation of the Hyp proteome.

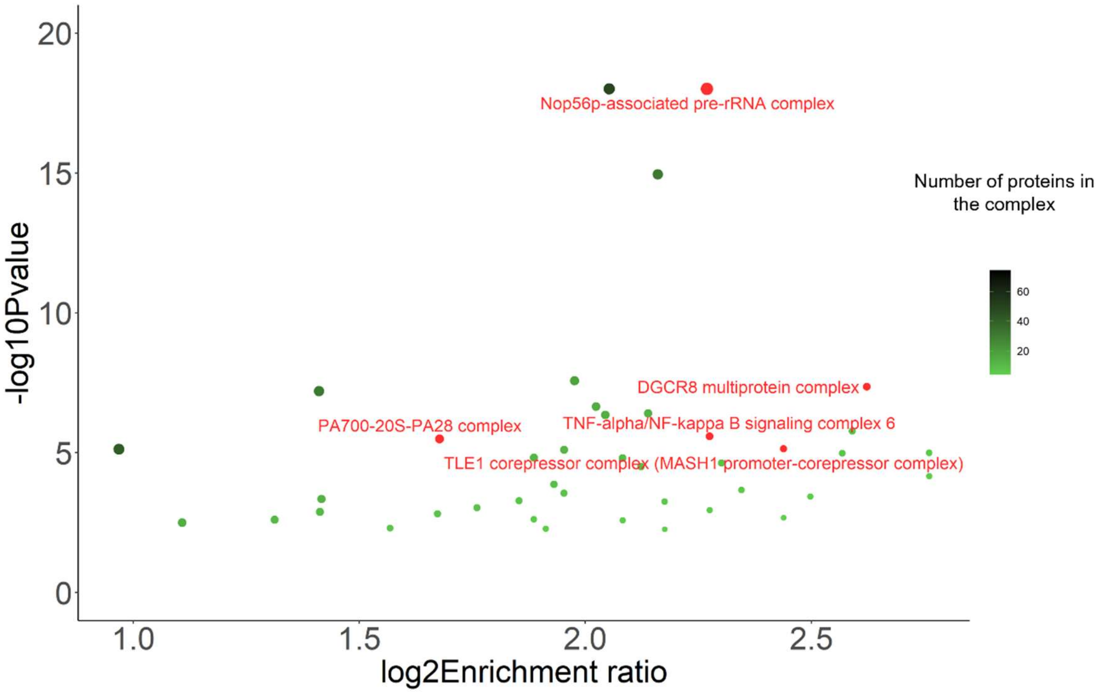


Proline Hydroxylase	Proteins with Class I sites	Class I Sites	Protein with Class I or Class II sites	Class I and Class II sites
EGLN1	13	27	18	58
EGLN2	5	7	11	19
EGLN3	186	267	402	813
P4HA1	33	69	57	154
P4HA2	13	219	31	288
P4HA3	9	17	26	66

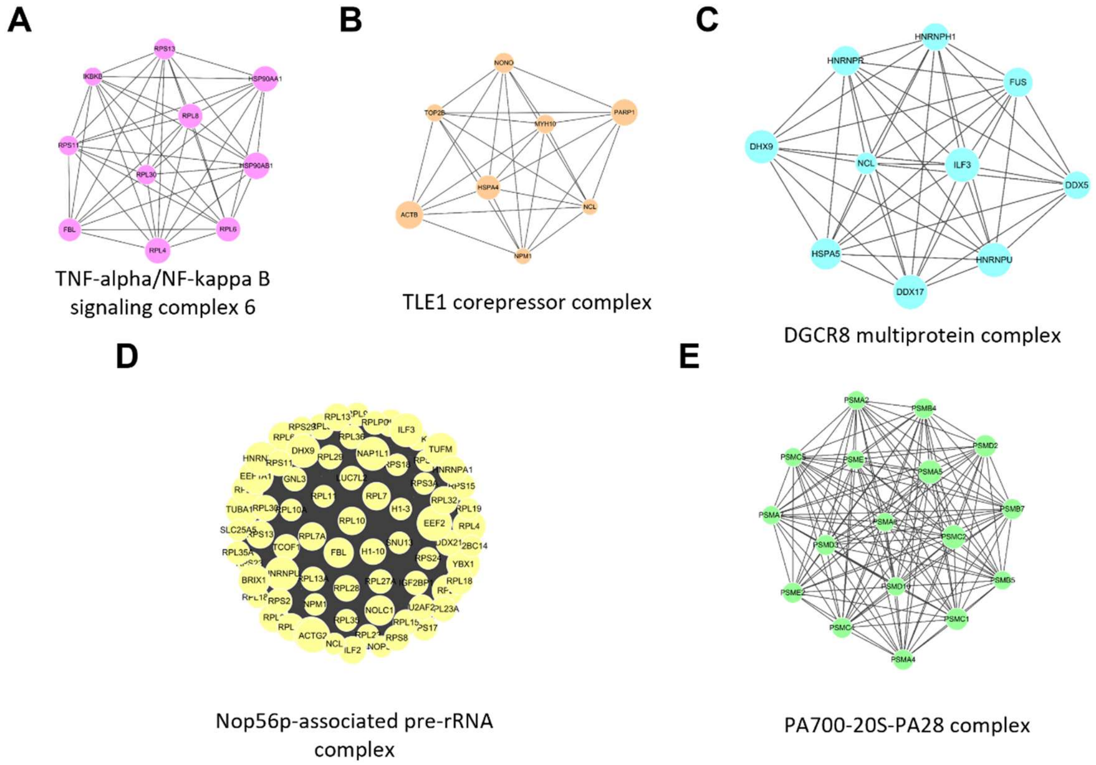
Supp Figure 3-4 The total numbers of Hyp sites and corresponding proteins that are known to interact with prolyl hydroxylase.



Supp Figure 3-5 Solvent accessibility, evolutionary conservation, and protein–protein interaction analysis of the Hyp sites.



Supp Figure 3-6 Enrichment of the CORUM protein complexes among the Hyp proteins.



TNF-alpha/NF-kappa B
signaling complex 6

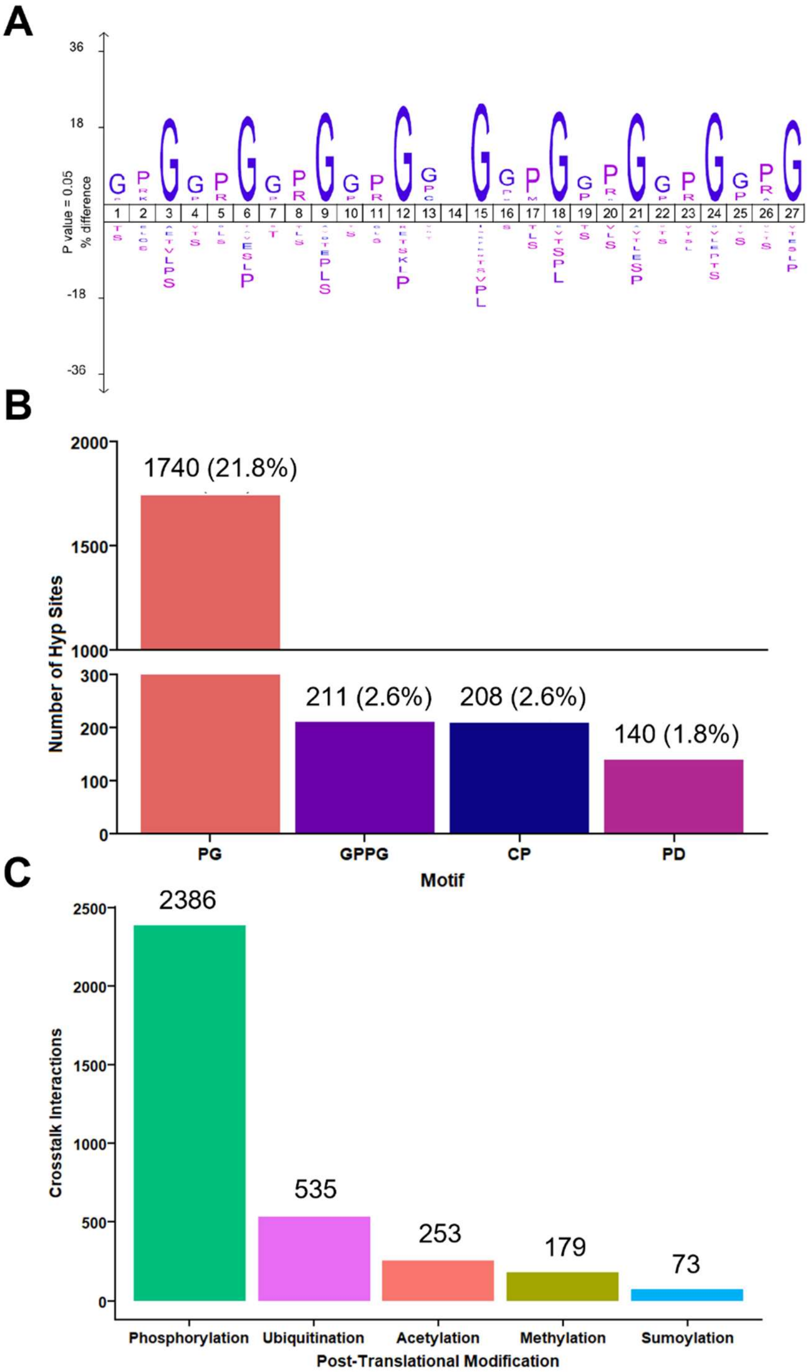
TLE1 corepressor complex

DGCR8 multiprotein complex

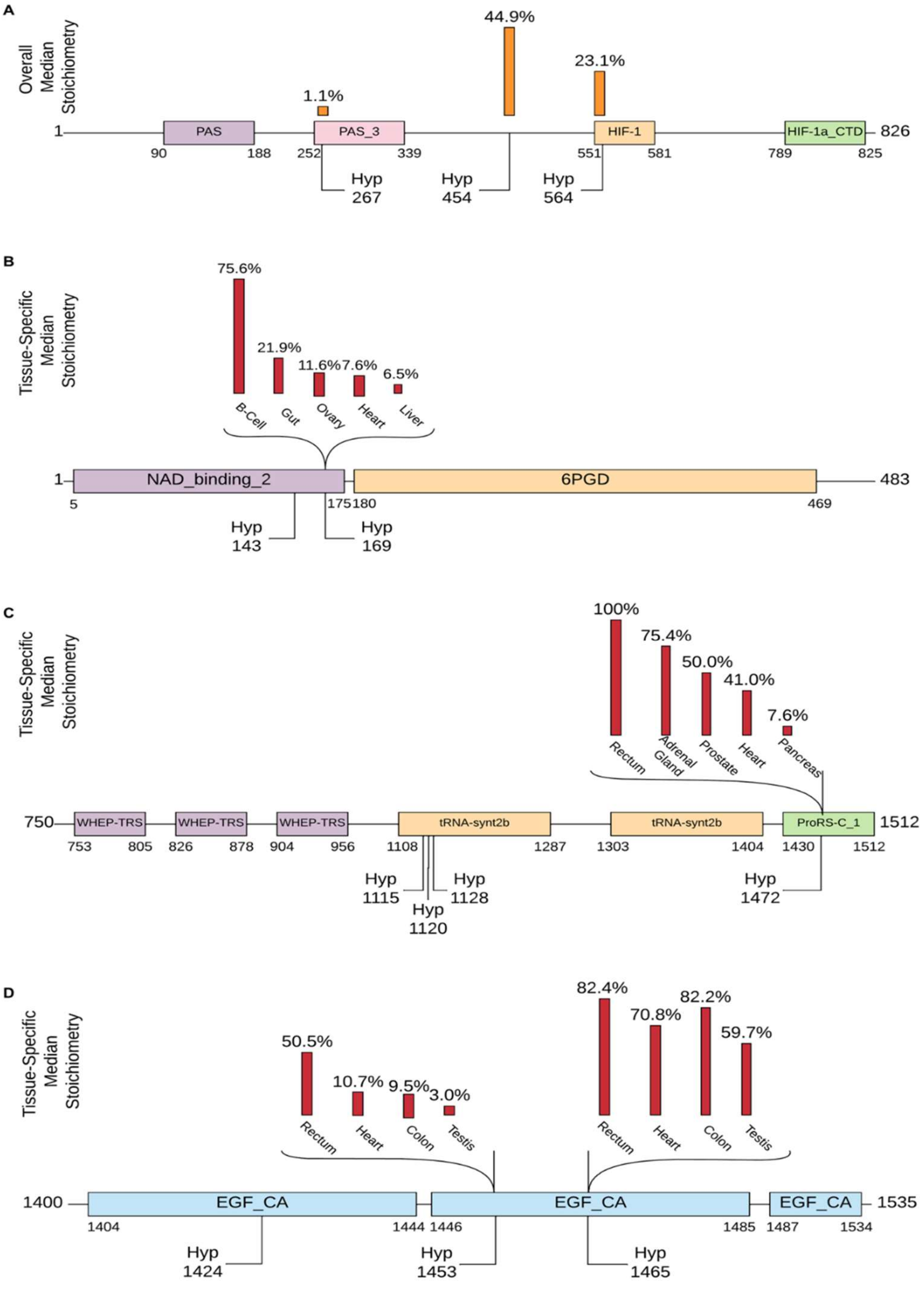
Nop56p-associated pre-rRNA
complex

PA700-20S-PA28 complex

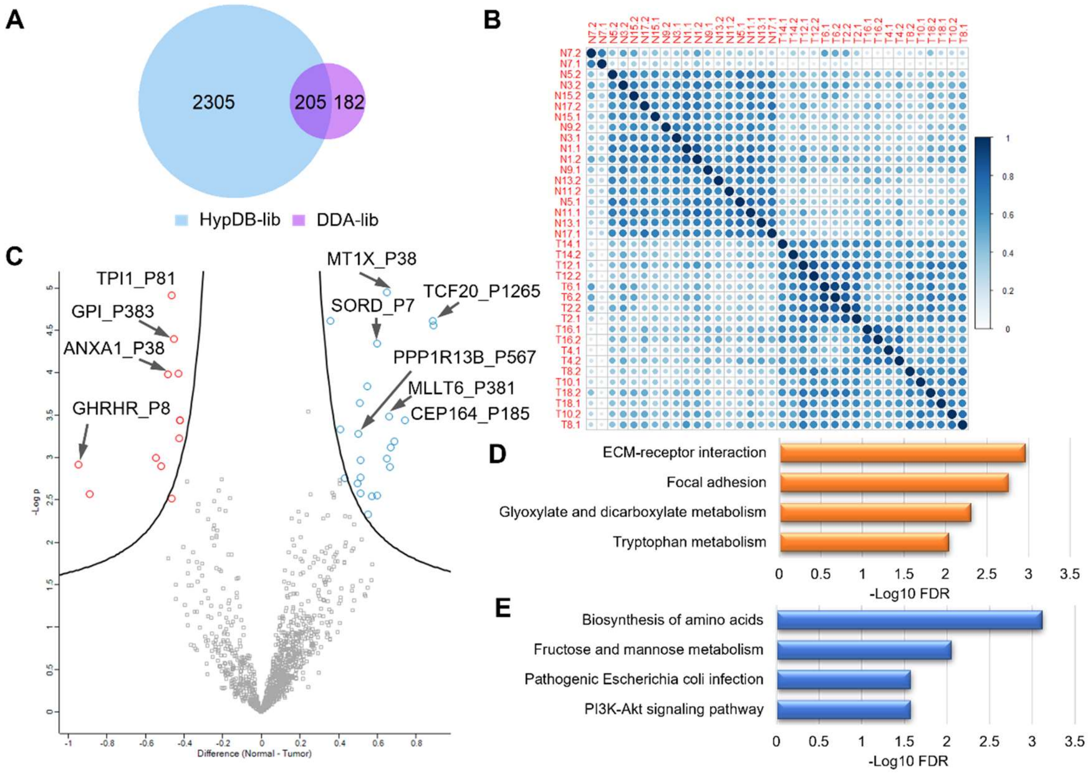
Supp Figure 3-7 Illustrations of Hyp protein networks that are enriched with various CORUM complexes.



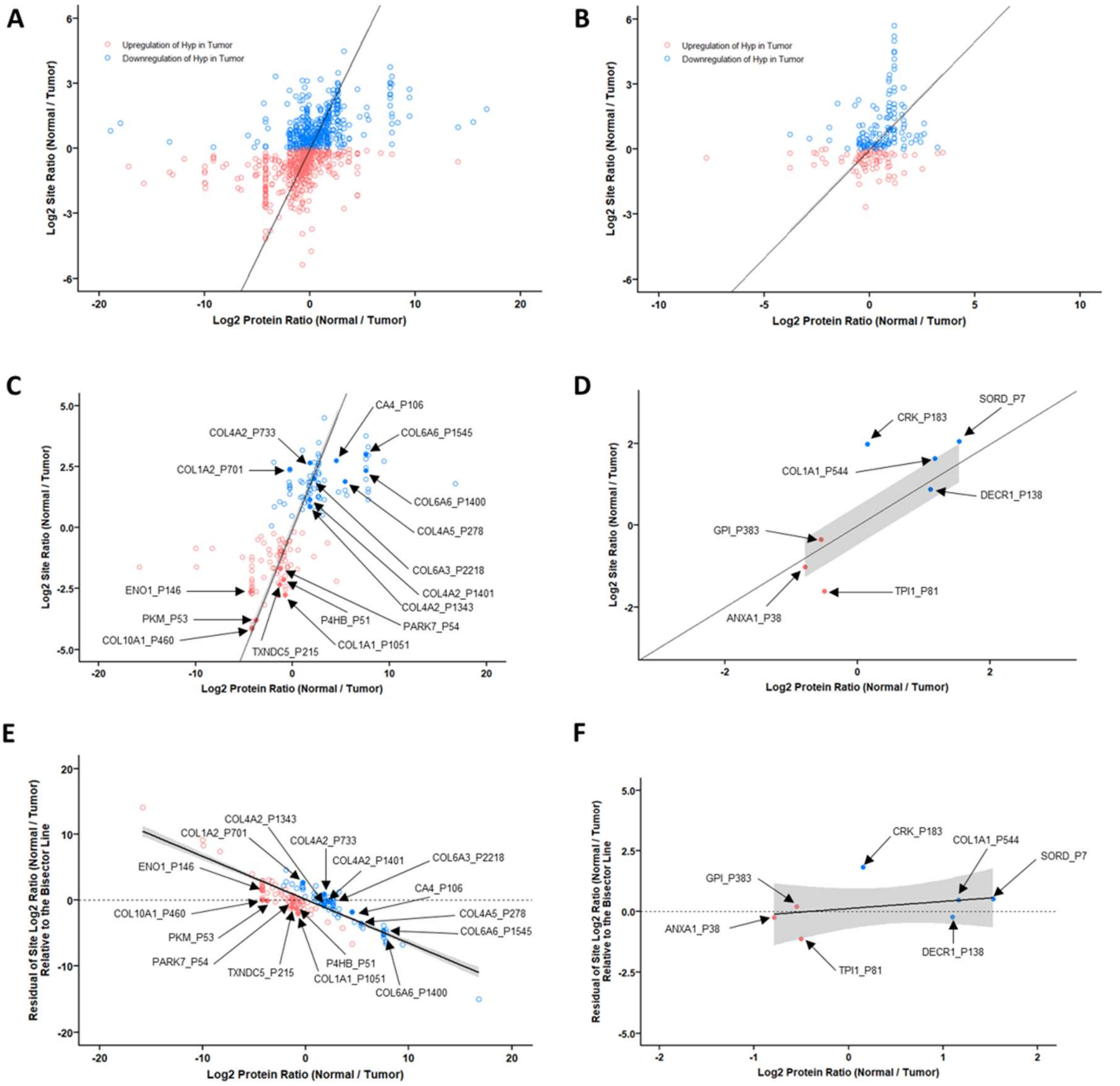
Supp Figure 3-8 Flanking sequence motif and neighboring protein modifications of Hyp sites.



Supp Figure 3-9 Examples of Hyp substrate proteins with site-specific Hyp stoichiometries in different tissues.



Supp Figure 3-10 DIA analysis of Guo and colleagues study of kidney cancer revealed differentially regulated Hyp substrates in tumor.



Supp Figure 3-11 Dynamics of the Hyp sites in correlation with corresponding protein abundance changes between normal and tumor tissue.

Chapter IV – Integration of Functional Proteomics Approaches and Functional Annotation of Hydroxyproline Proteome

Yao Gong[1,2], Yue Chen[1,2]

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

[2]Bioinformatics and Computational Biology Program, University of Minnesota at Twin Cities, Minneapolis, Minnesota, United States of America

This chapter contains an original research article in preparation to be published.

Yao Gong collected data, performed all experiments, analysis, and preparation of manuscript

Yue Chen formulated the concept, provide supervision, reviewed and edited the manuscript

4.1 Summary

Post-translational modifications (PTMs) play a vital role in regulating protein function and activity. Mass spectrometry (MS)-based proteomics has revolutionized PTM identification but lacks detailed functional information. To address this, we integrated functional approaches with MS-based proteomics to investigate proline hydroxylation, a critical PTM. Our study aimed to enhance understanding of the functional landscape of hydroxyproline sites and their impact on protein function.

We first explored the evolutionary conservation of hydroxyproline sites across 13 animal species. Certain ortholog groups, including Actins, Collagens, and Tubulins, showed high conservation, indicating their functional importance in structural roles. Functional enrichments and protein interactions provided insights into key biological processes and specific protein complexes associated with conserved hydroxyproline sites. Next, the turnover analysis revealed that hydroxyproline-containing peptides exhibited different half-lives than corresponding proteins in hepatocytes and monocytes, indicating stability affected by hydroxylation. Cell-specific regulation mechanisms and distinct functional associations were observed. Furthermore, we investigated hydroxyproline thermal profiles in childhood acute lymphoblastic leukemia (cALL) cell lines, revealing higher melting points for hydroxyproline-containing peptides, indicating their stabilizing effect on protein structure. Enrichment analysis uncovered diverse biological processes associated with hydroxyproline, regardless of its impact on thermal stability. Lastly, integrating all results above, PRDX3 emerged as a protein of interest, with conserved hydroxyproline sites and

interactions with proline hydroxylase EGLN3, highlighting their relationship with protein turnover, thermal stability, and cellular functions.

Our study provides insights into the functional landscape of proline hydroxylation and its impact on protein function. Further integration of proteomics data and functional approaches will contribute to the discovery of novel hydroxyproline targets, leading to potential diagnostic and therapeutic advancements.

4.2 Introduction

PTMs play a crucial role in regulating protein function and activity within biological systems. By introducing chemical groups to proteins, PTMs can effectively modify their properties, including stability, localization, and interactions with other molecules. Consequently, PTMs are key modulators of protein-protein interactions, enzymatic activity, and signaling pathways. Hence, it is imperative to identify and characterize PTMs to unravel the molecular mechanisms underlying biological processes and diseases. In this regard, MS-based proteomics has emerged as an indispensable tool for researchers, offering accurate identification and quantification of PTMs with exceptional specificity and sensitivity.

Recent years have witnessed significant advancements across the entire MS-based proteomics workflow, encompassing protein sample extraction and separation, the design and manufacture of mass spectrometers, and innovative approaches for analyzing MS data to precisely identify and quantify proteins. These improvements have substantially enhanced the effectiveness of this technique in identifying disease biomarkers, characterizing protein-protein interactions, and elucidating protein localization within cellular components through the identification and quantification of proteins, peptides, and PTMs. The applications of MS-based proteomics in the context of PTMs highlight its potential to revolutionize our comprehension of biological systems and pave the way for the development of novel therapies and treatments for a diverse array of diseases.

Despite the widespread use of MS-based proteomics in biological research, the field is still facing significant challenges. One of the major limitations of proteomics studies is the lack of mechanistic insights into the biological functions and activities of the identified proteins. While proteomics can accurately identify proteins, it often cannot provide detailed information on how these proteins interact with other molecules or how they function within a biological context. As a result, many proteomics studies remain correlative in nature, lacking the mechanistic insights that are necessary to fully understand biological processes.

Recently, novel functional approaches have been merged with proteomics experiments designated to overcome these challenges above. These designs combined proteome dynamics with different conditions that relate to protein functions, including temperature, time, and species.

Cross-species sequence comparison was one of the first approaches that discussed the protein functions in a global view. The initial usage of this approach helped build evolutional history and gathered proteins into superfamilies[267,268]. And there have been many phylogenetic studies discussing how the conserveness of protein sequences affects their functions in a protein family[269,270] and interactions between proteins[271,272] since then.

A large-scale PTM phylogenetic study from Krogan's lab explored about 200,000 PTM sites across 11 eukaryotic species, to illustrate that conservation of PTM sites with are related to carrying biological functions are more likely to be conserved and may indicate important regulatory regions[273] and a recent in-depth study by Müller *et al.* identified 340,000 proteins across 100 different organisms stringently with their advanced MS workflow[274], which provides rich resources for PTM evolution conservation and functional studies.

Turnover analysis was an emerging technique first used to compare turnover speed difference between messenger RNA and proteins translated from them[275]. Then this technique closely associated with the synthesis, regulation, and degradation of proteins, has become a powerful tool for filling the gap between finding proteins and understanding their functions. With the establishment of dynamic SILAC, or pulse SILAC, workflow based on stable isotopic amino acid feeding[154,155], turnover analysis has been applied to protein and PTM studies under various conditions in different perspectives[156–158].

Thermal profiling analysis was initially applied to study drug-protein binding dynamics[276], where researchers find it useful in exploring protein biological functions and protein-protein interactions. There has been a boom of thermal profiling studies recently, focusing on investigating the thermal profile of proteomes influenced by different aspects, such as solution concentration[160], PTMs[161], cell line difference, and proteoform groups[162].

As functional proteomics may provide more detailed insights into the function of proteins and PTM, there is a lack of studies regarding proline hydroxylation. Its most well-known function is performed through the proline hydroxylase, PHDs, and their target, HIF proteins, which are master transcription regulators in the oxygen sensing pathway[223,277]. Also, hydroxylated proline helped the formation of structural proteins like collagen[278,279] and actin[280], and many other metabolic pathways influenced by this PTM were found in the HypDB project[281]. However, there's little study about these functions associated with hydroxylated proline. To fill this knowledge gap, we would like to merge PTM identification with these emerging functional annotation approaches above to explore the functional proteome landscape influenced by individual proline hydroxylation sites. First, we explored how evolutional conserved hydroxyproline sites contribute to molecular function by filtering evolutional conserved sites among hydroxyproline sites collected across 13 animal species. Secondly, we investigated how proline hydroxylation may influence the turnover rate of different proteins which implies the relationship between these hydroxylation sites and protein functions. Thirdly, we studied proline hydroxylation sites that alter the thermal stability of the protein. We gathered all analyses from the three perspectives above and explored the potential association through the integration of the three analyses above. These results not only showed that the three methodologies we used were effective against finding functional relevant hydroxyproline sites, but also provided researchers in this field rich knowledge to facilitate their studies in this area.

4.3 Results

4.3.1 Construction of a Cross-Species Hydroxyproline Database for Conservation Analysis

Based on the idea that the conservation of phosphorylation sites is a better indicator of functional importance than the conservation of phosphorylation acceptor residues[273], we aimed to identify proline hydroxylation sites for further analysis of their evolutionary conservation using cross-species multiple sequence alignment. Our focus was on animal species, so we collected mass spectrometry (MS) raw data from 11 animal species as part of a cross-species proteomics study[274]. Additionally, MS raw data from a mouse proteome study[282] was also included. MS data of each species were searched using MaxQuant with variable modifications including N-terminal acetylation, methionine oxidation, and proline hydroxylation. Fasta files corresponding to different species were downloaded from UniProt and used as the library, while other parameters followed the settings from the original Nature study[274], with the remaining parameters set as default.

To compile a comprehensive set of hydroxyproline sites, we gathered the hydroxyproline site information from the search results of all species. We then removed reverse and potential contaminant sites from the dataset. In order to further enhance the coverage of hydroxyproline sites, we incorporated hydroxyproline sites identified in the HypDB project[281], which contained approximately 14,000 sites in humans. The combined sites obtained from these various sources constituted our cross-species hydroxyproline database, which would be utilized in the subsequent calculation of site evolutionary conservation. Figure 4-1 provides an overview of the cross-species hydroxyproline site collection.

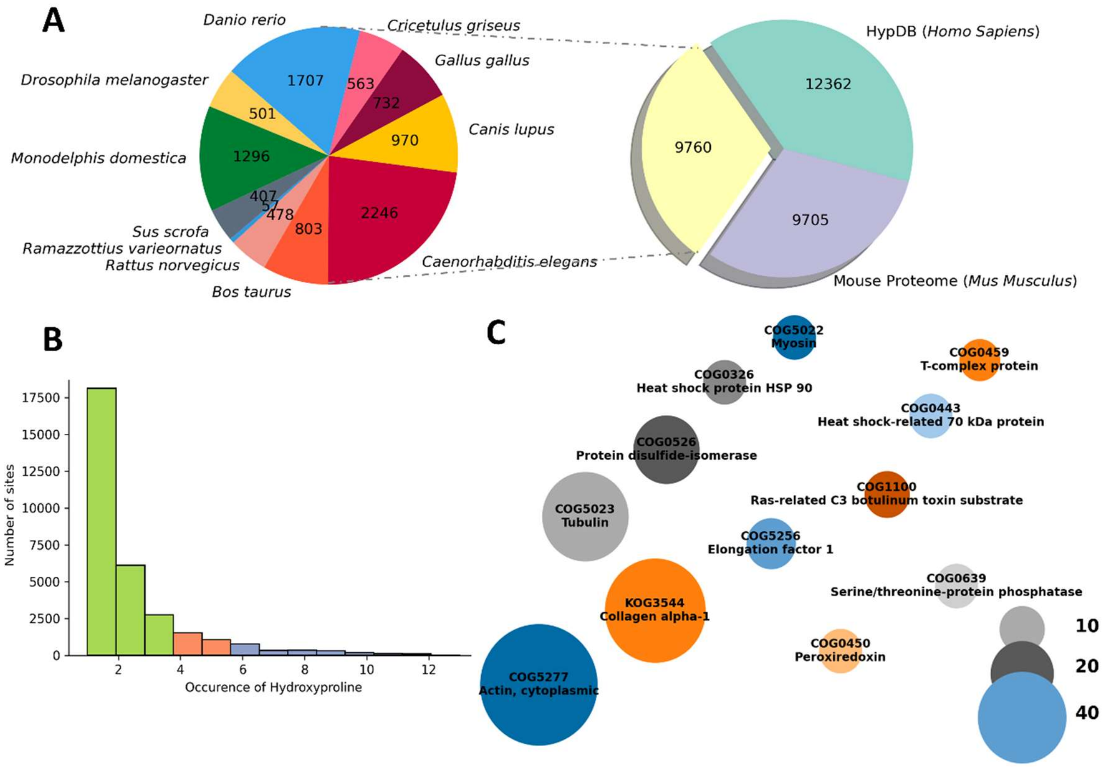


Figure 4-1 Overview of evolution conservation results across 13 species (A) Hydroxyproline sites identified in 13 species from different resources. (B) Distribution of occurrence of hydroxyproline among all sites analyzed. (C) Top metazoan orthologs with sites whose occurrence of hydroxyproline no less than six, size of circle showed total number of high conserved sites found among all human proteins in corresponding ortholog and protein names below ortholog ID are the protein representatives of orthologs.

4.3.2 Multiple Sequence Alignment and Evolutionary Conservation Analysis

To assess the conservation of proline hydroxylation sites across the 13 species, we performed multiple sequence alignments of the protein sequences corresponding to the collected sites. Initially, we employed the online Eggnog mapper tool to label the metazoan ortholog to which each protein belonged. Subsequently, we utilized MUSCLE[283], complemented by a custom Python code, to align the protein sequences within each metazoan ortholog, along with the associated hydroxyproline sites.

After aligning the protein sequences and hydroxyproline sites, we calculated the occurrence of protein, the occurrence of proline, and the occurrence of hydroxyproline for each site within its respective metazoan ortholog. By comparing the occurrence of proline to the occurrence of protein, we obtained a percentage indicating the conservation of the proline residue. Similarly, by comparing the occurrence of hydroxyproline to the occurrence of proline, we derived another percentage indicating the conservation of the hydroxylation modification, and its distribution among all analyzed sites was shown in Figure 4-1B. Additionally, we integrated the credit class labels and calculated stoichiometries from HypDB[281], and the interaction relationship with six proline hydroxylases provided by BioGRID[153] to these human conserved sites for more comprehensive information and prepared for the following analysis.

4.3.3 Identification of Metazoan Orthologs with Enriched Conserved Hydroxyproline Sites

Next, we focused on exploring the biological functions associated with the most evolutionarily conserved sites in Homo sapiens, which species set provided the most comprehensive information across the 13 species.

Specifically, we examined the human hydroxyproline sites that appeared in at least six different species. We first investigated the metazoan orthologs with the highest number of hydroxyproline sites, considering those with at least six occurrences of hydroxyproline.

The top metazoan orthologs, containing more than eight sites with six or more occurrences of hydroxyproline, were presented in Figure 4-1C.

We presumed that the function of proteins labeled with more evolutional conserved hydroxyproline sites may have a closer association with proline hydroxylation and not surprisingly, the top three metazoan orthologs with the most conserved sites found were Actins, Collagens, and Tubulins, all of which were structural proteins related with cytoskeleton and extracellular matrix organization where proline hydroxylation played crucial roles. Meanwhile, there were also protein orthologs that act as enzymes, like disulfide-isomerases, serine/threonine-protein phosphatases, and peroxiredoxins; together with proteins from other orthologs, they were closely related to protein folding, homeostatic process, and vesicle-mediated transport.

4.3.4 Investigation of Top Conserved Ortholog Groups

We conducted a detailed examination of the evolutionary conservation landscape within the top three ortholog groups enriched in hydroxylated proline sites. A comparison of the occurrence of proline and hydroxyproline across all positions of Actin, Tubulin, and Collagen is depicted in Figure 4-2A-F. Among these groups, both the Actin and Tubulin groups were present in all 13 species, and their proline residues remained conserved across most aligned positions. In contrast, the Collagen group exhibited distinct conservation patterns compared to the other two groups. Firstly, the metazoan ortholog KOG3544, encompassing Collagen proteins, was identified in only 11 out of the 13 species, excluding *Rattus norvegicus* and *Ramazzottius varieornatus,* and prolines were also less conserved in collagen sequences. Although hydroxylation was not consistently observed on all sites for

Actins and Tubulins across species, seven and five aligned positions, respectively, displayed hydroxylation in over half of the 13 species. The Collagen proline residues showed less conserveness across all aligned positions where most hydroxyproline sites were found in only one, two, or three species, with only a small portion present in more than three species, reaching a maximum of eight species out of the 11 species where Collagen proteins were found. Notably, hydroxyproline sites were predominantly found in three species: *Homo sapiens*, *Mus musculus*, and *C. elegans* illustrated in Figure 4-2F. Therefore, there is still significant potential for increasing the size of the data collected and improving the identification of hydroxyproline sites in other species, including *Rattus norvegicus* and *Ramazzottius varieornatus*.

We were also interested in if there were relationships between the varied occurrence of Collagen hydroxyproline sites and their stoichiometries. So, we plotted the box plot, Figure 4-2G, where ANOVA (analysis of variation) test and multiple comparisons showed that there were significant differences between the means of different groups ($p<10^{-15}$). For example, the group whose occurrence of hydroxyproline is one has a significantly different stoichiometry distribution compared to groups whose occurrence of hydroxyproline is six or eight. Also, there was a trend that as the occurrence of hydroxyproline increased, the range of site stoichiometries shrank towards the higher end, implying that more conserved collagen hydroxyproline sites tend to have more stable hydroxylation modification and functionality.

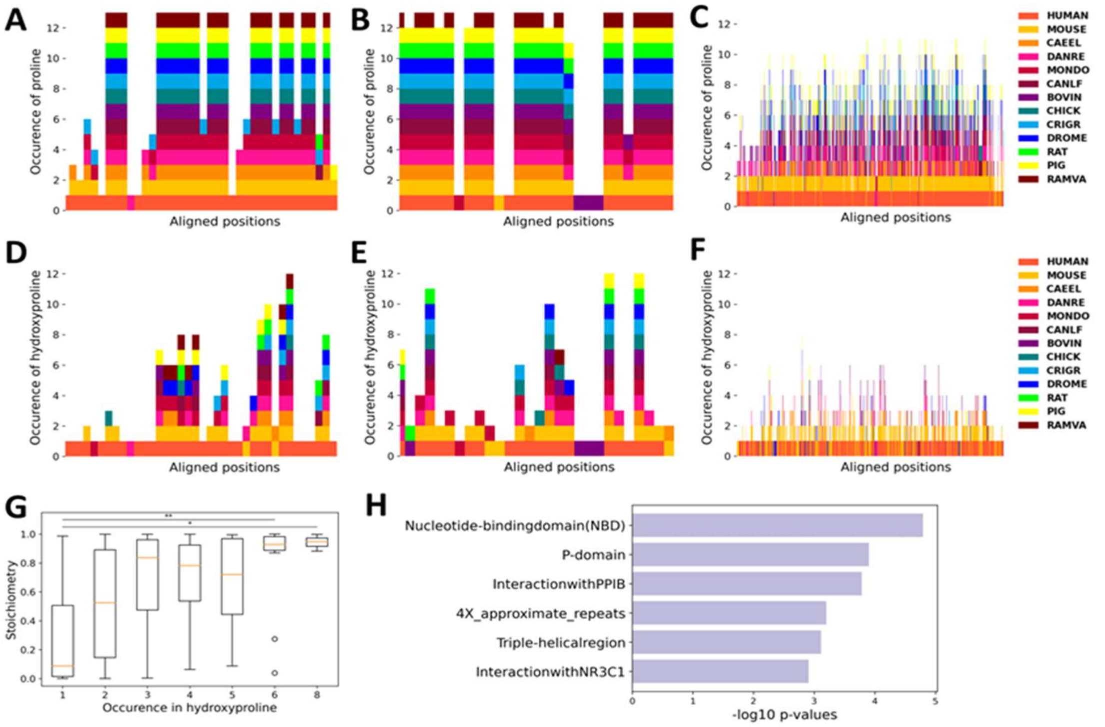


Figure 4-2 Profiles of evolutional conserved hydroxyproline sites (A, B, C) Occurrence of proline on all aligned positions of different metazoan orthologs, (A) COG5277, actins (B) COG5023, tubulins (C) KOG3544, collagens, in 13 species, different colors showed among which species this site was aligned with a proline. (D, E, F) Occurrence of hydroxyproline on all aligned positions of different metazoan orthologs, (D) COG5277, actins (E) COG5023, tubulins (F) KOG3544, collagens, in 13 species, different colors showed among which species this site was aligned with a hydroxyproline. (G) Boxplot of collagen sites stoichiometries grouped by their occurrence in hydroxyproline. (H) Top enriched regions of evolutional conserved sites. (** stands for adjusted p-value<0.01, * stands for adjusted p-value<0.05)

4.3.5 Enrichment Analysis of Conserved Hydroxyproline Sites

Motivated by the presence of potentially missed identification of evolutionarily conserved sites in certain species, we refined our analysis by increasing the occurrence restriction of hydroxyproline to four and reanalyzing the human proteins with conserved sites under this updated criterion. To mitigate the influence of potential false-positive hydroxyproline sites, we also excluded class III sites from the subsequent analysis. We performed a series of functional enrichment analyses using Metascape, and the results are presented in Figure 4-3A. The most significantly enriched terms included protein folding ($p<10^{-30}$), carbon metabolism ($p<10^{-28}$), cell adhesion molecule binding ($p<10^{-28}$), oxidoreductase activity ($p<10^{-24}$), and focal adhesion ($p<10^{-22}$).

Furthermore, we conducted protein complex enrichment analysis based on CORUM complexes. The enriched complexes included the Nop56p-associated pre-rRNA complex ($p<10^{-11}$), TNF-alpha/NF-kappa B signaling complex 6 ($p<10^{-5}$), and TLE1 corepressor complex ($p<10^{-4}$) (Figure 4-3B). Notably, these complexes were also enriched in the HypDB global enrichment results. Region and domain enrichment results were presented in Figure 4-2H and Supp Figure 4-1 Top enriched domains of evolutional conserved sites. The most significantly enriched region was the nucleotide-binding domain, followed by the triple helical region, while the most enriched domain was the collagen-like domain. Additionally, the thioredoxin domain showed a high level of enrichment.

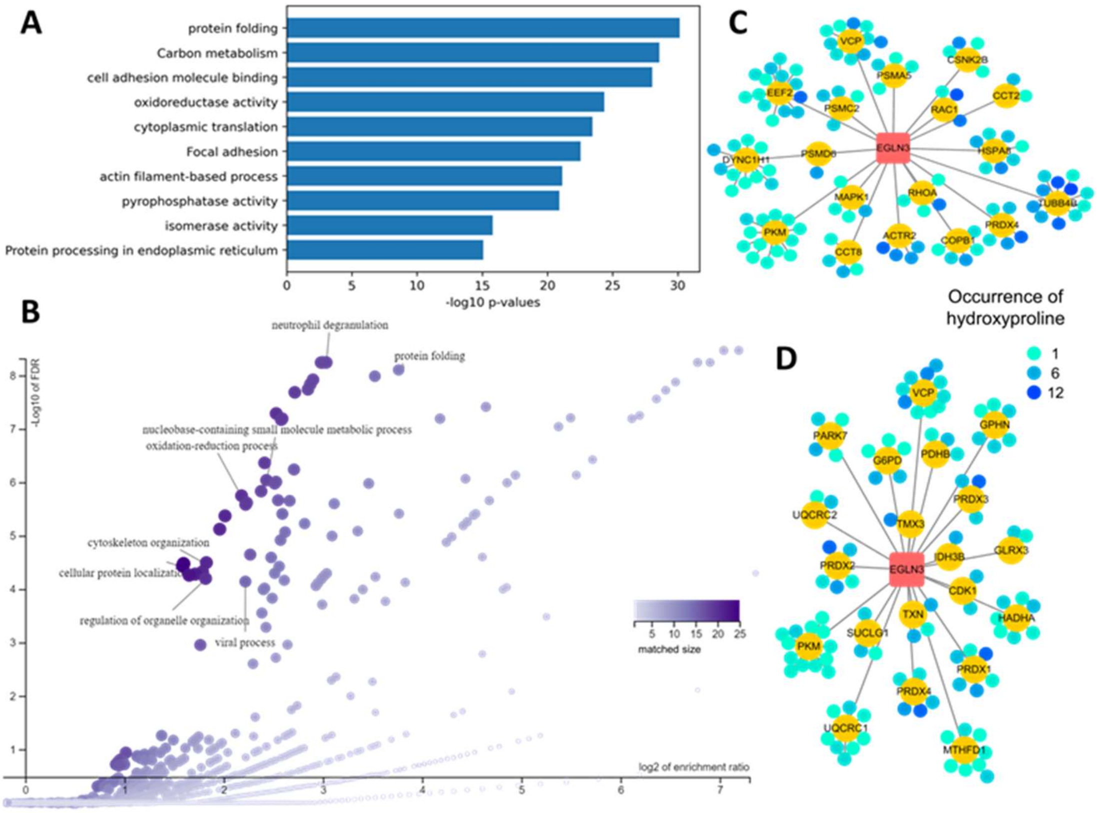


Figure 4-3 Enrichment analysis of evolutional conserved hydroxyproline sites (A)Top enriched terms of evolutional conserved sites. (B) CORUM complex enrichment of evolutional conserved sites. (C, D) Interaction between evolutional conserved sites enriched in (C) neutrophil degranulation, (D) oxidation-reduction process. Each yellow circle in (C) and (D) represents a protein, and each small circle surrounding yellow circles represents an evolutional conserved hydroxyproline site, its occurrence of hydroxyproline labeled in color gradient.

We further explored the interactions between proteins carrying evolutionarily conserved hydroxyproline sites and proline hydroxylases using information retrieved from BioGRID[284]. Among the six proline hydroxylases, EGLN3 displayed the highest number of interactions with hydroxyproline sites, with at least four occurrences of hydroxyproline. A Gene Ontology biological process enrichment analysis revealed that the overlap between proteins carrying these conserved sites and EGLN3 interactants was significantly enriched in processes such as neutrophil degranulation ($p<10^{-11}$), protein folding ($p<10^{-10}$), and nucleobase-containing small molecule metabolic process ($p<10^{-8}$) and oxidation-reduction process ($p<10^{-8}$). Interactions between EGLN3 and proteins containing conserved hydroxyproline sites enriched in neutrophil degranulation and oxidation-reduction process were shown in Figure 4-3C-D.

4.3.6 Collection and Preprocessing of Protein Turnover MS Data

To investigate the role of proline hydroxylation in reshaping the protein function landscape through protein turnover, we utilized data from a dynamic Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) study[285]. The dataset included MS raw data obtained from four human cell lines: b-cells, hepatocytes, monocytes, and NK cells. Each cell line was subjected to quantification at four different time points following a switch to culture media containing heavy isotope-labeled amino acids.

To analyze the turnover dynamics and functional implications of proline hydroxylation, we processed the MS raw data using the MaxQuant software. The parameter settings were aligned with those used in the original study. Additionally, we incorporated proline hydroxylation as a variable modification during the search process. By consolidating the search results from the four cell lines, we obtained comprehensive data for turnover time calculation and subsequent functional annotation. This allowed us to explore the temporal dynamics of protein turnover and investigate the impact of proline hydroxylation on protein function. The HypDB derived credit class labels and site-specific stoichiometries, together with proline hydroxylases interactant information from BioGRID were also integrated into the results, prepared for following analysis.

4.3.7 Calculation and Global Landscape of Protein Half-lives Across Various Cell Lines

Based on the methods derived from the paper[285] we collected turnover data from, we developed the formula that calculated the half-life of every protein we identified and every proline hydroxylation site we identified. We would like to explore if the half-life of hydroxyproline sites show similar patterns across four cell lines or not, and it is obvious that in the multi-scatter plot of Figure 4-4A-B, none of these four-cell lines tends to have similar half-life landscapes of peptides or proteins with hydroxylated prolines. The distributions of half-lives were not normal distributions, so we performed a Mann-Whitney U test between hydroxyproline peptide half-life distribution and the half-life distribution of corresponding proteins in each cell line. And as Figure 4-4C-D shown, the half-lives of peptides with hydroxyproline sites were significantly different than the half-lives of corresponding proteins in hepatocytes and monocytes, two cell lines with the most half-lives of hydroxyproline peptides calculated. However, the difference was not significant in the two other cell lines. And the number of hydroxyproline sites with half-life increase was greater than decreased ones consistently across four cell lines, suggesting that hydroxylation on proline may influence the half-lives of the protein. In this case, we would like to investigate how proline hydroxylation alters the turnover rates across the proteome, so we divided the site half-life by the corresponding protein half-life and named it half-life ratio. Although these four cell lines have distinct half-life changing patterns with proline hydroxylation, we also found clustering patterns around several hydroxyproline half-life ratios in these cell lines, so we clustered these hydroxyproline sites with K-means clustering with four clusters based on their hydroxyproline half-life ratio, naming them Cluster I to Cluster IV on ascending half-life ratios, shown in Figure 4-5A.

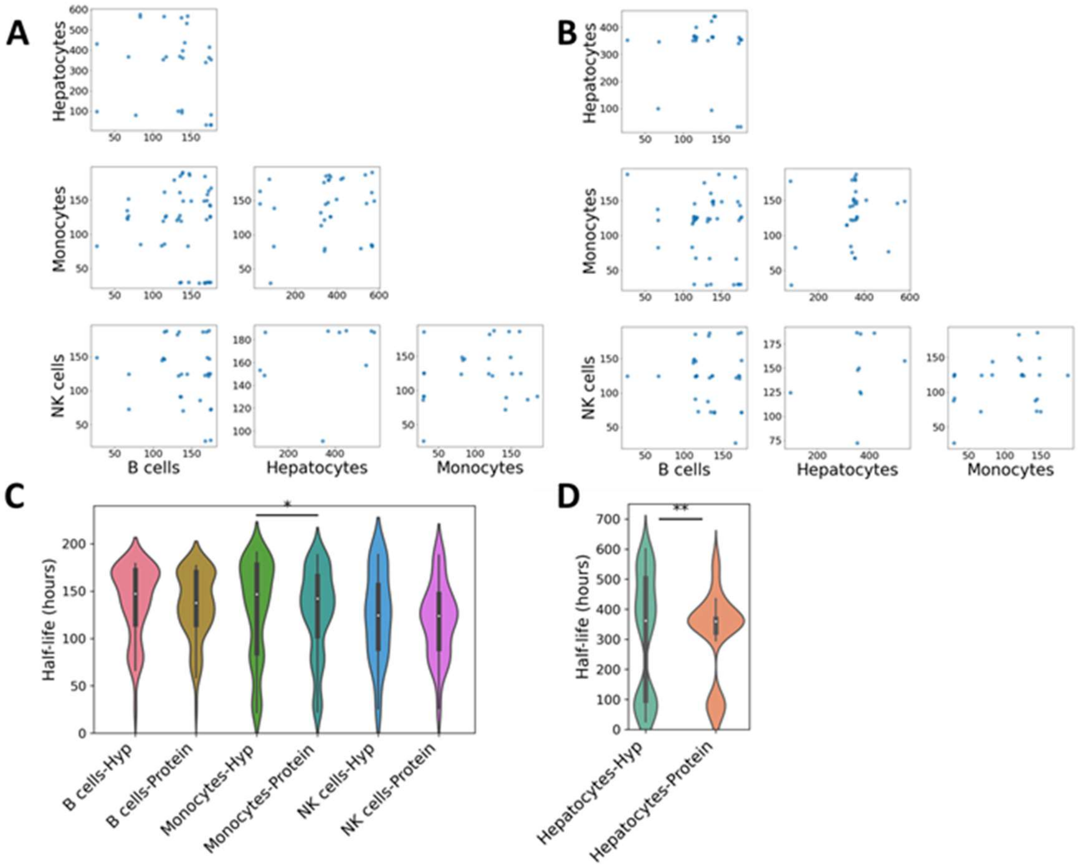


Figure 4-4 Global turnover profile among four cell lines (A) Multi-scatter plot for hydroxyproline sites half-lives in hours across four cell lines. (B) Multi-scatter plot for protein half-lives in hours with hydroxyproline sites in hours across four cell lines. (C) Violin plot for half-lives of peptides with hydroxyproline sites and half-lives of corresponding proteins in B cells, Monocytes, and NK cells. (D) Violin plot for half-lives of peptides with hydroxyproline sites and half-lives of corresponding proteins in Hepatocytes. (** stands for p-value<0.01, * stands for p-value<0.05)

Hepatocytes half-life ratios had the most explicit clustering pattern, as Figure 4-5A shown, it's obvious that half-life ratios clustered around four centers, 0.2, 1, 1.2, and 1.4, instead

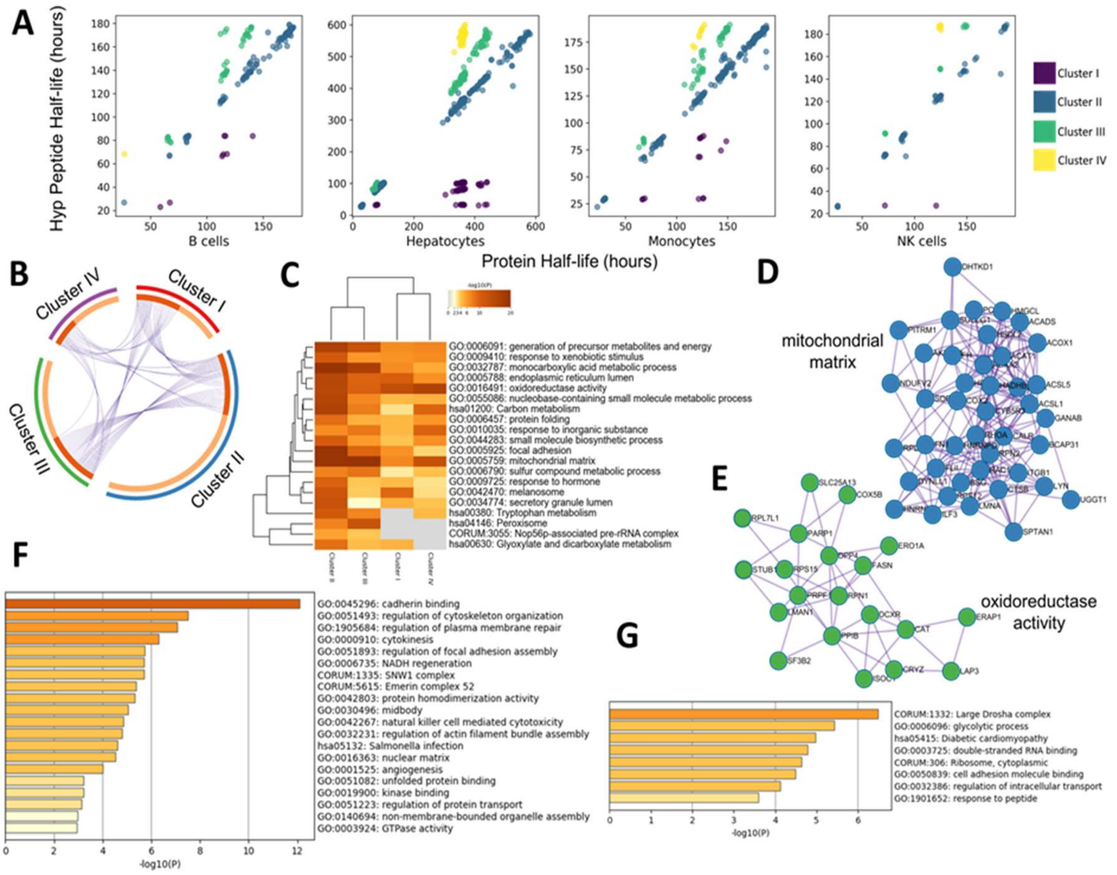


Figure 4-5 Half-life ratio-based clustering and enrichment analysis of clusters in four cell lines (A) Half-life scatter plot of hydroxyproline peptides to their corresponding proteins in four cell lines and k-means clustering results. (B) Pie chart of proteins with Clusters I-IV sites in hepatocytes and their overlaps. (C) Heatmap of top enriched terms of four clusters in hepatocytes. (D) Protein interaction network enriched in mitochondrial matrix of hydroxyproline proteins in hepatocytes. (E) Protein interaction network enriched in oxidoreductase activity of hydroxyproline proteins in hepatocytes. (F) Top enrichment terms of hydroxyproline proteins with Cluster III or IV sites in monocytes. (G) Top enrichment terms of hydroxyproline proteins with Cluster III or IV sites in B cells.

of having a normal or continuous uniform distribution. And the highest half-life ratio of Cluster I sites is below 0.43, significantly lower than the other three clusters where the lowest ratio of them is about 0.77. Meanwhile, Cluster III and Cluster IV represent hydroxyproline peptides with slight and significant increases in half-lives respectively.

4.3.8 Enrichment Analysis of Protein Clusters Based on Half-life Ratios in Different Cell Lines

With Class III sites removed, we performed GO biological enrichment analysis upon hepatocytes proteins in four clusters respectively and altogether. Although each cluster has more than half of the protein unique, it turned out that the enrichment results of the four clusters had little difference, shown in Figure 4-5B-C. And mitochondrial matrix ($p<10^{-41}$) was one of the most enriched terms for altogether results, followed by oxidoreductase activity ($p<10^{-38}$), monocarboxylic acid metabolic process($p<10^{-33}$), and cellular respiration ($p<10^{-27}$). Two protein interaction networks related to the top two enriched terms were shown in Figure 4-5D-E. Although terms like focal adhesion ($p<10^{-29}$) and protein folding ($p<10^{-13}$) were also enriched, hydroxyproline sites in hepatocytes had more focus on metabolisms. When mapped to the proline hydroxylase interactant list we fetched from BioGRID database[284], these proteins enriched in term oxidoreductase activity, showed close interaction with proline hydroxylase. Eleven of them interact with P4HA1, ten of them interact with EGLN3, and P4HA2 was also on the list.

Then we explored the turnover landscape of monocytes and B cells, both of which had fewer sites with half-life ratios changed compared to protein half-lives, especially their

Cluster I and Cluster IV only have less than twelve sites respectively. So, we only performed enrichment analysis with their hydroxyproline sites clusters with half-life ratios greater than one, their Cluster III and Cluster IV together in each cell line respectively. The heatmap of enrichment results is shown in Figure 4-5F-G. It was obvious that the functions carried by hydroxyproline sites in monocytes and B cells differed significantly from hepatocytes. Sites with half-life increased in monocytes were most enriched in cadherin binding ($p<10^{-12}$), focal adhesion ($p<10^{-11}$), regulation of cytoskeleton organization ($p<10^{-7}$), regulation of plasma membrane repair ($p<10^{-7}$), and cytokinesis ($p<10^{-6}$), all of which related with the irregular cell shape and cytoplasmic vesicles of monocytes. shape while those sites in hepatocytes served metabolisms more. There were interesting terms enriched, NADH regeneration ($p<10^{-5}$) bridged structural proteins with metabolism, and natural killer cell mediated cytotoxicity ($p<10^{-4}$) may suggest the potential influence of hydroxyproline sites in the cellular response to immune processes. When it came to B cells enrichments, large Drosha complex ($p<10^{-6}$), Nop56p-associated pre-rRNA complex ($p<10^{-5}$), glycolytic process (p<10^-4), oxidoreductase complex ($p<10^{-4}$), double-stranded RNA binding ($p<10^{-4}$), cell adhesion molecule binding ($p<10^{-4}$), and regulation of intracellular transport ($p<10^{-4}$) were the top enriched terms. Compared to hydroxyproline sites from hepatocytes and monocytes, B cell ones had closer relation with RNA binding and processing. Also, position 382 on IGHM was found in Cluster I of B cells with a half-life ratio of 0.39, which implied that there might be hydroxylation regulation sites on Immunoglobulin. Based on the global correlation analysis and enrichment analysis of individual cell lines, it seems that proline sites were not consistently hydroxylated across different cell lines, where its distribution was highly associated with the biological role of the corresponding cell.

4.3.9 Peptide Quantification and Thermodynamic Analysis of Temperature Series Mass Spectrometry Data

In this study, we collected MS data from a recent study[162] that explored the protein thermal profiles of childhood acute lymphoblastic leukemia (cALL) cell lines at different stages of B cell development. The thermal stabilities of proteins were measured at eight different temperatures ranging from 41-63°C for each cell line. To ensure comprehensive analysis, we labeled eight samples from two cell lines with one TMT16plex set.

Using MaxQuant, we analyzed the TMT16plex data for six cell lines and obtained the corrected intensity of reporter ions for each spectrum. We summed up the eight reporter ion intensities corresponding to the same peptide across all spectra. Subsequently, we fitted the ion intensity ratio at different temperatures for each peptide, considering both modified and unmodified forms, into a sigmoid curve based on protein thermodynamics [159,286]. Peptides with less than six valid reporter ion intensities at different temperatures, peptides with melting curves that did not fit the sigmoid function, and peptides without a valid melting point through calculation were excluded from further analysis.

4.3.10 Global Thermal Profiles of Hydroxyproline Peptides in cALL Cell Lines

On a global scale, we examined the melting points in the thermal profiling study. Firstly, we compared the melting points of hydroxyproline sites across six cALL cell lines and observed a consistent pattern shown in Figure 4-6A, indicating a similar hydroxyproline thermal stability landscape across cell lines of different genders and developmental stages.

Then we performed statistical tests between the melting points of different peptide groups across six cell lines. As Figure 4-6B illustrates, peptides with hydroxyproline sites showed significantly higher melting points compared to all other peptides in the same cell line across the six cell lines we analyzed. Moreover, in proteins with hydroxyproline, peptides with hydroxyproline also exhibited significantly higher melting point temperatures compared to peptides without hydroxyproline. These results suggested that hydroxylation of proline residues contributes to the stabilization of protein structure.

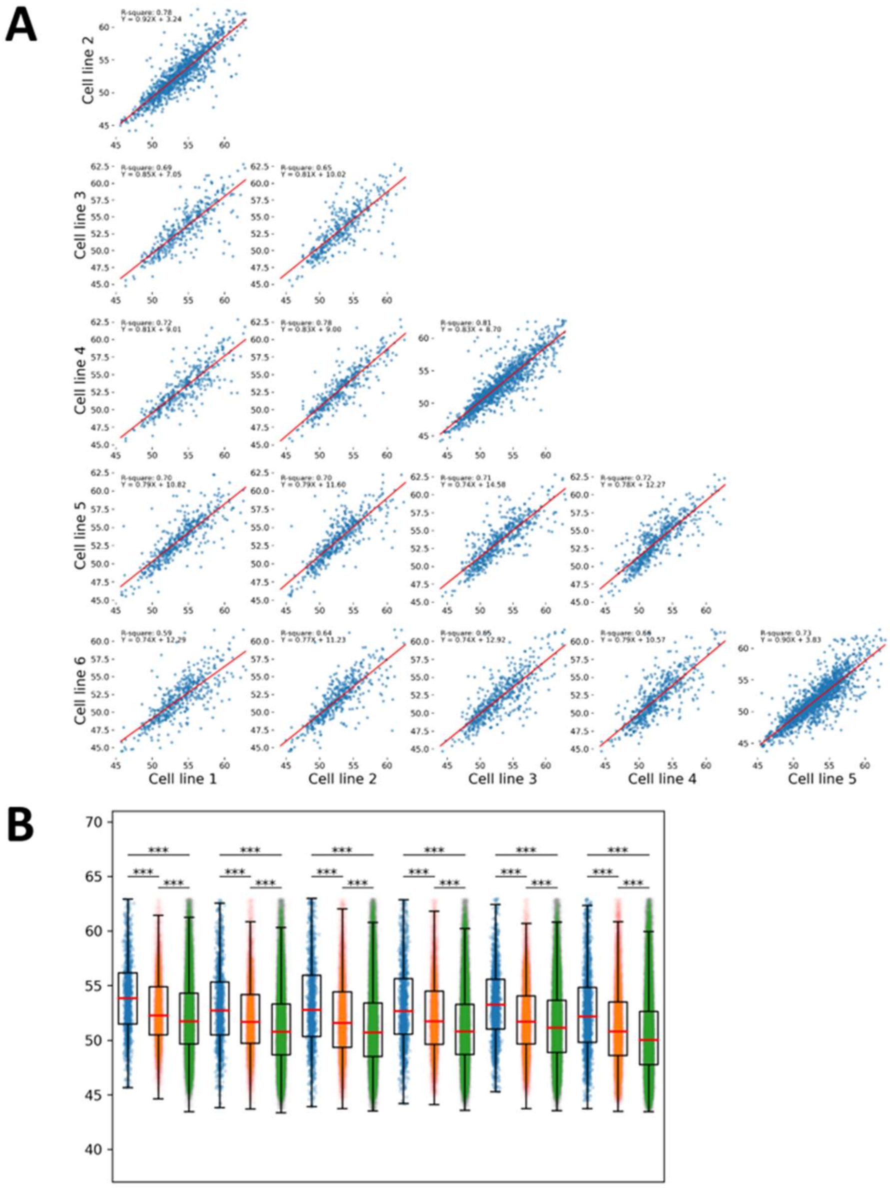


Figure 4-6 Global thermal profile of hydroxyproline peptides across six cell lines (A)Multi-scatter plot of hydroxyproline peptide melting points across six cell lines. (B) Boxplot of melting points of hydroxyproline peptide (blue), melting points of peptide in hydroxyproline proteins (orange), melting points of all peptides (green) and across six cell lines. (*** stands for adjusted p-value<0.001)

For each hydroxyproline peptide with a valid melting point, we performed a one-sample t-test between its melting temperature and all other valid melting point temperatures of the same protein to reveal if there was a significant difference between the melting temperatures of these peptides with proline hydroxylation and corresponding peptides. The average increase in melting temperature across six cell lines ranged from 0.11-0.29°C, and over 40 percent of peptides with hydroxyproline had a significantly different melting temperature compared to other peptides of the same protein when the p-value cutoff was set to 0.01. To ensure the selection of hydroxyproline peptides with significantly changed melting points, we set the p-value threshold to 0.001. Additionally, t-tests with less than five melting point samples were removed to reduce false positives.

4.3.11 Enrichment Analysis of Peptides with Altered Thermal Profiles in the Presence of Hydroxyproline

We explored the enrichment results separately for each of the six cell lines. In both groups where hydroxyproline peptide melting points significantly increased or decreased, we found that the majority of the proteins overlapped between cell lines, and they exhibited similar enrichment patterns (Figure 4-7A-B). We also investigated the enrichment results for all proteins with hydroxylation peptides from the six cell lines combined. In the group where proteins were more thermally stabilized, terms such as cadherin binding ($p<10^{-34}$), carbon metabolism ($p<10^{-24}$), nucleocytoplasmic transport ($p<10^{-23}$), ATP-dependent activity ($p<10^{-22}$), cytoplasmic vesicle lumen ($p<10^{-20}$), nucleobase-containing small molecule metabolic process ($p<10^{-19}$), and amide metabolic process ($p<10^{-17}$) were significantly enriched. In the group where proteins were destabilized, many of these terms remained top enriched, including cadherin binding ($p<10^{-26}$), carbon metabolism ($p<10^{-23}$), amide metabolic process ($p<10^{-21}$), cytoplasmic vesicle lumen ($p<10^{-18}$), nucleocytoplasmic transport ($p<10^{-17}$), and nucleobase-containing small molecule metabolic process ($p<10^{-15}$). This suggests that regardless of whether the protein thermal stability increased or decreased, proline hydroxylation played similar roles in many biological processes. Notably, in the protein-protein interaction enrichment analysis, we found interesting clusters from different groups, including secretory granule lumen ($p<10^{-6}$) and mRNA splicing via spliceosome ($p<10^{-13}$) in the increased thermal stability group and ubiquitin protein ligase binding ($p<10^{-8}$) and HIF-1 signaling pathway ($p<10^{-6}$) in the decreased group. All the interaction networks of enriched proteins in these terms were shown in Figure 4-7C-F.

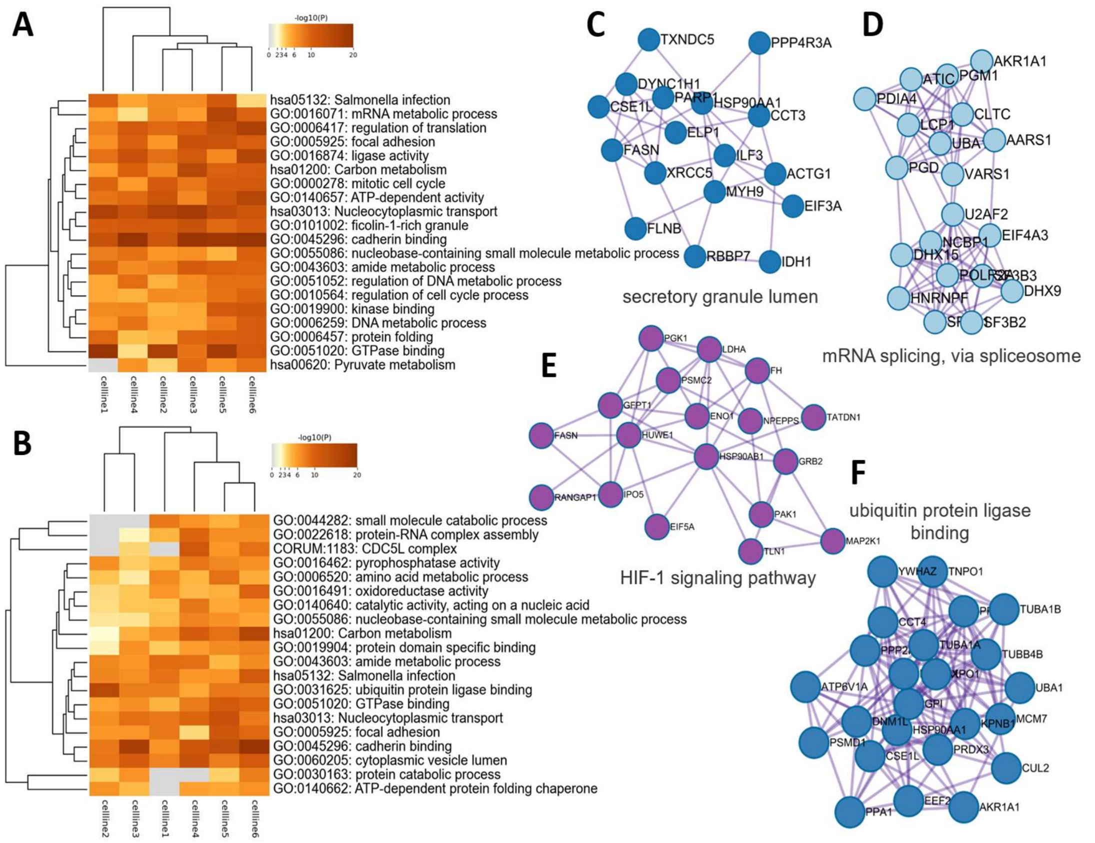


Figure 4-7 Enrichment results of proteins with melting point significantly changed peptides. (A) Top enriched terms of protein with sites whose thermal stability significantly increased when hydroxyproline present across six cell lines. (B) Top enriched terms of protein with sites whose thermal stability significantly decreased when hydroxyproline present across six cell lines. (C, D, E, F) Interaction network of proteins with thermal stability significantly changed sites enriched in different terms. (C) increased stability, secretory granule lumen (D) increased stability, mRNA splicing via spliceosome (E) decreased stability, HIF-1 signaling pathway (F) decreased stability, ubiquitin protein ligase binding

4.3.12 Integrated Analysis of Thermal Profiling, Evolution Conservation, and Protein Turnover

We were also interested in exploring how the thermal profiles of hydroxyproline sites related to their evolutionary conservation and turnover rates. For each peptide with proline hydroxylation, we calculated the mean difference between the peptide's melting point and the average melting point of all other peptides on the same protein across the six cell lines we investigated. We merged the thermal profiling data with the evolutionarily conserved data of human sites based on their protein names and positions and plotted a violin plot depicting the occurrence of hydroxyproline and the mean difference in melting point (Figure 4-8A). And through the ANOVA test, there were no significant differences between the means of groups. Then we performed the Shapiro-Wilk testing whether the data were normally distributed, which stands for symmetrical, and if not, followed by a skewness measurement to decide the tail on which end was longer. It turns out that ten groups out of twelve were not normal distribution, excepting groups where the occurrence of hydroxyproline was eight and eleven. Within these ten groups that were not normal distribution, seven of them had longer tails on the melting point decreasing end. We also generated a merged table of the B cell turnover landscape with thermal profiles, incorporating the data from the six B lymphocyte cell lines mentioned above. However, we observed very few sites with valid values in both analyses and the scatter plot (Figure 4-8B) between the half-life ratio and mean melting point difference showed little correlation.

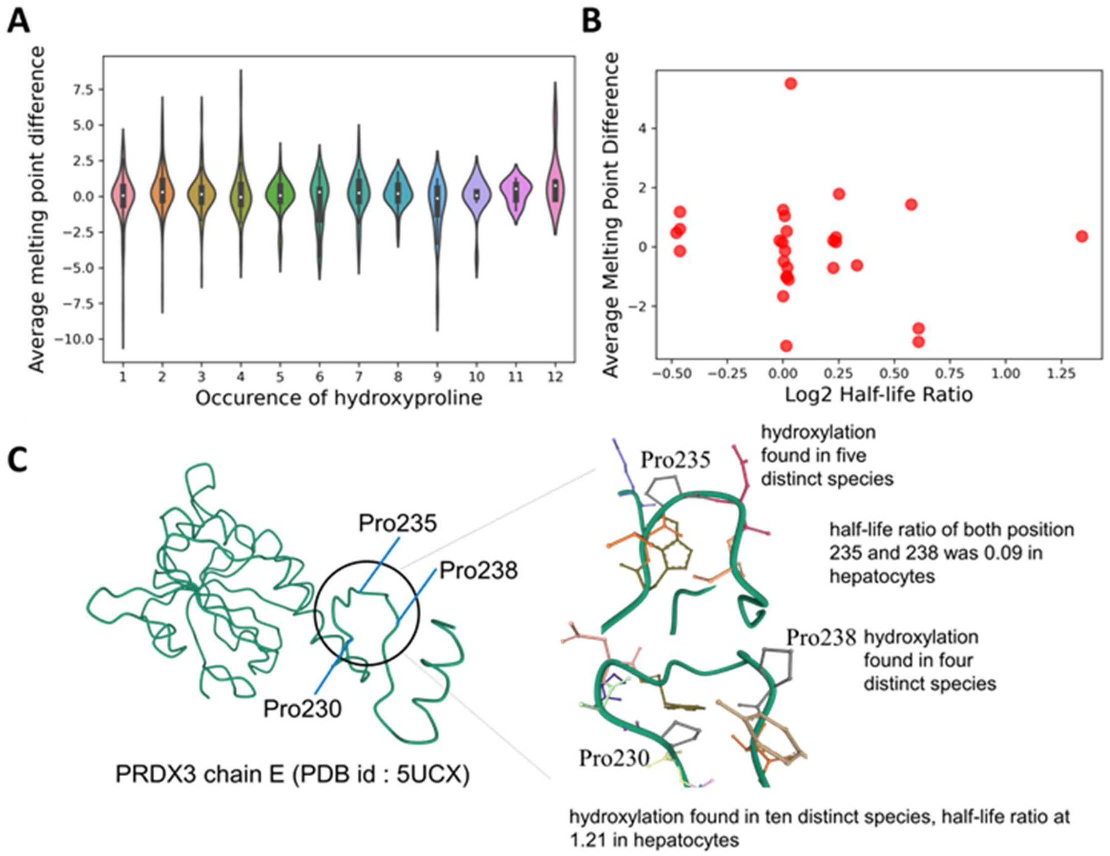


Figure 4-8 Integrated analysis of thermal profiling, evolution conservation, and protein turnover (A) Violin plot of mean difference of melting points across six cell lines of hydroxyproline sites with different occurrence of hydroxyproline. (B) Scatter plot of mean difference of melting point to log2 half-life ratio in B cells. (C) Three hydroxyproline sites on PRDX3 with high conservation and significantly changed half-lives in presence of hydroxylation.

Out of proteins with valid data in all three sets of analysis, PRDX3, an interactant of proline hydroxylase EGLN3, drew our attention. Not only were its positions 235 and 238 both Cluster I hydroxyproline sites, but both sites were also found in five and four distinct species in our evolutional conserved results. Interestingly, hydroxylation on proline position 230 of PRDX3 was found in ten species out of eleven species where proteins from the same ortholog were found, showing more conserveness, while having a much higher half-life ratio at 1.21 compared to 0.09 of the hydroxylated peptides with both position 235 and 238. Detailed evolutional conservation and turnover results of PRDX3 sites were shown in Figure 4-8C. Also, all three sites were found on the same peptide from position 218 to 241 in thermal profiling analysis across six cell lines. Peptides with zero to three hydroxyproline sites were found, and peptides with three hydroxyproline sites had melting points higher than all other peptides. The average melting point of peptides with three hydroxyprolines across all cell lines was 1.75°C higher than the average of peptides with no hydroxyproline across all cell lines. These results suggested that different hydroxyproline positions had different regulation effects and provide clues to reveal mechanisms on how their changes further influence the biological processes.

4.4 Discussion

Post-translational modifications (PTMs) play a pivotal role in regulating protein function and activity within biological systems. These modifications alter protein properties, such as stability, localization, and interactions, making PTMs critical modulators of protein-protein interactions, enzymatic activity, and signaling pathways. The advent of mass spectrometry (MS)-based proteomics has revolutionized our understanding of biological systems by enabling the accurate identification and quantification of PTMs. This powerful tool has unraveled the complexities of PTMs, facilitating the development of novel therapies aimed at modulating these modifications for therapeutic purposes.

While MS-based proteomics has significantly advanced the field of PTM identification with its high accuracy, it often lacks detailed information about protein interactions and functional context. To overcome this limitation, recent studies have sought to integrate functional approaches with proteomics experiments, combining proteome dynamics with conditions relevant to protein functions. These integrated approaches aim to provide a more comprehensive understanding of PTMs and their functional implications. For instance, evolution conservation analysis through cross-species sequence comparison, protein dynamics analysis through measuring turnover rates, and protein stability analysis through thermal profiling experiments have emerged as valuable strategies. However, limited studies have explored the functional annotation of specific PTMs, such as proline hydroxylation, using these integrated approaches in conjunction with MS-based identification.

To fill this gap, we integrated functional approaches with MS-based proteomics for PTM identification and quantification in this study. The integration of these approaches enhanced our understanding of the functional landscape of proline hydroxylation, a post-translational modification crucial for protein regulation. Through evolutionary conservation analysis, turnover rate analysis, and thermal stability analysis, we gained insights into the functional implications of hydroxyproline sites and their impact on the molecular function of proteins.

Firstly, we aimed to investigate the evolutionary conservation of hydroxyproline sites across animal species. By leveraging mass spectrometry data from 13 animal species and incorporating information from the HypDB project, we constructed a cross-species hydroxyproline database. Through multiple sequence alignment and comparative analysis, we identified the conservation patterns of hydroxyproline sites within metazoan orthologs. We observed that certain ortholog groups, such as Actins, Collagens, and Tubulins, exhibited high conservation of proline residues, indicating their functional importance in structural roles. Additionally, we identified enzymes involved in protein folding and transport that carried conserved hydroxyproline sites. Further exploration of conserved hydroxyproline sites allowed us to uncover functional enrichments and protein interactions associated with these sites. Through enrichment analysis, we identified key biological processes related to conserved hydroxyproline sites, such as protein folding, carbon metabolism, and cell adhesion. Additionally, protein complex analysis revealed the involvement of specific complexes, including the Nop56p-associated pre-rRNA complex and TNF-alpha/NF-kappa B signaling complex 6, which may play crucial roles in cellular functions. Furthermore, our investigation of interactions between proteins carrying conserved hydroxyproline sites and proline hydroxylases highlighted EGLN3 as a prominent player, demonstrating its involvement in processes such as neutrophil-mediated immunity and protein localization. These findings provide valuable insights into the functional landscape and regulatory potential of hydroxyproline sites in animal proteomes.

Secondly, turnover analysis shed light on the role of proline hydroxylation in protein function reshaping. We found that the half-lives of peptides containing hydroxyproline sites were significantly longer than the half-lives of corresponding proteins in hepatocytes and monocytes, indicating that hydroxylation on proline may increase protein stability. By calculating the half-life ratios, we observed distinct clustering patterns of hydroxyproline sites in different cell lines, suggesting cell-specific regulation mechanisms. Enrichment analysis revealed that hydroxyproline sites in hepatocytes were associated with metabolic processes, while those in monocytes and B cells exhibited enrichment in cellular adhesion, cytoskeleton organization, and immune response-related functions. These findings highlight the intricate relationship between proline hydroxylation, protein turnover, and cellular functions, providing valuable insights into the regulatory mechanisms governing protein dynamics.

Thirdly, we investigated the hydroxyproline thermal profiles of childhood acute lymphoblastic leukemia (cALL) cell lines at different stages of B cell development. Our analysis revealed that peptides with hydroxyproline sites exhibited higher melting points, indicating the stabilizing role of hydroxylation in protein structure. This effect was consistent across different cell lines, suggesting a shared hydroxyproline thermal stability landscape. Enrichment analysis highlighted the involvement of proline hydroxylation in various biological processes, irrespective of whether it increased or decreased protein thermal stability. Our findings contribute to understanding the molecular mechanisms underlying cALL and provide insights into potential therapeutic strategies targeting protein stability.

Additionally, we explored the relationship between thermal profiling, evolution conservation, and turnover rate data through integration. We discovered and studied the case of PRDX3, containing three evolutional conserved hydroxyproline sites whose turnover rates and melting points changed significantly with and without hydroxylation.

Although the MS raw data we collected may not fully depict the functional landscape of hydroxyproline, we plan to incorporate more comprehensive proteomics data and integrate multiple functional approaches in the future. There could be more hydroxyproline sites collected, on Collagen in *Rattus norvegicus* and *Ramazzottius varieornatus* for example. And we may annotate hydroxyproline proteome with more diverse functional approaches, such as mRNA expression level and protein cellular localization. In this way, we may analyze not only a more complete collection of hydroxyproline sites but also integrate functional information in multiple dimensions. Further application of functional annotated hydroxyproline proteome will facilitate the discovery of novel hydroxyproline targets with crucial functional impact in human cell lines and patient samples, which will lead to the development of diagnostic and therapeutic solutions.

4.5 Materials and Methods

4.5.1 Data Collection and Preprocessing for Evolution Conservation Analysis

We obtained MS raw data from a large-scale cross-species proteomics study published in Nature. The collected data consisted of raw files corresponding to 11 animal species, which were sourced from the PRIDE consortium[274]. The MaxQuant software (Version 1.5.3.12) was utilized to analyze the MS data. We performed searches of the raw file sets from each species against the UniProt fasta databases specific to their respective species, as of February 2023. For *Didelphis didelphinae*, we didn’t find a corresponding fasta database, so we used the fasta file from *Monodelphis domestica* from the same sub-family and have over 36,000 entries on Uniprot instead.

The parameter settings employed closely followed those described in the original study[274]. Cysteine carbamidomethylation was set as the fixed modification, while trypsin and LysC were used for digestion enzymes, allowing for a maximum of two missed cleavages. Meanwhile, we set the false discovery rate (FDR) thresholds at 1% for both protein and peptide levels and imposed a minimum peptide length requirement of seven amino acids. Variable modifications considered in the search included proline hydroxylation, N-terminal acetylation, and methionine oxidation. Furthermore, all identified reverse sequences and potential contaminant sites were excluded from subsequent analyses.

To obtain a comprehensive collection of cross-species hydroxyproline sites, we augmented the dataset with additional data sources. Specifically, we integrated hydroxyproline sites identified through mass spectrometry data obtained from an extensive mouse proteome study[282]. MS data fetched from this source were also preprocessed with MaxQuant following similar parameters mentioned above, with the digestion enzyme changed to trypsin only. Additionally, we incorporated hydroxyproline sites from the HypDB[281] database, focusing on the human species.

4.5.2 Mapping, Alignment, and Evolutional Conservation Analysis

To establish a systematic framework for analysis, we mapped the collected proteins with hydroxyproline sites from the 13 species to eggNOG Orthologous Groups (OGs) using the online eggNOG mapper tool (http://eggnog-mapper.embl.de/). This facilitated comparative analysis across species.

To enable sequence alignment, we employed the MUSCLE[283] tool with a custom Python code. We aligned the collected protein sequence belonging to the same metazoan ortholog. Subsequently, we mapped the hydroxyproline sites onto the aligned sequences, allowing us to identify corresponding positions within the alignment.

In the aligned sequences, we evaluated the consistency of proline and hydroxyproline occurrences across different species at the mapped positions. For each metazoan ortholog, we calculated three key parameters: the occurrence of protein (i.e., the number of distinct species in which at least one protein with proline hydroxylation was found), the occurrence of proline (i.e., the number of distinct species in which a proline was present at a given aligned position of at least one protein with hydroxyproline sites within the metazoan orthologs), and the occurrence of hydroxyproline (i.e., the number of distinct species in which at least one hydroxyproline was present at a given aligned position within the metazoan orthologs).

4.5.3 Data Collection and Preprocessing of Protein Turnover Analysis

We obtained MS raw data from a pulse SILAC proteomics study[285] that involved four human cell lines: b-cells, hepatocytes, monocytes, and NK cells. Each cell line was subjected to heavy amino acid culture media, and samples were harvested at four different time points. The MS data from each cell line were processed using MaxQuant software (Ver 1.5.3.12) and searched against an updated human FASTA database provided by UniProt in February 2021. For the MaxQuant search, the multiplicity was set to two, with the heavy group consisting of Arg10 and Lys8. Fixed modifications included Cysteine carbamidomethylation, while variable modifications comprised proline hydroxylation, N-terminal acetylation, and methionine oxidation. Trypsin was used for protein digestion, allowing for a maximum of three missed cleavages. To ensure high-confidence results, the false discovery rate (FDR) for both proteins and peptides was set to 1%, and only peptides with a minimum length of seven amino acids were considered.

4.5.4 Calculation of Half-Lives

To estimate the half-life of proteins and peptides containing hydroxyproline sites, we employed a formula based on the protein decay rate method originally described by Schwanhäusser et al.[287] and modified by Mathieson T et al. [285]. The formula used is as follows:

$$T_{halflife} = \frac{ln2 \sum_{i=1}^{k} \mathrm{t}_i^2}{\sum_{i=1}^{k} \ln\left(R_{hl_i} + 1\right) * \mathrm{t}_i^2}$$

In this formula, t_i represents the time point at which the measurement was taken, and R_{hl_i} denotes the SILAC heavy-to-light ratio at the corresponding time point.

Before calculating SILAC ratios of proteins and peptides using the formula, we performed data filtering steps. Records corresponding to potential contaminants and reverse decoys were removed from the analysis. Additionally, sites with MaxQuant localization scores lower than 0.75 were filtered out. The half-lives of hydroxyproline sites were determined using SILAC ratios obtained from peptides containing the corresponding site. On the other hand, the half-lives of proteins were calculated using the SILAC ratios of the corresponding protein groups. As the half-life distributions were not normal distributions, we performed the Mann-Whitney U test between half-lives of hydroxyproline sites and their corresponding proteins for four cell lines respectively.

4.5.5 Clustering of Turnover Analysis Results

To gain further insights into the hydroxyproline sites, we performed clustering based on the ratio of hydroxyproline peptide half-life to the half-life of the corresponding protein. This analysis allowed us to identify distinct clusters of hydroxyproline sites with varying characteristics. For each cell line, we applied k-means clustering with a cluster size of four using a custom Python code.

The resulting clusters, labeled as Cluster I to IV, were determined based on the ascending order of the half-life ratio. Each cluster represents a group of hydroxyproline sites with similar temporal dynamics. To investigate the functional implications of these clusters, we conducted multiple enrichment analyses on the clustered results.

4.5.6 Melting Point Calculation of Identified Peptides

To determine the melting points of hydroxyproline peptides and protein groups, we employed data obtained from a functional proteoform study[162]. The dataset comprised measurements at eight different temperatures for each B lymphocyte cell line, with samples from two cell lines labeled using a TMT16plex set. Analysis of the collected data was performed using MaxQuant software (Version 2.2.0.0).

For the analysis, we utilized the "Reporter ion MS2" type and selected TMT16plex as the isobaric labeling method. Fixed modifications included carbamidomethylation on cysteine residues, while variable modifications encompassed hydroxylation on prolines, oxidation on methionines, and acetylation on protein N-terminals. Trypsin served as the protease, with a maximum of two missed cleavages allowed. False discovery rate (FDR) thresholds for proteins and peptides were set at 1%, and peptides were required to have a minimum length of seven amino acids.

To calculate the melting points, we first summed up all eight corrected reporter intensities in all spectra corresponding to each peptide identified respectively, where peptides with either modification difference or sequence difference were considered as different peptides. Then we selected the initial valid corrected reporter intensity from the ascending temperature points for each peptide identified. This intensity was designated as the reference with a ratio of one. Subsequently, we normalized the remaining valid corrected reporter intensities by dividing them by the reference intensity, thereby obtaining ratios corresponding to different temperatures for each hydroxyproline site or protein group. Ratio series with fewer than six valid values from different temperatures were excluded from further analysis.

To fit the temperature ratio series, we developed a custom Python code that employed a sigmoid function:

$$f(T) = f_p + \frac{1 - f_p}{1 + \mathrm{e}^{(b - a/T)}}$$

In this function, T represents the temperature variable, while f_p, a, and b are parameters determining the curve's shape[159,286]. We determined the melting point by solving the equation $f(T_m) = 1/2$. Melting curves that lacked a valid fit, did not yield a valid melting point solution or had a melting point beyond the temperature range of the experiments were excluded from the exported data.

For peptides with hydroxyproline sites that possessed a valid melting point, we conducted a one-sample t-test to compare this melting temperature with all other valid melting temperatures of the same protein. The p-value significance threshold on the one-sample t-test was set to 0.01 for sorting hydroxyproline peptides with significant melting temperate changes.

4.5.7 Functional Enrichment Analysis

We performed functional enrichment analysis on two online platforms, Webgesgalt[288] and Metascape[289]. Both online tools were used to perform enrichment analysis, including Gene Ontology (GO) biological process, GO molecular function, GO cellular component, KEGG pathway, CORUM complex, and protein-protein interaction. Analyzed results were downloaded from their website, containing both tables with enrichment statistics and figures generated corresponding to the results.

4.5.8 Motif Enrichment Analysis

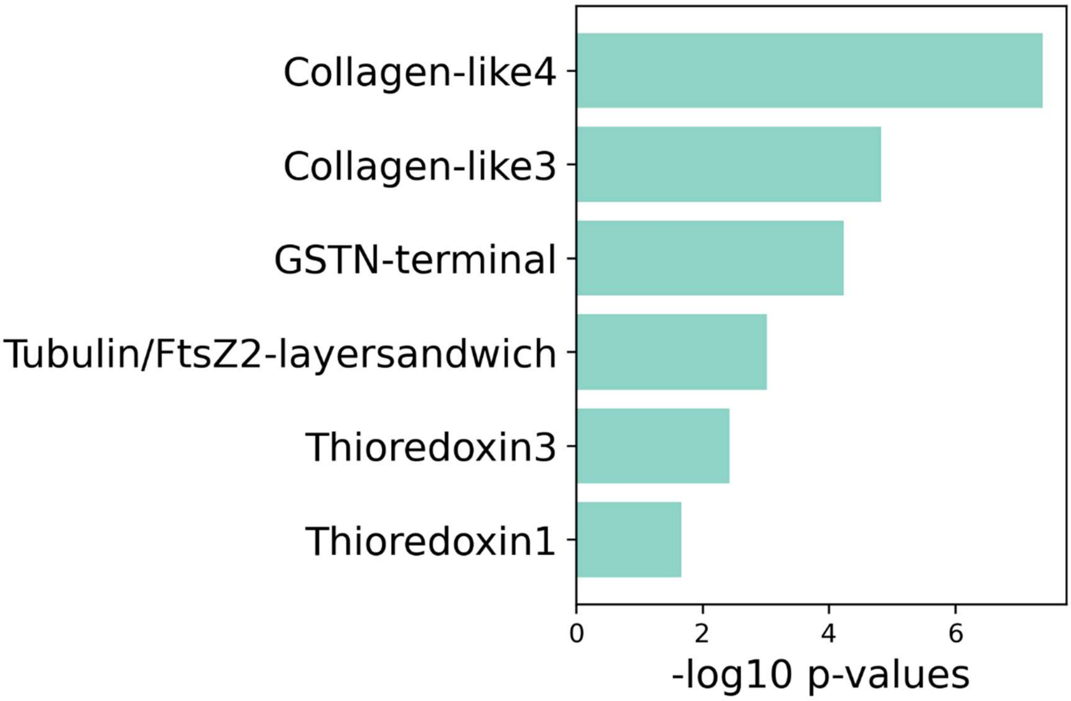
For a group of functionally important sites that we were interested in the motif of the neighboring sequence of the site, we performed motif analysis with homemade Python code and the MoMo tool[251] within the online MeMe suite[290]. Firstly, a +/-6 neighbor sequence list was extracted from interesting sites with homemade Python code, based on proteins and positions provided and proteins sequence fetched from UniProt API. Then the neighbor sequence list was submitted to MoMo online tool using the fasta of whole HypDB human sequences as context sequences in motif-x mode, and compressed analyzed results in figures and tables can be downloaded from the webpage when the job is done by the online server.

4.5.9 Region and Domain Enrichment Analysis

We gathered information related to regions and domains in the UniProt database for the protein list we were interested in. Then given the range of region and domain, we calculate the number of hydroxyproline sites and proline sites regardless of their modification state that resided in each type of region and domain respectively. Then we performed a hypergeometry test with the two numbers above, total hydroxyproline sites in the protein list and total proline sites in the protein list for each region and domain.

4.6 Acknowledgements

The authors thank Gaurav Behera and other members of the Chen lab for helpful suggestions and discussions.



Supp Figure 4-1 Top enriched domains of evolutional conserved sites

Chapter V Conclusion and Future Perspectives

Yao Gong wrote this chapter

5.1 Integrating Functional Information with PTM Identifications: Method Development and Applications

This work addresses the challenge of exploring and interpreting the incompletely annotated human proteome. By combining PTM identification and functional information, we not only identified and quantified diverse proteoforms with different modifications but also provided unbiased functional annotation from multiple perspectives.

Firstly, we developed UbE3-APA, a software that utilizes the enzyme-substrate interaction network to analyze the quantitative ubiquitylome. This approach revealed changes in E3 ligase activities and relationships under different environmental conditions.

Secondly, in the HypDB project, we collected, analyzed, and curated human proteome MS data from various sources to create a comprehensive collection of hydroxyproline sites. We integrated site identification, quantification, and functional annotation databases to reveal a global functional landscape of the hydroxyproline proteome across multiple levels, including site, protein, and tissue.

Thirdly, we gathered MS data from functional proteomics studies, which inherently contain functional information. We identified and quantified hydroxyproline sites, integrated functional annotation databases, and incorporated credibility profiles from HypDB to create detailed functional profiles for individual hydroxyproline proteomes. Integrated analysis across different functional studies highlighted hydroxyproline sites with functional changes in multiple studies, such as position 230, 235, and 238 on PRDX3.

The functional annotation approaches implemented in these projects have significantly advanced the field of functional proteomics and deepened our understanding of the complex molecular function network influenced by diverse PTMs. Furthermore, the integration of additional approaches and the incorporation of functional proteomes in future studies will continue to enhance our understanding of functional PTM proteomics. Collectively, these methods and accumulated information offer valuable insights into the impact of PTMs, particularly those within the dark proteome, on various biological processes within the intricate landscape of the human proteome.

5.2 Building an Integrated Platform and Website for Multi-dimensional PTM Functional Annotations and Future Expansion

Establishing a comprehensive database that consolidates annotation knowledge from various approaches is of paramount importance. Furthermore, it is crucial to create a platform for sharing novel discoveries with the scientific community and obtaining valuable feedback. With these principles in mind, we developed the HypDB platform to facilitate hydroxyproline research, a PTM known for its crucial roles in oxygen sensing and cellular structure but lacking a centralized database. Leveraging the sites we identified, verified, and quantified through spectra analysis, and integrating the results of diverse functional annotations, we constructed the HypDB platform, accessible through a user-friendly webpage supported by a MySQL database. To validate the utility of HypDB, we conducted analyses of DIA MS data from previous cancer proteomics studies, comparing the HypDB spectral library to the original spectral library obtained from DDA data of the same study. Additionally, the platform serves as a means to gather new datasets and continually enrich our collection. Our goal is not only to include sites from current workflows but also to incorporate results from uncovered species and functional approaches. This contains functional proteomics methods we have analyzed, including evolutional conservation across species, protein turnover rate with and without modification and thermal stability changes upon PTMs, and others that may deepen our understanding of the functional proteome. By expanding the number of collected sites, the number of covered species, and incorporating different dimensions of functional information, HypDB will serve as a comprehensive framework for interpreting the functions of both studied and previously unexplored hydroxyproline sites, thereby advancing the field related to hydroxyproline research.

5.3 Multidimensional Functional Proteomics Data: Unveiling the Complexities of Protein Interactions and PTM Functions

Proteins exhibit dynamic and intricate interactions, operating within a high-dimensional space that makes it challenging to comprehensively interpret their individual functions or pathways. To overcome this challenge, our research integrated functional information from various dimensions across all three projects. In the UbE3-APA project, we combined gene mutation data with the E3 ligase-substrate network. In the HypDB project, we incorporated stoichiometries of hydroxyproline sites, tissue specificity, and functional annotation databases. Additionally, we integrated results from three distinct types of functional proteomics studies. These multidimensional analyses provided a clearer understanding of the functional profiles of PTMs while narrowing down the number of PTM sites that may be functionally important in specific pathways or conditions.

The rapid expansion of proteomics data, coupled with the need for a global perspective on functional proteomics, suggests that integrating multiple dimensions of functional proteomics data is crucial. Existing projects like Galaxy-P have successfully integrated RNA expression levels, protein-protein interaction data, and proteomics analysis[291–293]. Furthermore, advancements such as AlphaFold, a model bridging the gap between protein sequence, structure, and function, showed promising potential[147,148,150]. When combined with emerging deep learning algorithms, these approaches may ultimately elucidate the entire functional landscape of the human proteome. They can explain regulatory pathways, uncover interaction residues, and reveal the enzymatic activity mechanisms of proteins with different PTMs. Such advancements will greatly contribute to genome engineering, enzyme discovery, drug development, and precision therapeutics.

In summary, by integrating multiple dimensions of functional proteomics information, identifying and quantifying PTMs, and employing computational models, we can significantly deepen our understanding of both well-studied and under-studied proteins and PTMs. These advancements will ultimately elevate the overall level of human health and well-being.

Bibliography

1. Nørregaard Jensen O. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol*. 2004;8(1):33-41. doi:https://doi.org/10.1016/j.cbpa.2003.12.009

2. Ayoubi TA, Van De Ven WJ. Regulation of gene expression by alternative promoters. *FASEB J Off Publ Fed Am Soc Exp Biol*. 1996;10(4):453-460.

3. Walsh CT, Garneau-Tsodikova S, Gatto Jr GJ. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chemie Int Ed*. 2005;44(45):7342-7372.

4. Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*. 2011;1. doi:10.1038/srep00090

5. Bakri Y, Sarrazin S, Mayer UP, et al. Balance of MafB and PU. 1 specifies alternative macrophage or dendritic cell fate. *Blood*. 2005;105(7):2707-2716.

6. Duan G, Walther D. The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput Biol*. 2015;11(2):e1004049.

7. IHARA Y, NUKINA N, MIURA R, OGAWARA M. Phosphorylated tau protein is integrated into paired helical filaments in Alzheimer's disease. *J Biochem*. 1986;99(6):1807-1810.

8. Grimsrud PA, Xie H, Griffin TJ, Bernlohr DA. Oxidative stress and covalent modification of protein with bioactive aldehydes. *J Biol Chem*. 2008;283(32):21837-21841. doi:10.1074/jbc.R700019200

9. Levene PA, Alsberg CL. The cleavage products of vitellin. *J Biol Chem*. 1906;2(1):127-133.

10. Vickery HB, Schmidt CLA. The History of the Discovery of the Amino Acids. *Chem Rev*. 1931;9(2):169-318.

11. Lipmann FA, Levene PA. Serinephosphoric acid obtained on hydrolysis of vitellinic acid. *J Biol Chem*. 1932;98(1):109-114.

12. Keenan EK, Zachman DK, Hirschey MD. Discovering the landscape of protein modifications. *Mol Cell*. 2021;81(9):1868-1878. doi:https://doi.org/10.1016/j.molcel.2021.03.015

13. Bergström S, Lindstedt S. On the isolation and structure of hydroxylysine. *Acta Chem Scand*. 1951;5:157-167.

14. Dixon GH, Dreyer WJ, Neurath H. The reaction of p-nitrophenyl acetate with chymotrypsin1. *J Am Chem Soc*. 1956;78(18):4810.

15. Rabinowitz M, Lipmann F. Reversible phosphate transfer between yolk phosphoprotein and adenosine triphosphate. *J Biol Chem*. 1960;235(4):1043-1050.

16. Wold F. In vivo chemical modification of proteins (post-translational modification). *Annu Rev Biochem*. 1981;50(1):783-814.

17. Uy R, Wold F. Posttranslational Covalent Modification of Proteins: Only 20 amino acids are used in. protein synthesis, yet some 140" amino acids" are found in various proteins. *Science (80-)*. 1977;198(4320):890-896.

18. Dole M, Mack LL, Hines RL, Mobley RC, Ferguson LD, Alice MB. Molecular beams of macroions. *J Chem Phys*. 1968;49(5):2240-2249.

19. Yamashita M, Fenn JB. Electrospray ion source. Another variation on the free-jet theme. *J Phys*

Chem. 1984;88(20):4451-4459.

20. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-)*. 1989;246(4926):64-71.

21. Tanaka K, Waki H, Ido Y, et al. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun mass Spectrom*. 1988;2(8):151-153.

22. Biemann K. Mass spectrometry of peptides and proteins. *Annu Rev Biochem*. 1992;61(1):977-1010.

23. Sutherland EW, Wosilait WD. Inactivation and activation of liver phosphorylase. *Nature*. 1955;175:169-170.

24. Krebs EG, Fischer EH. The phosphorylase b to a converting enzyme of rabbit skeletal muscle. *Biochim Biophys Acta*. 1956;20:150-157.

25. Allfrey VG, Faulkner R, Mirsky A. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci*. 1964;51(5):786-794.

26. Bensimon A, Heck AJR, Aebersold R. Mass spectrometry–based proteomics and network biology. *Annu Rev Biochem*. 2012;81:379-405.

27. Gouw JW, Krijgsveld J, Heck AJR. Quantitative proteomics by metabolic labeling of model organisms. *Mol Cell proteomics*. 2010;9(1):11-24.

28. Mann M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol cell Biol*. 2006;7(12):952-958.

29. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*. 2012;404:939-965.

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

31. Zhu W, Smith JW, Huang C-M. Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol*. 2009;2010.

32. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell proteomics*. 2014;13(9):2513-2526.

33. Ong S-E, Blagoev B, Kratchmarova I, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell proteomics*. 2002;1(5):376-386.

34. Krijgsveld J, Ketting RF, Mahmoudi T, et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat Biotechnol*. 2003;21(8):927-931.

35. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci*. 1999;96(12):6591-6596.

36. Wu CC, MacCoss MJ, Howell KE, Matthews DE, Yates JR. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal Chem*. 2004;76(17):4951-4959.

37. Miyagi M, Rao KCS. Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass Spectrom Rev*. 2007;26(1):121-136.

38. Stewart II, Thomson T, Figeys D. 18O labeling: a tool for proteomics. *Rapid Commun Mass Spectrom*. 2001;15(24):2456-2465.

39. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*. 1999;17(10):994-999.

40. Li J, Steen H, Gygi SP. Protein Profiling with Cleavable Isotope-coded Affinity Tag (cICAT) Reagents: The Yeast Salinity Stress Response* S. *Mol Cell Proteomics*. 2003;2(11):1198-1204.

41. Hsu J-L, Huang S-Y, Chow N-H, Chen S-H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem*. 2003;75(24):6843-6852.

42. Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc*. 2009;4(4):484-494.

43. Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics*. 2005;5(1):4-15.

44. Rauniyar N, Yates III JR. Isobaric labeling-based relative quantification in shotgun proteomics. *J Proteome Res*. 2014;13(12):5293-5309.

45. Ross PL, Huang YN, Marchese JN, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell proteomics*. 2004;3(12):1154-1169.

46. Choe L, D'Ascenzo M, Relkin NR, et al. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics*. 2007;7(20):3651-3660.

47. McAlister GC, Huttlin EL, Haas W, et al. Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem*. 2012;84(17):7469-7478.

48. Dayon L, Hainard A, Licker V, et al. Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Anal Chem*. 2008;80(8):2921-2931.

49. Thompson A, Schäfer J, Kuhn K, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem*. 2003;75(8):1895-1904.

50. Li J, Cai Z, Bomgarden RD, et al. TMTpro-18plex: the expanded and complete set of TMTpro reagents for sample multiplexing. *J Proteome Res*. 2021;20(5):2964-2972.

51. Zhang J, Wang Y, Li S. Deuterium isobaric amine-reactive tags for quantitative proteomics. *Anal Chem*. 2010;82(18):7588-7595.

52. Frost DC, Greer T, Li L. High-resolution enabled 12-plex DiLeu isobaric tags for quantitative proteomics. *Anal Chem*. 2015;87(3):1646-1654.

53. Frost DC, Feng Y, Li L. 21-plex diLeu isobaric tags for high-throughput quantitative proteomics. *Anal Chem*. 2020;92(12):8228-8234.

54. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. iTRAQ underestimation in simple and complex mixtures:"the good, the bad and the ugly." *J Proteome Res*. 2009;8(11):5347-5355.

55. Karp NA, Huber W, Sadowski PG, Charles PD, Hester S V, Lilley KS. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics*. 2010;9(9):1885-1897.

56. Wu Z, Xiang W, Huang L, Li S, Zhang X. Hyperplexing Approaches for up to 45-Plex Quantitative Proteomic Analysis. *Anal Chem*. 2023;95(12):5169-5175.

57. Neilson KA, Ali NA, Muralidharan S, et al. Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics*. 2011;11(4):535-553.

58. Collier TS, Sarkar P, Franck WL, Rao BM, Dean RA, Muddiman DC. Direct comparison of stable

isotope labeling by amino acids in cell culture and spectral counting for quantitative proteomics. *Anal Chem*. 2010;82(20):8696-8702.

59. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*. 2016;537(7620):347-355.

60. Plank MJ. Modern Data Acquisition Approaches in Proteomics Based on Dynamic Instrument Control. *J Proteome Res*. 2022;21(5):1209-1217. doi:10.1021/acs.jproteome.2c00096

61. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC− MS/MS. *J Proteome Res*. 2011;10(4):1785-1793.

62. Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*. 2018;15:440-448.

63. Stahl-Zeng J, Lange V, Ossola R, et al. High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol Cell Proteomics*. 2007;6(10):1809-1817.

64. van Bentum M, Selbach M. An introduction to advanced targeted acquisition methods. *Mol Cell Proteomics*. 2021;20.

65. Schmidt A, Claassen M, Aebersold R. Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr Opin Chem Biol*. 2009;13(5-6):510-517.

66. Rudomin EL, Carr SA, Jaffe JD. Directed Sample Interrogation Utilizing an A ccurate M ass Ex clusion-Based Data-Dependent Acquisition Strategy (AMEx). *J Proteome Res*. 2009;8(6):3154-3160.

67. Swaney DL, McAlister GC, Coon JJ. Decision tree–driven tandem mass spectrometry for shotgun proteomics. *Nat Methods*. 2008;5(11):959-964.

68. Webber JT, Askenazi M, Ficarro SB, Iglehart MA, Marto JA. Library dependent LC-MS/MS acquisition via mz API/L ive. *Proteomics*. 2013;13(9):1412-1416.

69. Venable JD, Dong M-Q, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods*. 2004;1(1):39-45. doi:10.1038/nmeth705

70. Gillet LC, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics*. 2012;11(6):O111.016717. doi:10.1074/mcp.O111.016717

71. Zhang F, Ge W, Ruan G, Cai X, Guo T. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. *Proteomics*. 2020;20(17-18):e1900276. doi:10.1002/pmic.201900276

72. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol*. 2018;14(8):e8126. doi:10.15252/msb.20178126

73. Purvine S, Eppel J-T, Yi EC, Goodlett DR. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics*. 2003;3(6):847-850. doi:10.1002/pmic.200300362

74. Carvalho PC, Han X, Xu T, et al. XDIA: improving on the label-free data-independent analysis. *Bioinformatics*. 2010;26(6):847-848. doi:10.1093/bioinformatics/btq031

75. Plumb RS, Johnson KA, Rainville P, et al. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom*. 2006;20(13):1989-1994. doi:10.1002/rcm.2550

76. Bond NJ, Shliaha P V, Lilley KS, Gatto L. Improving qualitative and quantitative performance for MS(E)-based label-free proteomics. *J Proteome Res*. 2013;12(6):2340-2353. doi:10.1021/pr300776t

77. Panchaud A, Scherl A, Shaffer SA, et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem*. 2009;81(15):6481-6488. doi:10.1021/ac900888s

78. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012;366(10):883-892. doi:10.1056/NEJMoa1113205

79. Guo T, Kouvonen P, Koh CC, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med*. 2015;21(4):407-413. doi:10.1038/nm.3807

80. Rosenberger G, Liu Y, Röst HL, et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat Biotechnol*. 2017;35(8):781-788. doi:10.1038/nbt.3908

81. Egertson JD, Kuehn A, Merrihew GE, et al. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods*. 2013;10(8):744-746. doi:10.1038/nmeth.2528

82. Moseley MA, Hughes CJ, Juvvadi PR, et al. Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. *J Proteome Res*. 2018;17(2):770-779. doi:10.1021/acs.jproteome.7b00464

83. Messner CB, Demichev V, Bloomfield N, et al. Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol*. 2021;39(7):846-854.

84. Koopmans F, Ho JTC, Smit AB, Li KW. Comparative analyses of data independent acquisition mass spectrometric approaches: DIA, WiSIM-DIA, and untargeted DIA. *Proteomics*. 2018;18(1):1700304.

85. Cai X, Ge W, Yi X, et al. PulseDIA: in-depth data independent acquisition mass spectrometry using enhanced gas phase fractionation. *bioRxiv*. Published online 2019:787705.

86. Meier F, Brunner A-D, Frank M, et al. diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat Methods*. 2020;17(12):1229-1236.

87. Jones AR, Eisenacher M, Mayer G, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*. 2012;11(7):M111.014381. doi:10.1074/mcp.M111.014381

88. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994;5(11):976-989. doi:10.1016/1044-0305(94)80016-2

89. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551-3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

90. Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. *J Proteome Res*. 2004;3(5):958-964. doi:10.1021/pr0499491

91. Fenyö D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem*. 2003;75(4):768-774. doi:10.1021/ac0258709

92. Verheggen K, Raeder H, Berven FS, Martens L, Barsnes H, Vaudel M. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev*. 2020;39(3):292-306. doi:10.1002/mas.21543

93. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res*. 2011;10(4):1794-1805.

doi:10.1021/pr101065j

94. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26(12):1367-1372. doi:10.1038/nbt.1511

95. Tabb DL. The SEQUEST family tree. *J Am Soc Mass Spectrom*. 2015;26(11):1814-1819.

96. Lin A, Howbert JJ, Noble WS. Combining high-resolution and exact calibration to boost statistical power: A well-calibrated score function for high-resolution ms2 data. *J Proteome Res*. 2018;17(11):3644-3656.

97. Wan Y, Yang A, Chen T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *Anal Chem*. 2006;78(2):432-437.

98. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Proteome Bioinforma*. Published online 2010:55-71.

99. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4(3):207-214.

100. Gonnelli G, Stock M, Verwaeren J, et al. A decoy-free approach to the identification of peptides. *J Proteome Res*. 2015;14(4):1792-1798.

101. Kim H, Lee S, Park H. Target-small decoy search strategy for false discovery rate estimation. *BMC Bioinformatics*. 2019;20(1):1-6.

102. Gao J, Sheng X, Du J, et al. Identification of 113 new histone marks by CHiMA, a tailored database search strategy. *Sci Adv*. 2023;9(14):eadf1416. doi:10.1126/sciadv.adf1416

103. Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics*. 2006;5(5):935-948.

104. Perchey RT, Tonini L, Tosolini M, et al. PTMselect: Optimization of protein modifications discovery by mass spectrometry. *Sci Rep*. 2019;9(1):4181.

105. Li Q, Shortreed MR, Wenger CD, et al. Global post-translational modification discovery. *J Proteome Res*. 2017;16(4):1383-1390.

106. Han X, He L, Xin L, Shan B, Ma B. PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications. *J Proteome Res*. 2011;10(7):2930-2936.

107. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014;11(11):1114-1125.

108. Solntsev SK, Shortreed MR, Frey BL, Smith LM. Enhanced global post-translational modification discovery with MetaMorpheus. *J Proteome Res*. 2018;17(5):1844-1851.

109. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. Published online 2008. doi:10.1038/nbt.1511

110. Khan Z, Bloom JS, Garcia BA, Singh M, Kruglyak L. Protein quantification across hundreds of experimental conditions. *Proc Natl Acad Sci*. 2009;106(37):15544-15548.

111. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol*. 2001;19(10):946-951.

112. Shadforth IP, Dunkley TPJ, Lilley KS, Bessant C. i-Tracker: For quantitative proteomics using iTRAQ^{TM}. *BMC Genomics*. 2005;6(1):1-6.

113. Arntzen MØ, Koehler CJ, Barsnes H, Berven FS, Treumann A, Thiede B. IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT. *J Proteome Res*. 2011;10(2):913-920.

114. Pedrioli PGA, Raught B, Zhang X-D, et al. Automated identification of SUMOylation sites using mass spectrometry and SUMmOn pattern recognition software. *Nat Methods*. 2006;3(7):533-539.

115. Deutsch EW, Mendoza L, Shteynberg D, et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*. 2010;10(6):1150-1159.

116. MacLean B, Tomazela DM, Shulman N, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010;26(7):966-968.

117. Sturm M, Bertsch A, Gröpl C, et al. OpenMS–an open-source software framework for mass spectrometry. *BMC Bioinformatics*. 2008;9(1):1-11.

118. Leung K, Lescuyer P, Campbell J, et al. A novel strategy using MASCOT Distiller for analysis of cleavable isotope-coded affinity tag data to quantify protein changes in plasma. *Proteomics*. 2005;5(12):3040-3044.

119. Monroe ME, Tolić N, Jaitly N, Shaw JL, Adkins JN, Smith RD. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics*. 2007;23(15):2021-2023.

120. Ishihama Y, Oda Y, Tabata T, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein* s. *Mol Cell Proteomics*. 2005;4(9):1265-1272.

121. Mallick P, Schirle M, Chen SS, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol*. 2007;25(1):125-131.

122. Olsen J V, Vermeulen M, Santamaria A, et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal*. 2010;3(104):ra3-ra3.

123. Wu R, Haas W, Dephoure N, et al. A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat Methods*. 2011;8(8):677-683. doi:10.1038/nmeth.1636

124. Hornbeck P V, Kornhauser JM, Latham V, et al. 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res*. 2019;47(D1):D433-D441. doi:10.1093/nar/gky1159

125. Nühse TS, Bottrill AR, Jones AME, Peck SC. Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. *Plant J*. 2007;51(5):931-940.

126. Mayya V, Lundgren DH, Hwang S-I, et al. Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Signal*. 2009;2(84):ra46-ra46.

127. Wang Z, Ma J, Miyoshi C, et al. Quantitative phosphoproteomic analysis of the molecular substrates of sleep need. *Nature*. 2018;558(7710):435-439. doi:10.1038/s41586-018-0218-8

128. Nukarinen E, Nägele T, Pedrotti L, et al. Quantitative phosphoproteomics reveals the role of the AMPK plant ortholog SnRK1 as a metabolic master regulator under energy deprivation. *Sci Rep*. 2016;6(1):31697. doi:10.1038/srep31697

129. Robitaille AM, Christen S, Shimobayashi M, et al. Quantitative Phosphoproteomics Reveal mTORC1 Activates de Novo Pyrimidine Synthesis. *Science (80-)*. 2013;339(6125):1320-1323. doi:10.1126/science.1228771

130. Huang H, Lin S, Garcia BA, Zhao Y. Quantitative proteomic analysis of histone modifications. *Chem Rev*. 2015;115(6):2376-2418.

131. Kim SC, Sprung R, Chen Y, et al. Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Mol Cell*. 2006;23(4):607-618.

132. Choudhary C, Weinert BT, Nishida Y, Verdin E, Mann M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat Rev Mol cell Biol*. 2014;15(8):536-550.

133. Dolgin E. The most popular genes in the human genome. *Nature*. 2017;551(7681):427-432.

134. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep*. 2018;8(1):1362.

135. Wood V, Lock A, Harris MA, Rutherford K, Bähler J, Oliver SG. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol*. 2019;9(2):180241.

136. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol*. 2018;16(9):e2006643.

137. Oprea TI, Bologa CG, Brunak S, et al. Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov*. 2018;17(5):317-332.

138. Sinha S, Eisenhaber B, Jensen LJ, Kalbuaji B, Eisenhaber F. Darkness in the human gene and protein function space: widely modest or absent illumination by the life science literature and the trend for fewer protein function discoveries since 2000. *Proteomics*. 2018;18(21-22):1800093.

139. Kustatscher G, Collins T, Gingras A-C, et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods*. 2022;19(7):774-779. doi:10.1038/s41592-022-01454-x

140. Dunham I. Human genes: Time to follow the roads less traveled? *PLoS Biol*. 2018;16(9):e3000034.

141. Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science (80-)*. 2009;325(5948):1682-1686.

142. Beltrao P, Trinidad JC, Fiedler D, et al. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol*. 2009;7(6):e1000134.

143. Beltrao P, Albanèse V, Kenner LR, et al. Systematic functional prioritization of protein posttranslational modifications. *Cell*. Published online 2012. doi:10.1016/j.cell.2012.05.036

144. Nguyen Ba AN, Moses AM. Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol*. 2010;27(9):2027-2037.

145. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. *Trends Genet*. 2009;25(5):193-197.

146. Tan CSH, Bodenmiller B, Pasculescu A, et al. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal*. 2009;2(81):ra39-ra39.

147. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706-710. doi:10.1038/s41586-019-1923-7

148. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

149. Bagdonas H, Fogarty CA, Fadda E, Agirre J. The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat Struct Mol Biol*. 2021;28(11):869-870.

150. Bludau I, Willems S, Zeng W-F, et al. The structural context of posttranslational modifications at a proteome-wide scale. *PLOS Biol*. 2022;20(5):e3001636. https://doi.org/10.1371/journal.pbio.3001636

151. Huttlin EL, Bruckner RJ, Navarrete-Perea J, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*. 2021;184(11):3022-3040.e28. doi:10.1016/J.CELL.2021.04.011

152. Schweppe DK, Huttlin EL, Harper JW, Gygi SP. BioPlex Display: An Interactive Suite for Large-Scale AP–MS Protein–Protein Interaction Data. *J Proteome Res*. 2018;17(1):722-726. doi:10.1021/acs.jproteome.7b00572

153. Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30(1):187-200. doi:10.1002/pro.3978

154. Doherty MK, Hammond DE, Clague MJ, Gaskell SJ, Beynon RJ. Turnover of the Human Proteome: Determination of Protein Intracellular Stability by Dynamic SILAC. *J Proteome Res*. 2009;8(1):104-112. doi:10.1021/pr800641v

155. Wilkinson DJ. Historical and contemporary stable isotope tracer approaches to studying mammalian protein metabolism. *Mass Spectrom Rev*. 2018;37(1):57-80.

156. Liu Y, Mi Y, Mueller T, et al. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat Biotechnol*. 2019;37(3):314-322.

157. Wu C, Ba Q, Lu D, et al. Global and Site-Specific Effect of Phosphorylation on Protein Turnover. *Dev Cell*. 2021;56(1):111-124.e6. doi:https://doi.org/10.1016/j.devcel.2020.10.025

158. Wu Z-H, Zhong Y, Zhou T, Xiao H-J. miRNA biomarkers for predicting overall survival outcomes for head and neck squamous cell carcinoma. *Genomics*. 2021;113(1 Pt 1):135-141. doi:10.1016/j.ygeno.2020.12.002

159. Savitski MM, Reinhard FBM, Franken H, et al. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science (80-)*. 2014;346(6205):1255784. doi:10.1126/science.1255784

160. Sridharan S, Kurzawa N, Werner T, et al. Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nat Commun*. 2019;10(1):1155. doi:10.1038/s41467-019-09107-y

161. Huang JX, Lee G, Cavanaugh KE, Chang JW, Gardel ML, Moellering RE. High throughput discovery of functional protein modifications by Hotspot Thermal Profiling. *Nat Methods*. 2019;16(9):894-901. doi:10.1038/s41592-019-0499-3

162. Kurzawa N, Leo IR, Stahl M, et al. Deep thermal profiling for detection of functional proteoform groups. *Nat Chem Biol*. Published online 2023. doi:10.1038/s41589-023-01284-8

163. Pickart CM. Mechanisms underlying ubiquitination. *Annu Rev Biochem*. 2001;70(1):503-533.

164. Lee DH, Goldberg AL. Proteasome inhibitors: valuable new tools for cell biologists. *Trends Cell Biol*. 1998;8(10):397-403. doi:10.1016/S0962-8924(98)01346-4

165. Schnell JD, Hicke L. Non-traditional functions of ubiquitin and ubiquitin-binding proteins. *J Biol Chem*. 2003;278(38):35857-35860. doi:10.1074/jbc.R300018200

166. Ordureau A, Münch C, Harper JW. Quantifying Ubiquitin Signaling. *Mol Cell*. 2015;58(4):660.

doi:10.1016/J.MOLCEL.2015.02.020

167. Hoeller D, Dikic I. Targeting the ubiquitin system in cancer therapy. *Nat 2009 4587237*. 2009;458(7237):438-444. doi:10.1038/nature07960

168. McNaught KSP, Olanow CW, Halliwell B, Isacson O, Jenner P. Failure of the ubiquitin–proteasome system in Parkinson's disease. *Nat Rev Neurosci 2001 28*. 2001;2(8):589-594. doi:10.1038/35086067

169. Oddo S. The ubiquitin-proteasome system in Alzheimer's disease. *J Cell Mol Med*. 2008;12(2):363-373. doi:10.1111/J.1582-4934.2008.00276.X

170. Herrmann J, Ciechanover A, Lerman LO, Lerman A. The ubiquitin–proteasome system in cardiovascular diseases—a hypothesis extended. *Cardiovasc Res*. 2004;61(1):11-21. doi:10.1016/J.CARDIORES.2003.09.033

171. Hoeller D, Hecker C-M, Dikic I. Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. *Nat Rev Cancer 2006 610*. 2006;6(10):776-788. doi:10.1038/nrc1994

172. Nakayama KI, Nakayama K. Ubiquitin ligases: cell-cycle control and cancer. *Nat Rev Cancer 2006 65*. 2006;6(5):369-381. doi:10.1038/nrc1881

173. Bulatov E, Ciulli A. Targeting Cullin–RING E3 ubiquitin ligases for drug discovery: structure, assembly and small-molecule modulation. *Biochem J*. 2015;467(3):365-386. doi:10.1042/BJ20141450

174. Petroski MD. The ubiquitin system, disease, and drug discovery. *BMC Biochem 2008 91*. 2008;9(1):1-15. doi:10.1186/1471-2091-9-S1-S7

175. Olsen J V., Mann M. Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry *. *Mol Cell Proteomics*. 2013;12(12):3444-3452. doi:10.1074/MCP.O113.034181

176. Krug K, Mertins P, Zhang B, et al. A Curated Resource for Phosphosite-specific Signature Analysis. *Mol Cell Proteomics*. 2019;18(3):576-593. doi:10.1074/mcp.TIR118.000943

177. Mischnik M, Sacco F, Cox J, et al. IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics*. 2016;32(3):424-431. doi:10.1093/BIOINFORMATICS/BTV699

178. Yang P, Patrick E, Humphrey SJ, et al. KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics*. 2016;16(13):1868-1871. doi:10.1002/PMIC.201600068

179. Wiredja DD, Koyutürk M, Chance MR. The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics*. 2017;33(21):3489-3491. doi:10.1093/BIOINFORMATICS/BTX415

180. Kuleshov M V, Xie Z, London ABK, et al. KEA3: improved kinase enrichment analysis via data integration. *Nucleic Acids Res*. 2021;49(W1):W304-W316. doi:10.1093/NAR/GKAB359

181. Elia AEH, Boardman AP, Wang DC, et al. Quantitative Proteomic Atlas of Ubiquitination and Acetylation in the DNA Damage Response. *Mol Cell*. 2015;59(5):867-881. doi:10.1016/J.MOLCEL.2015.05.006

182. Kim W, Bennett EJ, Huttlin EL, et al. Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome. *Mol Cell*. 2011;44(2):325-340. doi:10.1016/J.MOLCEL.2011.08.025

183. Udeshi ND, Svinkina T, Mertins P, et al. Refined Preparation and Use of Anti-diglycine Remnant (K-ε-GG) Antibody Enables Routine Quantification of 10,000s of Ubiquitination Sites in Single Proteomics Experiments *. *Mol Cell Proteomics*. 2013;12(3):825-831.

doi:10.1074/MCP.O112.027094

184. Wagner SA, Beli P, Weinert BT, et al. A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles *. *Mol Cell Proteomics*. 2011;10(10):M111.013284. doi:10.1074/MCP.M111.013284

185. Li Y, Evers J, Luo A, Erber L, Postler Z, Chen Y. A Quantitative Chemical Proteomics Approach for Site-specific Stoichiometry Analysis of Ubiquitination. *Angew Chemie*. 2019;131(2):547-551. doi:10.1002/ange.201810569

186. Chen D, Liu X, Xia T, et al. A Multidimensional Characterization of E3 Ubiquitin Ligase and Substrate Interaction Network. *iScience*. 2019;16:177-191. doi:10.1016/j.isci.2019.05.033

187. Du Y, Xu N, Lu M, Li T. hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. *Database J Biol Databases Curation*. 2011;2011. doi:10.1093/DATABASE/BAR055

188. Han Y, Lee H, Park JC, Yi G-S. E3Net: A System for Exploring E3-mediated Regulatory Networks of Cellular Functions. *Mol Cell Proteomics*. 2012;11(4):1-14. doi:10.1074/MCP.O111.014076

189. Li Y, Xie P, Lu L, et al. An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. *Nat Commun*. 2017;8(1):1-9. doi:10.1038/s41467-017-00299-9

190. Li Z, Chen S, Jhong J-H, et al. UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions. *Database*. 2021;2021. doi:10.1093/database/baab010

191. Sarraf SA, Raman M, Guarani-Pereira V, et al. Landscape of the PARKIN-dependent ubiquitylome in response to mitochondrial depolarization. *Nature*. 2013;496(7445):372-376. doi:10.1038/nature12043

192. Theurillat JPP, Udeshi ND, Errington WJ, et al. Ubiquitylome analysis identifies dysregulation of effector substrates in SPOP-mutant prostate cancer. *Science (80-)*. 2014;346(6205):85-89. doi:10.1126/science.1250255

193. Tanaka A, Cleland MM, Xu S, et al. Proteasome and p97 mediate mitophagy and degradation of mitofusins induced by Parkin. *J Cell Biol*. 2010;191(7):1367. doi:10.1083/JCB.201007013

194. Koyano F, Yamano K, Kosako H, et al. Parkin-mediated ubiquitylation redistributes MITOL/March5 from mitochondria to peroxisomes. *EMBO Rep*. 2019;20(12):e47728. doi:10.15252/EMBR.201947728

195. Hansen FM, Tanzer MC, Brüning F, et al. Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *Nat Commun 2021 121*. 2021;12(1):1-13. doi:10.1038/s41467-020-20509-1

196. Bradley JR, Pober JS. Tumor necrosis factor receptor-associated factors (TRAFs). *Oncogene*. 2001;20(44):6482-6491.

197. Erber LN, Luo A, Gong Y, et al. Iron Deficiency Reprograms Phosphorylation Signaling and Reduces O-GlcNAc Pathways in Neuronal Cells. *Nutrients*. 2021;13(1):1-18. doi:10.3390/NU13010179

198. Hornbeck P V., Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*. 2015;43(Database issue):D512. doi:10.1093/NAR/GKU1267

199. Nguyen V-N, Huang K-Y, Weng JT-Y, Lai KR, Lee T-Y. UbiNet: an online resource for exploring the functional associations and regulatory networks of protein ubiquitylation. *Database J Biol Databases Curation*. 2016;2016:54. doi:10.1093/DATABASE/BAW054

200. Bateman A, Martin MJ, O'Donovan C, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45(D1):D158-D169. doi:10.1093/NAR/GKW1099

201. Chatr-Aryamontri A, Oughtred R, Boucher L, et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res*. 2017;45(D1):D369-D379. doi:10.1093/NAR/GKW1102

202. Steger M, Demichev V, Backman M, et al. Time-resolved in vivo ubiquitinome profiling by DIA-MS reveals USP7 targets on a proteome-wide scale. *Nat Commun*. 2021;12(1):5399.

203. Gorres KL, Raines RT. Prolyl 4-hydroxylase. *Crit Rev Biochem Mol Biol*. 2010;45(2):106-124. doi:10.3109/10409231003627991

204. Ivan M, Kaelin WG. The EGLN-HIF O2-sensing system: multiple inputs and feedbacks. *Mol Cell*. 2017;66(6):772-779.

205. Semenza GL. Oxygen sensing, hypoxia-inducible factors, and disease pathophysiology. *Annu Rev Pathol Mech Dis*. 2014;9:47-71.

206. Ratcliffe PJ. Oxygen sensing and hypoxia signalling pathways in animals: the implications of physiology for cancer. *J Physiol*. 2013;591(8):2027-2042.

207. PETERKOFSKY B, UDENFRIEND S. ENZYMATIC HYDROXYLATION OF PROLINE IN MICROSOMAL POLYPEPTIDE LEADING TO FORMATION OF COLLAGEN. *Proc Natl Acad Sci U S A*. 1965;53(2):335-342. doi:10.1073/pnas.53.2.335

208. Kivirikko KI, Prockop DJ. Enzymatic hydroxylation of proline and lysine in protocollagen. *Proc Natl Acad Sci U S A*. 1967;57(3):782-789. doi:10.1073/pnas.57.3.782

209. Halme J, Kivirikko KI, Simons K. Isolation and partial characterization of highly purified protocollagen proline hydroxylase. *Biochim Biophys Acta (BBA)-Enzymology*. 1970;198(3):460-470.

210. Yu F, White SB, Zhao Q, Lee FS. HIF-1alpha binding to VHL is regulated by stimulus-sensitive proline hydroxylation. *Proc Natl Acad Sci U S A*. 2001;98(17):9630-9635. doi:10.1073/pnas.181341498

211. Jaakkola P, Mole DR, Tian YM, et al. Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science*. 2001;292(5516):468-472. doi:10.1126/science.1059796

212. Ivan M, Kondo K, Yang H, et al. HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. *Science*. 2001;292(5516):464-468. doi:10.1126/science.1059817

213. Epstein AC, Gleadle JM, McNeill LA, et al. C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell*. 2001;107(1):43-54. doi:10.1016/s0092-8674(01)00507-4

214. Bruick RK, McKnight SL. A conserved family of prolyl-4-hydroxylases that modify HIF. *Science*. 2001;294(5545):1337-1340. doi:10.1126/science.1066373

215. Friedman L, Higgin JJ, Moulder G, Barstead R, Raines RT, Kimble J. Prolyl 4-hydroxylase is required for viability and morphogenesis in Caenorhabditis elegans. *Proc Natl Acad Sci*. 2000;97(9):4736-4741.

216. Holster T, Pakkanen O, Soininen R, et al. Loss of assembly of the main basement membrane collagen, type IV, but not fibril-forming collagens and embryonic death in collagen prolyl 4-hydroxylase I null mice. *J Biol Chem*. 2007;282(4):2512-2519. doi:10.1074/jbc.M606608200

217. Ishikawa Y, Bächinger HP. A molecular ensemble in the rER for procollagen maturation. *Biochim*

Biophys Acta (BBA)-Molecular Cell Res. 2013;1833(11):2479-2491.

218. Semenza GL. Hypoxia-inducible factor 1: oxygen homeostasis and disease pathophysiology. *Trends Mol Med*. 2001;7(8):345-350.

219. Semenza GL, Wang GL. A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. *Mol Cell Biol*. Published online 1992.

220. Wang GL, Jiang B-H, Rue EA, Semenza GL. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension. *Proc Natl Acad Sci*. 1995;92(12):5510-5514.

221. Ohh M, Park CW, Ivan M, et al. Ubiquitination of hypoxia-inducible factor requires direct binding to the β-domain of the von Hippel–Lindau protein. *Nat Cell Biol*. 2000;2(7):423-427.

222. Maxwell PH, Wiesener MS, Chang G-W, et al. The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature*. 1999;399(6733):271-275.

223. Wenger RH, Stiehl DP, Camenisch G. Integration of oxygen signaling at the consensus HRE. *Sci STKE*. 2005;2005(306):re12-re12.

224. Weidemann A, Johnson RS. Biology of HIF-1α. *Cell Death Differ*. 2008;15(4):621-627.

225. Dengler VL, Galbraith MD, Espinosa JM. Transcriptional regulation by hypoxia inducible factors. *Crit Rev Biochem Mol Biol*. 2014;49(1):1-15.

226. Wong BW, Kuchnio A, Bruning U, Carmeliet P. Emerging novel functions of the oxygen-sensing prolyl hydroxylase domain enzymes. *Trends Biochem Sci*. 2013;38(1):3-11. doi:10.1016/j.tibs.2012.10.004

227. Zhang J, Zhang Q. VHL and hypoxia signaling: beyond HIF in cancer. *Biomedicines*. 2018;6(1):35.

228. Rodriguez J, Pilkington R, Garcia Munoz A, et al. Substrate-Trapped Interactors of PHD3 and FIH Cluster in Distinct Signaling Pathways. *Cell Rep*. 2016;14(11):2745-2760. doi:10.1016/j.celrep.2016.02.043

229. Zhou T, Erber L, Liu B, Gao Y, Ruan H-B, Chen Y. Proteomic analysis reveals diverse proline hydroxylation-mediated oxygen-sensing cellular pathways in cancer cells. *Oncotarget*. 2016;7(48):79154-79169. doi:10.18632/oncotarget.12632

230. Arsenault PR, Heaton-Johnson KJ, Li L-S, et al. Identification of prolyl hydroxylation modifications in mammalian cell proteins. *Proteomics*. 2015;15(7):1259-1267. doi:10.1002/pmic.201400398

231. Stoehr A, Yang Y, Patel S, et al. Prolyl hydroxylation regulates protein degradation, synthesis, and splicing in human induced pluripotent stem cell-derived cardiomyocytes. *Cardiovasc Res*. 2016;110(3):346-358. doi:10.1093/cvr/cvw081

232. German NJ, Yoon H, Yusuf RZ, et al. PHD3 Loss in Cancer Enables Metabolic Reliance on Fatty Acid Oxidation via Deactivation of ACC2. *Mol Cell*. 2016;63(6):1006-1020. doi:10.1016/j.molcel.2016.08.014

233. Lee DC, Sohn HA, Park Z-Y, et al. A lactate-induced response to hypoxia. *Cell*. 2015;161(3):595-609. doi:10.1016/j.cell.2015.03.011

234. Heir P, Srikumar T, Bikopoulos G, et al. Oxygen-dependent Regulation of Erythropoietin Receptor Turnover and Signaling. *J Biol Chem*. 2016;291(14):7357-7372. doi:10.1074/jbc.M115.694562

235. Liu X, Simon JM, Xie H, et al. Genome-wide Screening Identifies SFMBT1 as an Oncogenic Driver in Cancer with VHL Loss. *Mol Cell*. 2020;77(6):1294-1306.e5. doi:10.1016/j.molcel.2020.01.009

236. Casciello F, Al-Ejeh F, Kelly G, et al. G9a drives hypoxia-mediated gene repression for breast cancer cell survival and tumorigenesis. *Proc Natl Acad Sci U S A*. 2017;114(27):7077-7082. doi:10.1073/pnas.1618706114

237. Rodriguez J, Herrero A, Li S, et al. PHD3 Regulates p53 Protein Stability by Hydroxylating Proline 359. *Cell Rep*. 2018;24(5):1316-1329. doi:10.1016/j.celrep.2018.06.108

238. Zheng X, Zhai B, Koivunen P, et al. Prolyl hydroxylation by EglN2 destabilizes FOXO3a by blocking its interaction with the USP9x deubiquitinase. *Genes Dev*. 2014;28(13):1429-1444. doi:10.1101/gad.242131.114

239. Qi HH, Ongusaha PP, Myllyharju J, et al. Prolyl 4-hydroxylation regulates Argonaute 2 stability. *Nature*. 2008;455(7211):421-424. doi:10.1038/nature07186

240. Jiang W, Zhou X, Li Z, et al. Prolyl 4-hydroxylase 2 promotes B-cell lymphoma progression via hydroxylation of Carabin. *Blood*. 2018;131(12):1325-1336. doi:10.1182/blood-2017-07-794875

241. Guo J, Chakraborty AA, Liu P, et al. pVHL suppresses kinase activity of Akt in a proline-hydroxylation-dependent manner. *Science*. 2016;353(6302):929-932. doi:10.1126/science.aad5755

242. Luo W, Hu H, Chang R, et al. Pyruvate kinase M2 is a PHD3-stimulated coactivator for hypoxia-inducible factor 1. *Cell*. 2011;145(5):732-744. doi:10.1016/j.cell.2011.03.054

243. Kuznetsova A V, Meller J, Schnell PO, et al. von Hippel-Lindau protein binds hyperphosphorylated large subunit of RNA polymerase II through a proline hydroxylation motif and targets it for ubiquitination. *Proc Natl Acad Sci U S A*. 2003;100(5):2706-2711. doi:10.1073/pnas.0436037100

244. Mikhaylova O, Ignacak ML, Barankiewicz TJ, et al. The von Hippel-Lindau tumor suppressor protein and Egl-9-Type proline hydroxylases regulate the large subunit of RNA polymerase II in response to oxidative stress. *Mol Cell Biol*. 2008;28(8):2701-2717. doi:10.1128/MCB.01231-07

245. Hu L, Xie H, Liu X, et al. TBK1 Is a Synthetic Lethal Target in Cancer with VHL Loss. *Cancer Discov*. 2020;10(3):460-475. doi:10.1158/2159-8290.CD-19-0837

246. Erber L, Luo A, Chen Y. Targeted and Interactome Proteomics Revealed the Role of PHD2 in Regulating BRD4 Proline Hydroxylation. *Mol Cell Proteomics*. 2019;18(9):1772-1781. doi:10.1074/mcp.RA119.001535

247. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47(D1):D529-D541.

248. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolym Orig Res Biomol*. 1983;22(12):2577-2637.

249. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins Struct Funct Bioinforma*. 1994;20(3):216-226.

250. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47(D1):D309-D314.

251. Cheng A, Grant CE, Noble WS, Bailey TL. MoMo: discovery of statistically significant post-translational modification motifs. *Bioinformatics*. 2019;35(16):2774-2782. doi:10.1093/bioinformatics/bty1058

252. Zhou T, Chung Y, Chen J, Chen Y. Site-specific identification of lysine acetylation stoichiometries in mammalian cells. *J Proteome Res*. 2016;15(3):1103-1113.

253. Kitata RB, Choong W-K, Tsai C-F, et al. A data-independent acquisition-based global

phosphoproteomics system enables deep profiling. *Nat Commun*. 2021;12(1):2539.

254. Bekker-Jensen DB, Bernhardt OM, Hogrebe A, et al. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun*. 2020;11(1):787.

255. Liang Y, Xia W, Zhang T, et al. Upregulated collagen COL10A1 remodels the extracellular matrix and promotes malignant progression in lung adenocarcinoma. *Front Oncol*. 2020;10:573534.

256. Robinson AD, Chakravarthi BVSK, Agarwal S, et al. Collagen modifying enzyme P4HA1 is overexpressed and plays a role in lung adenocarcinoma. *Transl Oncol*. 2021;14(8):101128. doi:10.1016/j.tranon.2021.101128

257. Kim M-S, Pinto SM, Getnet D, et al. A draft map of the human proteome. *Nature*. 2014;509(7502):575-581.

258. Bekker-Jensen DB, Kelstrup CD, Batth TS, et al. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst*. 2017;4(6):587-599.e4. doi:10.1016/j.cels.2017.05.009

259. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.

260. Consortium TU, Bateman A, Martin M-J, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480-D489. doi:10.1093/NAR/GKAA1100

261. Velankar S, Alhroub Y, Best C, et al. PDBe: protein data bank in Europe. *Nucleic Acids Res*. 2012;40(D1):D445-D452.

262. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412-D419.

263. Luo A, Chen Y. Label-Free Interactome Analysis Revealed an Essential Role of CUL3-KEAP1 Complex in Mediating the Ubiquitination and Degradation of PHD2. *J Proteome Res*. 2020;19(1):260-268. doi:10.1021/acs.jproteome.9b00513

264. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods*. 2020;17(1):41-44.

265. Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat Methods*. 2016;13(9):731-740.

266. Wang S, Li W, Hu L, Cheng J, Yang H, Liu Y. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res*. 2020;48(14):e83-e83.

267. Dayhoff MO, Eck R V. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.; 1972.

268. Dayhoff M, Schwartz R, Orcutt B. 22 a model of evolutionary change in proteins. *Atlas protein Seq Struct*. 1978;5:345-352.

269. Valencia A, Chardin P, Wittinghofer A, Sander C. The ras protein family: evolutionary tree and role of conserved amino acids. *Biochemistry*. 1991;30(19):4637-4648.

270. Rojas AM, Fuentes G, Rausell A, Valencia A. The Ras protein superfamily: evolutionary tree and role of conserved amino acids. *J Cell Biol*. 2012;196(2):189-201.

271. Goh C-S, Cohen FE. Co-evolutionary analysis reveals insights into protein–protein interactions. *J Mol Biol*. 2002;324(1):177-192.

272. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng*. 2001;14(9):609-614.

273. Beltrao P, Albanèse V, Kenner LR, et al. Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell*. 2012;150(2):413-425. doi:https://doi.org/10.1016/j.cell.2012.05.036

274. Müller JB, Geyer PE, Colaço AR, et al. The proteome landscape of the kingdoms of life. *Nature*. 2020;582(7813):592-596. doi:10.1038/s41586-020-2402-x

275. Schwanhäusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342. doi:10.1038/nature10098

276. Martinez Molina D, Jafari R, Ignatushchenko M, et al. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science*. 2013;341(6141):84-87. doi:10.1126/science.1233606

277. Wenger RH. Cellular adaptation to hypoxia: O2-sensing protein hydroxylases, hypoxia-inducible transcription factors, and O2-regulated gene expression. *FASEB J*. 2002;16(10):1151-1162.

278. Hutton Jr JJ, Kaplan A, Udenfriend S. Conversion of the amino acid sequence gly-pro-pro in protein to gly-pro-hyp by collagen proline hydroxylase. *Arch Biochem Biophys*. 1967;121(2):384-391.

279. Hutton Jr JJ, Tappel AL, Udenfriend S. A rapid assay for collagen proline hydroxylase. *Anal Biochem*. 1966;16(3):384-394.

280. Luo W, Lin B, Wang Y, et al. PHD3-mediated prolyl hydroxylation of nonmuscle actin impairs polymerization and cell motility. *Mol Biol Cell*. 2014;25(18):2788-2796. doi:10.1091/mbc.E14-02-0775

281. Gong Y, Behera G, Erber L, Luo A, Chen Y. HypDB: A functionally annotated web-based database of the proline hydroxylation proteome. *PLOS Biol*. 2022;20(8):e3001757. https://doi.org/10.1371/journal.pbio.3001757

282. Giansanti P, Samaras P, Bian Y, et al. Mass spectrometry-based draft of the mouse proteome. *Nat Methods*. 2022;19(7):803-811. doi:10.1038/s41592-022-01526-y

283. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340

284. Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30(1):187. doi:10.1002/PRO.3978

285. Mathieson T, Franken H, Kosinski J, et al. Systematic analysis of protein turnover in primary cells. *Nat Commun*. 2018;9(1):689. doi:10.1038/s41467-018-03106-1

286. Childs D, Bach K, Franken H, et al. Nonparametric Analysis of Thermal Proteome Profiles Reveals Novel Drug-binding Proteins. *Mol Cell Proteomics*. 2019;18(12):2506-2515. doi:10.1074/mcp.TIR119.001481

287. Schwanhüusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342. doi:10.1038/nature10098

288. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199-W205. doi:10.1093/nar/gkz401

289. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523. doi:10.1038/s41467-019-09234-6

290. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Res*. 2015;43(W1):W39-W49. doi:10.1093/nar/gkv416

291. Sheynkman GM, Johnson JE, Jagtap PD, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics*. 2014;15:1-9.

292. Stewart PA, Kuenzi BM, Mehta S, et al. The Galaxy platform for reproducible affinity proteomic mass spectrometry data analysis. *Mass Spectrom Proteins Methods Protoc*. Published online 2019:249-261.

293. Afgan E, Nekrutenko A, Grüning BA, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*. 2022;50(W1):W345-W351.