UNIVERSITY OF LIVERPOOL

Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation and Phosphosite Conservation

A thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Anton Kalyuzhnyy

Institute of Systems, Molecular and Integrative Biology

March 2023

Thesis Abstract

Protein phosphorylation is a fundamental post-translation modification (PTM) that regulates protein function and is well-studied in relation to cell signalling pathways and disease. The development of high-throughput proteomics pipelines such as tandem mass spectrometry has led to the discovery of large numbers of specific phosphorylated protein motifs and sites, focussing primarily on the phosphorylation of serine, threonine and tyrosine amino acids. However, there is no database-level control for the false discovery of sites, likely leading to the overestimation of true phosphosites reported in phosphorylation resources. In addition, the vast majority of phosphosite discoveries are made in humans, with many other species only having a few reported phosphosites. Furthermore, only a small fraction of the currently characterised human phosphoproteome has an annotated functional role and the studies focusing on predicting the functional relevance of phosphosites on a large scale using techniques such as conservation analysis are scarce. As a result, this Thesis profiled the human phosphoproteome to estimate the true extent of protein phosphorylation and understand the evolutionary and functional trends of phosphosites.

First, in Chapter 2, we developed and validated an accessible Python pipeline which can determine the conservation of specific amino acid sites such as PTMs and perform several steps of a typical conservation analysis in a single step. In particular, for each query protein sequence, the pipeline identifies its likely homologous sequences from the selected species using the BLAST algorithm, generates multiple sequence alignments and calculates the conservation of target amino acid sites. In Chapter 3, we profiled the human phosphoproteome and developed a method of independent phosphosite FDR estimation in large datasets. We ranked all reported human phosphosites into sets according to the amount of identification evidence they had in public databases and analysed the sets in terms of conservation across 100 species, sequence properties and functional annotations. We demonstrated significant differences between the sets and estimated that around 62,000 Ser, 8,000 Thr and 12,000 Tyr phosphosites in the human proteome were likely to be true, which is lower than most published estimates. Furthermore, our analysis estimated that 86,000 Ser, 50,000 Thr and 26,000 Tyr phosphosites were likely false positive identifications, highlighting the significant potential of false positive data in phosphorylation databases. In Chapter 4, we analysed the evolutionary conservation of human phosphosites across different groups of eukaryotic species and linked their conservation patterns to diverse protein functions. Finally, we applied the conservation analysis to predict over 1,000,000 potential phosphosites in eukaryotes by using confident human phosphosites as a reference set. Our results highlighted the relevance of conservation analysis in studying phosphosites and can ultimately be used to improve proteome annotations of several species.

In Memory of

Marina Kalyuzhnaya, my Mum

and

Valentina Zaytseva, my Grandmother,

This Was Always Our Dream

Acknowledgements

I would first like to express my sincere gratitude to my primary supervisor Prof Andy Jones who had faith in me from the start and provided expert guidance throughout this PhD project. I am also thankful for his emotional support and understanding during some of the most difficult times of my life. Andy has always inspired me to become a better academic and provided many opportunities for me to learn new skills, attend exciting events and work on interesting research projects. I really could not have asked for a better mentor for this journey.

I would also like to thank my secondary supervisor Prof Claire Eyers and Dr Eric Deutsch for contributing to this research and introducing me to several exciting projects which broadened my research perspectives. I am extending my gratitude to my assessors, Prof Sonia Roche, Prof Rob Beynon and Dr Niall Kenneth, for reviewing my progress and offering valuable feedback which significantly improved this PhD research.

I am also thankful for all the support I received from the staff at Computational Biology Facility. In addition, this research would not have been possible without the financial help from the University of Liverpool which generously offered to waive the tuition fees for me.

I must thank my lovely fiancée Rebecca Owens for her never-ending support and patience throughout this journey. Finally, I am grateful for all the help I received from my family. In particular, I would like to thank my sister, Yana Kalyuzhnaya and my dad, German Kalyuzhnyy for their encouragement and love.

Table of Contents

Publications and Conferences

Publications

The most relevant publication mentioning the research described in this Thesis is:

- **Kalyuzhnyy, A**., Eyers, P. A., Eyers, C. E., Bowler-Barnett, E., Martin, M. J., Sun, Z., Deutsch, E. W., Jones, A. R. (2022). Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation. *J Proteome Res.* **21**(6), 1510-1524.

Additional publications contributed to:

- Johnson, P. J., Pinato, D. J., **Kalyuzhnyy, A**., Toyoda, H. (2022). Breaking the Child-Pugh Dogma in Hepatocellular Carcinoma. *J Clin Oncol*., **40**(19):2078-2082.

- Johnson, P. J., Innes, H., Hughes, D. M., **Kalyuzhnyy, A**., Kumada, T., Toyoda, H. (2022). Evaluation of the aMAP score for hepatocellular carcinoma surveillance: a realistic opportunity to risk stratify. *Br J Cancer*. **127**(7):1263-1269.

- Campbell, A. E., Ferraz Franco, C., Su, L. I., Corbin, E. K., Perkins, S., **Kalyuzhnyy, A**., Jones, A. R., Brownridge, P. J., Perkins, N. D., Eyers, C. E. (2021). Temporal modulation of the NF-κB RelA network in response to different types of DNA damage. *Biochem J*., **478**(3):533-551.

- Byrne, D. P., Clarke, C. J., Brownridge, P. J., **Kalyuzhnyy, A**., Perkins, S., Campbell, A., Mason, D., Jones, A. R., Eyers, P. A., Eyers, C. E. (2020). Use of the Polo-like kinase 4 (PLK4) inhibitor centrinone to investigate intracellular signalling networks using SILAC-based phosphoproteomics. *Biochem J*., **477**(13):2451-2475.

- Hardman, G., Perkins, S., Brownridge, P. J., Clarke, C. J., Byrne, D. P., Campbell, A. E., **Kalyuzhnyy, A**., Myall, A., Eyers, P. A., Jones, A. R., Eyers, C. E. (2019). Strong anion exchange-mediated phosphoproteomics reveals extensive human non-canonical phosphorylation. *EMBO J*., **38**(21):e100847.

Conferences

- Poster presentation at HUPO ReCONNECT 2021 Online Conference, November 15th-19th: "*Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation*" by **Kalyuzhnyy, A**., Eyers, P. A., Eyers, C. E., Sun, Z., Deutsch, E. W. & Jones, A. R.

Key Abbreviations

ACES	Analysis of Conservation with an Extensive List of Species
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
CID	Collision-Induced Dissociation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DNA	Deoxyribonucleic Acid
ETD	Electron Transfer Dissociation
FDR	False Discovery Rate
FLR	False Localisation Rate
GH	Growth Hormone
GO	Gene Ontology
HPRD	Human Protein Reference Database
HTP	High-Throughput
IDE	Integrated Development Environment
IMAC	Immobilised Metal Ion Affinity Chromatography
LC-MS/MS	Liquid Chromatography-Tandem Mass Spectrometry
MEGA	Molecular Evolutionary Genetic Analysis
MS	Mass Spectrometry
MSA	Multiple Sequence Alignment
PA	PeptideAtlas
PDB	Protein Data Bank
PSM	Peptide Spectrum Match
PSP	PhosphoSitePlus
PTM	Post-Translational Modification
RBH	Reciprocal Best Hits
SDS-PAGE	Sodium Dodecyl Sulfate–Polyacrylamide Gel Electrophoresis
Ser	Serine
Thr	Threonine
TP	True Positive
TPP	Trans-Proteomic Pipeline
Tyr	Tyrosine

Supplementary Information

All the supplementary information (SI) relating to supplementary figures S1-S7 and Table S3 is summarised in the **SI file** "*AntonK_Supplementary_Information.pdf*". The remaining supplementary tables can be accessed at https://figshare.com/s/12ca201b9d99d4092713

Chapter 1

Introduction and Thesis Aims

1.1 Abstract

Proteins are complex molecular structures which have essential roles across all forms of life. Therefore, understanding the structure and function of proteins is an integral part of any biological research. This introductory Chapter begins by providing an overview of protein structure to highlight key structural elements involved in protein folding which is an essential first step in facilitating protein function. The Chapter then summarises various biological functions regulated by proteins, such as providing structural support to cells, tissues and organs, transporting molecules, facilitating immune response, coordinating signal transduction and catalysing various chemical reactions such as phosphorylation, which is an important post-translational modification of proteins. The exact mechanism and the functional role of phosphorylation is then explored in more detail. In addition, this Chapter describes how novel phosphorylation peptides and sites are discovered in proteomics studies using traditional biochemical approaches and more recent high-throughput techniques based on mass spectrometry (MS). Several important statistical methods applied in the analysis of MS-generated phosphoproteomics data to control the rate of false positive identifications are also discussed before providing an overview of key bioinformatics resources that store phosphorylation data. After that, the Chapter highlights how evolutionary conservation analysis can be applied to study the functional relevance and evolution of novel phosphosites. Several crucial steps of a typical conservation analysis are also discussed, such as homologue sequence searching with BLAST and generating multiple sequence alignments to identify conserved protein regions. Finally, the Chapter introduces the basic principle of a functional enrichment analysis used to study protein function and concludes by outlining the main aims of this Thesis.

1.2 The Fundamentals of Protein Structure

Proteins are complex molecular structures which have essential roles across all forms of life. They are synthesised in a process known as translation which takes place in cell's ribosomes. Proteins are incredibly diverse and perform a variety of important functions depending on their structure[1]. The structure of proteins is made up of four key levels. The first level is the protein's primary structure which is a chain of different amino acids joined together by peptide bonds. There are 20 different amino acids which can make up the primary structure. The overall amino acid sequence of the polypeptide chain is encoded by its underlying genetic code, or deoxyribonucleic acid (DNA), and it ultimately determines the protein's folding properties, function and molecular interactions[2]. In particular, a protein's primary structure contains specific short and recognisable amino acid sequences known as motifs which are characteristic of protein's structural or functional properties[2]. These sequences are often conserved between species or proteins from the same protein family that share a common evolutionary ancestor[2]. The secondary structure of a protein is formed when the amino acid chain folds into regular repeated patterns through the formation of hydrogen bonds between the backbone chemical groups of the chain (carbonyl and amide groups)[2]. This ensures structural stability and maintains a protein's folded state needed for its function[3]. The most common secondary protein structures are alpha helices and beta sheets. They are eventually folded further into a more complex, three-dimensional tertiary structure which represents the overall arrangement of the protein's polypeptide chain[2]. The formation of a stable tertiary structure under an organism's physiological conditions is critical for a protein to be able to perform its biological function as it creates a surface with specific physical and chemical properties required for molecular interactions[4]. It is also possible for multiple folded polypeptide chains to bond together by association through chemical interactions such as hydrogen bonds, disulphide bonds and van der Waals forces to create a protein's oligomeric (i.e., containing multiple units) quaternary structure, although not all proteins have this structural level[2, 4]. An example of a protein with a quaternary structure is haemoglobin which is made up of two alpha-globin and two beta-globin polypeptide chains that play a role in oxygen transport[5]. Additional complex structural protein motifs can be formed, which often contribute to protein's function and mechanical stability. For example, coiled coils are combinations of multiple alpha helices wrapped around each other as a coiled structure involved in facilitating protein-protein interactions and providing structural support to the cell[6]. Finally, regions of protein's polypeptide chains can fold independently from the rest of the protein's structure to form

compact and stable units known as protein domains which can contribute to protein's overall function[2]. An example of a protein domain is SH3 which is made of 50-60 amino acids and is found in many proteins involved in cell signalling[7]. The presence or absence of specific protein domains can be used by researchers to classify proteins into families and infer their evolutionary relationship[8].

The structure of proteins can be determined and analysed using a variety of experimental techniques including traditional approaches such as x-ray crystallography[9] and nuclear magnetic resonance spectroscopy[10], as well as more recent techniques based on mass spectrometry[11]. Newly established protein structures, their annotations and underlying amino acid sequences are stored in several publicly accessible bioinformatics resources. For example, Protein Data Bank (PDB) provides tools for the analysis and visualisation of experimentally determined, three-dimensional protein structures and contains over 200,000 entries[12]. Another important computational database is UniProt which is a comprehensive protein resource that contains data for over 220 million protein sequences from a wide range of species and offers high-quality, manually curated structural and functional sequence annotations[13]. For each protein entry, UniProt provides tools to explore its sequence, domains and families, subcellular location, post-translational modifications (PTMs), evolutionary relationships and its functional significance in biological processes and disease[13]. In addition, UniProt provides links to relevant literature and other bioinformatics resources including PDB to allow further analysis of target proteins, making it a valuable resource for protein analysis[13].

1.3 Functional Significance of Proteins

The final folded structure of a protein defines how it interacts with other molecules and this interaction forms the basis of how proteins function. In fact, most proteins function by binding to another molecule, known as a ligand, using a complementary binding site made of a specific amino acid arrangement which is brought together during protein folding[8]. Examples of ligands include nucleic acids, small molecules (adenosine triphosphate (ATP), amino acids, lipids, carbohydrates, etc.), ions and other proteins[8]. Ligand specificity and affinity varies between proteins and depends on their structure, function and metabolic conditions, with some proteins also having multiple binding sites which allows them to interact with different ligands simultaneously[8].

In general, proteins are involved in the structural and functional maintenance of organism's cells, tissues and organs by performing a variety of important roles. For example, the

importance of proteins with structural roles can be highlighted by actin proteins which bind together to form microfilament networks that make up a part of a cell's cytoskeleton and therefore provide structural support to the cell, establish its shape and regulate cell division and movement[14]. In addition, actin can interact with filaments made of myosin motor proteins to assist the development of muscle structure and allow muscle contraction[15]. Furthermore, a group of structural proteins known as intermediate filaments (for example, keratin and vimentin), which are typically thicker than microfilaments, provide additional mechanical strength and support within a cell's cytoskeleton and are prevalent in tissues affected by physical stress such as bone and epidermis[16, 17].

Proteins can also facilitate the transport of molecules and ions to regulate homeostasis (maintenance of stable internal conditions) and assist nutrient uptake, toxin removal and cell signalling[8]. For example, ABC transporter proteins are membrane-bound, ATP-powered pumps which transport a variety of molecules across cell membranes including metal ions, amino acids, carbohydrates, lipids and toxins[8, 18]. In addition, proteins called aquaporins are channel-forming structures which facilitate transmembrane diffusion of water and ensure tissue hydration, waste removal and nutrient uptake across all forms of life[19]. Some proteins act as hormones which are secreted chemical messengers that coordinate molecular processes by binding to specific cell surface receptors and ultimately eliciting a physiological response[20]. For example, insulin and glucagon are protein hormones which are secreted by the pancreas to regulate carbohydrate metabolism[20]. Another example of a protein hormone is growth hormone (GH) which plays a key role in regulating body growth and development in humans and other animals. In particular, GH promotes tissue and organ growth while also regulating nutrient metabolism[20]. Proteins also play an important role in the immune system. Antibodies are proteins with high ligand specificity, which are produced by the immune system in response to an infection by foreign molecules. They work by recognising and binding to its target molecule called an antigen and the formation of this antigen-antibody complex either marks the foreign molecule for destruction by other cells or directly neutralises it[8].

Finally, proteins can act as biological catalysts known as enzymes, the role of which is to speed up and organise chemical reactions within various essential biological pathways involved in functions such as cell development, metabolism, protein synthesis and many others[8]. In any chemical reaction catalysed by an enzyme, the binding of the involved enzyme to one or more of its specific ligand molecules, also known as substrates, modifies the molecules which lowers the activation energy needed for the reaction to occur, ultimately accelerating its speed[21]. In

general, enzymes act by transferring a functional group from one molecule onto another, changing the shape of a substrate or repositioning a substrate to optimise its functional interactions[21]. The overall rate of a chemical reaction catalysed by an enzyme depends on substrate and enzyme concentrations, as well as temperature and pH levels, with different enzymes having their own optimal levels[21]. Enzymes can be classified according to similarities between chemical reactions that they catalyse. For example, hydrolase enzymes act by catalysing the hydrolysis of covalent bonds (i.e., by adding water molecules to break the bonds) between amino acids or nucleic acids[8]. This reaction is useful in various biological processes such as ATP hydrolysis which releases energy and regulates metabolism[22], or protein breakdown which plays a key role in digestion[23] and repair mechanisms[24]. Another major class of enzymes called kinases modify proteins by catalysing the addition of a phosphate chemical group[8]. This chemical process is known as phosphorylation which is an important post-translational modification (PTM) of proteins that plays a significant role in regulating various signalling pathways involved in cell cycle maintenance, metabolism, immune response and disease progression[25, 26]. In fact, the research described in this Thesis is centred around protein phosphorylation and aims to investigate its scope, functional relevance and evolution.

1.4 Exploring the Mechanism and the Roles of Protein Phosphorylation

The exact mechanism of protein phosphorylation involves the transfer of a phosphate chemical group (PO_4) from an ATP molecule onto a particular amino acid residue in the target protein[26]. The reaction is catalysed by a kinase enzyme which recognises and binds to a specific sequence motif within a target protein that contains the phosphorylated amino acid site. The spontaneous hydrolysis of covalent bonds in ATP which occurs in the presence of water and a kinase enzyme releases the phosphate group (forming ADP) and the energy required to drive phosphorylation. As a result, the kinase enzyme is able to transfer the phosphate onto the target amino acid site to modify the protein[26]. Phosphorylation is a reversible reaction in which enzymes called phosphatases can remove the added phosphate from the target protein by catalysing the hydrolysis of the phosphate bond[26, 27] (Fig. 1). This reversible nature of phosphorylation allows protein function to be readily regulated in response to stimuli[26, 27]. Cells contain many different kinase and phosphatase enzymes responsible for regulating phosphorylation of numerous proteins[8]. For example, several genome sequencing studies identified more than 500 different kinase-encoding genes in humans alone[28]. In fact, it is estimated that in a typical mammalian cell, more than one-third of proteins are phosphorylated at any given time, which further highlights the great extent and the importance of protein phosphorylation[8].

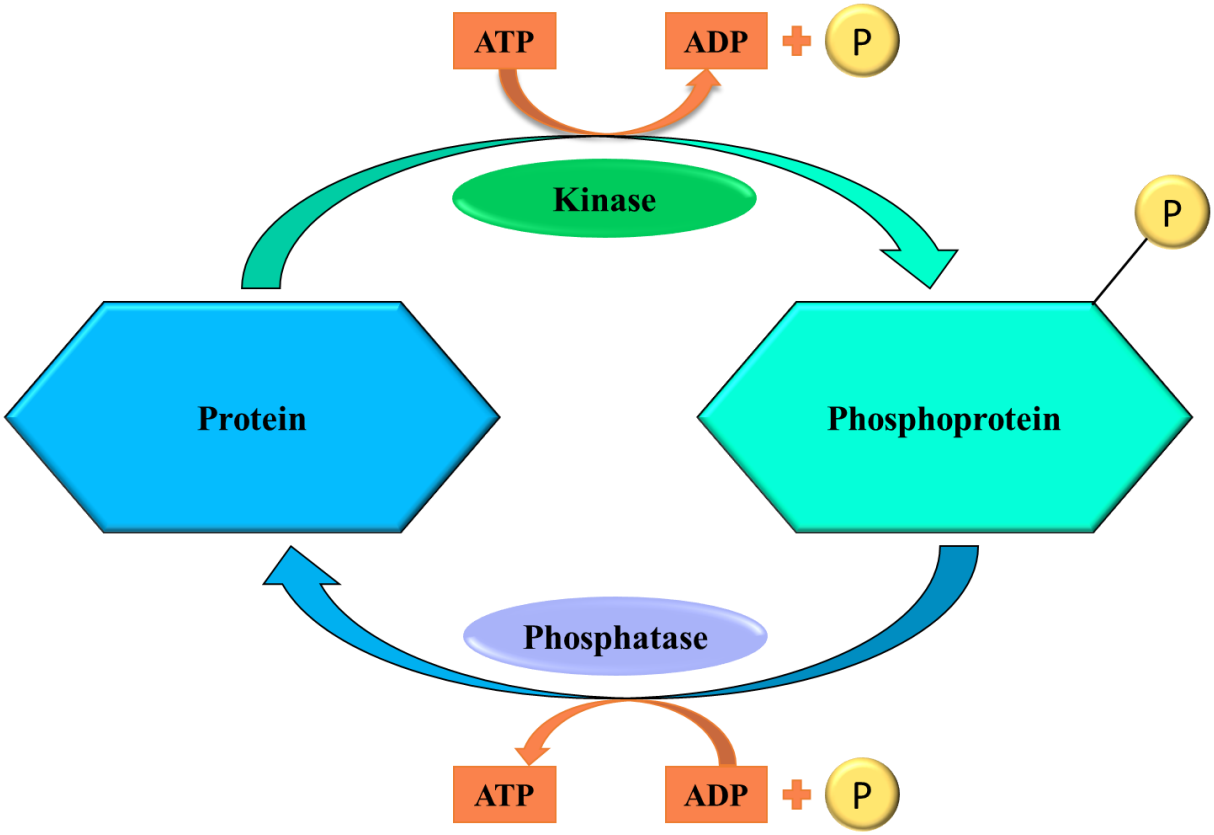


Figure 1. A graphical overview of protein phosphorylation.

The most frequently observed and studied phosphorylatable amino acids are serine (Ser), threonine (Thr) and tyrosine (Tyr), which are polar amino acids phosphorylated on the hydroxyl groups of their side chains through the formation of phosphodiester bonds with phosphate groups[29-33]. In particular, the more abundant Ser and Thr phosphorylation plays a significant role in many conserved biological pathways involved in signal transduction, cell proliferation, metabolism and disease progression. For example, various cell cycle checkpoints including mitosis are regulated by serine/threonine kinase enzymes such as cyclin-dependent kinase (CDK) and polo-like kinase (PLK) which phosphorylate multiple target proteins and promote the activation of downstream molecular pathways involved in chromosome segregation, spindle assembly and cytokinesis[34, 35]. As a result, the incorrect regulation of those enzymes can lead to uncontrollable cell growth and the development of cancer, thus making them promising targets for novel therapeutic drugs[34, 35]. Furthermore, a conserved family of serine/threonine kinases called mitogen-activated protein kinases (MAPK) are involved in complex functional cascades which regulate cell proliferation and differentiation, apoptosis, stress responses and growth factor signalling[36]. Similarly, tyrosine phosphorylation plays an important role in various molecular functions conserved across many species[37]. For

example, tyrosine kinase Src is made of several functional domains including SH2 and SH3 which help regulate the phosphorylation of transcription factors, adaptor proteins and other kinase enzymes involved in cell proliferation and differentiation, neuronal signalling and bone metabolism[38]. In addition, tyrosine phosphorylation is involved in adaptive immune responses in animal organisms, where, for example, an enzyme called spleen tyrosine kinase (SYK) initiates a cascade of signalling pathways which ultimately lead to the activation and proliferation of B cells necessary for the release of antibodies[39]. Protein phosphorylation can also occur on less studied, non-canonical amino acids such as arginine (Arg), lysine (Lys), histidine (His), cysteine (Cys), aspartate (Asp) and glutamine (Glu) which regulate various protein signalling functions in eukaryotic and prokaryotic systems[40]. In fact, a recent large-scale proteomics study discovered many non-canonical phosphorylation sites (phosphosites) in human cells and established their involvement in functions such as transcription and cell signalling, further highlighting the general scope of phosphorylation and its functional significance[41].

In addition to phosphorylation, the regulation of protein function and stability can depend on other PTMs, with most proteins having multiple amino acid sites which can be modified[42, 43]. Moreover, phosphorylation can interplay with other PTMs as part of PTM crosstalk, where the occurrence of one PTM affects the function and localisation of another[44]. For example, the activation of tumour suppressor protein p53 which facilitates cellular response to DNA damage is regulated by phosphorylation on serine residues which in turn increases the affinity of p53 for acetyltransferase enzymes that catalyse acetylation (addition of an acetyl group) of a lysine residue[45]. Furthermore, the activity of histone H3 protein, which is involved in chromatin formation and gene expression, is regulated by multiple PTMs and there is also evidence of crosstalk between phosphorylation and methylation (addition of a methyl group) during mitosis[46].

As highlighted throughout this Thesis, protein phosphorylation is an essential post-translational modification involved in regulating many vital conserved molecular functions. As a result, the studies of phosphoproteomes are frequent and often focus on discovering, quantifying and annotating new phosphorylation sites, motifs and associated kinase enzymes across various species[47].

1.5 Discovering Novel Phosphorylation Sites

The discovery of phosphorylated proteins and their associated phosphosites has been a constantly evolving field in proteomics. One of the earliest analytical techniques used to detect the presence of phosphorylation in a protein sample is based on radioactive labelling of target proteins with ^{32}P (phosphorus-32 isotope), followed by their analysis and visualisation with biochemical techniques such as SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) and autoradiography[48-50]. To increase the accuracy of detecting specific phosphorylation events, it has been possible to incorporate the radiolabelling method with a technique called Edman sequencing which is used to determine specific amino acid sequences of target proteins or peptides and identify phosphorylated amino acids[51, 52]. In addition, phosphorylation motifs and sites can be detected using highly specific antibodies that only bind to target proteins or known sites in their phosphorylated state[53-55]. The use of phosphorylation-specific antibodies has allowed the identification and analysis of many novel Ser and Thr phosphosites[56-58], as well as less commonly studied Tyr phosphosites[59]. Although these techniques can detect phosphorylated protein regions with high accuracy and be used to validate phosphosite predictions from proteomics studies, they are only able to process a small number of protein samples (i.e., low-throughput) and are often time-consuming[60, 61]. With the development of more advanced methods of detecting phosphorylated proteins, the use of low-throughput approaches is now rare as it lacks the depth of coverage needed in modern proteomic workflows.

Several high-throughput (HTP) methods have been developed to increase the scale of phosphosite detection and establish a more comprehensive view of the phosphoproteome[61]. Those methods are primarily based on mass spectrometry (MS) which can be optimised to detect and characterise phosphorylated peptides in a protein sample[60, 61]. At its simplest, MS is a powerful analytical technique which works by identifying and quantifying molecules in complex biological samples such as protein mixtures through the analysis of ions with specific mass-to-charge (m/z) ratios that correspond to those molecules[62, 63]. A significant breakthrough in phosphoproteomics was the development of a technique called liquid chromatography-tandem MS (LC-MS/MS) which is routinely used for the accurate identification and quantification of phosphopeptides with canonical Ser, Thr and Tyr phosphosites[60, 63, 64]. LC-MS/MS includes essential steps during protein sample preparation and analysis which enable phosphopeptide isolation and phosphosite detection (Fig. 2).

A standard workflow of LC/MS-MS begins with sample preparation which includes the extraction of a protein sample from cells or tissues through cell lysis techniques such as homogenisation[65]. The proteins are then purified based on their physical properties such as mass, charge and hydrophobicity using chromatography techniques[66]. After this, the extracted protein mixture is digested into smaller peptides either by protease enzymes such as trypsin or through chemical means such as the addition of formic acid[66]. The next key step which allows to selectively isolate phosphorylated peptides from the overall peptide mixture in the sample is phosphopeptide enrichment[67]. One method of phosphopeptide enrichment is immobilised metal ion affinity chromatography (IMAC) which relies on high binding affinity of certain metal ions towards phosphorylated amino acid residues[68]. An alternative enrichment method involves the use of titanium dioxide (TiO_2) which is more selective than IMAC for binding phosphorylated peptides and has a lower preparation time, making it ideal for the analysis of large and complex protein samples[69]. Furthermore, immunoaffinity-based enrichment using antibodies can be applied in combination with LC-MS/MS to specifically isolate phosphopeptides with tyrosine phosphosites[70, 71]. Before being introduced into the mass spectrometer, the enriched phosphopeptides are separated by their hydrophobicity in solution using reversed-phase liquid chromatography (LC) technique which is directly coupled to the MS instrument[66]. In a typical mass spectrometer, the eluted peptides, starting from the most hydrophilic ones, are first ionised, usually by electron ionisation or electrospray ionisation, and the resulting ions are separated based on their m/z value using an electric or magnetic field produced by a mass analyser[62]. An example of a commonly used mass analyser is Orbitrap in which the ions are orbited around a spindle-like electrode at a very characteristic frequency and this frequency of oscillations is converted into m/z values[72]. The separated ions are then detected and quantified with a detector to establish their abundance[62]. The resulting data from the detector is processed with specialised software such as Proteome Discoverer[73] or MaxQuant[74] to generate a mass spectrum which visualises the abundance of peptide ions relative to their m/z value, where each peak in the spectrum represents an ion with a certain m/z value and the height of the peak corresponds to the abundance of that ion in the sample[62].

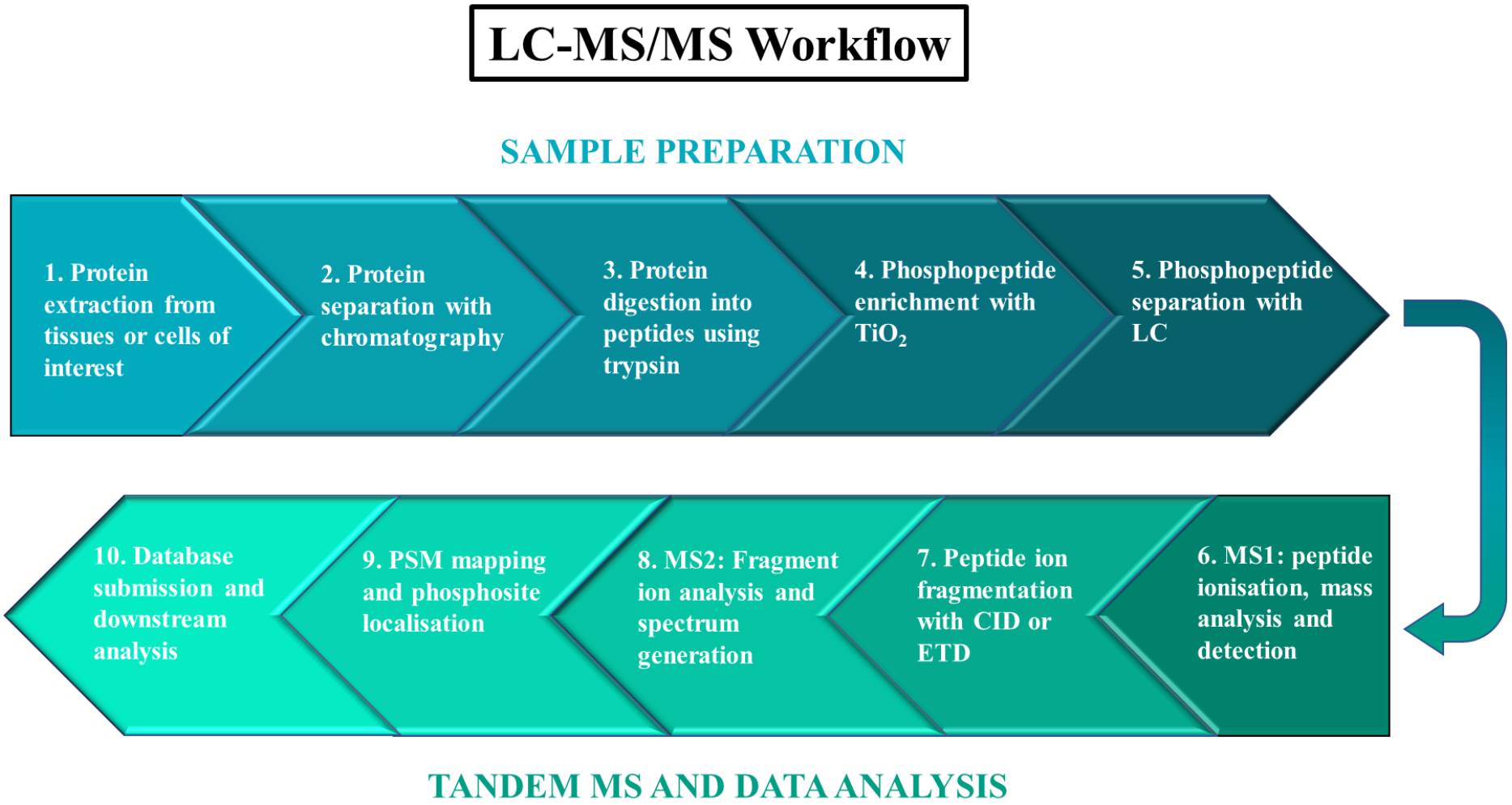


Figure 2. A standard workflow of LC-MS/MS in phosphoproteomics analysis. Several examples of analytical techniques and computational methods involved in sample preparation, MS analysis and data processing are provided.

Peptide ions with the most intense signals from the first MS step are selected and fragmented into smaller ions by techniques such as collision-induced dissociation (CID)[75] and electron transfer dissociation (ETD)[76]. For each selected peptide ion (also referred to as the precursor ion), its resulting set of fragment ions is analysed by the second mass analyser (MS2) to create a fragmentation spectrum which presents their abundance and m/z values within the precursor. After that, the experimental fragmentation spectrum from the MS2 step is searched against a sequence database to identify the most likely corresponding peptide sequences, also known as peptide spectrum matches (PSMs)[67]. To perform the database search, relevant software algorithms first pre-process known protein sequences from the sequence databases such as UniProt[13] and Reference Sequence Database (RefSeq)[77] *in silico*, which involves sequence digestion and fragmentation to predict their perfect theoretical MS2 spectra[78]. The resulting database is searched using the precursor mass to find matching hits with a similar mass. The experimental MS2 spectrum from the selected precursor is then compared against the theoretical ones to obtain a scored candidate list of likely associated PSMs[78]. Several algorithms have been developed to perform database searches and assign PSMs including Mascot[79], Comet[80] and Andromeda[81] which vary in terms of PSM scoring approaches and search parameters[78]. Confidently identified PSMs are analysed with site localisation algorithms to identify potential phosphorylation events on selected amino acids, where each target site is assigned a probability score that indicates the likelihood of that site being phosphorylated[82-84]. In general, to assign a local probability score, site localisation algorithms analyse fragmentation patterns and m/z values of fragment ions, their originating precursor ions and the number of potential phosphosites in a given sequence[82-84]. Examples of site localisation software and algorithms used in the analysis of PSM data include Ascore[82], phosphoRS[85], LuciPHOr[86], Andromeda's PTM Score[81] and the recently developed PTMProphet[87].

The method of fragmenting specific precursor peptide ions in LC-MS/MS as described above is referred to as data-dependent acquisition (DDA)[78]. Although powerful, this method tends to be biased towards selecting precursor ions with the highest peak intensity which limits the analysis and quantification of peptides with lower abundance. To increase the accuracy of peptide quantification, data-independent acquisition (DIA) methods are emerging in proteomics which are based on fragmenting all the eluting peptides within a defined m/z window during the second MS stage[88, 89]. The DIA methods generate larger amounts of data, tend to have higher precision and result in better reproducibility compared to the DDA

methods[88, 89]. However, because the precursor ions are unknown in the DIA approach, the methods used for database searching and estimating false positive matches are different compared to the DDA method. In addition, due to the novelty of the DIA approach, most of the phosphorylation data analysed in this Thesis was acquired using DDA experiments. Therefore, although the DIA approach is an important development in proteomics, it is currently beyond the scope of this research.

1.6 Evaluating the Reliability of Phosphosite Identifications

The development and the use of LC/MS-MS in proteomics studies has led to the discovery of large numbers of phosphorylation motifs and sites[29-33, 90]. However, researchers must be aware that the data resulting from PSM mapping approaches and site localisation algorithms may contain incorrect peptide mappings or site identifications (i.e., false positive identifications). It is important to reduce the number of resulting incorrect phosphosite predictions that end up in downstream analysis as they can lead to inaccurate biological conclusions regarding the relevance of phosphosites in specific signalling pathways or their role in kinase-substrate relationships. Several statistical processes are therefore applied, either within the search engine used for PSM mapping, or in a downstream site localisation software, to calculate additional statistics which reflects the probability that the mapped PSMs or localised phosphosites are true identifications[91-93].

For example, the false discovery rate (FDR) of PSMs reflects the ratio between false positive PSM identifications and the total number of identified PSMs above a certain score threshold[92]. A common target-decoy approach can estimate FDR in a resulting set of mapped peptides following a PSM database search[92]. This approach involves searching an artificial negative control decoy database which contains protein sequences that are similar to the sequences in the actual searched database in terms of their length and amino acid composition[92]. The decoy database is usually generated by randomising or reversing the amino acid sequences from the database used in PSM searching[92, 94]. Assuming that the likelihood of obtaining incorrect PSM mappings is equal between searching the decoy and the real target database, it is possible to estimate FDR by taking into account the number of detected decoy PSMs above a certain score threshold which would infer the number of expected false positive identifications from searching the target database[92]. As a result, the use of FDR and FDR-related statistical scores such as q-value (used for global FDR estimation across all mapped PSMs) and posterior error probability (used to estimate local FDR for each individual PSM) allows researchers to detect

highly confident PSMs and control the rate of false positive PSM mappings in their analysis which is usually set to 1%[95, 96].

When it comes to assessing the reliability of phosphosite localisation in peptide sequences, additional statistical scoring methods are applied by site localisation algorithms to estimate the rate of incorrectly localised phosphosites, also referred to as the false localisation rate (FLR)[93]. Site localisation algorithms vary in terms of mapping their localisation probability scores to FLR estimation, with the overall performance also depending on the fragmentation mode and resolution used in the MS pipeline[97]. For example, PhosphoRS suggests that obtaining a local probability score of 0.99 for a particular PSM would reflect an FLR of 1%[85]. However, it does not predict global phosphosite FLR across all PSMs. In addition, LuciPHOr estimates FLR by using decoy phosphopeptides in which random amino acids are artificially phosphorylated[86]. This creates a decoy fragment ion distribution with random patterns which is then compared to the ion distribution of the candidate peptide and the differences between the two populations are used to estimate FLR, where larger differences indicate lower FLR[86]. Additional methods have been proposed and validated to independently estimate global FLR for site localisation including using decoy amino acids which cannot be phosphorylated in nature such as alanine and leucine[98, 99]. However, there is no general agreement in terms of how FLR should be estimated[93].

To conclude, although LC-MS/MS and computational analysis is generally recognised as very effective and reliable for phosphosite detection, from each study it is likely that there is some element of remaining false discovery of peptides and sites wrongly localised, depending on the statistical thresholds applied[93]. Furthermore, the guidelines for dealing with FLR are still evolving and unfortunately, FLR estimation methods are not consistently applied in reported phosphoproteomics studies that identify phosphorylated peptides and sites. As a result, it is likely that published studies contain considerable numbers of falsely localised phosphosites which can lead to overestimation of the total number of known true phosphosites if database providers do not control for false discovery rate across multiple datasets[100]. The issue of phosphosite FDR across large datasets is explored further in Chapter 3, where the whole human phosphoproteome is profiled to independently estimate the true extent of protein phosphorylation and the proportion of false positive phosphosite identifications.

1.7 Phosphorylation Databases

Following confident identification of phosphopeptides and localisation of target phosphosites in proteomics studies, the resulting data is deposited and organised in relevant phosphorylation databases which can be accessed by researchers to assist their projects. There are currently over 60 available phosphorylation databases which differ in terms of species covered, the sources of phosphorylation data and the number of reported phosphosites[101]. In addition, global databases such as the previously mentioned UniProt provide relevant online tools to promote further exploration of reported phosphorylation data, including phosphosite conservation, predicting kinase-substrate relationships and establishing phosphosite relevance and interactions in protein functions and disease development[13]. UniProt reports phosphorylation data for many species and identifies phosphosites using several methods including manual assertion based on published experimental evidence, inference from sequence similarity evidence obtained using a related experimentally characterised and annotated protein, or the use of phosphorylation data imported from another database[13]. Therefore, UniProt is one of the most accessible resources containing high-quality phosphorylation data.

Another well-cited, comprehensive and publicly available phosphorylation resource is PhosphoSitePlus (PSP) which contains experimental mammalian PTM data, primarily focusing on human phosphorylation[57]. PSP provides structural and functional information about specific modification sites and offers tools for further data interpretation in the context of biological pathway regulation, disease development and cellular localisation of target sites[57]. In addition, a recent update incorporates a powerful tool for predicting and exploring kinase-substrate relationships based on a given peptide sequence[102]. As of February 2023, PSP reported nearly 300,000 Ser, Thr and Tyr phosphosites in over 20,000 non-redundant protein sequences from several species. Phosphorylation data integrated in PSP comes from several sources including manually curated reviews of literature describing high-throughput tandem MS studies and low-throughput experiments, as well as from in-house, unpublished MS studies[57]. However, the vast majority (>95%) of phosphosite identifications in PSP comes from HTP studies alone and many sites were identified in just a single reported PSM[57]. As discussed previously, phosphosite identifications from MS studies are stochastic by nature, especially when phosphosite false discovery rate is not controlled for. As a result, PSP recommends that researchers should be cautious when accepting phosphosites from MS studies alone as true identifications. Nevertheless, PSP datasets are widely used in proteomics studies

and PSP remains one of the most cited phosphorylation resources with its latest release having over 2,000 citations[57].

Large amounts of phosphorylation data can also be found in PeptideAtlas (PA) which is a compendium of phosphopeptide observations and their associated annotations obtained from large tandem MS datasets[103]. To ensure high and consistent quality of reported phosphopeptide identifications, the input MS data in PA is typically processed through the Trans-Proteomic Pipeline (TPP) which is a collection of software tools specifically designed for validating LC-MS/MS data[104]. For example, PeptideProphet is a statistical tool which is used in TPP to validate the results of a search engine search during PSM assignment by applying the previously mentioned target-decoy approach for FDR estimations[94, 105]. PeptideProphet is applied in combination with iProphet which is a tool designed to further validate peptide matches by combining the evidence from multiple assignments of the same PSM by different search engines, as well as collating evidence from multiple PSMs for the same target peptide[106]. Furthermore, the resulting PSMs are statistically mapped to their potential protein candidates using ProteinProphet[107]. The latest PA builds also incorporate the use of the PTMProphet algorithm for phosphosite localisation where each potential phosphosite within an observed PSM is assigned a probability score between 0 and 1 of being phosphorylated[87]. Phosphorylation evidence from PSP and PA is further evaluated in Chapter 3, where FDR is independently estimated for the sets of human Ser, Thr and Tyr sites with varying amounts of cumulative evidence available in those resources.

Additional phosphorylation databases provide peptide and site identifications for specific sets of species. For example, Human Protein Reference Database (HPRD) contains experimentally verified PTMs and their annotations for over 15,000 human proteins[108]. HPRD can also be used to explore kinase-substrate relationships by integrating the PhosphoMotif Finder tool which contains a collection of phosphorylation motifs and their associated kinase enzymes from published literature[90]. Another database, Plant PTM Viewer, is a central resource for investigating PTMs in plant species which stores comprehensive phosphorylation data for model plant organisms such as *A. thaliana* (Arabidopsis), *O. sativa* (rice) and *Z. mays* (maize) and provides the details of relevant experiments and validation techniques used to identify the target sites[109]. In addition, the iProteinDB database provides information on PTMs in several *Drosophila* species and offers a comparative analysis of identified sites with other model organisms[110].

The general scope of phosphorylation studies is mostly limited to a specific set of species, with humans and commonly studied model organisms (e.g., mouse, flies, worms and plants) having the most phosphosite annotations, and other species (e.g., several vertebrates, invertebrates and lower eukaryotes) having very few or no identified phosphosites at all[13, 57, 111]. The process of phosphosite identification and proteome annotation can be expensive and time-consuming. Therefore, proteomics studies are often designed around benefiting human life and improving the understanding of human biology and disease[112]. Despite this, only a small fraction of the currently characterised human phosphoproteome has an annotated functional role[13, 57, 113]. This is likely because the rate of phosphosite discovery is far greater than the rate at which each individual site or motif can be analysed experimentally and validated in terms of functional annotations. In addition, several studies suggested that a significant portion of phosphosites may have no clear functional relevance[114, 115]. The difficulty in distinguishing functionally significant phosphorylated regions from those that do not contribute to protein function is exacerbated by the added complexity of proteins having multiple phosphorylated sites within their sequence, as well as several kinase enzymes being able to phosphorylate multiple sites[30, 31]. The identification and validation of a protein's functionally significant phosphosites is therefore a crucial step in predicting its biological function, understanding its molecular interactions and directing downstream analysis[116]. As a result, one of the main aims of this Thesis is to analyse the functional relevance of human phosphorylation sites and use characterised confident human phosphosites with strong identification evidence as a reference set to predict phosphosites in eukaryotic species which may not have comprehensive phosphoproteome annotations. This is achieved in Chapter 4 by applying conservation analysis to study protein sequences from the reference human proteome and analyse the evolutionary and functional trends of their associated phosphosites which are expected to have diverse conservation patterns across groups of eukaryotic species.

1.8 The Basic Principle of Conservation Analysis

The conservation analysis in proteomics involves comparing the amino acid sequence of a protein in question to the sequences of its homologues (i.e., proteins that have a common evolutionary ancestor) and identifying local regions of similarity which are likely to have a common functional implication amongst the compared proteins[116]. It is possible, however, that common protein ancestry may not necessarily infer functional similarities, especially when only a small portion of a target sequence is conserved, or if known functional annotations of conserved homologues are inaccurate[117-119]. As a result, a careful examination of common

protein features such as protein domains, their presence and conservation across homologous sequences in addition to the overall sequence conservation could be considered to enhance the accuracy of predictions surrounding functional relevance[118]. Nevertheless, conservation plays a central role in research surrounding model organisms and how they can be used to study human biology and disease[112, 120]. This is highlighted by various studies of organisms such as flies[121], worms[122], yeast[123] and mammals[122, 124] that uncovered novel molecular pathways and demonstrated a direct functional connection of those pathways to human biology by analysing conservation of the involved proteins[120, 122].

When it comes to studying the evolution and function of phosphorylated protein regions, it is generally hypothesised that functionally significant phosphosites would be highly conserved because their mutations to non-phosphosites would alter protein function and ultimately hinder evolutionary selection[125, 126]. Several studies demonstrated that Ser, Thr and Tyr phosphosites are indeed significantly more conserved compared to non-phosphosites in general[113, 126]. This observation is examined further in Chapter 3, where the conservation is assessed between human phosphosites and characterised non-phosphosites in the same protein sequences. In addition, assuming that phosphosites are generally highly conserved, Chapter 3 applies the conservation analysis to separate confident human phosphosites from likely false positive identifications.

1.9 Identifying Homologous Protein Sequences

A typical conservation analysis of a query protein sequence begins with identifying its homologous sequences which can be done by performing a sequence search against protein sequence databases and establishing regions of excess sequence similarity[127]. An example of a popular protein sequence database is UniProt's Swiss-Prot which contains hundreds of thousands of manually curated entries and provides comprehensive sequence annotations[13]. Other widely used protein resources include RefSeq which contains sets of well-annotated reference sequences[77], and the Protein Data Bank (PDB) which also links protein sequences to their relevant three-dimensional structures[128].

One of the most reliable and widely used methods for searching sequence databases and identifying local regions of similarity between two amino acid sequences is Basic Local Alignment Search Tool (BLAST) and in particular, its blastp algorithm which analyses query protein sequences against a protein database[129, 130]. The query sequences are typically provided in standardised FASTA format where a sequence of amino acids is preceded by a ">"

symbol and contains relevant protein ID tags and descriptive sequence information. The BLAST algorithm begins by processing the query protein sequence to filter out any low-complexity regions (i.e., sequence regions comprised of a limited set of amino acids, often containing repeats[131]) if required. The query sequence is then divided into shorter sequences of fixed length (also referred to as "*words*") to make a sequence list which is to be searched against the database. Because identical matches are not always made, the initial list of "*words"* is also expanded to include neighbouring "*words*" of the same length which have a certain similarity score to their initial "*word"*. The similarity score of the neighbours is reflected by a combination of individual amino acid matches and differences relative to the original "*word"*. Those scores are defined by a selected substitution matrix, where the exact amino acid matches typically produce the highest score, favourable (or likely) amino acid substitutions have positive values and unlikely substitutions have negative values. An example of a commonly used scoring matrix in protein alignments is BLOSUM (BLOcks SUbstitution Matrix), with BLOSUM62 being a default option in blastp, which is based on the relative frequencies and probabilities of substitutions between amino acid pairs[132]. The resulting lists of "*words"* containing the "*word*" itself and its high-scoring neighbours, which have a similarity above a certain threshold (also known as the neighbourhood word-score threshold[133]), are searched against the sequence database to find the exact (conserved) matches. Once a match is made and aligned (seeded), the alignment is extended in both directions by first using a gap-free extension method which continues as long as the alignment score relative to the composition of the query sequence increases and does not fall below a certain value (i.e., the drop-off value), or until a maximum query length is reached. The resulting high-scoring segment pairs are further extended by the introduction of gaps which consider insertion and deletion mutations. The final pairwise alignment is assigned a Maximum score which reflects the overall quality of the alignment by taking into account the length of the alignment, any rewards for matched amino acids between the compared sequences and any penalties for mismatches. BLAST also incorporates a statistical parameter called an Expectation value (*E*-value) which represents the number of times an alignment with a given score can be obtained by chance when the query sequence is searched against a random protein database of the same size. Therefore, alignments with high Maximum scores and low *E*-values are unlikely to occur by chance and are instead likely to represent homologous sequences[134]. BLAST is applied throughout this Thesis as part of the conservation analysis, where it is used to identify potential homologues of the target human proteins in different species (orthologues) to help establish how conserved human phosphosites are across eukaryotes.

1.10 Comparing Multiple Protein Sequences

Once the likely homologue candidates of the query protein sequence are selected, the next step of a conservation analysis involves a sequence comparison, which is done by generating and examining a multiple sequence alignment (MSA) between the full query sequence and the sequences of its matched homologues[135, 136]. MSAs provide information about the similarity of the primary structure between the compared proteins, which can in turn be used to make predictions about their phylogenetic relationship and functional similarity by identifying conserved functional amino acid sites and domains[136]. In a traditional MSA, the compared protein sequence entries are aligned as rows and the amino acid sites are split into columns, with indels (gaps) being added to account for any insertion and deletion mutations between the sequences. The degree of conservation can be assessed for each location within an MSA either as a column-wise score or at a global sequence level with several different conservation scores and scoring algorithms available[116, 137]. Additional software tools are applied for the analysis, interpretation and annotation of MSAs such as MEGA (Molecular Evolutionary Genetic Analysis)[138] and Jalview[139].

There are many published computational algorithms available for generating MSAs which differ in terms of their speed, accuracy and approaches used for sequence comparisons[140, 141]. For example, ClustalW[142] is one of the earliest and most widely used MSA methods which applies a progressive, matrix-based alignment approach, where scored pairwise alignments are first made between all sequences to generate a distance matrix, which represents the divergence of each sequence pair. The distance matrix is then used to calculate a phylogenetic tree with the neighbour-joining method[143] where the branch lengths are proportional to the estimated evolutionary divergence. Finally, the branching order of the resulting tree is used to guide the process of the sequence alignment in a progressive manner, starting from the two closest sequences and then gradually adding more distant ones. Since the release of ClustalW, many other MSA algorithms with superior performance have been developed and benchmarked[140, 141]. The popular notable examples of such algorithms include MAFFT[144], MUSCLE[145], Clustal Omega[146] and T-COFFEE[147]. The choice of an MSA tool generally depends on the research goal, the desired accuracy and execution time, the available computational power and the dataset size[148]. In particular, MUSCLE has proven to be more suitable for large-scale alignments (i.e., alignments involving longer sequences and/or many sequences at once) when compared to other methods and its application can be easily integrated into computational pipelines and programming languages[145]. As in ClustalW, the algorithm of MUSCLE is also

based on a progressive approach, but it includes additional iterative refinement steps which increase the accuracy of the alignment and gap placement[145]. MUSCLE refines the alignment until a point of convergence is reached where the alignment cannot be improved further or if a defined maximum number of iterations is met.

An example of an alignment generated by MUSCLE and visualised in Jalview is demonstrated in Figure 3, where the sequence of a human SH2 domain-containing protein 3C (SH2D3C) is compared against its potential orthologues in eukaryotic species. SH2D3C plays a role in regulating cell adhesion and migration[149] by using a conserved protein domain SH2 which is typically involved in a variety of signalling pathways[150]. In addition, SH2D3C has known functional Tyr278 and Tyr283 sites in its sequence which are phosphorylated during cell signalling pathways[13, 151]. The example MSA in Figure 3 was able to identify the conserved SH2 domain, which spans between positions 220 and 319 in the human sequence, and distinguish it from the less conserved disordered protein regions which are found earlier in the sequence. In addition, the alignment successfully detected the conserved functional tyrosine phosphosites. This highlights how such analysis can be used to make evolutionary and functional conclusions about this protein and its relevant sites.

Overall, MSAs play an integral role in conservation analysis when comparing multiple protein sequences, understanding evolutionary relationships and inferring protein regions of functional relevance. This is further highlighted by a range of recent conservation studies which, for example, examined SARS-CoV-2 spike protein to understand the transmission of COVID-19 between species and its emergence in humans[152, 153], identified liver cancer-relevant mutations in a lipid-binding protein domain called START[154] and explored conserved anti-fungal defence mechanisms in plant species[155]. In addition, several examples of MSAs are demonstrated in Chapter 4 where they are used to visualise different conservation patterns of human phosphosites across various groups of eukaryotic species.

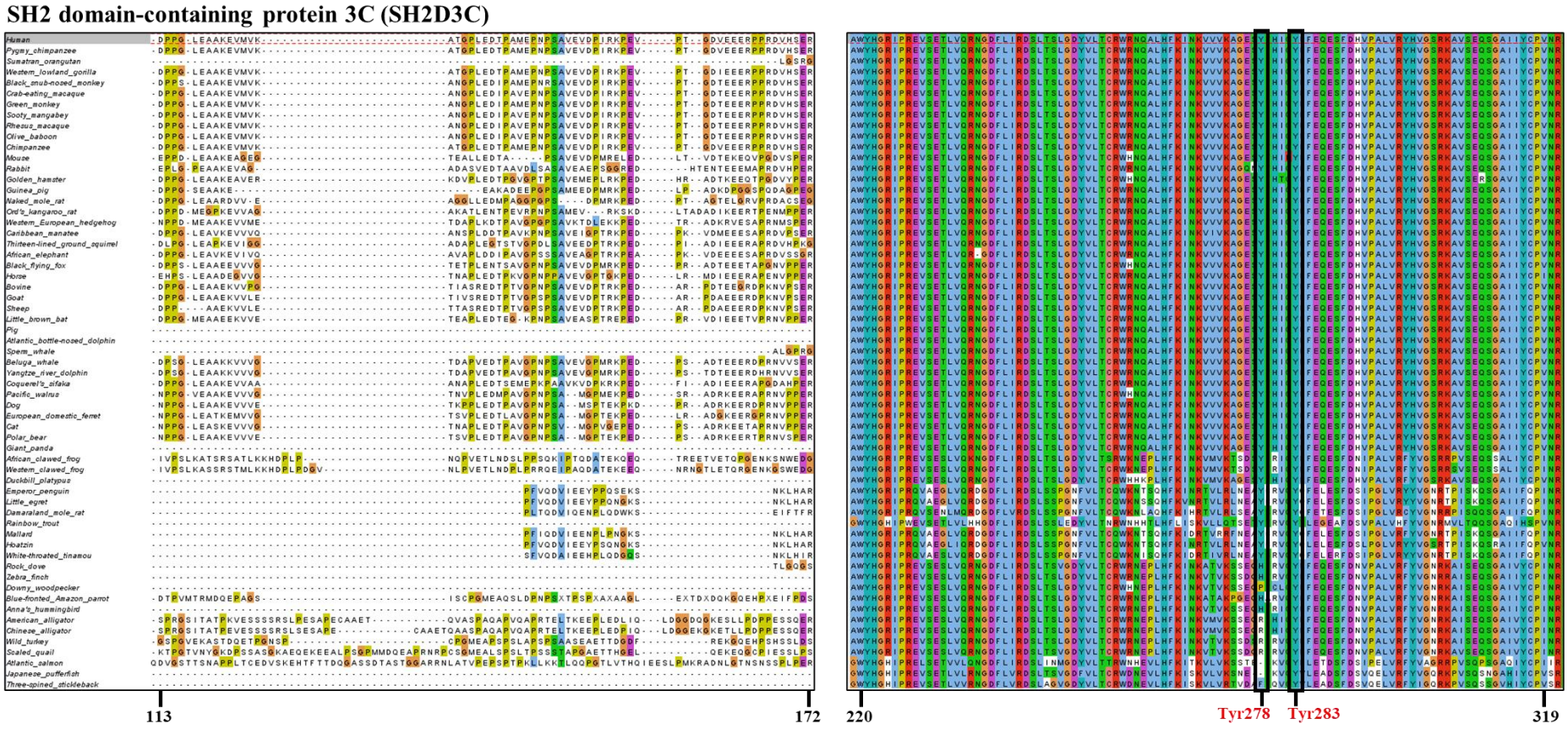


Figure 3. Sections of a multiple sequence alignment between human SH2 domain-containing protein 3C (shown at the top) and its potential orthologues from 60 eukaryotic species. The alignment was generated using MUSCLE and visualised with Jalview. Functional sites Tyr278 and Tyr283 are highlighted by black boxes and red labels underneath. The flanking positions of each alignment section are also provided which correspond to the site positions in the human sequence.

The described steps of a typical conservation analysis can be automatically performed by several publicly available computational pipelines, which can identify homologues, generate MSAs and assess conservation in a single step. The key examples of such pipelines include ConSurf[156], ACES[157], PANTHER[158] and Ensembl Compara[159], some of which are further discussed in Chapter 2. However, it is important to note that none of the existing conservation pipelines can easily and efficiently determine the conservation of multiple protein targets at once across several species on a large scale. One of the aims of this Thesis is to establish global evolutionary and functional trends of human phosphosites. As a result, it is important to have a reliable methodology in place which can determine the conservation of every target site in the human phosphoproteome across a large number of eukaryotic species. The development and validation of such method is described in Chapter 2 where a computational conservation pipeline is built for large-scale evolutionary analysis of multiple protein sequences and species.

1.11 Functional Enrichment Analysis of Proteins

As previously highlighted, the analysis of a protein's function is crucial in understanding its associated biological pathways and molecular interactions. In addition to conservation analysis, a commonly used technique to infer the functions of specific protein groups is functional enrichment analysis[160]. The basic principle of the functional enrichment analysis is to identify functional annotations which are significantly overrepresented (enriched) in a target list of proteins when compared to a control background set. A standard vocabulary system used for classifying and describing functional annotations is Gene Ontology (GO) which organises the annotations, referred to as GO terms, into a hierarchical order according to the three main groups: cellular component, molecular function and biological process[161]. Each GO term represents a specific biological concept and is assigned to proteins based on experimental evidence or computational predictions from the conservation analysis of similar proteins with a known function[160]. The terms are also commonly linked together using parent-child relationships which highlight functional relationships[161]. For example, a parent term "*protein kinase activity*" has a child term "*protein serine/threonine kinase activity*" which in turn has a child term "*MAP kinase activity*".

There are many bioinformatics tools available for performing functional enrichment analysis of protein targets[162]. In a typical workflow, the software assigns GO terms for each entry in a specific background set. Once the user uploads their list of target proteins, the software performs a statistical analysis to compare the number of targets associated with a particular GO term to the number of proteins in the background set linked to that term. The significance of

the enrichment is assessed with a p-value which reflects the probability of observing this enrichment by chance. Additional multiple testing correction may be applied to adjust the p-value. One of the most commonly used tools for functional enrichment of protein sets is DAVID (Database for Annotation, Visualization and Integrated Discovery)[163] which applies Fisher's exact statistical test[164] to select enriched functional annotations in the user's protein set. DAVID is suitable for large-scale analysis involving thousands of target proteins and searches additional functional annotations in addition to GO terms, including KEGG pathways[165], UniProt keywords[13] and protein domain annotations from databases such as SMART[166] and InterPro[167]. Another useful functional enrichment tool is clusterProfiler[168] which is an open-source, user-friendly R package that offers comprehensive analysis and visualisation of enriched functional terms. Functional enrichment analysis is applied in Chapter 3 to compare the functional annotations of proteins containing phosphosites with different levels of identification evidence. In addition, Chapter 4 includes functional enrichment analysis with DAVID and clusterProfiler to investigate functional trends of human phosphosites and link them to various phosphosite conservation patterns.

1.12 Thesis Aims and Chapter Outline

The main aim of this Thesis is to profile the human phosphoproteome to estimate the true extent of protein phosphorylation and analyse the evolutionary conservation of phosphosites. To achieve this aim, the Thesis is logically split into three chapters which investigate phosphosites from the human proteome in terms of their phosphorylation evidence in relevant databases, conservation across 100 eukaryotic species and functional relevance in associated proteins.

Chapter 2 - Developing a Computational Pipeline for Predicting Amino Acid Conservation Across Multiple Species

Several publicly available computational pipelines automatically perform various steps of a typical conservation analysis, often focusing on identifying potential homologues of target proteins, generating multiple sequence alignments and assessing site conservation in a single convenient step. However, none of the existing methods can easily and efficiently determine site conservation in multiple protein targets across several species, which is essential for understanding the evolutionary trends of human phosphosites on a large scale. As a result, Chapter 2 describes the development and optimisation of an accessible computational pipeline which determines phosphosite conservation by performing several steps of a conservation analysis in a single step. Our novel conservation pipeline can search multiple protein query sequences against reference proteomes of selected species using BLAST, extract a top hit (likely orthologue) from each species for each query, generate MSAs between each query sequence and its top hits, calculate site conservation and map the results to phosphorylation data.

Chapter 3 - Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation

Public phosphorylation databases such as PSP and PA compile results from published papers or openly available MS data. However, there is no database-level control for false discovery of sites, likely leading to the overestimation of true phosphosites. Chapter 3 describes the profiling analysis of the human phosphoproteome to estimate the FDR of phosphosites and predict a more realistic count of true identifications. In particular, human Ser, Thr and Tyr phosphosites are ranked into sets according to the strength of their identification evidence and analysed in terms of their conservation across 100 species, sequence properties and functional annotations. The analysis demonstrates significant differences between the sets and independently estimates phosphosite FDR in the human phosphoproteome, highlighting the extent of false positive identifications in large datasets.

Chapter 4 - Discovering Evolutionary and Functional Trends of Human Phosphorylation Sites

The vast majority of phosphosite discoveries are made in humans, with many other species only having a few experimentally confirmed or computationally predicted phosphosites. In addition, only a small fraction of the currently characterised human phosphoproteome has an annotated functional role. The analysis in Chapter 4 investigates the conservation of human phosphosites across various groups of eukaryotic species and establishes their relevance in several protein functions. In addition, by using conservation analysis and confident human phosphosites as a reference set, we predict over 1,000,000 potential phosphosites in eukaryotic species ranging from primates and other mammals to fungi and plants.

Chapter 5 - Thesis Conclusion and Future Research Directions

This Chapter will conclude this Thesis by summarising key results, highlighting their significance and offering directions for further research.

Chapter 2

Developing a Computational Pipeline for Predicting Amino Acid Conservation Across Multiple Species

2.1 Abstract

The analysis of amino acid conservation is an efficient method of predicting significant functional sites within a protein sequence. Several publicly available computational pipelines automatically perform various steps of a typical conservation analysis, often focusing on identifying potential homologues of target proteins, generating multiple sequence alignments and assessing site conservation in a single convenient step. However, none of the existing methods can easily and efficiently determine the conservation of multiple protein targets at once across several species. In this Chapter, we developed and optimised an accessible Python pipeline which can determine the conservation of specific amino acid sites such as PTMs and perform the following steps of a conservation analysis in a single step:

- Search multiple query sequences against proteomes of selected species using BLAST.
- For each query sequence, extract a top-matched protein sequence from each species.
- Generate MSAs between each query sequence and its matches.
- Determine the conservation of a specified amino acid across the aligned sequences at each of its positions in a query sequence.
- Produce a comprehensive summary output which can be mapped to target sites of interest and used to predict their conservation patterns across the selected species.

We demonstrated that the pipeline is robust, easy to use and can be readily applied to analyse thousands of protein targets at once from different species. The pipeline generates multiple useful outputs which permit an in-depth downstream analysis, such as BLAST results, FASTA sequences of top hits from each species, multiple sequence alignments and percentage conservation of target and adjacent sites. In addition, the pipeline is supported by a guide containing detailed installation and running instructions, explanations of any inputs and outputs, a troubleshooting guide and links to example inputs and results. Finally, the user does not need any prior knowledge of Python programming to run the pipeline. Therefore, the pipeline is ideal for studying the evolutionary conservation of any biological sites of interest such as post-translational modifications across multiple species.

2.2 Introduction

2.2.1 The relevance of conservation analysis in proteomics

The analysis of a protein's amino acid sequence can be used to predict its function and structure. However, not all amino acid residues within a protein sequence are of equal functional relevance and often only a few sites are involved in important protein domains, with others being readily interchangeable without having a significant effect on protein function. Therefore, the identification of a protein's functionally significant sites is a crucial step in predicting its biological function, understanding its molecular interactions and directing downstream analysis[116].

One efficient and commonly applied method of predicting significant protein sites is the analysis of sequence conservation. At its simplest, an amino acid sequence of a protein in question is compared to the sequences of its homologues, with a central concept being that proteins which share common evolutionary ancestry are also similar in terms of function and structure[137]. The identification of highly conserved sites or motifs can therefore be used to derive their primary structural patterns and functional significance when compared to groups of annotated protein homologues with a known shared function. It is possible, however, that common protein ancestry may not necessarily infer functional similarities, especially when only a small portion of a target sequence is conserved, or if known functional annotations of conserved homologues are inaccurate[117-119]. As a result, a careful examination of common protein features such as protein domains, their presence and conservation across homologous sequences in addition to the overall sequence conservation should be considered to ensure accurate functional assignments for the proteins in question[118].

The analysis of evolutionary conservation and protein homology has been successfully applied in studying protein interactions[169-171], detecting and verifying motifs involved in substrate binding[172-174], and uncovering the origins and functions of proteins implicated in modern diseases such as Ebola virus[175], COVID-19[153] and cancer[176, 177]. Furthermore, the importance of conservation analysis has been highlighted in a comparative computational study which revealed that sequence conservation is the most accurate individual predictor of protein function when compared to other properties such as amino acid identity, catalytic properties, and relative site position on a protein surface[178]. However, it was also concluded that using a combination of those attributes would maximise the overall accuracy of functional predictions[178].

2.2.2 Identifying homologous protein sequences

The first logical step in a typical conservation analysis of a new protein sequence involves searching and selecting its homologous sequences. This is done by performing sequence similarity searches against protein sequence databases. For example, UniProt's Swiss-Prot database is an expertly curated resource which not only contains hundreds of thousands of protein sequences, but also provides comprehensive annotations describing sequence features, protein isoforms, post-translational modifications (PTMs) and potential protein interactions, allowing more accurate functional predictions to be made from the identified homologues[13]. Other widely used protein resources include Reference Sequence Database (RefSeq)[77] and the Protein Data Bank (PDB)[128]. The sequence searching itself can be performed by using publicly available algorithms such as HMMER[179], the Smith-Waterman local similarity algorithm[180, 181] and BLAST[129, 130], with the latter being considered the most reliable and widely used method for identifying local regions of similarity between two sequences. The presence or absence of homology between the query sequence and the resulting matched sequences can be inferred from excess sequence similarity, conservation of active sites and available sequence annotations[127]. To assess sequence similarity between two matched sequences, a search algorithm returns a similarity score which generally allows to distinguish between potentially homologous sequences and the ones which are unrelated to the query sequence, based on the principle that unrelated sequences would have a similar score to the sequences matched at random[127]. For example, BLAST runs pairwise alignments between a query sequence and its matches to generate a Maximum Score (or bit score) which takes into account any rewards for matched amino acids between the compared sequences and penalties for any mismatches, and which does not depend on database size (i.e. finding the exact match from a database of a different size would generate same score)[130]. To further evaluate sequence similarity and add statistical significance within a search algorithm, BLAST and other tools incorporate an Expectation value (*E*-value) which is a statistical parameter that reflects the number of times a sequence match with a given score can be obtained by chance when searched against a random protein database of the same size[130]. Therefore, lower *E*-values and higher similarity scores are assigned to sequence matches which are unlikely to occur by chance and which, in fact, are likely to represent homologous sequences. In protein sequence comparisons, homology can be reliably and statistically inferred by a commonly accepted *E*-value threshold of <0.001 and a bit score of >50, although additional less sensitive metrics are also available, including percentage identity[127].

Typical search algorithms such as BLAST are easily accessible locally or through various online resources and allow efficient identifications of homologues for individual protein sequences. However, several computational tools have been developed to also detect protein homology on a greater scale from large sets of target sequences with a focus on identifying groups of significant orthologues (genes which evolved from speciation events) and paralogues (genes which evolved from duplication events)[182]. Accurate orthologue identification is essential in computational biology for improving genome annotations, studying phylogenetic relationships between species and predicting the functional significance of unknown protein sequences[183]. One of the most widely used and simplest methods for predicting an orthologous relationship between identified homologues is by applying the reciprocal best hits (RBH) principle, which is based on selecting the best-scoring match out of all homologues[184, 185]. RBH assumes that if protein *X* from database *x* is an orthologue of protein *Y* from database *y*, then when protein *X* is searched against database *y*, the best match would be protein *Y*, whereas a reciprocal search of protein *Y* against database *x* would return protein *X* as the best match[184, 185]. In addition to RBH, orthologue identification and grouping can be performed using various algorithms which are implemented within specialised tools[182, 186]. For example, OrthoMCL allows users to execute a global BLAST analysis where complete proteomes of multiple species are compared against each other to identify potential orthologues for each given sequence and cluster them into specified orthologue groups, thus allowing a functional prediction to be made based on common annotations present within that group[187]. Such large-scale approaches are generally computationally intensive and demanding in terms of memory and CPU usage, although multiple efforts have been made to build upon existing algorithms and improve their overall efficiency[188, 189].

2.2.3 Generating multiple sequence alignments

Once the homologous sequences are selected, the next step in a conservation analysis involves generating and examining a multiple sequence alignment (MSA) between a query sequence and the sequences of its homologues[135, 136]. Each protein MSA has protein sequences aligned as rows and amino acids as columns, with gaps inserted at certain positions if needed to improve alignment accuracy and account for any insertion or deletion mutations between the aligned sequences[136]. An MSA is usually assessed to identify areas or columns of interest in which amino acids are conserved across the aligned sequences and may therefore be functionally significant since they have fewer mutations compared to the remaining areas of the alignment, likely due to stronger evolutionary constraints[136, 190]. It is also possible to

pinpoint the conservation of any sequence areas already known to be functionally significant such as active sites or modified residues, to determine if those features are potentially present in another species. The degree of conservation can be assigned for each location within an MSA either as a column-wise score or at a global sequence level with several different conservation scores and scoring algorithms available[116, 137]. In addition, it is feasible to calculate the percentage conservation of a certain amino acid at a given aligned position out of all sequences to predict how conserved a specific site of interest from a query sequence is. A traditional method of generating multiple sequence alignments is ClustalW[191], although many other published alignment tools are available[140, 141]. These include widely used iterative approaches such as MAFFT[144] and MUSCLE[145] which are more suitable for generating large alignments, show superior performance in terms of speed and accuracy, and the application of which can be easily incorporated within computational workflows.

2.2.4 Computational pipelines for conservation analysis

As discussed in Chapter 1, one of the main aims of this Thesis is to predict the conservation of all known Ser, Thr and Tyr human phosphorylation sites (phosphosites) in multiple eukaryotic species. The conservation data can then be used to assess the likelihood of target phosphosites being "*real*" and to understand their evolutionary and functional patterns across eukaryotes. To achieve this, it would be necessary to search every protein sequence in the human proteome against the proteomes of selected species to identify top matching orthologues, perform MSAs and assign conservation scores to target human phosphosites at their known sequence positions.

Several publicly available computational pipelines automatically perform various steps of a conservation analysis, often focusing on identifying homologues, generating MSAs and assessing conservation in a single convenient step. For example, ConSurf is a computational tool which was ultimately built for identifying functional sequence regions by analysing the evolutionary rates of amino or nucleic acids across homologues[156]. In brief, given a query sequence of amino acids or nucleotides, ConSurf searches its homologous sequences using BLAST or HMMER in a selected database such as Swiss-Prot, generates a multiple sequence alignment using MUSCLE, MAFFT or ClustalW, builds phylogenetic trees and predicts evolutionary rates at each position within the sequence[156]. The rates of evolution are inferred by applying the Rate4Site algorithm which is based on the maximum likelihood principle and considers branch length and overall topology of generated phylogenetic trees[192]. The conservation pipeline integrated within the ConSurf tool has been successfully used to identify

functionally significant protein sites from large alignments[116, 193]. However, it has also been demonstrated that ConSurf applications can be time-consuming and that the Rate4Site algorithm decreases in performance when dealing with less than 50 aligned sequences[116, 137]. In addition, ConSurf does not allow running sequence searches against whole proteomes and it is also not possible to automatically select the species of interest from the resulting homologues (although manual filtering can be done for each individual query search), with a focus rather being made on closest homologue matches from a specified protein database such as Swiss-Prot[156]. Therefore, although ConSurf offers an efficient conservation analysis pipeline, it is not suitable for the aims outlined in this Thesis.

Another publicly available pipeline for conservation analysis is ACES (Analysis of Conservation with an Extensive list of Species) which analyses the conservation of a query sequence in relation to its orthologous sequences from the reference genomes of selected species[157]. In particular, ACES allows to run a global BLAST search of query sequences against whole genomes, identify top matching orthologues from each species, build phylogenetic trees, perform basic sequence annotations, generate multiple sequence alignments using MUSCLE and determine conservation patterns at specific sequence regions[157]. However, the obvious limitation of ACES is that it can only be applied to nucleotide sequences. Nevertheless, the general principles behind the ACES pipeline can prove useful in the conservation analysis of human protein sequences that is aimed in this Thesis.

Extensive pre-generated collections of genomics and proteomics data related to evolutionary relationships between sequences are also readily accessible from various databases and can be linked to query sequences. For example, the Ensembl comparative genomics resource (Ensembl Compara) is a comprehensive database which contains pairwise all-vs-all BLAST alignments between various genomes, predictions and groupings of orthologues, pre-generated phylogenetic trees and sequence annotations for over 70 species[159]. Ensembl Compara also offers a range of tools for each of the steps in conservation analysis but analysing multiple sequences is complicated and the list of available species is mostly limited to vertebrates[159].

2.2.5 Aims

As discussed, there are plenty of computational workflows available for the analysis of conservation in genomics and proteomics, but to our understanding, none of them were perfectly suitable for achieving the goals outlined in this Thesis. Therefore, the main aim of the work described in this Chapter was to develop a novel and convenient computational

pipeline for the analysis of query protein sequences, which would ultimately generate data that researchers can use to determine the conservation of target protein sites in a selected set of species. In particular, it was proposed that the conservation pipeline could be applied in this Thesis to predict the conservation of human phosphosites from the reference human proteome across multiple eukaryotic species. To achieve this, the pipeline was optimised to perform the following analytical functions in a single step:

- Search multiple query sequences against reference proteomes of selected species.
- For each query sequence, extract a top-matched protein sequence from each species.
- Generate multiple sequence alignments between each query sequence and its matches.
- Determine the conservation of a specified amino acid across the aligned sequences at each of its positions in a query sequence.
- Produce a comprehensive summary output which can be mapped to target sites of interest and used to predict their conservation patterns across the selected species.

The development of the conservation pipeline was done using Python which is a high-level, dynamic and easily accessible programming language that offers readable, clean code as well as interplay with other programming languages[194]. Furthermore, Python encompasses many useful libraries for data analysis and visualisation such as NumPy, SciPy, pandas and Matplotlib, making it an increasingly popular choice for scientific programming[195, 196]. Several specialised Python libraries have also been developed to assist bioinformatics research. For example, Biopython is a library which allows accessing major databases such as Swiss-Prot and PDB, processing multiple sequence alignments, analysing 3D molecular structures and interacting with various commonly used tools such as BLAST and ClustalW[197]. In fact, multiple previous studies have successfully implicated Python and its analytical libraries as a key part of their methodology[198-200]. As a result, Python was an optimal choice of a programming language for the development of our conservation pipeline.

2.3 Method

A comprehensive computational pipeline was written in Python (ver. 3.7.3) while using Spyder (ver. 3.3.6) as an integrated development environment (IDE). The pipeline was optimised to combine multiple steps of a typical conservation analysis (Fig. 4).

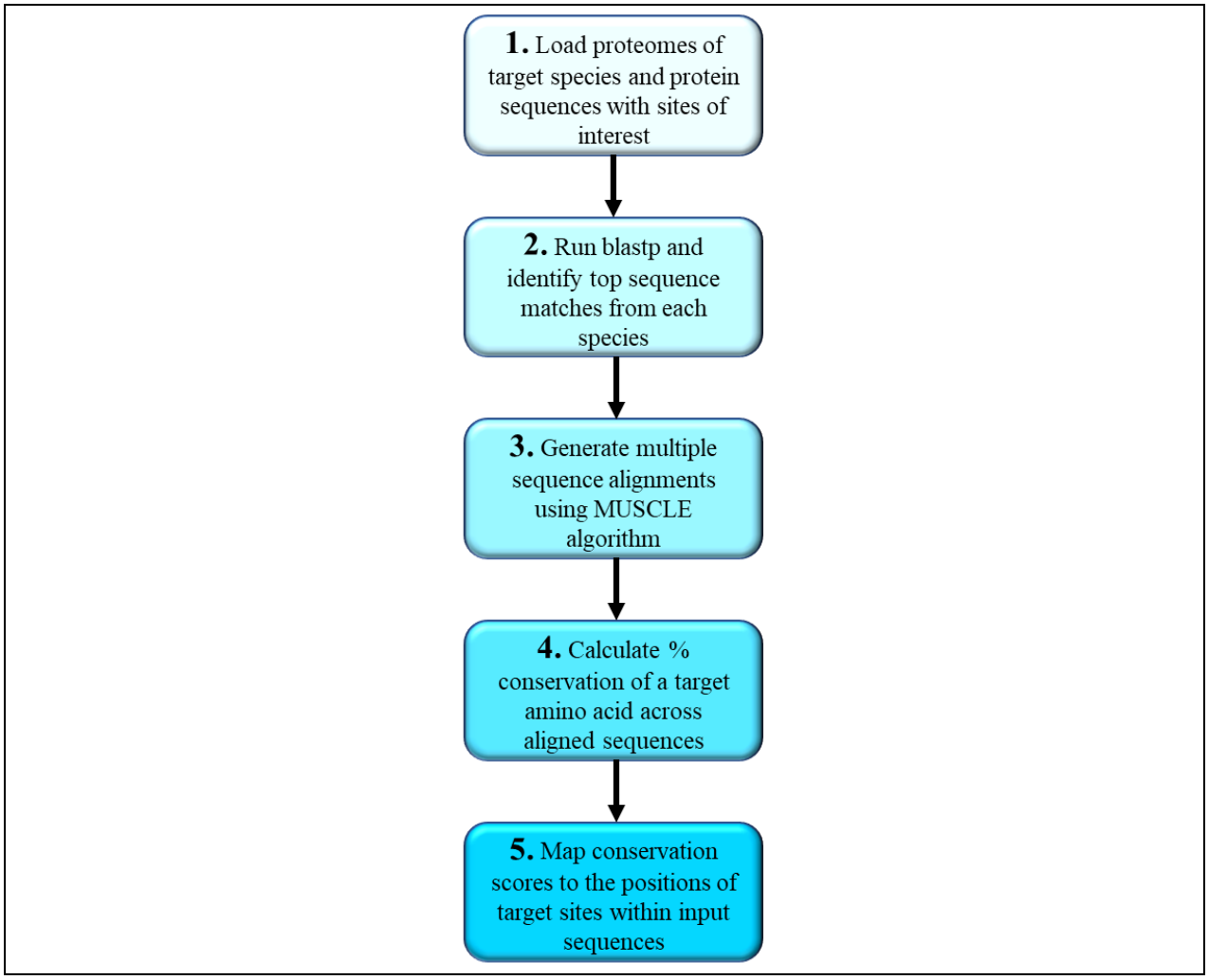


Figure 4. A flowchart of the conservation analysis pipeline.

The first step of the conservation pipeline is to load a FASTA file "*targets.fasta*" containing user-selected protein sequences from a particular species with target sites of interest for which a conservation score has to be calculated. By default, the target sequences must come from human, although it is possible to analyse targets from any species with an available reference proteome. Another FASTA file "*proteomes.fasta*" is also uploaded which contains all selected reference proteomes of target species and the proteome of species from which the sites of

interest originate. All reference proteomes and target protein sequences can be readily downloaded from UniProt in FASTA format[13].

Each target protein sequence is used as a query in a BLASTp search against every selected proteome using default settings (BLAST+ ver. 2.10.0)[129, 130, 201]. For each target sequence, its matched sequences in the resulting BLAST output are filtered by applying a user-selected *E*-value threshold (default set at ≤0.00001). A single top matching protein sequence is then selected from each of the species based on a given *E*-value, thus identifying the most likely orthologue candidate from that species, although a true orthologous relationship cannot be guaranteed without further evaluation. A FASTA output is also generated per target protein which contains its sequence and the sequences of all its matches to be aligned.

A multiple sequence alignment is then generated between a target sequence and its matches using the MUSCLE algorithm (ver. 3.8.31) with default settings, applied via a locally installed executable file[145]. If any sequences to be aligned (either the target sequence or any of the matched sequences) are ≥2,000 amino acids long, 2 iterations of the algorithm are run using recommended settings for large alignments (`-maxiters 2` option)[145].

The resulting MSAs are processed using Biopython (ver. 1.74)[197] to calculate percentage conservation of a user-specified amino acid at each of its positions within a target protein sequence. The conservation of neighbouring amino acids at -1 and +1 positions around each target amino acid is also given. In addition, the pipeline can consider user-specified frequent amino acid substitutions when calculating percentage conservation scores. For example, if a target amino acid is serine, any matched sequence with a threonine (a common mutation of serine) at the aligned position can be included in the conservation calculation and vice versa. The conservation scores are finally cross-referenced with a separate file containing positions of target sites of interest to determine the conservation of those sites across the selected species. Conservation percentage scores are provided out of the total number of analysed target species and out of the number of sequences which were matched and aligned with the query sequence.

The pipeline was designed to account for potential errors by identifying any query proteins which cannot be analysed, extracting them into a separate output "*targets_not_analysed.csv*" and eliminating them from further processing. A protein target can be excluded from the pipeline's analysis for one of the following reasons:

- A target produces no hits in a BLAST search at all, likely due to its sequence being too short or poorly annotated.
- A target produces no significant hits in a BLAST search that meet the set *E*-value threshold, likely due to its sequence being unique to its origin species.
- A target is not found in the reference proteome of its origin species, likely due to poor sequence annotation or because the target is an isoform of a canonical protein from the reference proteome. This is detected when an identical match to the query protein is not found in a BLAST search when searched against the reference proteome of its origin species.
- Multiple sequence alignments are not generated between the target sequence and its top hits. This is likely to happen when very large sequences of ≥30,000 base pairs are being aligned.
- A target protein does not have a specific amino acid in its sequence.

2.4 Application

All the inputs necessary for running the conservation pipeline including the main Python script and the "*README.md*" file with detailed instructions can be accessed through a GitHub repository via link:

https://github.com/antonk-liv/conservation_pipeline

The user has to ensure that their system has Python installed and that the main script can be accessed through a relevant IDE or via Windows/MacOS/Linux command line. In order to run the BLAST step of the pipeline, the user must download BLAST+ executable files from the National Institutes of Health (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Furthermore, the user must ensure that several Python libraries and modules are installed (Table 1).

Table 1. Python modules and libraries incorporated into the conservation pipeline.

Python module/library	Version used	Link for more details and installation instruction
NumPy	1.21.5	https://numpy.org/
Biopython	1.74	https://biopython.org/
Csv	1.0	https://docs.python.org/3/library/csv.html
Re	2.2.1	https://docs.python.org/3/library/re.html
Pandas	1.3.5	https://pandas.pydata.org/

The following inputs must be placed into the working directory before running the pipeline:

- Main Python pipeline ("*conservation_code_run_with_config_ver250922.py*").
- Configuration file with user-specified settings ("*configurations.ini*").
- Linker Python file ("*link_to_config.py*") which connects the main conservation code with the configuration file.
- User-prepared file "*targets.fasta*" with FASTA sequences of target proteins containing the sites of interest.
- User-prepared file "*proteomes.fasta*" with complete reference proteomes of target species (including the origin species of the target proteins) in FASTA format across which the conservation of target sites would be assessed.
- User-prepared CSV file "*sites.csv*" which must contain UniProt accession numbers of target protein sequences in the first column and positions of the sites of interest in the second column.
- MUSCLE executable file ("*muscle.exe*").
- Dictionary file "*Mapped_Uniprot_Species_Names.tsv*" containing UniProt codes for all available species as well as their common and scientific names.

If required, any relevant BLAST+ source files must also be placed into the directory.

The user has to then access the configuration file ("*configurations.ini*") using any appropriate text editor and specify the following parameters for the conservation pipeline:

- Origin species of the target protein sequences ("*species_of_targets"* parameter). The species name must be entered using a relevant UniProt species code (if the species code is unknown, the user would refer to the UniProt database or search the pre-processed dictionary input "*Mapped_Uniprot_Species_Names.tsv*").
- Amino acid identity of target sites ("*target_amino_acid"* parameter) using single-letter amino acid code. The pipeline calculates the conservation of a target amino acid at each of its positions within every query protein sequence.
- A most likely substitution of a target amino acid ("*sub"* parameter) which may not influence the site function, and which is therefore included in the conservation calculation if found instead of a target amino acid within aligned sequences. If no substitution is available, the user must enter any amino acid other than the target amino acid and ignore any columns in the output referring to the substitution.
- *E*-value of the BLASTp search ("*eval_thres"* parameter). Any resulting BLAST hits equal to or less than the specified *E*-value threshold are accepted by the pipeline.

Once the parameters are specified and the configuration file is saved, the user can then run the conservation pipeline either by using a relevant Python IDE or through command line, making sure that the location of the working directory containing all the necessary inputs is specified. It is recommended to use high-throughput computing or parallelisation when analysing more than 1,000 targets at a time to increase the speed and efficiency of the pipeline.

2.5 Results and Discussion

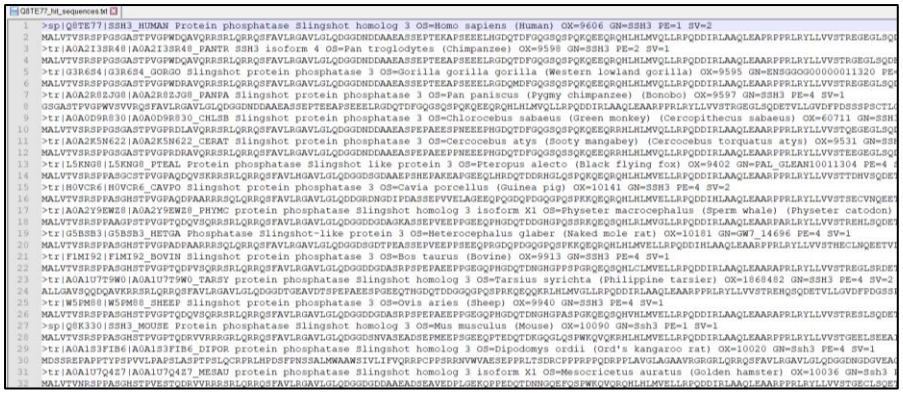
The conservation pipeline was successfully tested (Fig. 5) using 1,000 randomly selected query protein sequences from the UniProt's reference human proteome (UniProt release 2019_10). Each query sequence was searched against the proteomes of 100 example eukaryotic species using BLASTp to identify top matching hits from each of the species (Fig. 5A), extract all sequences of the matches (Fig. 5B), generate multiple sequence alignments (Fig. 5C) and calculate percentage conservation of target sites within the query sequence (Fig. 5D). For this test, the target amino acid was serine and the goal was to identify the conservation of potential serine phosphorylation sites within the query sequences. Phosphorylation data for the query sequences (if available) was extracted from the phosphorylation database PhosphoSitePlus (PSP)[57]. The data was pre-filtered to extract human sequences and positions of phosphorylated sites (11/03/20 build; https://pgb.liv.ac.uk/~antonk/PhosphoSitePlus_build/). All example inputs and outputs from the described test including the "*proteomes.fasta*" file containing 100 proteomes of eukaryotic species can be found in a separate online repository: https://pgb.liv.ac.uk/~antonk/. This version of the "*proteomes.fasta*" file also contains human reference proteome and is therefore suitable for a large-scale conservation analysis of human protein targets which is described in Chapters 3 and 4 of this Thesis.

In the test of 1,000 targets, it was possible to obtain conservation data for 991 protein targets and successfully map it to their phosphorylation data. For the remaining targets, it was not possible to calculate site conservation either due to the protein having no matches in BLAST (1 protein), no significant matches in BLAST which met a set *E*-value threshold (5 proteins), no serine in its sequence (1 protein) or due to failed alignments (2 proteins) (Fig. 5E).

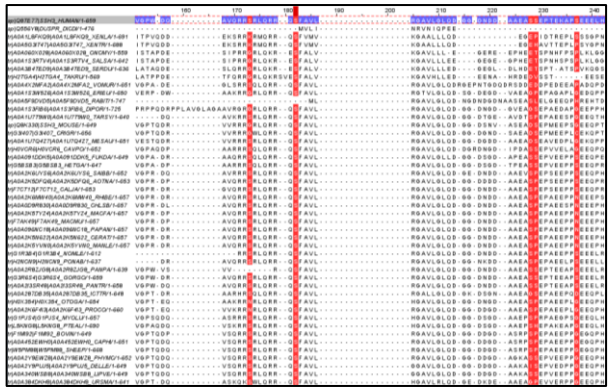
A

Target	Species	Hit	%Identity	E-value	Max. Score
Q8TE77	HUMAN	Q8TE77	100	0	1336
Q8TE77	PANTR	A0A2I3SR48	98.8	0	1312
Q8TE77	GORGO	G3R6S4	98.2	0	1311
Q8TE77	PANPA	A0A2R8ZJG8	96.8	0	1256
Q8TE77	CHLSB	A0A0D9R830	94.4	0	1155
Q8TE77	CERAT	A0A2K5N622	94.4	0	1155
Q8TE77	PTEAL	L5KNG8	79.9	0	1011
Q8TE77	CAVPO	H0VCR6	81.3	0	1011
Q8TE77	PHYMC	A0A2Y9EWZ8	83.8	0	1008
Q8TE77	HETGA	G5BSB3	82.9	0	1006
Q8TE77	BOVIN	F1MI92	81.8	0	979
Q8TE77	TARSY	A0A1U7T9W0	82.9	0	977
Q8TE77	SHEEP	W5PM88	80.6	0	972
Q8TE77	MOUSE	Q8K330	80.2	0	965
Q8TE77	DIPOR	A0A153FIB6	80.8	0	964
Q8TE77	MESAU	A0A1U7Q4Z7	81.5	0	941
Q8TE77	TRIMA	A0A2Y9RR33	74.9	0	921
Q8TE77	ERIEU	A0A153W5Z6	76.7	0	917
Q8TE77	URSMA	A0A384DKH8	74.7	0	903
Q8TE77	VOMUR	A0A4X2MFA2	68.9	0	849
Q8TE77	RABIT	A0A5F9DVD5	73.3	0	715
Q8TE77	ANOCA	G1KHA2	54.4	0	620
Q8TE77	ALLSI	A0A1U7SAM9	62	0	550
Q8TE77	XENTR	A0A5G3IT47	53	3.46E-163	489
Q8TE77	SERDU	A0A3B4TED9	50.3	5.23E-163	486
Q8TE77	XENLA	A0A11BFKQ9	45.6	6.22E-162	486
Q8TE77	ONCMY	A0A060X028	50.4	3.12E-160	477

B



C



D

Protein	Site	Peptide sequence	Pos. in aln.	Position in prot.	No. of species analysed	%Cons. out of 100	%Cons. out of 100 inc. subs.	%Cons. in matched seqs. only	%Cons. in matched seqs. inc. subs.	No. of species aligned	Species aligned (UP codes)	Species aligned (common or sci names)	-1 site	+1 site	-1 site position in aln.	+1 site position in aln.	%Cons. of -1 site in matched seqs.	%Cons. of +1 site in matched seqs.
Q8TE77	S	ALVTVsRsPPGsGAs	124	9	100	50	52	54.3	56.5	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		R	P	123	125	53.3	54.3
Q8TE77	S	VsRsPPGsGAsTPVG	145	13	100	43	46	46.7	50	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		G	G	144	146	17.4	48.9
Q8TE77	S	sPPGsGAsTPVGPWD	148	16	100	43	46	46.7	50	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		A	T	147	149	15.2	46.7
Q8TE77	S	sRLQRRQsFAVLRGA	182	37	100	69	69	75	75	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		Q	F	181	183	53.3	78.3
Q8TE77	S	EPTEKAPsEEELHGD	237	70	100	28	29	30.4	31.5	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		P	E	236	238	40.2	46.7
Q8TE77	S	QTDFGQGsQsPQKQE	255	85	100	14	14	15.2	15.2	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		G	Q	254	256	50	45.7
Q8TE77	S	DFGQGsQsPQKQEEQ	257	87	100	43	43	46.7	46.7	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		Q	P	256	258	45.7	40.2
Q8TE77	S	LEsLRPPsAEPGGss	722	259	100	37	42	40.2	45.7	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		P	A	721	723	79.3	32.6
Q8TE77	S	LDVsDLEsVTsKEIR	792	290	100	45	46	48.9	50	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		E	V	791	833	67.4	59.8
Q8TE77	S	sDLEsVTsKEIRQAL	835	293	100	81	81	88	88	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		T	K	834	836	90.2	89.1
Q8TE77	S	VRLWDEEsAQLLPHW	1129	385	100	46	61	50	66.3	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		E	A	1128	1130	88	51.1
Q8TE77	S	EQKVGGVsPEEHPAP	1669	484	100	47	49	51.1	53.3	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		V	P	1668	1670	29.3	51.1
Q8TE77	S	sIsLLEPsLELEsTs	2062	547	100	30	35	32.6	38	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		P	L	2061	2064	48.9	22.8
Q8TE77	S	PALKsRQsVVTLQGs	2401	602	100	64	66	69.6	71.7	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		Q	V	2400	2402	54.3	50
Q8TE77	S	sVVTLQGsAVVANRT	2408	609	100	13	16	14.1	17.4	92	DELLE;LIPVE;URSMA;FELCA;Iangabey;Drill;Northern white-cheeked gibl		G	A	2407	2420	14.1	46.7

E

Target	Reason for exclusion
A0A286YF77	Target has no significant hits in BLAST
A0N4Z7	Target has no significant hits in BLAST
P0C864	Target has no significant hits in BLAST
P59052	Target has no significant hits in BLAST
Q9H8Q6	Target has no significant hits in BLAST
O75147	Failed alignment due to one of the sequences being too long (>30000bp)
Q8WZ42	Failed alignment due to one of the sequences being too long (>30000bp)
P0DOY5	No hits at all in BLAST
Q9NRI6	No target amino acid in sequence

Figure 5. Example outputs from the conservation pipeline for a human query protein Q8TE77 (Slingshot Protein Phosphatase 3). **(A)** Top sequence matches from each of the selected proteomes (1 hit per species) produced by a BLASTp search. **(B)** FASTA sequences of the top hits from each of the species. **(C)** A fragment of a multiple sequence alignment viewed in Jalview (version 2.11.2.3) between a query protein sequence and its top hits. The query sequence is on the top (highlighted in blue) and the positions of each target amino acid (serine) within the query sequence are highlighted in red. **(D)** Main output from the pipeline containing various conservation scores for each target amino acid per target protein, aligned species and conservation of adjacent sites. **(E)** Target proteins which were not analysed by the pipeline and the reasons for their exclusion.

To further highlight that our conservation pipeline can be tailored to study PTMs such as phosphosites across various species, a separate test was performed which focused on the analysis of known phosphosites from species other than human. A small number of proteins from different species and their phosphosites (*n*=4; Table 2) were randomly selected by searching public phosphorylation databases including The Plant PTM Viewer[109], PTM database iProteinDB[110] and PSP (mouse dataset)[57]. As before, selected targets were analysed across the proteomes of 100 other species to determine the conservation of their phosphosites.

Table 2. Target proteins from various species and their reported phosphosites used to test the conservation pipeline.

Protein	**Species**	**UniProt Accession**	**Phosphosite Positions**	**Evidence**
Cyclin-dependent kinase A-1	*Arabidopsis thaliana* (mouse-ear cress)	P24100	Tyr15; Ser141; Thr161	Plant PTM Viewer[109]
Nascent polypeptide-associated complex	*Zea mays* (maize)	K7VIA7	Ser124; Ser144; Ser148	Plant PTM Viewer[109]
Sphingosine kinase 2	*Mus musculus* (mouse)	Q9JIA7	Ser358; Ser364; Thr377; Ser379; Thr584	PSP[57]
Protein kinase shaggy	*Drosophila melanogaster* (fruit fly)	P18431	Ser8; Ser9; Ser17; Ser213; Tyr214; Ser217	iProteinDB[110]

The results from the pipeline, including multiple sequence alignments (Fig. 6), were used to assess the conservation of phosphosites in Table 2 in relation to different species groups. Interestingly, we observed different conservation patterns which can provide an insight into the evolution of the analysed proteins and their function. For example, the results showed that phosphorylated Tyr15 from *A. thaliana* in cyclin-dependent kinase A-1 (CDKA1), an enzyme which controls cell division and is known to be prevalent in many species[202], was indeed conserved across all species in the analysis including plants (Fig. 6A), further highlighting the importance of tyrosine phosphorylation in plants[203]. A similar conservation pattern was observed for protein kinase shaggy from *D. melanogaster* where all its target phosphorylation sites were conserved across all species (Fig. 6D). Most aligned sequences for this target were glycogen synthase kinase 3 (GSK3) enzymes. In fact, protein kinase shaggy from *Drosophila melanogaster* is a known orthologue of GSK3 which is responsible for cell-fate specification and is found in animals, plants and fungi[204, 205]. Our pipeline was successfully able to infer this orthologous relationship by identifying top sequence matches and producing multiple sequence alignments which correctly aligned key functional sites (Fig. 6D). However, the user must be careful when making assumptions about true orthologous relationships between target proteins and their resulting top hits from other species. For example, our test identified different conservation patterns for phosphosites in nascent polypeptide-associated complex from *Z. mays*, where Ser124 was only conserved in plants, Ser144 was conserved in plants and insects, and Ser148 was conserved primarily in vertebrates (Fig. 6B). This suggests that the aligned phosphosites

may be associated with different proteins depending on the species they are found in, even though there may be an overlap in terms of functional domains. Therefore, further analysis of specific domains is recommended before assuming a true orthologous relationship between the aligned sequences. Finally, our test revealed an example of where phosphosites in Sphingosine kinase 2 from *M. musculus* were conserved primarily in mammals (Fig. 6C), suggesting that their function is either unique to that species group or that the sequences where the sites were not conserved are poorly annotated. To summarise, our tests showed that the pipeline can successfully determine the conservation of known phosphosites across different species and is applicable for both large and small-scale studies.

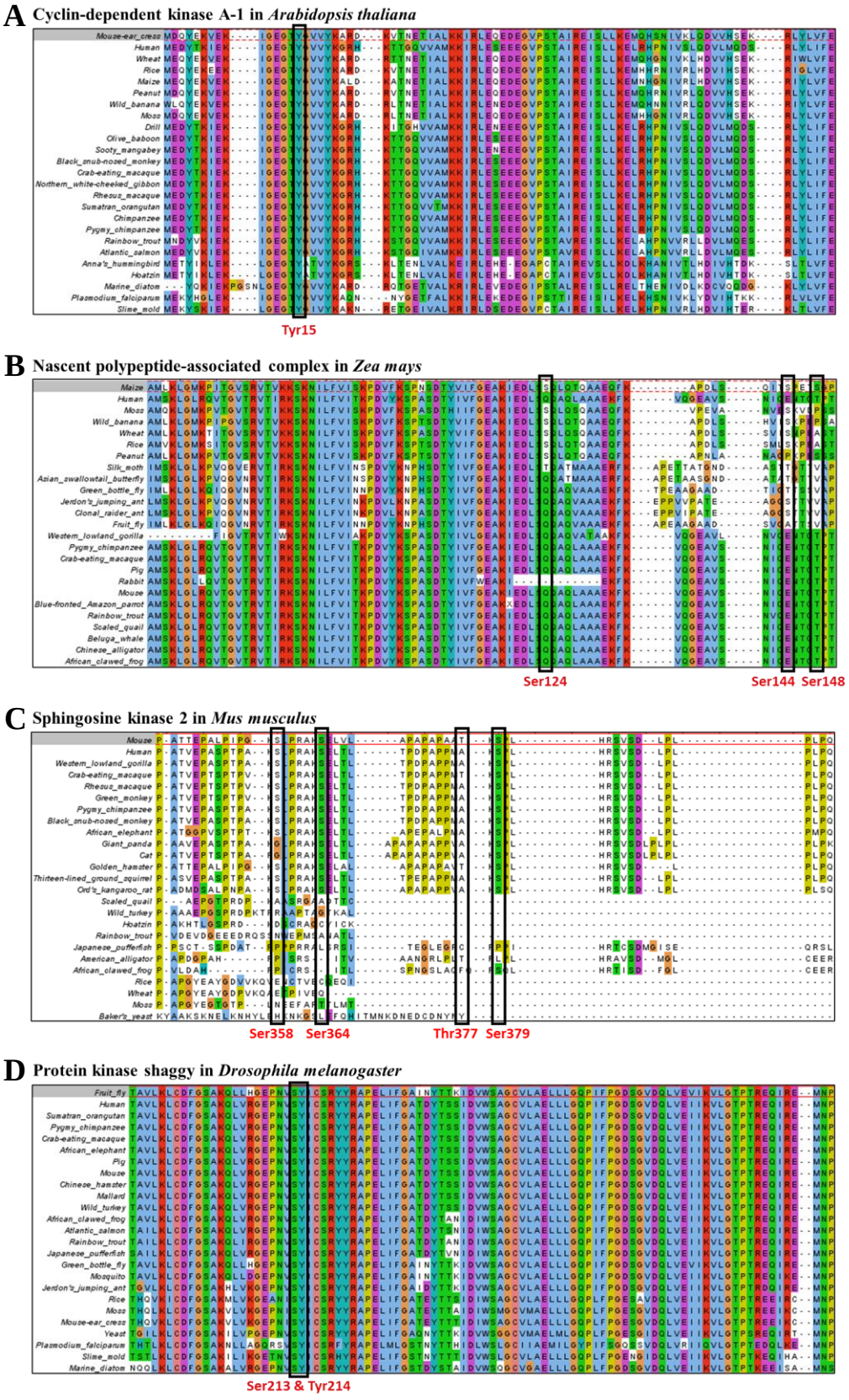


Figure 6. Sections of multiple sequence alignments of protein targets and their phosphorylated sites from **(A)** *Arabidopsis thaliana*, **(B)** *Zea mays*, **(C)** *Mus musculus* and **(D)** *Drosophila melanogaster* aligned with top-matched sequences from other species and analysed by conservation pipeline. For each alignment, the sequence of the target protein is located at the top. Phosphorylated amino acids are marked by black rectangles and their location within the sequence is provided.

2.6 Conclusion

In this Chapter, we developed a publicly available Python pipeline which can be used to study evolutionary conservation of target sites within a selected set of protein targets across multiple species. We demonstrated that the pipeline is robust, easy-to-use and can be readily applied to analyse thousands of protein targets at once from different species, something which is not easily permissible by the existing computational methods of conservation analysis such as ConSurf[156] or Ensembl Compara[159]. There is also no limit on the number of species which can be included, although naturally increasing the number of species would decrease the speed of the BLAST searches. In addition, the pipeline is supported by a guide written as "*README.md"* file containing detailed installation and running instructions, explanations of any inputs and outputs, a troubleshooting guide and links to example results. Finally, any user-defined parameters can be specified through "*configurations.ini*" file, thus ensuring that no prior knowledge of Python programming is needed to run the pipeline.

The pipeline has been designed so that users can analyse the conservation of any amino acids in any target protein sequences across any selected species. However, it is important to note that the pipeline was specifically tailored to study canonical protein targets from reference proteomes. If a target protein is not in the reference proteome of the species it originates from (i.e., the "*proteomes.fasta*" file), then it is still possible to analyse the conservation of its sites without affecting the core functionality of the pipeline by manually adding its FASTA sequence to the "*proteomes.fasta*" file. Furthermore, the pipeline was designed to analyse a single selected amino acid per run. To overcome this and analyse the conservation of additional amino acids without having to re-run the whole pipeline, a separate code is available ("*conservation_code_aln_section_with_config_ver250922.py*") in the GitHub repository which allows to process resulting multiple sequence alignments from the main pipeline and determine the conservation of any additional selected amino acids. The code is applied in a similar way to the main pipeline as described in the "*README.md*" file. In conclusion, for each query protein sequence, our conservation pipeline generates useful outputs such as BLAST results, FASTA sequences of top BLAST hits from each species, MSAs and percentage conservation of target sites and their adjacent sites at -1 and +1 positions. Therefore, the pipeline is ideal for studying the evolutionary conservation of any biological sites of interest, such as post-translational modifications and their short encompassing motifs. High efficiency of the pipeline is further highlighted in Chapters 3 and 4 where it is applied to study the conservation of all known human phosphosites across multiple eukaryotic species.

Chapter 3

Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation

The work described in Chapter 3 has been previously reviewed and published in Journal of Proteome Research, where A. Kalyuzhnyy is the lead author.

Citation: ***Kalyuzhnyy, A****., Eyers, P. A., Eyers, C. E., Bowler-Barnett, E., Martin, M. J., Sun, Z., Deutsch, E. W., Jones, A. R. (2022). Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation. J Proteome Res.* ***21****(6), 1510-1524.*

3.1 Abstract

Public phosphorylation databases such as PhosphoSitePlus (PSP) and PeptideAtlas (PA) compile results from published papers or openly available MS data. However, there is no database-level control for false discovery of sites, likely leading to the overestimation of true phosphosites. By profiling the human phosphoproteome, we estimate the false discovery rate (FDR) of phosphosites and predict a more realistic count of true identifications. We rank sites into phosphorylation likelihood sets and analyse them in terms of conservation across 100 species, sequence properties and functional annotations. We demonstrate significant differences between the sets and develop a method for independent phosphosite FDR estimation. Remarkably, we report an estimated FDR of 84%, 98% and 82% within sets of phosphoserine (pSer), phosphothreonine (pThr) and phosphotyrosine (pTyr) sites, respectively, that are supported by only a single piece of identification evidence - the majority of sites in PSP. We estimate that around 62,000 Ser, 8,000 Thr and 12,000 Tyr phosphosites in the human proteome are likely to be true, which is lower than most published estimates. Furthermore, our analysis estimates that 86,000 Ser, 50,000 Thr and 26,000 Tyr phosphosites are likely false-positive identifications, highlighting the significant potential of false positive data to be present in phosphorylation databases.

3.2 Introduction

Protein phosphorylation is a fundamental post-translation modification (PTM) that regulates protein function and is well-studied in relation to cell signalling pathways and disease[25, 26, 206]. Huge numbers of phosphorylated peptides and sites have been reported and characterized after isolation from human cells using approaches allied to tandem mass spectrometry (LC-MS/MS), focussing primarily on the phosphorylation of canonical (established) serine (Ser), threonine (Thr) and tyrosine (Tyr) residues[29-33]. However, large numbers of non-canonical phosphorylation sites have also been annotated on proteins from a variety of sources including human cells[41]. This additional complexity highlights the ongoing requirement for careful, evidence-based phosphosite identification from mass spectrometric datasets.

Historically, the focused analysis of phosphorylation sites in proteins tended to rely on biochemical analysis including, for example, chromatography and solid-state Edman sequencing[51, 52, 207]. However, while giving confidence in phosphosite identification, such low-throughput approaches are now rare, lacking the depth of coverage needed for most large-scale studies. The dominance of MS approaches has led to the development of multiple strategies to both understand and help mitigate the high levels of phosphopeptide false discovery rate (FDR), particularly in sets of mapped peptide spectral matches (PSMs) that result from LC-MS/MS and sequence database analysis[92, 96, 208]. The goal of such approaches is to separate true identifications from false ones. Even without considering non-canonical phosphorylation (which is likely to be absent in typical phosphoproteomics pipelines due to its acid-labile nature), many confidently identified phosphopeptides possess multiple Ser, Thr or Tyr residues that could be differentially modified in a given proteolytically-generated peptide[41]. Phosphosite occupancy is variable on any given protein under different biological conditions, such that analysis of a peptide containing, for example, two Ser residues that have the potential to be phosphorylated with different dynamics could present evidence for neither, only one, or both being modified, depending on the sample studied[31, 32, 93, 209]. Many phosphorylation events are also sub-stoichiometric, possibly falling below the limit of detection of certain analyses[32, 93]. As such, careful data handling and statistical processes should be applied, either within the search engine used for peptide mapping, or in a downstream software application to calculate additional statistics, such as a local *false localisation rate* (FLR) or conversely the probability that a given site within a peptide is correct or incorrect. Software/algorithms include phosphoRS[85], Ascore[82], Andromeda's PTM Score[81] and recently released PTMProphet[87]. We have previously benchmarked the performance of some instrumental parameters and software pipelines for phosphoproteomics[97], demonstrating that there is considerable variability in how such scores map to robust statistics, such as local or global FLR, depending on the instrument fragmentation mode and resolution.

Following confident identification of phosphopeptides and localisation of given sites, data tend to be compiled from within a single study or across multiple studies (meta-analysis) to determine the extent of evidence for a given site from multiple PSMs. In general, where there are independent observations of PSMs supporting a phosphosite, it can be reasonably assumed that the evidence for a site to be real increases, although to our knowledge there are no current statistical models to calculate this phenomenon accurately. Multiple PSMs can be observed per identified phosphosite as a result of either different peptide sequences containing that site, or the same peptide sequence being detected several times[91]. There are some caveats to this logic though, as it is possible for the same PSM to be wrongly assigned to a phosphopeptide multiple times. This can occur if the correct interpretation for the spectrum had a very similar peptide sequence and identical mass to the wrongly assigned phosphopeptide[93, 210]. Although LC-MS/MS and computational analysis is generally recognised as very effective and reliable for phosphosite detection, from each study it is likely that there is some element of remaining false discovery of peptides and sites wrongly localised, depending on the applied statistical thresholds. This is particularly problematic for studies that set relatively weak thresholds for phosphosite localisation (e.g., equating to site probability >0.75) in order to maximise sensitivity – more true positives may be identified, but at the expense of very large numbers of false positives passing the threshold. A multi-centre benchmarking study highlighted some of the challenges in practice, showing considerable variability in the number of true positive, false positive and false negative sites reported across different laboratories, with particular issues arising when a peptide carried multiple phosphate groups[211]. Methods and guidelines for FLR are still evolving and not consistently applied in phosphoproteome studies, and so it is likely that most published studies contain considerable numbers of falsely localised phosphosites[57, 103, 211-213]. This can lead to overestimation of the total number of known true human phosphosites if database providers do not control for FDR across multiple datasets[100].

One such database is PhosphoSitePlus (PSP) which represents a comprehensive, manually-curated and well-cited resource containing experimentally defined PTMs primarily focusing on phosphorylation[57]. As of March 2020, PSP encompassed phosphosite evidence across 17,830 human protein sequences which are defined as canonical in UniProt[13] (i.e., representing the most prevalent protein product per gene). The evidence for phosphorylation comes from manually-curated reviews of literature primarily describing tandem MS studies and also low-throughput experiments, or from in-house MS studies[57]. Interestingly, the majority of phosphosites in PSP only have a single piece of evidence associated with their identification (i.e., there is only one study identifying the phosphosite). As mentioned in the PSP documentation itself, researchers should be cautious when accepting such sites as true positives[57]. It is possible that many users of PSP are not aware of the

need for caution when reviewing or re-using data, and we are not familiar with any previous effort to assess phosphosite FDR within PSP. A second curated proteomics resource is PeptideAtlas (PA)[103] which is a repository of tandem MS datasets that have been processed through Trans-Proteomic Pipeline to ensure high and consistent quality of phosphopeptide identifications[104]. The latest PA builds incorporate the use of the PTMProphet algorithm for phosphosite localisation where each potential phosphosite within an observed PSM is assigned a probability score between 0 and 1 of being phosphorylated[87]. As with PSP, researchers should also be careful when accepting sites in PA with only a single piece of identification evidence (i.e., a single associated PSM) as positively identified phosphosites. Instead, phosphosites that not only have multiple PSM observations in PA, but also have high phosphorylation probability scores assigned within the majority of those PSMs are most likely to be true positive identifications. In addition to PSP and PA, other databases containing data on human phosphosites include UniProt, which collates mostly manually curated phosphosites from literature, but is planned to start incorporating high-throughput derived data in later releases[13]; dbPTM, a server importing data from other resources, but currently unavailable as of July 2021[214]; and PhosphoDB containing results from a set of studies on phosphopeptides derived from multiple proteases[215]. Even with easy access to these accumulated phosphorylation site resources, to our knowledge, no estimates have been made to predict the scale of phosphosite FDR across large datasets.

In this work, by profiling the reported human phosphoproteome, we aimed to estimate the false discovery rate of phosphosites with evidence in PSP and/or PA and use these estimates to predict the count of true phosphosites within the currently explored human phosphoproteome. We categorised the sites into sets of various predicted phosphorylation likelihood based on the amount of positive identification evidence reported in PSP and PA, properties not readily available in other databases. By using orthogonal features of phosphosites assigned to these sets, such as evolutionary conservation, sequence properties and functional annotations, we aimed to demonstrate significant differences between the sets and develop an improved method for independent FDR estimation.

3.3 Method

3.3.1 Processing and categorising phosphorylation data in PSP and PA

Phosphorylation data in PeptideAtlas (PA) (2020 build)[103] was filtered to only include human Ser/Thr/Tyr sites from canonical UniProt protein sequences with at least one PSM observation (1,069,709 sites across 63,616 sequences) (Table S1). The sites were categorised according to the number of PSM observations with a certain phosphorylation probability score assigned by PTMProphet[87]. The counts of observations with a probability of >0.95 were used as positive evidence for site phosphorylation. The counts at a probability threshold of ≤0.19 were used as negative evidence in favour of a site being a non-phosphosite. The total number of PSM observations per site was considered to distinguish sites for which ≥10% of all associated PSMs had a PTM probability >0.95, from sites where a small minority (<10%) of associated PSMs had this probability. Based on this, selected confidence categories were applied to predict site phosphorylation likelihood in PA ("*High*": ≥5 positive observations which accounted for ≥10% of total observations across all probabilities; "*Medium*": ≥5 positive observations which accounted for <10% of total observations or 2-4 positive observations; "*Low*": 1 positive observation; "*Not phosphorylated*": 0 positive observations and ≥5 negative observations; "*Other*": site did not fall into any described categories). PhosphoSitePlus (PSP) data (11/03/20 build; Phosphorylation_site_dataset.gz)[57] was filtered to only include human Ser/Thr/Tyr sites from canonical protein sequences labelled by UniProt identifiers (231,607 sites across 17,830 sequences) (Table S2). The sites were ranked based on the number of times they have been characterised in low/high-throughput studies. The sum of observations across all studies was used to predict site phosphorylation likelihood in PSP ("*High*": ≥5 observations; "*Medium*": 2-4 observations; "*Low*": 1 observation).

3.3.2 Evolutionary conservation analysis

The conservation pipeline developed in Chapter 2 was applied to determine the cross-species conservation of all Ser, Thr and Tyr sites in the reference human proteome[13] which had phosphorylation evidence in PSP and PA. The human reference proteome (20,605 sequences, UniProt ID: UP000005640) and the proteomes of 100 eukaryotic species (50 mammals, 12 birds, 5 fish, 4 reptiles, 2 amphibians, 11 insects, 4 fungi, 7 plants and 5 protists; Table S3) were downloaded from UniProt (UniProt release 2019_10). Each sequence in the human proteome was used as a query in a BLASTp search (BLAST+ 2.10.0 version)[129, 130, 201] against all 100 eukaryotic proteomes. The BLAST output was processed to extract a top matching significant orthologue (*E*-value of ≤0.00001) from each species for each human target. Human targets were then aligned with their matched orthologues using the MUSCLE algorithm (version 3.8.31)[145] with default settings if all sequences

to be aligned were <2,000 amino acids long. If any sequences to be aligned (either the human sequence or any of the orthologue sequences) were ≥2,000 amino acids long, 2 iterations of the algorithm were run using settings for large alignments (`--maxiters 2` option)[145]. From the alignments, percentage conservation scores were calculated for every Ser, Thr and Tyr site within each human target out of 100 (all eukaryotic proteomes) and out of the number of aligned orthologues. Conservation percentages were calculated considering any Ser/Thr substitutions in orthologues, whereby an orthologue was included in the count if, for example, a Thr in its sequence was aligned with a Ser in the target human sequence and vice versa. Conservation data was then cross-referenced with PSP/PA datasets to identify sites in the human proteome with phosphorylation evidence in PSP/PA and determine their conservation. To ensure consistency in terms of proteins and sites used, any human protein target for which it was not possible to calculate site conservation either due to the protein having no matches in BLAST (14 proteins), no significant matches in BLAST (236 proteins), no Ser/Thr/Tyr sites in its sequence (1 protein) or due to failed alignments (10 proteins), was excluded from any further analysis (Table S4). Any human targets labelled with the same UniProt identifier in the reference human proteome, PSP and PA, but which corresponded to different protein sequences across the datasets (73 proteins; Table S4) were also excluded. Conservation was assessed for the remaining targets (Table S5) by linear regression models with non-assumed intercept for simpler interpretation of slope between phosphosites and non-phosphosites. Average conservation of likely phosphosites (sites ranked "*High*" or "*Medium*" in PSP and/or PA) was plotted against average conservation of likely non-phosphosites (sites in "*Not phosphorylated*" and "*Other*" sets) within each target protein that had at least 3 likely phosphosites and 3 likely non-phosphosites. Conservation scores (%) were also compared across all sites within phosphorylation likelihood sets using box plots.

3.3.3 Analysis of amino acids adjacent to phosphosites

Target protein sequences (20,271 sequences; Table S6) were processed to identify amino acids at the -1 and +1 proximal positions adjacent to every Ser, Thr and Tyr site. If a target sequence ended with a Ser, Thr or Tyr site then its +1 amino acid was marked as "*Not found*". For each amino acid, its frequency at each proximal position was first normalised to 1,000 and then to its frequency in the pre-filtered human reference proteome (expected distribution). Proximal amino acid frequencies around target Ser, Thr and Tyr in "*High in PSP and PA*" set were compared to those in the "*Not phosphorylated*" set, and to the expected amino acid distribution. The comparisons were assessed by Fisher's exact statistical test[164] performed using scipy module in Python[196] with Bonferroni corrections to generate adjusted p-values. For each amino acid, any significant difference (Bonferroni corrected p-value <0.001) between the compared sets was used to estimate phosphosite false discovery rate across all phosphorylation likelihood sets. FDR estimates assumed that all sites in the

highest phosphorylation likelihood set "*High in PSP and PA*" set were true positive phosphosite identifications, whereas all sites with the weakest phosphorylation confidence (either the "*Not phosphorylated*" or the "*Other*" set) were non-phosphosites. Therefore, the observed count of a certain proximal amino acid in the "*High in PSP and PA*" (nPos) corresponded to its expected count at 0% FDR, whereas its observed count in the "*Not phosphorylated*" or "*Other*" set (nNeg) corresponded to its expected count at 100% FDR. To estimate % FDR in any other phosphorylation likelihood set based on the observed count of the compared proximal amino acid in that set (nObs), we used the following equation:

$$\% \text{ FDR} = \left(1 - \frac{\text{nObs} - \text{nNeg}}{\text{nPos} - \text{nNeg}}\right) \times 100$$

The equation has the effect of estimating what proportion of the observed count (nObs) is explained by assumed false positives (nNeg) and what proportion by true positives (nPos). For example, if amino acid X was found at +1 position next to 500 Ser sites in the highest phosphorylation confidence set (0% FDR set; nPos = 500) compared to 10 Ser sites in the "*Not phosphorylated*" set (100% FDR set; nNeg = 10), and next to 350 sites in the set of interest (nObs = 350), then pSer FDR within the set of interest would be 31%. This would suggest that 31% of sites in that set behave like false positive pSer in terms of X amino acid frequency at +1 position, whereas 69% of those sites behave like sites in the highest phosphorylation likelihood set (true pSer). An average FDR with 95% confidence intervals (CI) was calculated per each likelihood set using all significantly enriched amino acids around a particular target phosphosite and which had an enrichment of >1.5 relative to the expected distribution. Final FDR estimates were used to derive the total number of true positive (TP) phosphosite identifications across phosphorylation likelihood sets.

To compare FDR/TP estimates between individual PSP and PA sets, the method was replicated using alternative phosphorylation likelihood sets, where sites were categorised according to the highest amount of positive phosphorylation evidence from one database (at least one observation at PTM probability >0.95 in PA; at least one observation in PSP), without taking into account any evidence in the other. Phosphosite FDR estimates within "*High*" sets in each database were presented as a weighted average between FDR estimates in sites ranked "*High*" in that database only and sites ranked "*High*" in both PSP and PA. For example, the FDR in "*High in PA*" set was a weighted average of FDR estimates in "*High in both*" set and "*High in PA only*" set.

To analyse phosphosites in UniProt, phosphorylation data for the reference human proteome was downloaded directly from UniProt (June 2021 version; release 2021_04) and processed to split phosphosites according to associated evidence codes from the Evidence and Conclusion Ontology (ECO:0007744 - combinatorial computational and experimental evidence imported from large-scale

experiments; ECO:0000269 - manually annotated experimental evidence; ECO:0000250 – similarity evidence based on orthologous sequence). Any target proteins removed earlier (Table S4) were also removed from this analysis. The resulting protein sequences (*n*=9,481) and sets of Ser, Thr and Tyr phosphosites were analysed in terms of adjacent amino acids and conservation using the above method. Phosphosite FDR was calculated for the large-scale study set (ECO:0007744) using "*High in PSP and PA*" as 0% FDR set and "*Not phosphorylated*" as the 100% set.

3.3.4 Functional enrichment analysis

All protein sequences in the filtered reference human proteome (Table S6) were categorised into sets according to what their highest ranked Ser, Thr and Tyr site was in terms of phosphorylation evidence ("*High in PSP and PA*", "*High in PSP or PA*", "*Medium in PSP and/or PA*", "*Low in PSP and/or PA*", "*Other in PA*", "*Not phosphorylated"* and "*No evidence in PSP or PA*"). Each protein set within Ser, Thr and Tyr datasets was analysed with DAVID (version 6.8)[163] using all proteins in filtered proteome with any Ser, Thr or Tyr evidence in PSP or PA (16,296, 14,565 and 12,912 proteins respectively) as control background. Protein sets containing no reported evidence in PSP or PA were searched against a background of all proteins in the filtered reference proteome to determine any differences in their functional enrichment compared to proteins with PSP/PA evidence. Per each set searched, the top 10 (where possible) significant (Benjamini–Hochberg corrected p-value <0.05) functional terms with the highest percentage of proteins mapped were identified, replacing any near synonymous terms with additional terms from outside the initial top 10. All target protein sets were also searched in UniProt (release 2020_04) to determine percentage of proteins mapped to UniProt keywords "*Phosphoprotein*", "*Alternative splicing*", "*Nucleus*", "*Transcription*", "*Acetylation*", "*Membrane*", "*Glycoprotein*", "*Signal*" and "*Disulfide bond*".

3.3.5 Secondary structure analysis

Categorised Ser, Thr and Tyr sites in filtered reference human proteome were mapped to protein structures (beta strand, helix, turn and coiled coil) described for those proteins in UniProt (release 2020_04) (Table S5; Table S7). Any target proteins searched in UniProt which were marked as obsolete (15 proteins) or represented different sequences despite being labelled with the same identifier (25 proteins) were removed from further and marked as "*NA*" (Table S5). Normalised (to 1,000) counts of target amino acids within protein structures were assessed with Fisher's exact statistical test[164] using the scipy module in Python[196] to generate p-values and indicate any significant enrichment (p <0.05) between "*High in PSP and PA*" set and the "*Not phosphorylated*" set. The method was also applied separately for Ser sites with phosphorylation evidence in UniProt, and which were mapped to the described phosphorylation likelihood sets based on PSP/PA evidence.

3.4 Results and Discussion

3.4.1 Categorising all Ser, Thr and Tyr annotated phosphosites in the human proteome

We first ranked all Ser, Thr and Tyr phosphosites in PA and PSP in the filtered reference human proteome according to the amount of accumulated identification evidence (Fig. 7; Table 3; Table S5). The majority of Ser, Thr and Tyr sites (50.1%, 63.3% and 54.3% respectively) with phosphorylation evidence in PSP were placed into the "*Low*" phosphorylation likelihood set, meaning that there was only a single piece of evidence supporting their positive identification (Fig. 7A). Furthermore, out of all analysed Ser, Thr and Tyr sites with at least one observation at PTM probability >0.95 in PA (suggesting a positive phosphosite identification), 21.7%, 34.0% and 33.5% respectively were placed in the "*Low*" set (Fig. 7B; Table 3), highlighting that a considerable amount of potential phosphosites only had one piece of positive identification evidence across both databases. Interestingly, we found that in the human proteome there were more Tyr sites assigned to "*High*" set in PSP (5+ observations) than Thr sites (Fig. 7A; Table 3), indicating a higher initial proportion of pTyr compared to pThr in PSP. High prevalence of likely true Tyr phosphosites in the PSP dataset could have been a result of in-house studies which identified large numbers of pTyr sites using immunoaffinity strategies not suitable for pSer/pThr discovery[70, 216], and studies which have not been officially published[57].

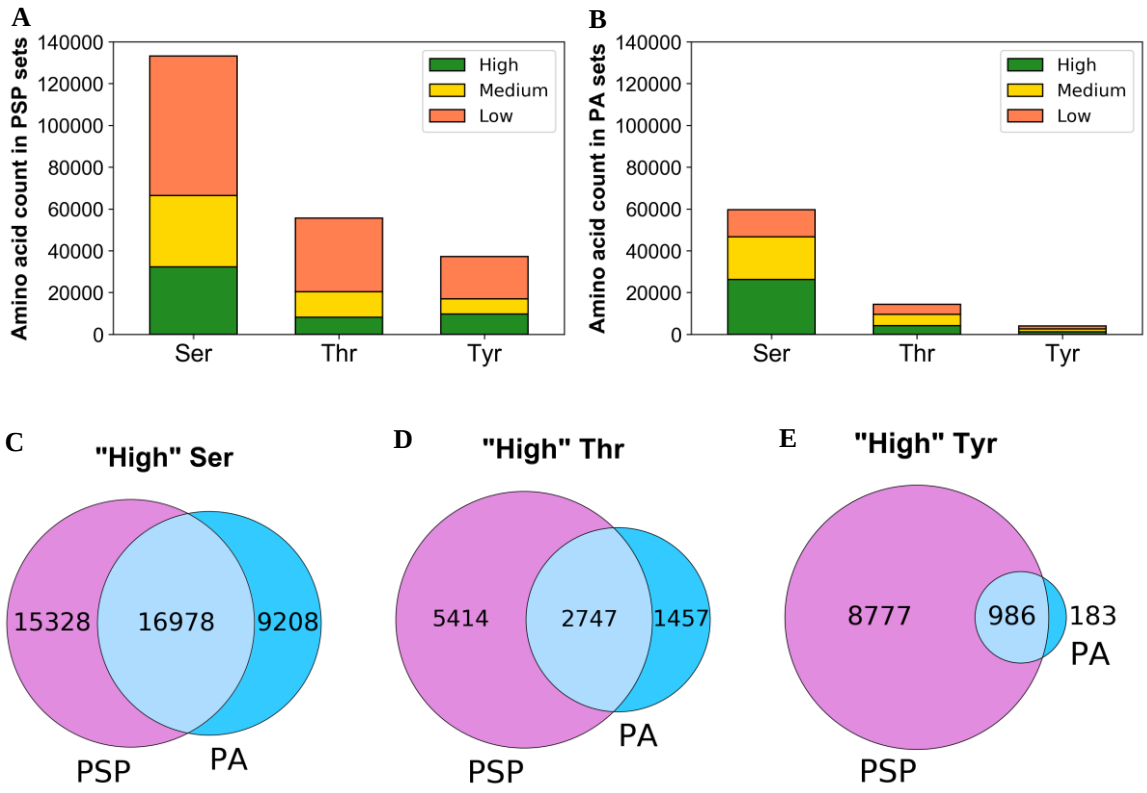


Figure 7. Distribution of serine (Ser), threonine (Thr) and tyrosine (Tyr) phosphosites from UniProt's reference human proteome that have any positive identification evidence in **(A)** PhosphoSitePlus (PSP) or **(B)** PeptideAtlas (PA) based on established phosphorylation likelihood sets (see "*Methods*"). Venn diagrams provide the counts of **(C)** Ser, **(D)** Thr and **(E)** Tyr sites ranked "*High*" in PSP (left), PA (right) and both resources (overlap).

From PA, it is possible to identify sites for which covering phosphopeptides are observed, but for which the modifications are only localised to other sites in the same peptides, thus providing strong evidence for likely non-phosphosites. Sets of potential Ser, Thr and Tyr non-phosphosites were therefore established based initially on evidence in PA (Table S8). Those sets were then cross-referenced with data in PSP to determine whether PSP contained any sites ranked as non-phosphosites in PA. Interestingly, we found that 2,489 Ser, 1,341 Thr and 891 Tyr sites assigned to the "*Not phosphorylated*" set in PA were found to have evidence in PSP (Table S9). In fact, out of those potential PA non-phosphosites, 146 Ser, 97 Thr and 293 Tyr sites were placed into "*High*" phosphorylation likelihood set according to PSP evidence (Table S9). This strongly indicated the presence of potential false positives in PSP and/or false negatives in PA. For example, Ser42 in protein P17066 (HSPA6) and Ser59 in Q8N488 (RYBP) had 8 and 6 phosphosite identification references in PSP respectively (mostly from in-house MS studies) but had no positive identification evidence in PA or any mention in UniProt[13] (Table S5). On the other hand, Ser4 in P15927 (RPA2) had 33 phosphosite identification references in PSP and was also mentioned in UniProt's annotations, but has never been positively localised in any of its 127 associated PSMs in PA (Table S5). To eliminate potential false assignments when considering evidence in both PSP and PA, a site was only categorised as a non-phosphosite if it had no evidence in PSP in addition to having negative phosphorylation evidence in PA (Table 3). As a result, we established final negative control sets containing 13,892 Ser, 8,462 Thr and 2,184 Tyr sites. Similar adjustments were made to the "*Other*" PA set (sites in that set must have no evidence in PSP) which contained the majority of analysed PA sites (Table 3).

Table 3. Categorising serine (Ser), threonine (Thr) and tyrosine (Tyr) sites from UniProt's reference human proteome into phosphorylation likelihood sets based on available phosphorylation evidence in PhosphoSitePlus (PSP) and PeptideAtlas (PA).

Phosphorylation likelihood set	Phosphorylation evidence per site	Ser count	Thr count	Tyr count
High in PSP	5+ pieces of evidence	32,306	8,161	9,763
Medium in PSP	2-4 pieces of evidence	34,154	12,197	7,228
Low in PSP	1 piece of evidence	66,777	35,173	20,191
High in PA	5+ observations at PTM score >0.95 which is ≥10% of total observations	26,186	4,204	1,169
Medium in PA	5+ observations at PTM score >0.95 which is <10% of total observations OR 2-4 observations at PTM score >0.95	20,517	5,297	1,460
Low in PA	1 observation at PTM score >0.95	12,950	4,895	1,324
Not phosphorylated	0 observations at PTM score >0.19 AND 5+ observations at PTM score ≤0.19 AND no evidence in PSP	13,892	8,462	2,184
Other sites	At least 1 observation in PA but does not fall into any other PA categories AND no evidence in PSP	60,221	35,000	10,009

Having further cross-referenced sets of sites of various phosphorylation likelihood between PSP and PA (Table S9), we established a "*gold standard*" set of phosphosites, all of which had "*High*" phosphorylation likelihood according to both PSP and PA evidence (Table S10). This set contained 16,978 Ser, 2,747 Thr and 986 Tyr highly likely true phosphosites (Fig. 7C-E; Table S10). As for the general agreement between PSP and PA in terms of phosphorylation evidence, we found that 37.7% of Ser, 20.5% of Thr and 9.10% of Tyr sites with PSP evidence also had at least one observation at PTM probability >0.95 in PA (Table S9). This variation in phosphosites observed between the two databases can be explained by the likely use of different methods for phosphosite detection and localisation between PA and the sources referenced in PSP, as well as due to a considerable presence of random false positives in both datasets before thresholding has been applied.

3.4.2 Evolutionary conservation analysis

Phosphoproteomes from all species are constantly evolving, although many ancient phosphosites are conserved across species and taxa, increasing the likelihood of them being functionally relevant[126, 217, 218]. In our analysis, we determined the conservation of all potential Ser, Thr and Tyr phosphosites and non-phosphosites in UniProt's filtered reference human proteome across 100 eukaryotic species (Table S5), weighed towards vertebrates, but also including examples of insects, plants and unicellular eukaryotes (Table S3). In our first analysis, we explored the mean conservation of phosphosites and non-phosphosites per protein (at least three of each present per protein) and performed a correlation analysis across all proteins (Fig. 8). We fitted linear regression models through the origin, under the theory that proteins unique to humans would have zero conservation for both phosphosites and non-phosphosites. We found great variation between the conservation of both site types, ranging from near zero to 100%, which was mostly dependent on the overall conservation of the protein sequence. However, based on the generated linear regression models, we concluded that on average, Ser, Thr and Tyr phosphosites ("*High*" or "*Medium*" in PSP and/or PA) were around 4.6%, 5.4%, and 2.0% respectively more conserved across all 100 eukaryotes than corresponding likely non-phosphosites (sites in "*Not phosphorylated*" and "*Other*" sets) within analysed proteins when allowing Ser/Thr substitutions towards the conservation score (Fig. 8). Similar results were obtained when assessing phosphosite conservation only across found orthologues for each protein (Fig. S1). The results (Fig. 8; Fig. S1) provide additional evidence that phosphosites are generally more conserved than non-phosphosites[126, 217, 219]. The difference in conservation is thus subtle and variable, but statistically robust. Furthermore, in our analysed sets of proteins which had at least 3 likely phosphosites and 3 likely non-phosphosites, we found 104, 88 and 19 proteins where conservation of Ser, Thr and Tyr likely phosphosites respectively was at least 20% higher than conservation of likely non-phosphosites (Table S11).

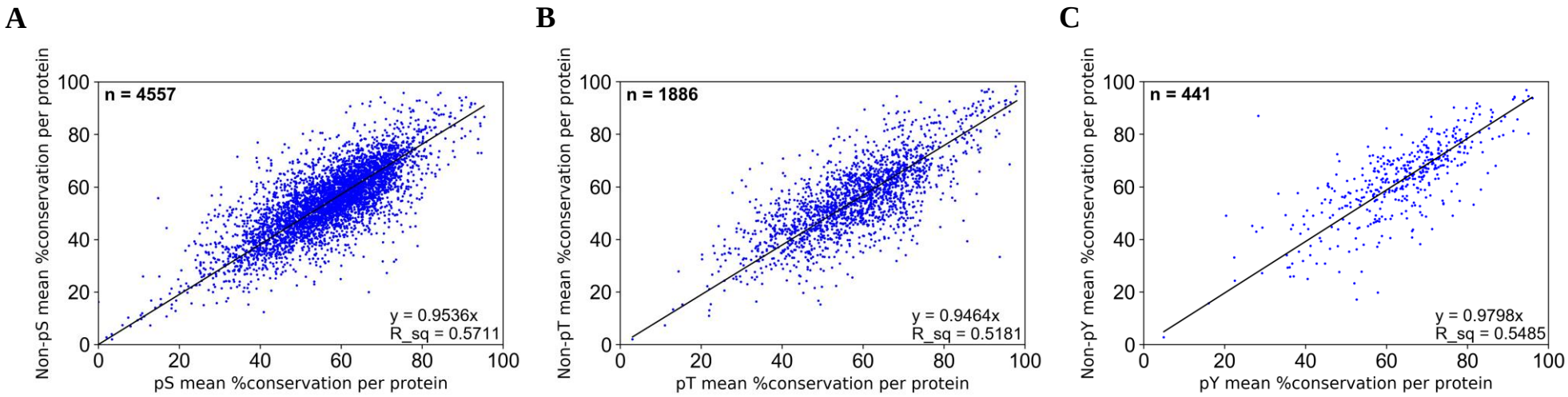


Figure 8. Mean % conservation across 100 eukaryotic species of likely **(A)** Ser, **(B)** Thr, **(C)** Tyr phosphosites and corresponding likely non-phosphosites within each target protein (*n*=number of proteins analysed). The regression coefficient (R^2) is given by "*R_sq"*.

We next compared the conservation of all sites split by phosphorylation likelihood sets (Fig. 9) and observed that sites in the highest phosphorylation likelihood set ("*High in both PSP and PA*") had the highest average conservation across all 100 eukaryotic proteomes considering Ser/Thr substitutions (average conservation of 58.4%, 58.6% and 69.4% across 16,978 Ser, 2,747 Thr and 986 Tyr sites respectively) (Fig. 9; Table S12). In comparison, the sites in "*Low in PSP and/or PA*" set had slightly lower average conservation scores of 54.3%, 55.4% and 64.0% in 35,126 Ser, 13,253 Thr and 7,471 Tyr sites respectively (Fig. 9; Table S12). Assuming that high conservation is a property of true phosphosites, that property was observed more frequently in higher phosphorylation likelihood sets compared to lower ones suggesting higher FDR in sets with less phosphorylation evidence.

There were numerous cases in our analysis of likely non-phosphosites and sites with "*Low*" phosphorylation likelihood where amino acid conservation was also high compared to likely phosphosites, indicative of a conserved function for these amino acids in, for example, catalysis or a biomolecular interaction that is unrelated to phosphorylation. Furthermore, we found 64, 30 and 6 proteins in which the average conservation across 100 eukaryotes of Ser, Thr and Tyr likely non-phosphosites respectively was at least 20% higher than the conservation of corresponding likely phosphosites (Table S11). It is possible that the predicted phosphosites within those proteins were either false positives or were non-functional true phosphosites, explaining the comparative weaker selective pressure. In fact, previous reports estimated that as many as 65% of known phosphosites may be non-functional as individual sites (although may have a more general structural role) due to limited kinase specificity and therefore have similar evolution rates compared to non-phosphosites which would explain the observed trends[114, 115]. It is also possible that some proteins were formed by recent gene fusion events leading to regions containing phosphorylation sites only found in a few closer related orthologues (low conservation), with other protein domains being more highly conserved. In addition, higher evolutionary rates in closely related species (primates, for example) could lead to new protein functions unique to that group of species, further explaining low conservation of some phosphosites in our analysis. We further note that Tyr sites in the highest phosphorylation likelihood set ("*High in PSP and PA*") had a higher mean conservation (69.4%) compared to Ser/Thr sites in that set (58.4% and 58.6% respectively) (Fig. 9; Table S12). There are several possible explanations for this result, including the idea that pTyr is under stronger conservation pressure (i.e., mutations cannot easily be tolerated) in animals which make up the vast majority (84/100) of the species analysed (Table S3). It is also possible that there is a degree of experimental bias due to the comparison of the much larger set of pSer/pThr to pTyr. The typically higher data quality for pTyr, enhanced by the availability of epitope-specific monoclonal antibodies may also contribute to this phenomenon.

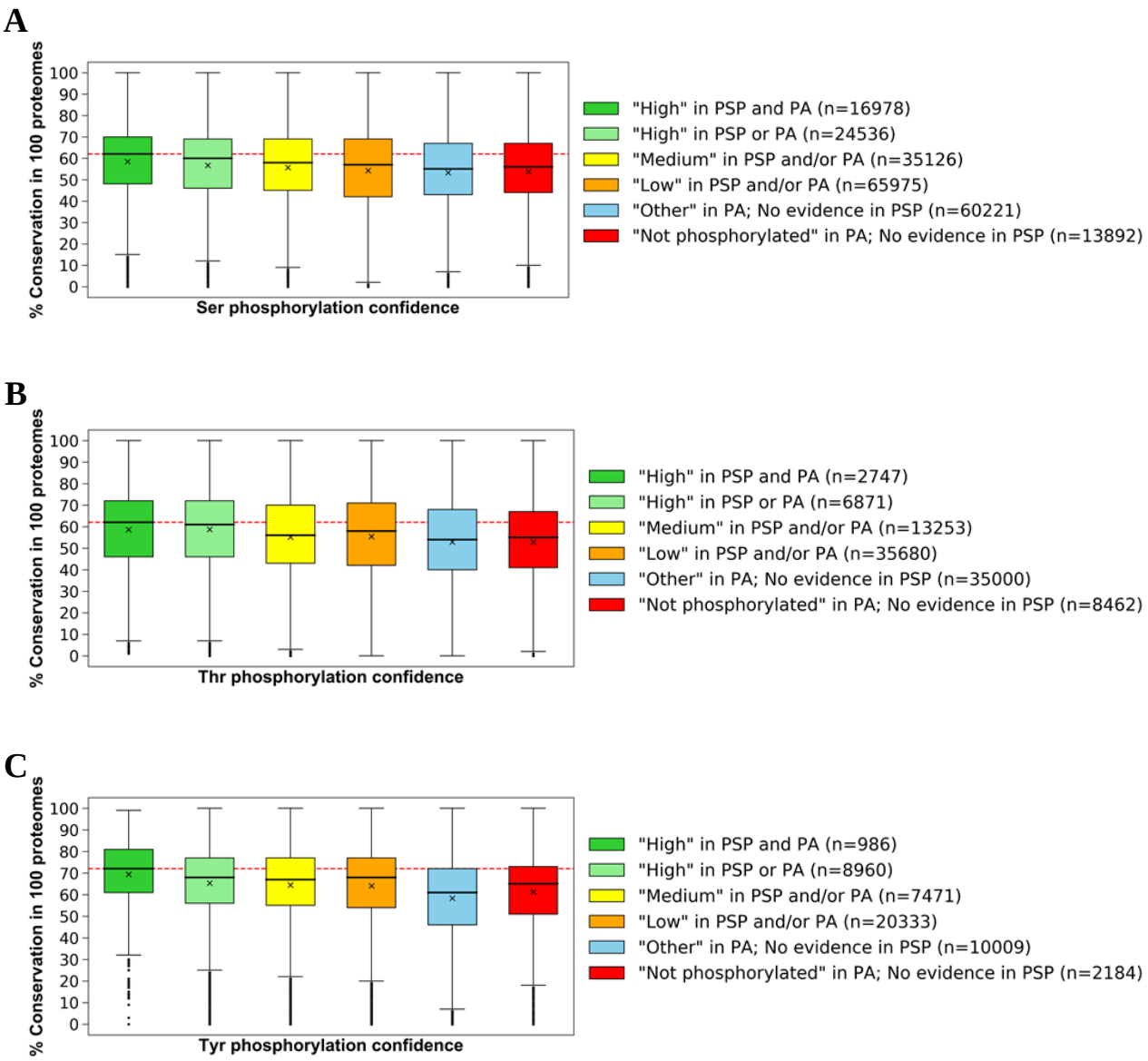


Figure 9. Box plots of conservation percentages (%) across 100 eukaryotic species of human **(A)** Ser, **(B)** Thr and **(C)** Tyr sites categorised across established phosphorylation confidence sets based on PSP and PA evidence. Within each box, a horizontal line represents median % conservation, an (x) symbol represents mean % conservation per group. Each box extends from the 25th to the 75th percentile of each set's distribution of conservation % values. Vertical lines extending from the boxes correspond to adjacent values. Dots (•) represent outlier values. Red line shows median % conservation in "*High in PSP and PA*" set for visual comparison.

3.4.3 Analysis of amino acids adjacent to phosphosites

Amino acids directly adjacent to known phosphorylation sites are often involved in optimising substrate capture for subsequent phospho-transfer by the kinase enzymatic machinery[220-222]. Multiple reports specifically highlight the importance of proline (Pro) in the mechanism of phosphorylation for families of kinases such as the cyclin-dependent kinases, mitogen-activated protein kinases and, more recently, the centrosomal kinase PLK4[220, 223-228]. Consequently, there is a high prevalence of Pro in numerous phosphorylation motif sequences as part of Ser/Thr-Pro combinations[90, 229].

In our analysis, we identified the frequency of -1 and +1 amino acids relative to a possible phosphosite and compared it across different sets of sites ranked by the relative strength of phosphorylation evidence in Table 3. We found a strong enrichment of Pro at the +1 position next to Ser and Thr sites in the reference human proteome that were placed in the set with the most phosphorylation evidence ("*High in PSP and PA*") (Fig. 10A, B; Table S13). In fact, Pro was observed at the +1 position next to 44.3% and 74.9% of all Ser and Thr sites respectively in that set (Table S13). The enrichment of Pro at +1 position around those sites was significant (adj. p-value <0.001) in relation to the normalised distribution of Pro in the human proteome, where it is, in fact, only the sixth most observed amino acid (Table S13). The normalised number of observations of Pro at +1 relative to Ser and Thr sites in the highest phosphorylation likelihood set was also significantly (adj. p-value <0.001) higher than around Ser/Thr sites in the "*Not phosphorylated*" set (Fig. 10A, B), where only 2.68% of Ser and 5.67% of Thr sites had Pro at +1 position (Table S13). Therefore, the enrichment of Pro around highly likely Ser and Thr phosphosites suggests that this feature, amongst others, can be used as a differentiating characteristic for phosphosites compared to non-phosphosites.

We also found a significant enrichment of Asp at +1 position next to Ser sites in the highest phosphorylation likelihood set (Fig. 10A). To explain this, we linked the sequences containing those sites to phosphorylation motifs which commonly feature Ser-Asp combinations, including those phosphorylated by Casein kinase II[90, 230]. At the -1 positions around target Ser, we found significant enrichment (adj. p-value <0.001) of Asp and Gly in the highest phosphorylation likelihood set compared to "*Not phosphorylated*" set (Fig. 10D). It is possible that the observed enrichment was due to the presence of those amino acids within substrate motifs of Casein Kinase II, CDK5, PKC and MEKK[90], suggesting high prevalence of potential true Ser phosphosites. Similar conclusions were made for the enrichment of Gly at -1 around Thr sites in the highest phosphorylation likelihood set (Fig. 10E) which was linked to possible Gly-Thr combinations within PKA, ERK1 and ERK2 kinase substrate motifs[90].

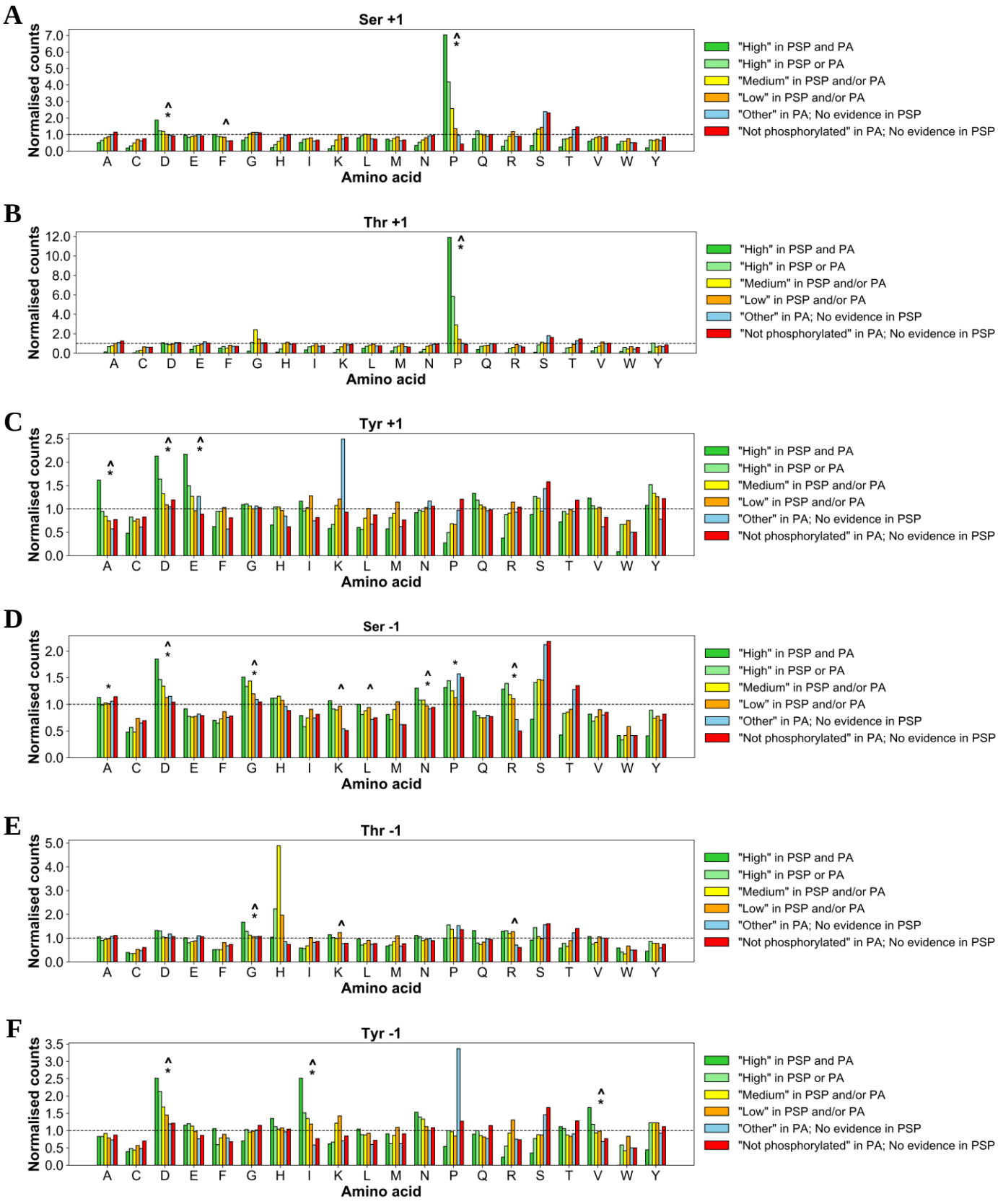


Figure 10. Counts of proximal amino acids positioned at **(A)** +1 around Ser; **(B)** +1 around Thr; **(C)** +1 around Tyr; **(D)** -1 around Ser; **(E)** -1 around Thr; **(F)** -1 around Tyr sites of various phosphorylation likelihood based on evidence in PSP and PA, normalised to observed distribution of those amino acids in human proteome (represented by dotted baseline fixed at 1). Significant (Bonferroni corrected p-value <0.001) enrichment of proximal amino acids in the "*High in PSP and PA*" set is highlighted by the caret symbol (^) when compared against the "*Not phosphorylated*" set, and an asterisk symbol (*) when compared to the expected amino acid distribution.

We compared the frequency of significantly enriched amino acids (Bonferroni corrected p-value <0.001; enrichment >1.5) across the sites within different phosphorylation likelihood sets and used the comparison to estimate phosphosite false discovery rate across those sets. Using the counts of all four enriched amino acids (Asp and Pro at +1; Asp and Gly at -1) around Ser sites of various phosphorylation likelihood (Fig. 10A, D) and working under the assumption of FDR = 0% in set 1 "*High in PSP and PA*", we estimated average Ser phosphosite FDR = 49% (CI ± 12%) in set 2 "*High in PSP or PA"*; FDR = 54% (CI ± 25%) in set 3 "*Medium in PSP and/or PA*"; FDR = 84% (CI ± 11%) in set 4 "*Low in PSP and/or PA*" and FDR = 91% (CI ± 4%) in the "*Other*" set. Similarly, by using the enrichment of Pro at +1 and Gly at -1 around target Thr sites, we estimated Thr phosphosite FDR = 59% (CI ± 7%) in set 2; FDR = 86% in set 3 (CI ± 8%); FDR = 98% (CI ± 5%) in set 4 and FDR = 99% (CI ± 1%) in the "*Other*" set (Table 4A; Table S14). Our FDR estimates clearly highlight that the majority of Ser and Thr sites with just one piece of phosphosite identification evidence are likely false positive identifications, and users of these databases can reasonably assume that if a site does not have multiple levels of evidence, then it is unlikely to represent a true phosphorylation site.

In our analysis of proximal sites around target Tyr, we found a significant enrichment (adj. p-value <0.001) of Ala, Glu and Asp at +1 positions, in addition to enriched Ile, Val and Asp at -1 in "*High in PSP and PA*" set compared to "*Not phosphorylated*" set (Fig. 10C, F). We were able to link the enrichment of those proximal sites to their possible involvement in various phosphorylation motifs including EGFR and Abl kinase substrate motifs; PTP1B and PTPRJ phosphatase substrate motifs, and multiple SH2 domain binding motifs[90, 231], therefore indicating higher frequency of true Tyr phosphosites in the highest confidence set compared to other sets. By using the frequencies of all six enriched proximal amino acids around target Tyr in "*High in PSP and PA*" (Fig. 10C, F), we estimated FDR = 49% (CI ± 9%) in set 2 "*High in PSP or PA*", FDR = 69% (CI ± 5%) in set 3 "*Medium in PSP and/or PA*", FDR = 82% (CI ± 9%) in set 4 "*Low in PSP and/or PA*", and FDR = 98% (CI ± 4%) in the "*Other*" set (Table 4A; Table S14).

Our FDR estimates varied depending on the selected enriched proximal amino acid in the highest phosphorylation likelihood set (Table S14), and thus the FDR estimates obtained with our method should be seen as approximate indicators of the extent of false positives in a set of sites with quantifiable phosphorylation evidence.

Based on our Ser, Thr and Tyr phosphosite FDR estimates, we predicted that there were around 62,000 Ser, 8,000 Thr and 12,000 Tyr true positive (TP) phosphosite identifications in the human proteome that were supported by evidence in PSP and/or PA (Table 4A). Furthermore, the results suggested that 86,000 Ser, 50,000 Thr and 26,000 Tyr sites with positive phosphorylation evidence

in PSP and/or PA (sites in "*High*", "*Medium*", "*Low*" sets) were false positives (Table 4A). Interestingly, the estimated count of Tyr TPs was higher than the count of Thr TPs which goes against the general understanding of threonine phosphorylation being more prevalent than tyrosine[31], although it is difficult to estimate the underlying true distributions, given experimental biases due to availability of different tools and methods. Our results are influenced because there are initially more Tyr sites with "*High*" or "*Medium*" phosphorylation evidence than Thr sites, particularly in PSP (Fig. 7A; Table 4A), where there has been a strong focus to identify Tyr sites using in-house methods. The ratio of count of sites that have been recorded as "*High*" in both databases is however 16,978 (pSer), 2,747 (pThr) and 986 (pTyr), following more closely previously reported estimates of phosphorylation site frequency. It thus remains to be seen if the pTyr sites reported in PSP, but without independent evidence are true or false.

Using the same method, we compared phosphosite FDR between PSP and PA sets by considering positive phosphorylation evidence ("*High*", "*Medium*" or "*Low*" sets) in one database without taking into account any evidence in the other (Fig. S2, Table S15). The analysis revealed a generally lower FDR per each set in PA compared to the respective set in PSP, overall suggesting that a higher proportion of analysed sites in PA are true phosphosites compared to the analysed sites in PSP (Table 4B; Table S15).

Table 4. Counts of estimated true positive (TP) serine (Ser), threonine (Thr) and tyrosine (Tyr) phosphosites within sets of various phosphorylation likelihood based on **(A)** combined evidence and **(B)** individual positive identification evidence in PhosphoSitePlus (PSP) or PeptideAtlas (PA). Per each set, TP counts were derived from the FDR estimates within the set and the overall count of target amino acids in the set.

A

Phosphorylation likelihood	Ser count	Ser % FDR (95% CI)	Ser TP count	Thr count	Thr % FDR (95% CI)	Thr TP count	Tyr count	Tyr % FDR (95% CI)	Tyr TP count
High in PSP and PA	16,978	0	16,978	2,747	0	2,747	986	0	986
High in PSP or PA	24,536	49 (± 12)	12,513	6,871	59 (± 7)	2,817	8,960	49 (± 9)	4,570
Medium in PSP and/or PA	35,126	54 (± 25)	16,158	13,253	86 (± 8)	1,855	7,471	69 (± 5)	2,316
Low in PSP and/or PA	65,975	84 (± 11)	10,556	35,680	98 (± 5)	714	20,333	82 (± 9)	3,660
Other in PA; No evidence in PSP	60,221	91 (± 4)	5,420	35,000	99 (± 1)	350	10,009	98 (± 4)	200
Not phosphorylated in PA; No evidence in PSP	13,892	100	0	8,462	100	0	2,184	100	0
Total excl. "Not phosphorylated"	202,836		61,625	93,551		8,483	47,759		11,732

B

Phosphorylation likelihood	Ser count	Ser % FDR (95% CI)	Ser TP count	Thr count	Thr % FDR (95% CI)	Thr TP count	Tyr count	Tyr % FDR (95% CI)	Tyr TP count
High in PA	26,186	9 (± 7)	23,829	4,204	7 (± 1)	3,910	1,169	4 (± 3)	1,122
High in PSP	32,306	29 (± 4)	22,937	8,161	46 (± 7)	4,407	9,763	44 (± 8)	5,467
Medium in PA	20,517	42 (± 18)	11,900	5,297	70 (± 25)	1,589	1,460	44 (± 8)	818
Medium in PSP	34,154	49 (± 28)	17,419	12,197	83 (± 9)	2,073	7,228	60 (± 11)	2,891
Low in PA	12,950	45 (± 32)	7,123	4,895	88 (± 19)	587	1,324	57 (± 15)	569
Low in PSP	66,777	81 (± 11)	12,688	35,173	97 (± 5)	1,055	20,191	71 (± 11)	5,855
Total in PA	59,653		42,852	14,396		6,086	3,953		2,509
Total in PSP	133,237		53,044	55,531		7,535	37,182		14,213

As noted in the "*Introduction*", manually curated evidence for phosphorylation sites is also collated in UniProt. However, while this resource provides information pertaining to the publication providing this evidence, the numbers of individual observations are not reported, preventing a matched analysis being performed with PSP/PA. Nevertheless, we were able to extract all phosphorylation data from the human reference proteome in UniProt and separate phosphosites into sets according to the type of manually curated phosphosite evidence (experimental evidence, combinatorial computational and experimental evidence from large-scale experiments, sequence similarity with an orthologous protein). As before, the sets were analysed in terms of adjacent amino acids around target phosphosites (Fig. S3A-F). Due to potential phosphosite differences and biases associated with different discovery methods (motif frequency, for example), we suggest that our method should only be used to analyse sites from high-throughput studies because it was built primarily using sites of similar evidence type. This is further evident from the conservation analysis of UniProt sites (Fig. S3H, I) which revealed different conservation patterns between the set of sites identified by large-scale studies and the other UniProt sets. As a result, we were able to estimate FDR for a set of UniProt sites with evidence from large-scale proteomics studies (Fig. S3G). In that set, we estimated average pSer FDR = 7% (CI ± 8%); pThr FDR = 22% (CI ± 14%) and pTyr FDR = 6% (CI ± 7%) (Fig. S3G), suggesting that there is a much higher proportion of true positive phosphosites in UniProt compared to PSP or PA datasets. The FDR difference between pSer and pThr follows the statistical expectation from analyses of large data sets with unbalanced counts of true positives for different residues. For example, if a study reported 1,200 phosphosites at 5% FDR, of which 1,000 are pSer and 200 are pThr, the false positives (~60) would assort approximately equally across pSer (~30 out of 1,000 i.e. 3% FDR on pSer) and pThr (~30 out of 200 i.e. 15% on pThr), meaning that in the vast majority of studies (which do not correct for this issue) the general pThr FDR will be significantly higher than for pSer. For pTyr, the majority of sites comes from separate studies that specifically enrich for pTyr via antibodies, which likely accounts for the pTyr FDR being similar to the pSer FDR.

3.4.4 Functional enrichment analysis

In our analysis, we categorised all 20,271 proteins in the filtered human reference proteome (Table S5) according to what their highest ranked Ser, Thr and Tyr site was based on phosphorylation likelihood sets in Table 3. The resulting sets (Table S16) were analysed in DAVID[163] to compare functional enrichment patterns between phosphorylation likelihood sets. First, we found that across all datasets (Ser, Thr and Tyr) the protein sets containing sites ranked "*High in both PSP and PA*" were associated with the most significant (Benjamini–Hochberg adj. p-value <0.05) functional groups (Fig. S4) suggesting their functional coherence i.e., sharing mappings to keywords, ontology terms or pathways. Interestingly, proteins with sites from "*Low in PSP and/or PA*" set as their highest

ranked site and proteins which did not have any evidence phosphorylation evidence ("*No evidence in PSP or PA*" set) were also enriched for numerous functional categories suggesting that they too share some functional properties (Fig. S4). Proteins containing sites from the "*Not phosphorylated*" set as their highest ranked Ser/Thr/Tyr site were enriched for one significant functional group in the case of Tyr dataset and no functional groups in the case of Ser/Thr datasets, which was likely due to small protein sample size in those sets.

To investigate this further, we compared the top 10 enriched functional groups between the protein sets and found that proteins containing Ser, Thr and Tyr sites with most phosphorylation evidence ("*High in PSP and PA*" set) were significantly enriched for categories and terms associated with phosphorylation such as "*Phosphoprotein*", "*Transcription*", "*Nucleus*" and "*Alternative splicing*" (Fig. 11) suggesting that those proteins were true phosphoproteins. There is a risk of generating circular evidence here, as the enriched term "*Phosphoprotein*" is a UniProt keyword, and will have been annotated based on literature evidence, potentially shared with PSP. UniProt does not yet load phosphorylation evidence from high-throughput data sets, and so classifications of phosphoproteins are generally independent of evidence used in PA. Other enriched keywords have also likely been determined based on independent evidence, and thus we believe are unbiased observations of our sets. Overall, 92.3%, 93.9% and 88.2% of proteins containing Ser, Thr and Tyr sites of the highest phosphorylation likelihood respectively were enriched for the term "*Phosphoprotein*", which, as per description in UniProt, is a term assigned to a "*protein which is post-translationally modified by the attachment of either a single phosphate group, or of a complex molecule, such as 5'-phospho-DNA, through a phosphate group*"[13]. Furthermore, those proteins were enriched for "*Acetylation*" (Fig. 11) which in some cases might indicate phosphorylation since crosstalk between acetylation and phosphorylation has been frequently reported[232, 233], alongside other modifications such as O-glycosylation[234]. Another enriched function is "*Alternative splicing*" (Fig. 11) which is known to be controlled by reversible phosphorylation[235], further indicating that those proteins likely contain functional phosphosites. However, it is possible that this enrichment could correlate with the depth of analysis of the mentioned proteins rather than their phosphorylation likelihood, since extensively studied gene products (and abundant proteins with more easily detectable phosphosites) are likely to have better quality data associated with isoform identification and be consequently linked to "*Alternative splicing*".

In comparison, proteins that only had sites from "*Low in PSP and/or PA*" set as their highest ranked Ser, Thr and Tyr sites (i.e., proteins which did not have sites with strong phosphorylation evidence) were not enriched for clear phosphorylation-associated terms and were instead enriched for categories

such as "*Glycoprotein*", "*Signal*" and "*Disulfide bond*" and "*Membrane*" (Fig. 11), suggesting that the majority of those proteins were likely non-phosphoproteins and their associated phosphosites with weak evidence were therefore likely false positives. Assuming that sites with no phosphorylation evidence in PSP or PA are likely non-phosphosites (although it is possible that phosphorylation has not been investigated or localised yet), potential high FDR in the "*Low in PSP and/or PA*" set was further supported by proteins with no phosphorylation evidence being enriched for similar functional groups (Fig. 11). In fact, we observed a clear decrease in the proportion of proteins enriched for phosphorylation-associated functional groups (where a set was enriched for at least 10 functional groups) going across our established sets suggesting higher phosphosite FDR in lower confidence sets (Fig. S5).

Our investigation of UniProt terms linked to protein sets revealed that the enrichment for term "*Phosphoprotein*" and other terms likely to be associated with phosphorylation ("*Alternative splicing*", "*Nucleus*", "*Acetylation*", "*Transcription*") generally decreased across the sets of reduced confidence, which suggested higher FDR in sets with less phosphorylation evidence (Fig. 12). For example, only 13.0%, 31.7% and 36.3% of all proteins, which had Ser, Thr and Tyr sites respectively from "*Low in PSP and/or PA*" phosphorylation likelihood set as their most confident site, were marked as phosphoproteins in UniProt (Fig. 12; Table S17), suggesting that most proteins in those sets were not phosphoproteins, and further highlighting that the associated sites with only a single piece of evidence are likely false positive identifications.

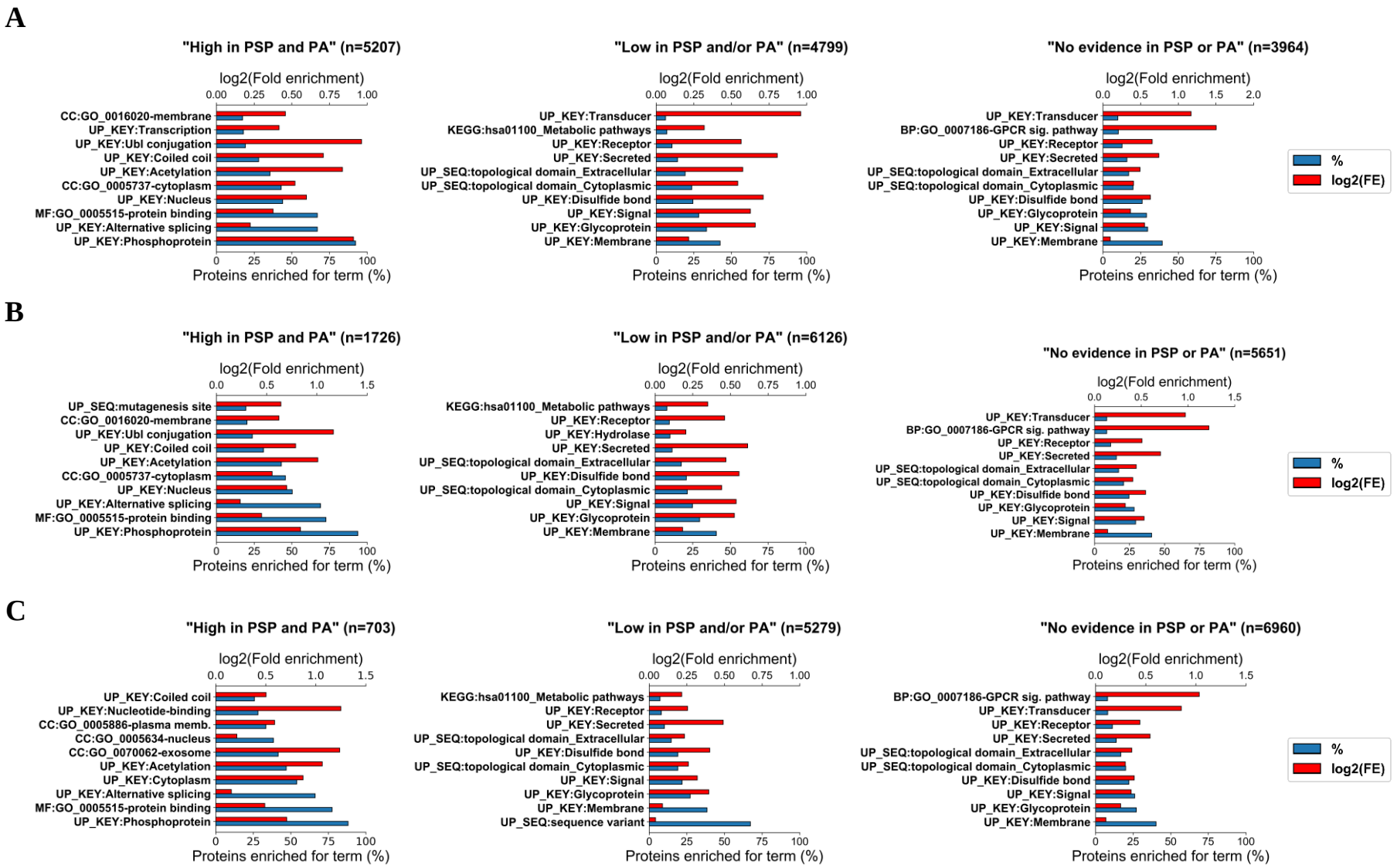


Figure 11. Top 10 functional categories for which protein sets containing various highest ranked **(A)** Ser, **(B)** Thr, **(C)** Tyr sites based on the amount of available phosphorylation evidence ("*High in PSP and PA*", "*Low in PSP and/or PA*", "*No evidence in PSP or PA*") were significantly enriched in DAVID (Benjamini–Hochberg corrected p-value <0.05). For each protein set, the % of proteins enriched for a particular functional category is given as well as the log2(fold enrichment) for that set. The number of proteins in each set is presented by *n*.

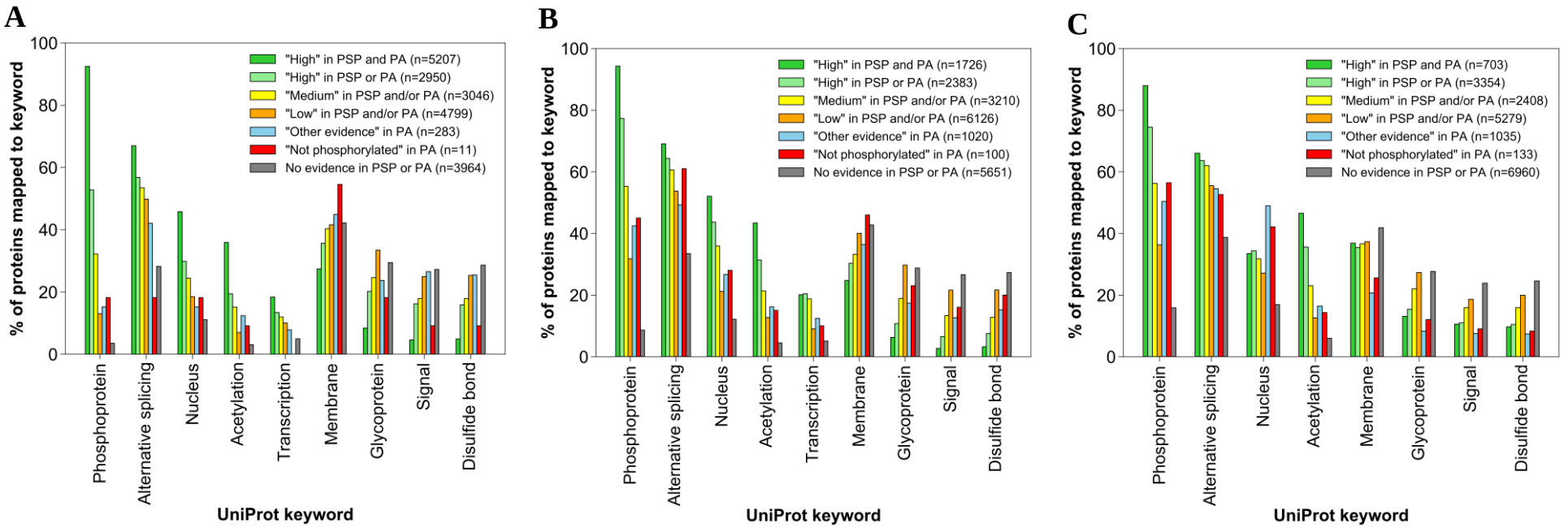


Figure 12. The percentage of proteins within sets containing **(A)** Ser, **(B)** Thr, **(C)** Tyr sites of various phosphorylation likelihood as their highest ranked site, annotated with specific UniProt keywords. The number of proteins in each set is presented by *n.*

3.4.5 Secondary structure analysis

We also investigated whether Ser, Thr and Tyr sites with strong phosphorylation evidence were located more frequently within specific protein secondary structures, when compared to sites with less evidence. For example, previous analysis of thousands of phosphosites from multiple species identified hotspots within domain families of proteins, particularly near domain interfaces and adjacent to catalytic residues, where they presumably regulate enzymatic output[113, 236]. We found that significantly more (Fisher's test p-value <0.05) Ser, Thr and Tyr sites with the strongest phosphorylation evidence ("*High in PSP and PA*" set) were localised within coiled coils compared to sites in the "*Not phosphorylated*" set (Fig. 13). This might readily be explained by coiled coils being frequently found in transcription factors, the activity or subcellular location of which is often dependent on phosphorylation[237-239]. Therefore, the results in Figure 13 further indicated that there were more potential true Ser, Thr and Tyr phosphosites in "*High in PSP and PA*" set than in other sets. In terms of other analysed protein structures (beta strand, turn, alpha helix), there was no significant enrichment of sites from the highest phosphorylation confidence set within those structures when compared to the "*Not phosphorylated*" set (Fig. 13). In fact, our current reading of the literature suggests that it is still unclear whether phosphorylation sites are found on average to be localised more or less frequently within beta strands, turns or alpha helices, though clear evidence for localisation of PTMs at functionally important loci in proteins has been previously presented[240].

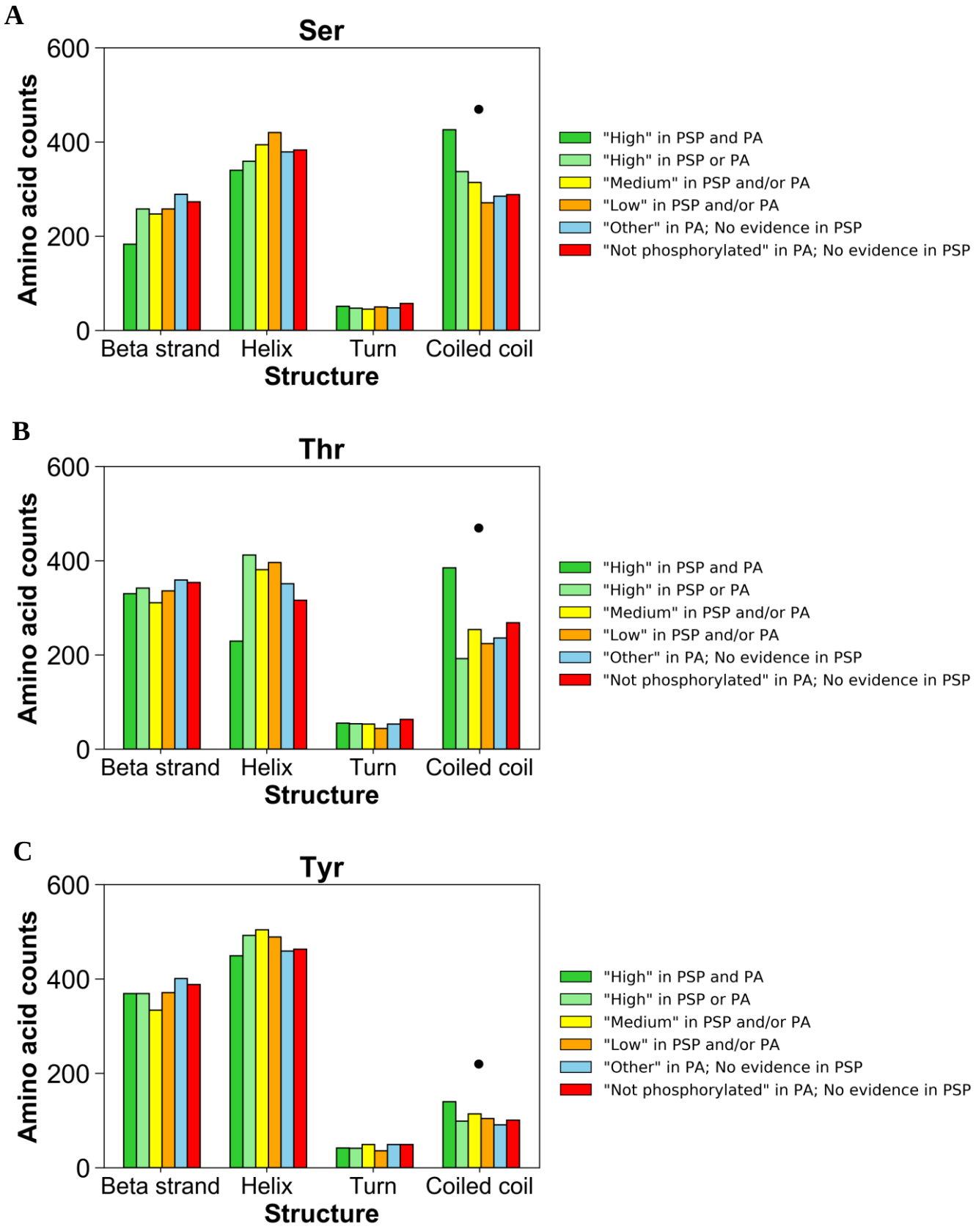


Figure 13. Normalised counts of **(A)** Ser; **(B)** Thr; **(C)** Tyr amino acids of various phosphorylation likelihood based on evidence in PSP and PA which are found within protein structures (beta strand, alpha helix, turn, coiled coil). Significant (Fisher's test p-value <0.05) enrichment of amino acids from "*High in PSP and PA*" set within protein structures is highlighted by the dot symbol (•) when compared against the "*Not phosphorylated*" set.

3.5 Conclusion

In our analysis, we ranked all potential Ser, Thr and Tyr phosphosites in UniProt reference human proteome according to how much quantitative and qualitative phosphorylation evidence they were assigned in PSP and PA databases. Having analysed the sites and the proteins that contain them in terms of conservation, proximal site patterns, functional enrichment and structural properties, we established that Ser, Thr and Tyr sites with weak phosphosite identification evidence, particularly sites with a single piece of supporting evidence, were likely to be false positive identifications. This finding was further confirmed by FDR estimations across the established phosphorylation likelihood sets which revealed phosphosite FDR of 84%, 98% and 82% in sets of Ser, Thr and Tyr sites respectively where only one piece of identification evidence was present. Since there is a considerable presence of such sites in PSP and PA datasets, our results implied high FDR in both those datasets, although PSP was predicted to have a generally higher proportion of false positive phosphosites compared to PA. This is potentially a cause for concern since many potential false positives are presented to scientists as true phosphosites, without clear explanation of the likelihood of such claims. Nevertheless, using our FDR estimates we predicted that there are around 62,000 Ser, 8,000 Thr and 12,000 Tyr true positive phosphosites in the human proteome that are supported by evidence in PSP and/or PA. These estimated counts are lower than other published estimates[57, 100, 241, 242] particularly for Ser/Thr sites, presumably due to the previous inclusion of false positives and subsequent overestimation of the number of true phosphosites. We conclude that researchers must be aware of the potential for false positive sites in both public and self-generated databases. As a general rule, phosphorylation sites with <5 independent observations should be treated with caution, and those with only one observation in a database are likely to be false positives. In a recent phosphoproteomics study from our group, we demonstrated the utility of the classification presented here, by matching the sites identified by LC-MS/MS to their evidence categories from PSP and PA[243]. For new phosphoproteomics studies, it will be common for some ambiguity to remain regarding phosphosite localisation, and many sites will be observed that would not pass a 1% or 5% false localisation rate cut-off from a single dataset, but for which there may be some supporting evidence. By evaluating new datasets in combination with all the evidence collated from the large number of previous studies, greater confidence can be assigned to "*borderline*" significant phosphosites that may indeed be correct, or conversely, sites with weak evidence that have never been reported can be rejected. Here, we have provided a methodological framework for estimating global FDR in large-scale phosphorylation data sets, which does not rely on native scores from search engines or site localisation software. Methods for estimating global FDR in meta-analyses of phosphosites are not yet robust, and thus we would recommend that other groups profile orthogonal properties of ranked sets, as we have done here, to estimate the real distribution of true and false phosphosites in their data.

Chapter 4

Discovering Evolutionary and Functional Trends of Human Phosphorylation Sites

4.1 Abstract

Protein phosphorylation is the most important and frequently observed post-translational modification which is well-studied in relation to cell signalling pathways and disease across all life. The development of high-throughput proteomics pipelines has led to the discovery of large numbers of specific phosphorylated protein motifs and sites (phosphosites) across many eukaryotic species. Despite this, the vast majority of phosphosite discoveries are made in humans on serine (Ser), threonine (Thr) and tyrosine (Tyr) amino acids, with many other species only having a few experimentally confirmed or computationally predicted phosphosites. In addition, only a small fraction of the currently characterised human phosphoproteome has an annotated functional role. A common method of predicting functionally relevant phosphosites is conservation analysis which can identify conserved protein sequence regions and infer their functional relevance. However, extensive evolutionary studies which investigate phosphosite conservation across large numbers of species are scarce. In this Chapter, we explore the conservation of human phosphosites across 100 eukaryotic species and establish various phosphosite conservation patterns within specific species groups ranging from primates and other mammals to plants, fungi and protists. We link the observed evolutionary patterns to the functional relevance of phosphosites in those groups and also investigate the evolution of protein domains that encompass the target phosphosites. We identify several protein functions regulated by phosphorylation which range from ancient functions conserved across all life (cell cycle regulation, cytoskeleton formation, stress response, etc.) to relatively new functions only conserved in primates and other closely related species to humans (neuronal development, tissue and organ formation, immune response, etc.). We clearly demonstrate the importance of conservation analysis in predicting the functional significance of phosphosites and identifying organisms which can be used as biological models to study conserved signalling pathways relevant to human biology and disease. Furthermore, we apply conservation analysis to predict over 1,000,000 potential Ser, Thr and Tyr phosphosites in the analysed eukaryotes by using human protein sequences with confident phosphosites as a reference. Our results can ultimately be used to improve proteome annotations of species with few identified phosphosites and direct further downstream research surrounding the evolution and functional relevance of phosphosites in eukaryotes.

4.2 Introduction

4.2.1 The extent of protein phosphorylation in humans and other eukaryotes

In proteomics, kinase-regulated protein phosphorylation is the most important and frequently observed post-translational modification[47] which is well-studied in relation to cell signalling pathways and disease across all life[25, 26, 206]. The extent of protein phosphorylation in various eukaryotic species has been highlighted by several genome sequencing studies which revealed around 500 kinase-encoding genes in humans[28, 244], 1,000 in *Arabidopsis thaliana*[245], 240 in *Drosophila melanogaster*[246] and 120 genes in *Saccharomyces cerevisiae*[247]. Furthermore, the development and optimisation of high-throughput proteomics pipelines such as tandem mass spectrometry (LC-MS/MS) has led to the discovery of large numbers of specific phosphorylated protein motifs and sites, focussing primarily on the phosphorylation of canonical (established) serine (Ser), threonine (Thr) and tyrosine (Tyr) residues[29-33, 90]. Newly identified phosphorylation sites (phosphosites) are characterised and compiled in numerous publicly available resources which contain post-translational modification data from humans and other eukaryotic species (Table 5).

However, as discussed in Chapter 3 and emphasised in similar studies, databases may not always account for phosphosite false discovery rate across large proteomics datasets which results in the accumulation of false positive identifications[100, 248]. Therefore, the number of "*real*" phosphosites in those resources may be lower than reported and researchers are advised to carefully evaluate the evidence behind phosphosites involved in their research. Nevertheless, our in-depth analysis of phosphorylation data from commonly used resources such as PhosphoSitePlus (PSP)[57] and PeptideAtlas (PA)[103] revealed that the count of "*real*" phosphosites in humans alone may exceed 80,000[248]. As such, phosphorylation remains an active area of research, with new phosphosites being discovered all the time.

Table 5. Examples of phosphorylation databases with eukaryotic data.

Database name (version)	Count of reported phosphorylation sites	Species covered
PhosphoSitePlus (v6.7.0.1)[57]	>350,000	Human (63%); Mouse (28%); Rat (8%); Other eukaryotes (1%)
dbPAF (Version 1.0)[249]	>450,000	Human (51%); Mouse (25%); Yeast species (10%); Rat (8%); Fruit fly (4%); Roundworm (2%)
Human Protein Reference Database (Release 9)[108]	>50,000	Human (100%)
iProteinDB (v1.0.1)[110]	>100,000	Six fruit fly species (around 17% per species)
The Plant PTM Viewer (2021 build)[109]	>300,000	Mouse-ear cress (38%); Maize (33%); Rice (13%); Wheat (12%); Algae (4%)
PhosphoGRID (2.0)[250]	>20,000	Baker's yeast (100%)

4.2.2 Predicting conservation and functional relevance of phosphosites

Despite numerous phosphosites being continuously reported and the extensive ongoing research surrounding the role of phosphorylation in proteomics, only a small fraction of the currently characterised human phosphoproteome has an annotated functional role[57, 113]. This is likely because the rate of phosphosite discovery is far greater than the rate at which each individual site or motif can be analysed experimentally. Furthermore, it has been proposed that a significant portion of phosphosites may have no regulatory function at all[114, 115]. The difficulty in distinguishing functionally significant phosphorylated regions from those that do not contribute to protein function is exacerbated by the added complexity of proteins having multiple phosphorylated sites within their sequence, as well as several kinase enzymes being able to phosphorylate multiple sites[30, 31].

A common and effective method for predicting functionally significant phosphorylated protein regions is conservation analysis. At its simplest, conservation analysis works by comparing the amino acid sequence of a protein in question to the sequences of its homologues and identifying local regions of similarity which may have a common functional implication amongst the compared proteins[116]. Therefore, identifying conserved regions between protein homologues from different species (orthologues) may predict a common function and provide an insight into its mechanism and evolution. In fact, conservation analysis of genes and proteins often plays a central role in research surrounding model organisms and how they can be used to study human biology and disease[112, 120]. This is highlighted by various studies of organisms such as flies[121], worms[122], yeast[123] and mammals[122, 124] that uncovered novel molecular pathways and demonstrated a direct functional connection of those pathways to human biology by analysing conservation of the involved proteins[120, 122].

When it comes to studying the evolution and function of phosphorylated protein regions, it is generally hypothesised that functionally significant phosphosites would be highly conserved because their mutations to non-phosphosites would alter protein function and ultimately hinder evolutionary selection[125, 126]. Several studies demonstrated that Ser, Thr and Tyr phosphosites are indeed significantly more conserved compared to non-phosphosites in general[113, 126, 213]. Our large-scale profiling analysis of the reference human phosphoproteome assessed the conservation of human phosphosites across 100 eukaryotic species and also revealed that confident phosphosites are on average more conserved than sites with no phosphorylation evidence in the same protein sequences (Fig. 8 & Fig. 9)[248]. Therefore, identifying highly conserved phosphosites and exploring their functions could broaden our understanding of biological mechanisms and their evolution.

Multiple cross-species studies of phosphosites successfully characterised their conservation and linked it to functions such as cell cycle maintenance and metabolism[172, 174, 251]. However, extensive evolutionary studies which investigate phosphosite conservation across large numbers of species are scarce[111, 113]. One of the major aims of this Chapter is to therefore provide a further insight into general evolutionary and functional trends surrounding human phosphosites. This is achieved by calculating and analysing phosphosite conservation within specific groups of eukaryotic species (vertebrates and invertebrates, mammals only, primates only, etc.) and understanding their functional relevance within those groups. We also analyse functional enrichment of human protein sets with different phosphosite conservation patterns by using clusterProfiler[168] which is an open-source, user-friendly R package that offers comprehensive analysis and visualisation of enriched functions. Additional functional annotations are mapped using DAVID online tool[163] which allows to extract functional terms from bioinformatics databases such as UniProt[13], KEGG[165], SMART[166] and InterPro[167]. Furthermore, we infer conservation patterns of prevalent human protein domains by linking phosphosite conservation to the data from the database Pfam[252] which offers accessible large-scale bulk downloads of accurate domain annotations within protein sequences.

4.2.3 Applying conservation analysis to predict phosphosites in eukaryotes

The studies of phosphorylation sites and kinases are mostly limited to a specific set of species, with the most frequent phosphosite discoveries being made in humans, followed by several model organisms such as mouse, flies, worms, yeast and Arabidopsis (Table 5). As a result, the number of reported phosphosites varies between species, with most eukaryotes having little to no evidence of either experimentally confirmed or computationally predicted phosphosites[13, 57, 111]. For example, in the January 2023 build of PhosphoSitePlus database, there are only 24 characterised eukaryotes with any reported phosphorylation data in low or high-throughput studies, where species such as ferret, starfish and buffalo only have a single known phosphosite (Table 6)[57]. The process of identifying phosphosites with LC-MS/MS and subsequent proteome annotation can be expensive and time-consuming, and its complexity also depends on the size of the proteome being annotated, as well as the availability of accurate experimental and genomic data for the species from which the proteome originates[253, 254]. As a result, phosphoproteomics studies are not attempted for the majority of the species. In fact, most experiments are designed and funded with the purpose of benefiting human life, as well as improving the understanding of human biology and diseases, which is often achieved by analysing a specific set of model organisms that have conserved functions in humans or the studies of which may bring benefit to human life[112, 120, 122].

Table 6. Approximate counts of phosphosites reported for eukaryotes in PhosphoSitePlus.

Species	Phosphosite count
Human	<240,000
Mouse	<110,000
Rat	<35,000
Cow, Chicken, Rabbit, Pig	<1,000
Dog, Hamster, Frog, Sheep, Fruit fly	<50
Goat, Horse, Quail, Monkey, Guinea pig, Turkey, Cat, Marmoset, Torpedo, Ferret, Starfish, Buffalo	<10

As discussed previously, the analysis of conserved sequence regions between proteins can predict functionally relevant motifs and sites such as phosphosites when experimental validation is not readily available. In fact, various computational tools such as ConSurf[156], ACES[157], Ensembl Compara[159], NetPhos[255] and our own computational Python pipeline described in Chapter 2 incorporate algorithms which analyse sequence conservation to predict functionally relevant sites within given protein sequences. Furthermore, some phosphosite annotations in UniProt are propagated from sequence similarity between a query sequence and a well-annotated homologous sequence with experimentally confirmed phosphosites, provided that the phosphorylated residue and the surrounding motif is conserved in the homologous sequence[13]. The propagations are usually limited between closely related species from the same taxonomic group and can be further validated if the kinase responsible for modifying the target is also conserved between the species[13]. In this Chapter, we expand the scope of phosphosite predictions using conservation analysis by mapping conserved phosphosites from the reference human proteome that have plenty of identification evidence to aligned sites in potential homologue sequences from 100 eukaryotic species.

4.2.4 Aims of the Chapter

- Expand on the conservation analysis of confident human phosphosites across 100 eukaryotic species introduced in Chapter 3 and establish phosphosite conservation patterns within specific groups of species.
- Establish functional enrichment patterns within the defined clusters of proteins with varying phosphosite conservation to understand the evolution of functions regulated by phosphosites.
- Link phosphosite conservation patterns to associated protein domains, thus providing an insight into their evolution between the analysed species and supporting the results of functional analysis.
- Analyse multiple sequence alignments between reference human proteins containing confident phosphosites and their potential homologues from 100 eukaryotic species to predict new phosphosites in those species.

4.3 Method

4.3.1 Establishing human phosphosite conservation patterns within eukaryotic species

Conservation percentage scores across 100 eukaryotic species (Table S3) for all Ser, Thr and Tyr amino acids in the reference human proteome with any phosphorylation evidence were obtained by using Python conservation pipeline developed in Chapter 2 and following the methods described in Chapter 3. To ensure that only the most confident phosphosites were utilised in this Chapter's analysis, the overall conservation data and multiple sequence alignments from Chapter 3 were extracted exclusively for our previously characterised "*gold standard*" set of phosphosites (i.e., phosphosites which had at least 5 pieces of identification evidence in PSP and PA databases). In total, the analysed set contained 16,978 Ser, 2,747 Thr and 986 Tyr sites across 5,709 proteins (Table S10).

To establish general phosphosite evolutionary patterns within eukaryotes, the data was processed to determine percentage conservation scores within specific groups of species such as primates (*n*=18), other mammals (*n*=32), birds (*n*=12), fish (*n*=5), reptiles (*n*=4), amphibians (*n*=2), insects/invertebrates (*n*=11), fungi (*n*=4), plants (*n*=7) and protists (*n*=5). In addition, conservation scores were calculated for broader groups such as animals (*n*=84), vertebrates (*n*=73) and mammals (*n*=50). To allow downstream protein-level functional analysis, the average conservation of all Ser/Thr and Tyr phosphosites across each described species group was calculated for each protein in the "*gold standard*" set (Table S10).

Taxonomic relationships between the selected 100 eukaryotic species were displayed with a phylogenetic tree built with NCBI's Taxonomy Browser tool[256] using relevant UniProt proteome ID numbers as inputs (Table S3). The resulting phylogenetic tree was annotated and visualised using MEGA (version 10.2.2)[138] and iTOL (version 5)[257]. Additional silhouette images within the resulting tree were obtained from PhyloPic database (https://www.phylopic.org/).

All proteins in the analysis (*n*=5,709) were then grouped into ten conservation clusters based on similarities in their Ser/Thr and Tyr phosphosite conservation patterns across the described species groups. The clustering was performed automatically using pheatmap package (version 1.0.12)[258] in R programming software[259], which applied the Euclidean distance method to assess similarity in phosphosite conservation patterns between target proteins and group them into specific clusters. The resulting protein clusters were presented as heatmaps and each cluster was manually named with an appropriate descriptive label which corresponded to the most observed conservation pattern within the cluster (i.e., at least 50% of proteins within the cluster had to match the label description in terms of their conservation patterns). The assigned labels characterised phosphosite conservation across species groups as "*High*" (phosphosites are ≥75% conserved within described species group) or

"*Medium*" (phosphosites are ≥50% conserved within described species group). Similar clustering analysis was performed on individual sites, where target Ser/Thr and Tyr phosphosites were grouped according to their percentage conservation across the species groups. To highlight the diversity of conservation patterns identified for target phosphosites, multiple sequence alignments for randomly selected proteins that accurately matched the description of the corresponding clusters of interest and had a characterised function in UniProt were annotated and visualised in Jalview (version 2.11.2.3)[139].

4.3.2 Functional enrichment analysis of conservation clusters

Each protein cluster was analysed with R package clusterProfiler (version 4.4.1)[168] to determine the functional enrichment of proteins with certain Ser/Thr and Tyr phosphosite conservation patterns against a control background of all analysed proteins (*n*=5,709) with at least one phosphosite from the "*gold standard*" set. For each cluster, a maximum of the 10 most enriched Gene Ontology (GO) terms were selected and visualised as dot plots, where the number of enriched proteins for a given GO term was provided and the statistical significance of the enrichment was measured with adjusted p-values. Clusters were counted as enriched for a particular GO term if their Benjamini–Hochberg adjusted p-value for the corresponding enrichment was <0.1. The functional enrichment analysis of the target protein clusters was extended by utilising DAVID online tool (version 6.8)[163] and again using all analysed proteins (*n*=5,709) as a control background. For each cluster, the top 10 (where possible) significant (Benjamini–Hochberg adjusted p-value <0.1) functional terms with the highest percentage of proteins mapped were identified, with any near synonymous terms being filtered out. In addition to mapping target clusters to GO terms, DAVID analysis was also used to determine whether the clusters were enriched for any KEGG pathways[165], UniProt keywords[13] and annotations from domain databases such as SMART[166] and InterPro[167].

4.3.3 Linking phosphosite conservation data with protein domains

In order to link protein domains which encompass target Ser, Thr and Tyr phosphosites with their conservation patterns across the species groups, domain data was extracted from Pfam database (version 35.0; November 2021 release)[252] and cross-referenced with the site-level conservation data by protein's UniProt ID tag and site position within the protein sequence. In order to identify domains for which phosphosites with specific conservation patterns were enriched, fold enrichment (i.e., enrichment factor) was calculated for each domain against a control background of all phosphosites mapped to any Pfam domains. To calculate fold enrichment, a standard equation for enrichment analysis described in various enrichment pipelines such as DAVID (Database for Annotation, Visualization and Integrated Discovery)[163] was applied as follows:

$$\textbf{Domain fold enrichment} = \frac{a/b}{A/B}$$

a = Count of phosphosites with a specific conservation pattern mapped to domain X

b = Count of all phosphosites with that specific conservation pattern

A = Count of all phosphosites mapped to domain X

B = Count of all phosphosites in the background distribution (i.e., all phosphosites mapped to any protein domain)

The domain data was then filtered to only include domains with at least 2 mapped phosphosites. The top 10 domains with the highest percentage of mapped sites from each Ser/Thr and Tyr conservation pattern were visualised and $\log_2$(fold enrichment) was presented for easier interpretation of enrichment. The resulting data describing phosphosite conservation and mapped protein domains was summarised in an Excel spreadsheet.

4.3.4 Predicting phosphosites across eukaryotes

Multiple sequence alignments between target human proteins and their potential homologues from the selected species were processed with a separate Python code to identify all amino acids in the matched sequences which were aligned with target Ser, Thr and Tyr phosphosites in the human sequences. In addition, amino acids adjacent to the aligned sites at the +1 position in protein sequence were analysed. For every species in each alignment, if both the amino acid that is aligned with the human phosphosite and its +1 adjacent site were conserved in the human sequence (considering Ser/Thr substitutions), then that amino acid was predicted to be a phosphosite. In order to validate the resulting phosphosite predictions, phosphorylation data was extracted for mouse and Arabidopsis from PSP[57] and Plant PTM Viewer[109] databases, respectively, to determine how many of the predicted phosphosites in those species had experimental phosphorylation evidence. Additional validation of our phosphosite predictions was done by assessing the likelihood of a certain site in mouse being aligned with a human phosphosite and also having strong phosphorylation evidence (>5 pieces of evidence in PSP). Finally, all resulting phosphosite predictions across 100 eukaryotic species were summarised in a convenient, easily accessible Excel file.

4.4 Results and Discussion

4.4.1 Evolutionary and functional analysis of human phosphorylation sites

In the previous Chapter, we analysed the global conservation of phosphosites from the human proteome and concluded that individual Ser, Thr and Tyr phosphosites from the "*gold standard*" set with strong identification evidence (Table S10) were on average 5%, 6% and 8% more conserved than our characterised set of corresponding non-phosphosites, respectively (Table S12)[248]. In this Chapter, we expanded on that analysis and investigated the conservation of the "*gold standard*" phosphosites (Table S10) across 100 eukaryotic species ranging from primates and other mammals to plants and fungi (Table S3, Fig. 14). By analysing average Ser/Thr and Tyr phosphosite conservation from each target human protein, we were able to split the phosphoproteins into independent clusters according to their similarity in phosphosite conservation patterns within specific groups of eukaryotes. Overall, our analysis successfully identified distinct phosphosite conservation patterns in human proteins (Table 7; Fig. 15; Table S18). For example, cluster C contained 197 proteins which had at least one Ser or Thr phosphosite conserved across all analysed species, suggesting that these phosphosites were present in early eukaryotes such as fungi and were likely involved in regulating ancient protein functions (Table 7; Fig. 15A, cluster C). In comparison, cluster D contained 603 proteins in which Ser/Thr phosphosites were only conserved in primates, indicating their relatively recent evolution and potential significance in functions which are only relevant to primates (Table 7; Fig. 15A, cluster D). Various evolutionary patterns were also identified when conservation was assessed for individual phosphosites without taking into account the proteins they were found in (Fig. S6). For example, there were 162 Tyr phosphosites which were conserved primarily in animals which indicates their evolutionary divergence from lower eukaryotes (Fig. S6, cluster N1). The diversity of the established conservation patterns was further highlighted by multiple sequence alignments of the individual human protein examples which accurately matched the described phosphosite conservation pattern within the clusters (Fig. 16).

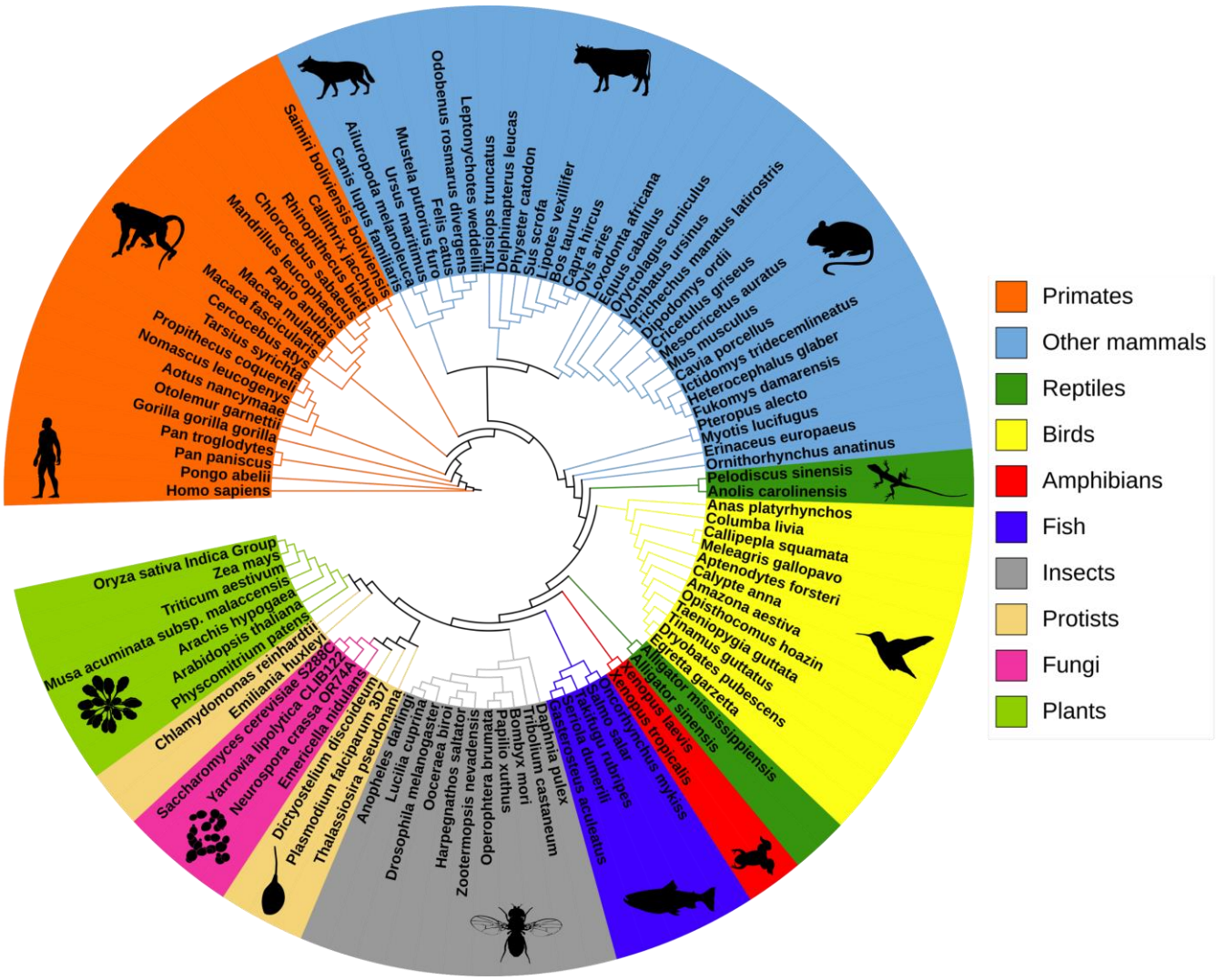
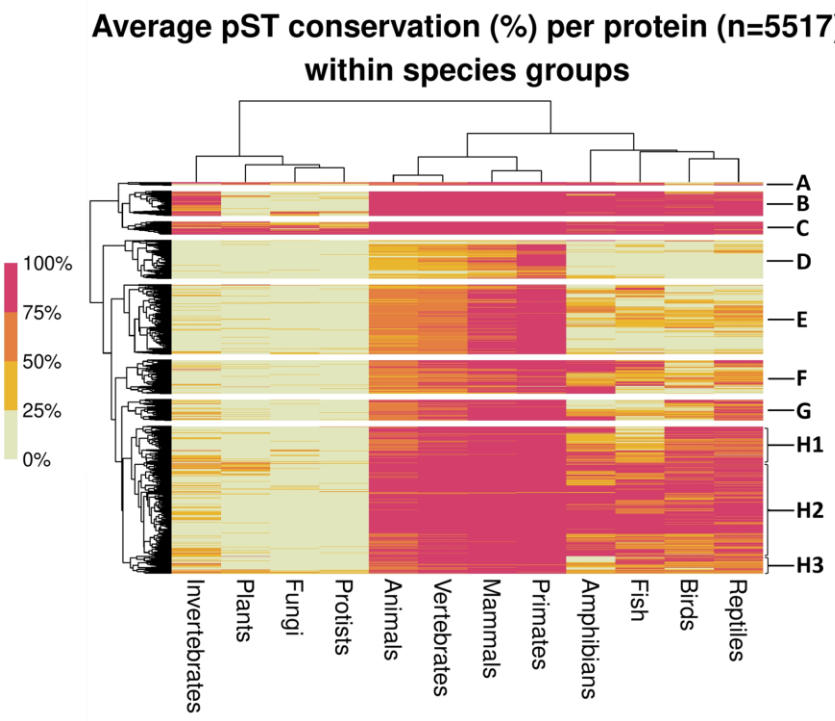


Figure 14. The phylogenetic relationship between groups of eukaryotic species involved in the conservation analysis of human phosphosites.

Table 7. Independent clusters of human phosphoproteins split according to their similarity in average Ser/Thr and Tyr phosphosite conservation patterns across established species groups. High and medium conservation indicates that phosphosites are ≥75% and ≥50% conserved across the species in each cluster, respectively. Cluster labels are also assigned for easier interpretation.

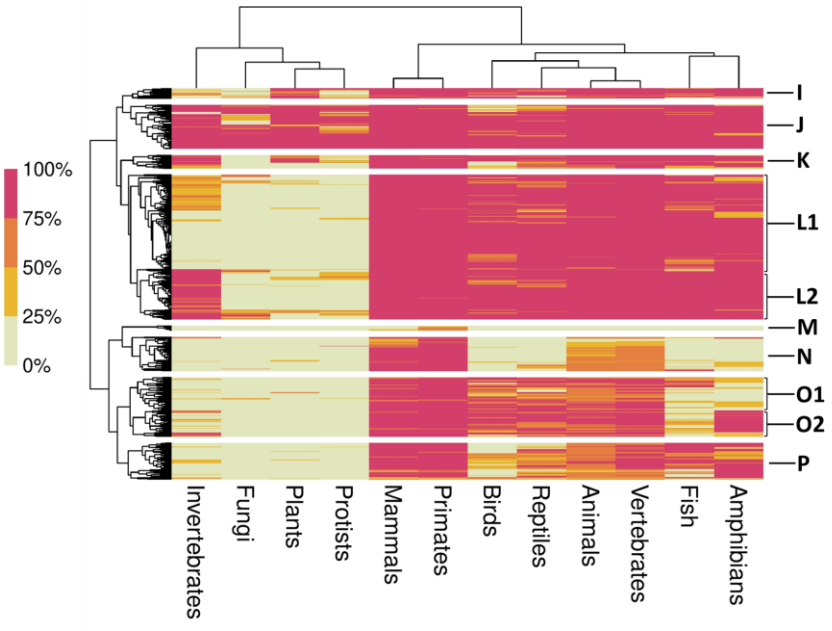
Conservation pattern	Phosphosite	Protein count	Phosphosite count	Cluster label
High in all vertebrates except birds	Ser/Thr	45	80	A
High in animals only	Ser/Thr	384	820	B
High or medium in all species	Ser/Thr	197	379	C
High in primates only	Ser/Thr	603	1,650	D
High in mammals only	Ser/Thr	1,100	4,573	E
Medium in all vertebrates except birds	Ser/Thr	528	1,827	F
High in mammals and medium in reptiles	Ser/Thr	327	993	G
High in all vertebrates except fish	Ser/Thr	837	3,636	H1
High in vertebrates only	Ser/Thr	1,222	5,041	H2
Medium in all vertebrates except amphibians	Ser/Thr	274	726	H3
High in all vertebrates and medium in plants	Tyr	21	24	I
High in all species	Tyr	89	102	J
Medium in all species except fungi and protists	Tyr	27	40	K
High in vertebrates only	Tyr	190	313	L1
High in animals only	Tyr	103	160	L2
Medium in primates only	Tyr	10	10	M
High in mammals only	Tyr	69	85	N
Medium in all vertebrates except fish and amphibians	Tyr	67	82	O1
High in all vertebrates except fish	Tyr	53	61	O2
Medium in all vertebrates except birds	Tyr	74	109	P

A



Cluster	Conservation pattern
A	High in all vertebrates except birds
B	High in animals only
C	High or Medium in all species
D	High in primates only
E	High in mammals only
F	Medium in all vertebrates except birds
G	High in mammals/Medium in reptiles
H1	High in all vertebrates except fish
H2	High in vertebrates only
H3	Medium in all vertebrates except amphibians

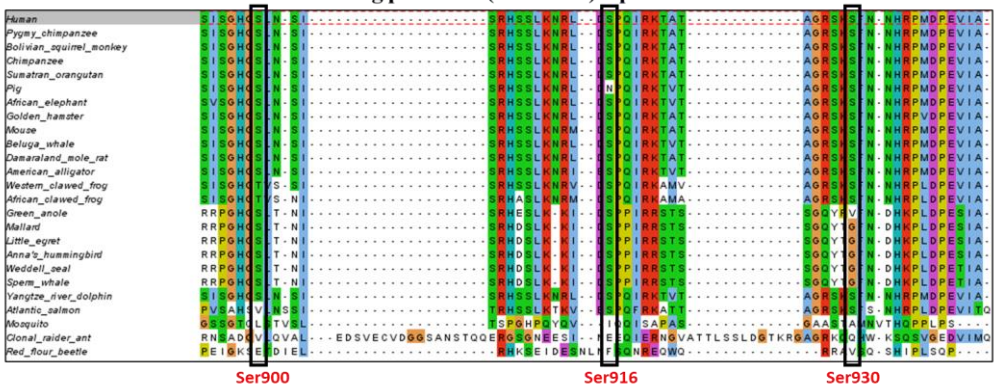
B



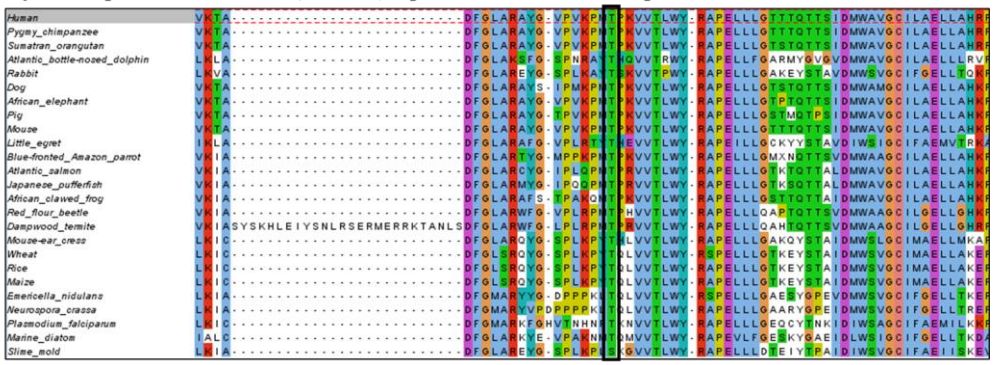
Cluster	Conservation pattern
I	High in all vertebrates/Medium in plants
J	High in all species
K	Medium in all species except fungi and protists
L1	High in vertebrates only
L2	High in animals only
M	Medium in primates only
N	High in mammals only
O1	Medium in all vertebrates except fish and amphibians
O2	High in all vertebrates except fish
P	Medium in all vertebrates except birds

Figure 15. Conservation patterns of **(A)** Ser/Thr and **(B)** Tyr phosphosites from human proteins across the groups of eukaryotic species. Each row in the heatmap represents an individual human protein and its phosphosite conservation across specific species groups which are separated into columns. Conservation is scored as a percentage out of all species per group and reflected by a colour gradient divided at quarterly intervals. The proteins were clustered based on their similarity in conservation patterns using the Euclidean distance method. For each cluster, a label is assigned which describes the most observed conservation pattern (i.e., at least 50% of proteins in the cluster follow the described phosphosite conservation pattern), where high and medium conservation refers to conservation scores of ≥75% and ≥50%, respectively. The total number of analysed proteins containing target phosphosites is given by *n*.

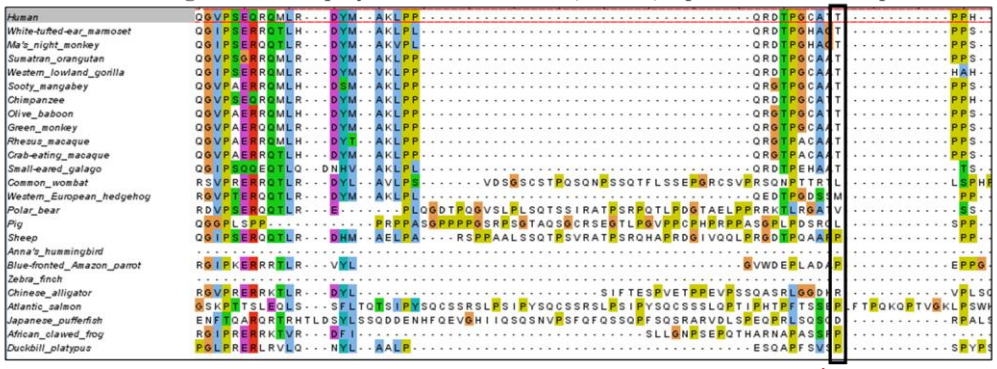
A SLIT-ROBO Rho GTPase-activating protein 2 (SRGAP2) – pSer conserved in vertebrates



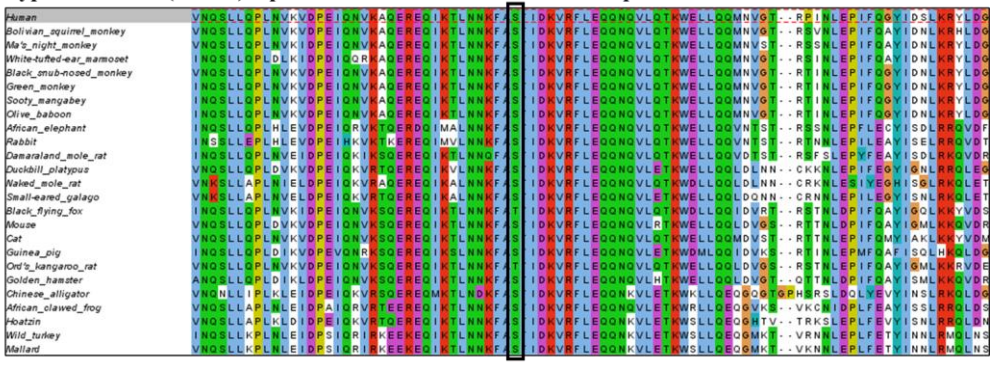
B Cyclin-dependent kinase 10 (CDK10) – pThr conserved in all species



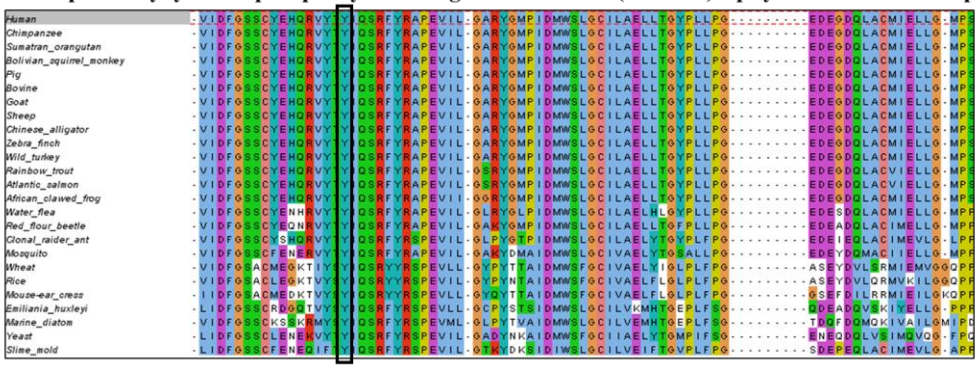
C TATA box-binding factor RNA polymerase I subunit C (TAF1C) – pThr conserved in primates



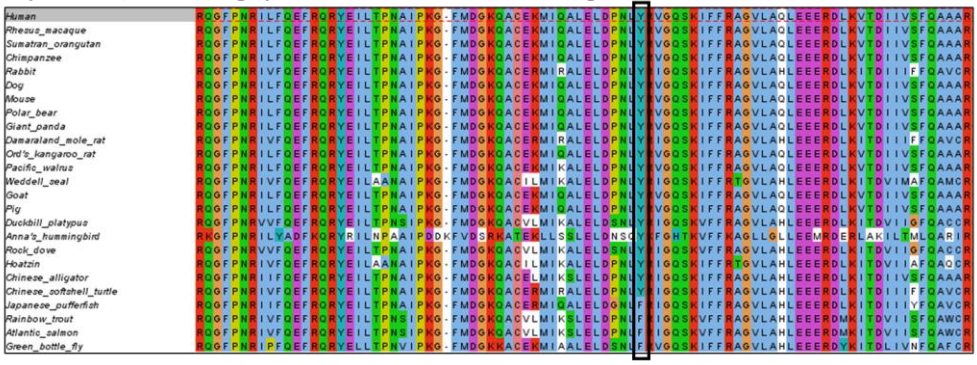
D Type II keratin (KRT2) – pSer conserved in vertebrates except fish



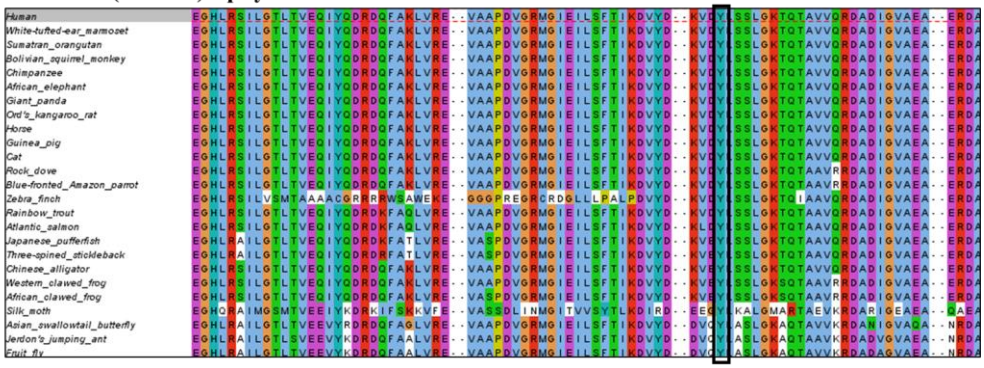
E **Dual specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2) – pTyr conserved in all species**



F **Myosin-14 (MYH14) – pTyr conserved in vertebrates except fish**



G **Flotillin-2 (FLOT2) – pTyr conserved in animals**



H **Macrophage-capping protein (CAPG) – pTyr conserved in mammals**

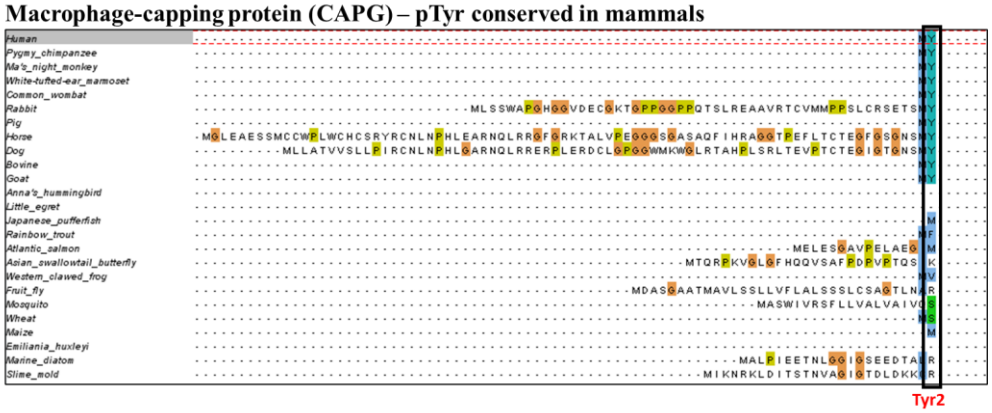
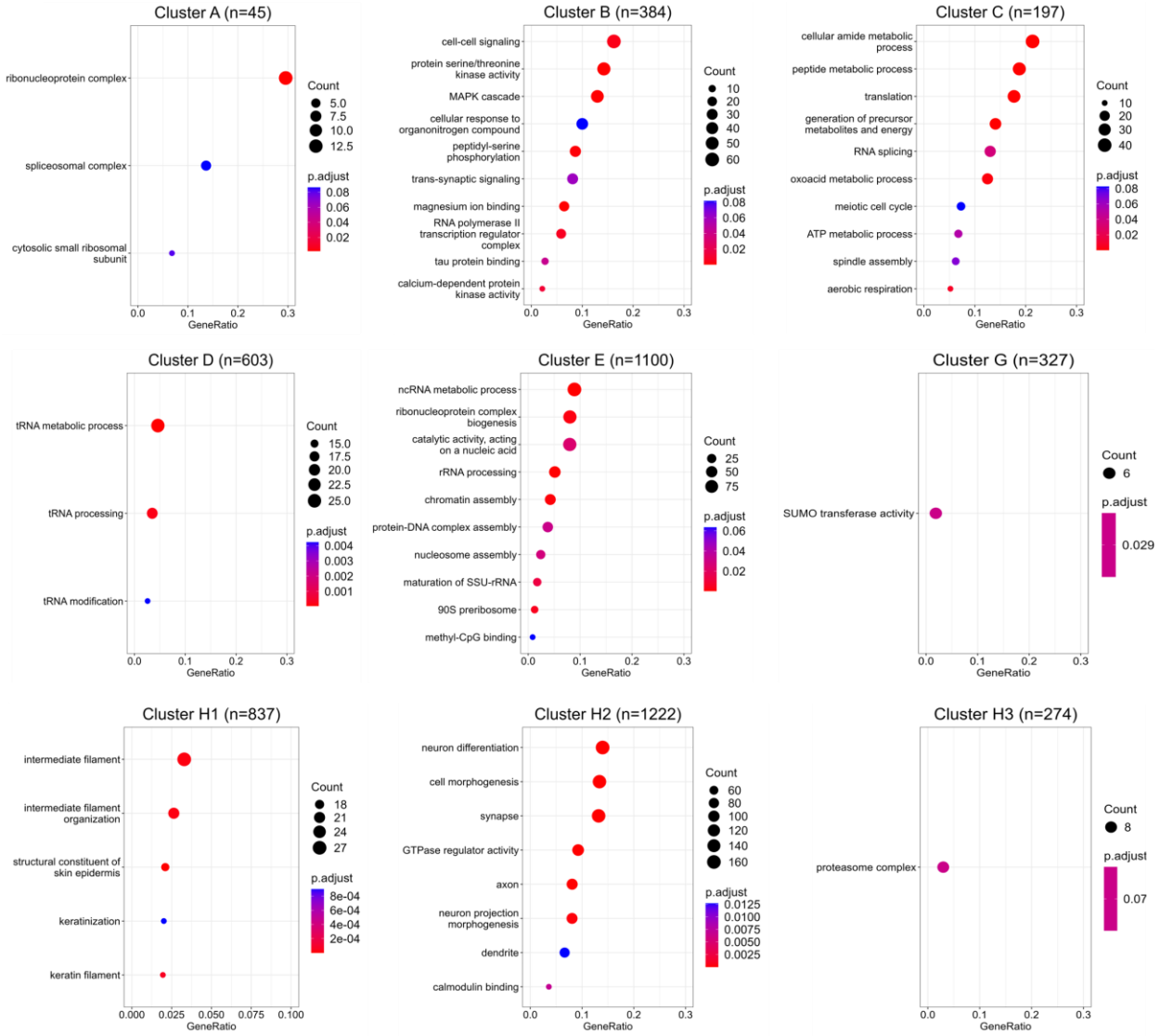


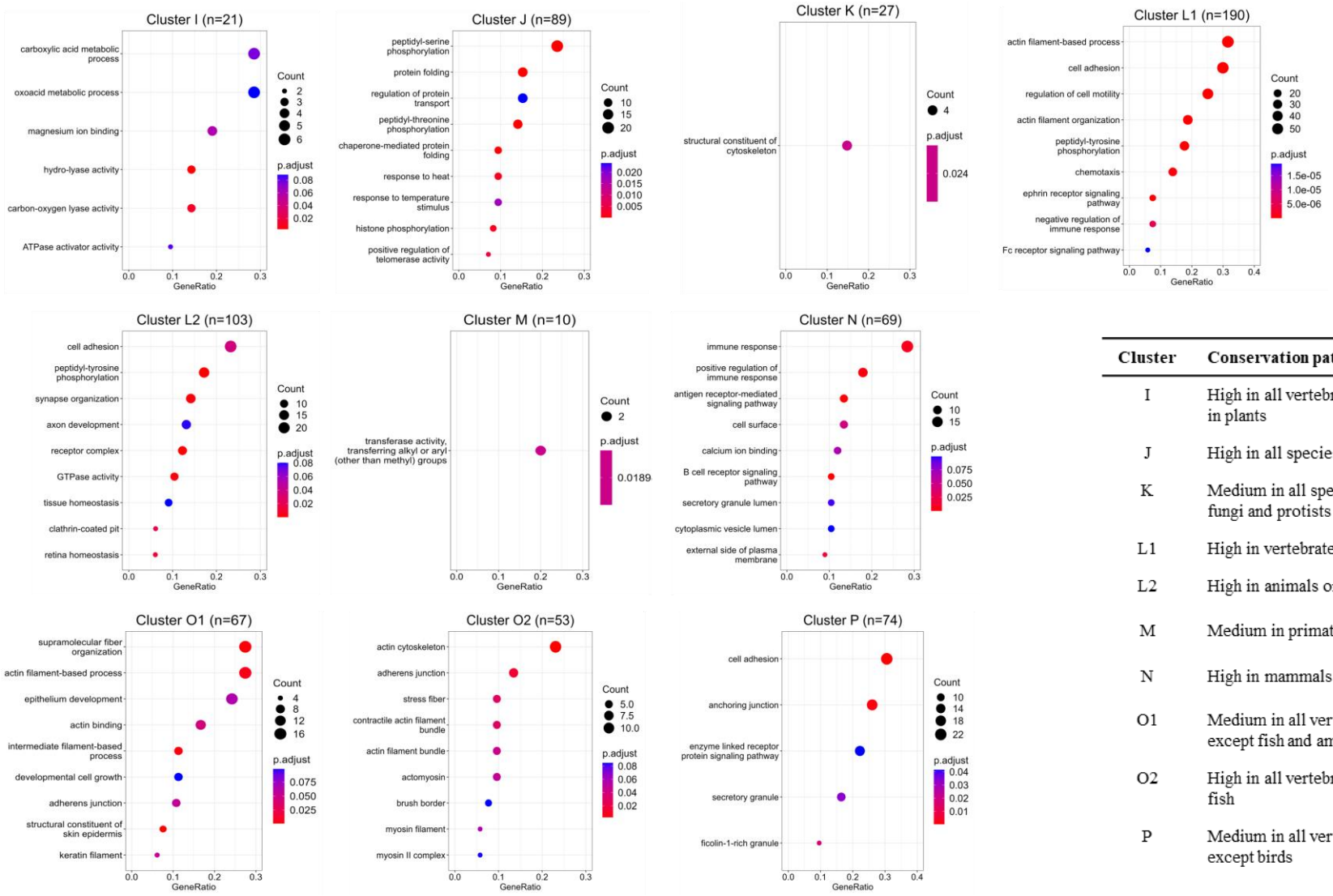
Figure 16. Sections of multiple sequence alignments of target human proteins with different conservation patterns of **(A-D)** Ser/Thr and **(E-H)** Tyr phosphosites across eukaryotic species. For each alignment, the sequence of the human protein containing the phosphosites is located at the top and the aligned protein sequences of potential orthologues from other species are given below. Phosphorylated sites are marked by black rectangle boxes and their location within the human sequence is provided underneath in red. Alignments were annotated in Jalview.

A



Cluster	Conservation pattern
A	High in all vertebrates except birds
B	High in animals only
C	High or Medium in all species
D	High in primates only
E	High in mammals only
F	Medium in all vertebrates except birds (no enrichment)
G	High in mammals/Medium in reptiles
H1	High in all vertebrates except fish
H2	High in vertebrates only
H3	Medium in all vertebrates except amphibians

B



Cluster	Conservation pattern
I	High in all vertebrates/Medium in plants
J	High in all species
K	Medium in all species except fungi and protists
L1	High in vertebrates only
L2	High in animals only
M	Medium in primates only
N	High in mammals only
O1	Medium in all vertebrates except fish and amphibians
O2	High in all vertebrates except fish
P	Medium in all vertebrates except birds

Figure 17. Functional enrichment for Gene Ontology terms of human phosphoproteins with different **(A)** Ser/Thr and **(B)** Tyr phosphosite conservation patterns. The enrichment is visualised with dot plots generated using clusterProfiler. In each dot plot, the dots represent protein sets enriched for a specific functional term described on the y-axis. The size of the dots reflects the number of proteins in the enriched set and the colour corresponds to the significance of the functional enrichment determined by Benjamini–Hochberg adjusted p-value. The position of the dots on the x-axis indicates the proportion of enriched proteins out of all analysed proteins in the protein set. The total number of proteins in each set is given by *n*.

The differences in conservation patterns of human phosphosites are a likely result of their functional relevance within groups of species in which they are conserved[113, 125, 126], with some phosphosites being involved in ancient protein functions relevant to all life forms, and others contributing to relatively novel functions which are only conserved in specific species groups more closely related to humans (mammals or primates, for example). To investigate this further, we performed a functional enrichment analysis of proteins with different phosphosite conservation patterns to identify conserved functions which are potentially regulated by those phosphosites (Fig. 17 and Fig. S7).

Firstly, we found that in 384, 1,222 and 1,100 human phosphoproteins, their phosphorylated Ser and Thr sites were on average conserved in broad species groups such as animals (Fig. 15A, cluster B), vertebrates (Fig. 15A, cluster H2) or mammals (Fig. 15A, cluster E), respectively (Table 7; Table S18). For example, Ser900, Ser916 and Ser930 sites in protein O75044 (SLIT-ROBO Rho GTPase-activating protein 2; SRGAP2) were conserved in most vertebrates, but not in insects (Fig. 16A). It appears that the aligned insect sequences are unlikely to be true orthologues of the human protein as indicated by the lack of any conserved motifs in the analysed phosphorylated region (Fig. 16A). For human SRGAP2, there were also no significant matches found in BLAST (*E*-value of $\leq$0.00001) from plant species, fungi and protists. This conservation pattern can be explained by the functional relevance of SRGAP2 in neuronal morphogenesis during the development of cerebral cortex necessary for complex brain functions, which are expectedly not present in insects, plants, fungi and protists[260, 261]. This connection was further highlighted by the functional enrichment analysis, which revealed a general enrichment of target proteins for functional terms related to brain development such as "*neuron differentiation*", "*synapse*" and "*axon*" (Fig. 17A, cluster H2). Most importantly, the conservation of the identified pathways regulated by proteins with Ser/Thr phosphosites characterised in our analysis suggests that the selected vertebrates can be used as model organisms to study human brain development pathways.

In addition, we identified 197 proteins in which Ser/Thr phosphosites were conserved in all species, ranging from primates and other mammals to plants and single-celled organisms (Fig. 15A, cluster C). We linked those proteins to ancient molecular functions relevant across all life such as translation, metabolism and cell cycle regulation (Fig. 17A, cluster C). For example, human Thr196 site in protein Q15131 (cyclin-dependent kinase 10; CDK10) was well-conserved across all species (Fig. 16B) and its phosphorylation plays a role in promoting cell proliferation and transcription regulation[262]. This provides further evidence for the functional significance and conservation of CDK enzymes across life, with many CDK enzymes having been previously characterised across all species[263]. Additional DAVID analysis of proteins with Ser/Thr phosphosites conserved in all species (Fig. S7A, cluster C) revealed enrichment for methylation and acetylation which suggests that those post-translational

modifications, along with phosphorylation, play a role in key biological functions across all life, potentially as part of crosstalk between PTMs[43]. However, as similarly discussed in Chapter 3, we cannot rule out some biases that different PTM sites might have been more heavily studied in conserved proteins involved in the cell cycle.

Interestingly, we also found that in around 11% of analysed proteins, their Ser/Thr phosphosites were only conserved in primates, indicating their potential relevance in relatively novel functions which diverged from other animals (Table 7; Fig. 15A, cluster D). For example, phosphorylated Thr834 in protein Q15572 (TATA box-binding factor RNA polymerase I subunit C; TAF1C) was conserved in primates but absent in other animals, and there were no BLAST matches found for that protein in lower eukaryotes (Fig. 16C). TAF1C is involved in functional pathways that regulate transcription[264] and our results suggest that those pathways may be exclusive to primates. Further functional enrichment analysis of proteins in cluster D inferred their involvement in tRNA processing and association with zinc finger-related terms including Kruppel-associated box (KRAB) zinc finger proteins (Fig. 17A; Fig. S7A, cluster D). In fact, previous studies also characterised groups of KRAB proteins which rapidly evolved in primates and adapted to regulate complex pathways involved in brain development[265, 266]. Furthermore, other proteins from cluster D were characterised as G-antigen (GAGE) proteins (Fig. S7A, cluster D). Those proteins have been known to play a regulatory role in primate germ cell development and were proposed as candidates for immunotherapy of cancer due to their expression in many cancer tissues[267]. This highlights that identifying human phosphosites which are conserved in closely related species such as primates and other mammals can extend the availability of genetically similar species for the development of preclinical models used in drug development and testing.

Another interesting conservation pattern was found for proteins in which Ser/Thr phosphosites were conserved in most vertebrates except fish (Fig. 15A, cluster H1). For example, phosphorylated Ser191 in human protein P35908 (Type II keratin; KRT2) was conserved across mammals, reptiles, birds and amphibians, but BLAST did not produce any significant matches for that protein in any of the 5 fish species we analysed (Fig. 16D). The functional enrichment analysis of proteins in cluster H1 revealed that the conserved phosphosites may be involved in the formation of intermediate filaments during keratinisation (Fig. 17A; Fig. S7A, cluster H1), a process required for the development of epidermis[268]. Therefore, the results suggest that those functional pathways are absent in fish, likely because they are not necessary for survival in the aquatic environment. This conservation pattern was further highlighted by evolutionary studies which similarly identified groups of keratin proteins only conserved in tetrapods (i.e., four-legged vertebrates which evolved later than fish) (Fig. 14) and linked

those proteins to biological pathways involved in the development of tissues and organs such as skin, hair and nails necessary for protection against the friction caused by terrestrial movement[269, 270].

In our analysis we also separately grouped human proteins according to the conservation of their Tyr phosphosites to further understand functional and evolutionary patterns of tyrosine phosphorylation in eukaryotes (Table 7; Fig. 15B; Table S18). In particular, we identified 89 proteins in which Tyr phosphosites were conserved in all species (Table 7) and linked this pattern to their potential involvement in ancient mechanisms that regulate protein folding, heat response and nucleotide-binding (Fig. 17B; Fig. S7B, Cluster J). For example, we found that Tyr382 site in protein Q92630 (Dual specificity tyrosine-phosphorylation-regulated kinase 2; DYRK2) was conserved across all species groups including animals and lower eukaryotes (Fig. 16E). DYRK2 is a serine-threonine kinase which plays an important role in regulating cell cycle, proliferation and apoptosis[271]. It is activated by autophosphorylation of the Tyr382 site which further explains the functional relevance and consequent conservation of that Tyr site across all species[271].

Furthermore, we identified 53 proteins in which Tyr phosphosites were on average conserved in most vertebrates except fish (Table 7; Fig. 15B, cluster O2; Table S18), which is a similar conservation pattern found in some proteins regulated by Ser/Thr phosphorylation (Fig. 15A, cluster H1). The proteins in cluster O2 were primarily enriched for functional terms describing cell motility pathways which are likely to be absent in fish (Fig. 17B; Fig. S7B, Cluster O2). One example of such protein is Q7Z406 (Myosin-14; MYH14), in which Tyr778 site was conserved across most vertebrates, but was mutated to phenylalanine in fish (Fig. 16F). This evolutionary pattern was also confirmed in another study which identified several myosin domains that emerged as a result of divergent evolution between fish and tetrapods[272]. Our results further highlight the functional relevance of tyrosine phosphorylation in tetrapods and help understand the evolution of established myosin-related functions such as cell signalling and motility[273, 274]. In addition, MYH14 (Fig. 16F) has been previously linked to hearing loss in humans[275] and therefore the species in which the functional Tyr phosphosite was found to be conserved can be used as potential clinical models to further study its involvement in human disease.

Finally, our conservation analysis characterised 103 human proteins in which Tyr phosphosites were mostly conserved across animals, indicating an evolutionary divergence from lower eukaryotes (Table 7; Fig. 15B, cluster L2). We linked those proteins to the development of animal features by identifying functionally enriched terms such as "*axon development*", "*synapse organisation*", "*tissue homeostasis*" and "*retina homeostasis*", in which tyrosine phosphorylation must play an important role (Fig. 17B, cluster L2). For example, one of the analysed proteins, Q14254 (Flotillin-2; FLOT2), which has phosphorylated Tyr163 site conserved in animals (Fig. 16G), is indeed known to be

regulated by tyrosine phosphorylation to facilitate axonal development and endocytosis[276, 277]. Furthermore, we were able to identify 69 proteins in which Tyr phosphosites were exclusively conserved in mammals (Table 7; Fig. 15B, cluster N) and linked those proteins to functions involved in immune response (Fig. 17B; Fig. S7B, Cluster N). For example, Tyr2 in protein P40121 (Macrophage-capping protein; CAPG) was found to be conserved in mammals but absent in all other investigated groups of eukaryotes (Fig. 16H). This indicated that tyrosine phosphorylation might play an important role in regulating immune responses which are unique to mammalian systems. In addition, the results suggest that the selected mammals in which Tyr phosphosites were found to be conserved can potentially be used to study specific mammalian molecular pathways also involved in human immune responses. Some of those pathways may also extend to other vertebrates as indicated by a similar functional enrichment for immune system-related functions of a small number of proteins with conserved Tyr phosphosites in cluster L1 (Fig. 17B; Fig. S7B, Cluster L1).

There were also other clusters of phosphosite conservation patterns identified in our analysis (Fig. 15). However, we were unable to link some of those clusters to any specific functions which might have helped explain the observed conservation patterns (Fig. 17). This is likely a result of a small sample size of proteins involved in those clusters or weaker enrichment which did not yield any conclusive results. It is also possible that there were sequencing or annotation errors in proteomes of certain species used in our analysis or within the specific protein sequences aligned with our human targets. This could have in turn led to errors in resulting multiple sequence alignments, causing consequent inaccuracies in the characterisation of phosphosite conservation patterns and functional enrichment predictions as demonstrated by several studies[278-280]. Nevertheless, we were able to successfully establish clear conservation patterns of human Ser/Thr and Tyr phosphosites across groups of eukaryotic species and highlight those patterns with specific protein examples. Furthermore, we linked most phosphosites to their functional relevance in specific groups of species and provided a clear insight into the evolution of several protein functions regulated by phosphorylation. We also demonstrated that the analysis of phosphosite conservation described in this Chapter can identify species which can be used as potential biological models to study functions involved in human biology and disease. As a result, our method can be readily applied to study phosphosites and even other PTMs in terms of their evolution and functional significance.

4.4.2 Linking phosphosite conservation to protein domains

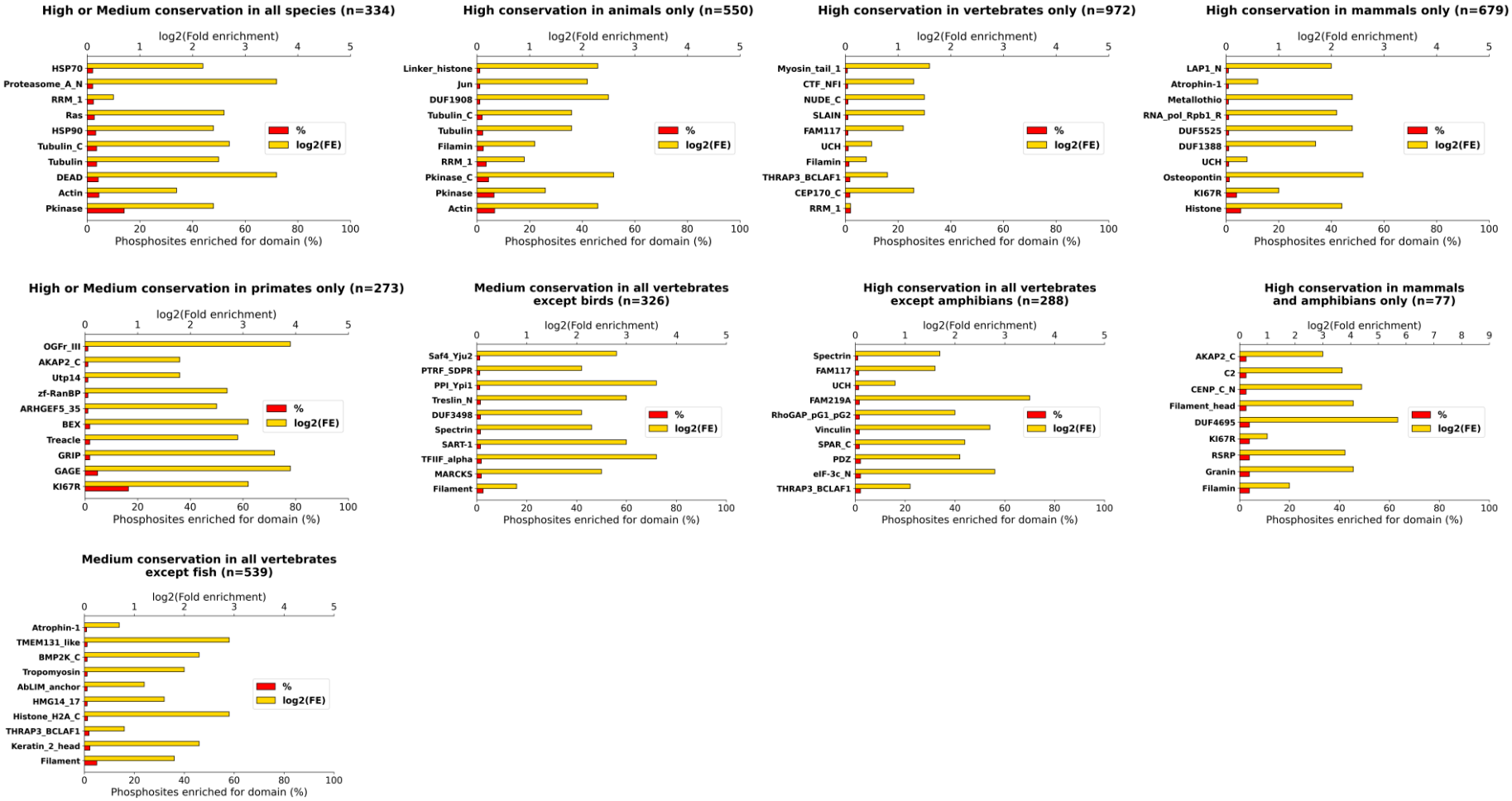
Having established specific conservation patterns of human phosphosites, we were able to identify relevant protein domains in which those phosphosites were found by linking the data from domain database Pfam[252]. In total, we successfully mapped 22% (4,576/20,711) of our analysed Ser, Thr and Tyr phosphosites to 1,419 different protein domains characterised in the Pfam database (Table S18).

It is likely that the unmapped phosphosites had an independent functional role outside of any given protein domain or belonged to a recently discovered protein which has not yet been annotated in Pfam. We also linked the domains with the conservation patterns of the associated phosphosites to further understand their evolution and functional relevance in eukaryotes (Table S18). By identifying frequently observed (enriched) domains per conservation cluster, we provided further evidence to support the functional relevance of conserved phosphosites within specific groups of eukaryotes (Fig. 17, Fig. S7).

For example, phosphosites conserved across most eukaryotic species were found within domains from protein kinases, heat shock proteins, proteins associated with apoptosis (indicated by the enrichment of "*DEAD*" domains[281]), as well as domains from actin and tubulin proteins (Fig. 18). As discussed previously, those phosphosites are therefore likely to be associated with ancient functions conserved in all life forms such as cell signalling, regulation of cell proliferation and death, stress response and cytoskeleton formation. Furthermore, phosphosites conserved in only animals were primarily linked with actin and kinase domains (Fig. 18), perhaps indicating their relevance in cell signalling pathways which evolved exclusively in animals to facilitate functions such as muscle development and contraction[15]. Interestingly, we also found an enrichment of Ser/Thr phosphosites conserved in mammals for the "*KI67R*" domain which is known to be associated with genome stability and mammalian immune response against highly proliferative cells[282]. This domain was further enriched in primates suggesting its potential involvement in additional, primate-exclusive signalling pathways (Fig. 18A). A similar pattern was found for Tyr phosphosites which were exclusively conserved in mammals or primates and enriched for protein domains known to be involved in immune response such as "*LIME1*"[283, 284], "*Annexin*"[285] and "*Defensin_1*"[286], further highlighting that Ser/Thr/Tyr phosphorylation plays a key role in immune system pathways of higher eukaryotes (Fig. 18B). Ser/Thr phosphosites conserved in primates were also linked to the "*GAGE*" domain (Fig. 18B) which reaffirmed our previous result that linked primate-specific phosphosites to GAGE-type proteins involved in germ cell line development and cancer (Fig. S7A, cluster D). Moreover, the domain analysis of phosphosites conserved in all vertebrates except fish revealed enrichment for the "*Filament*" domain (Fig. 18) involved in the formation of intermediate filaments[287]. This protein domain regulates cell signalling pathways likely to be specifically relevant to tetrapod survival such as movement and response to mechanical stress, which further highlights our results of the protein-level functional enrichment analysis (Fig. 17, clusters H1 and O2).

A

pST sites


High or Medium conservation in all species (n=334)
log2(Fold enrichment)
HSP70
Proteasome_A_N
RRM_1
Ras
HSP90
Tubulin_C
Tubulin
DEAD
Actin
Pkinase
%
log2(FE)
Phosphosites enriched for domain (%)
High conservation in animals only (n=550)
log2(Fold enrichment)
Linker_histone
Jun
DUF1908
Tubulin_C
Tubulin
Filamin
RRM_1
Pkinase_C
Pkinase
Actin
%
log2(FE)
Phosphosites enriched for domain (%)
High conservation in vertebrates only (n=972)
log2(Fold enrichment)
Myosin_tail_1
CTF_NFI
NUDE_C
SLAIN
FAM117
UCH
Filamin
THRAP3_BCLAF1
CEP170_C
RRM_1
%
log2(FE)
Phosphosites enriched for domain (%)
High conservation in mammals only (n=679)
log2(Fold enrichment)
LAP1_N
Atrophin-1
Metallothio
RNA_pol_Rpb1_R
DUF5525
DUF1388
UCH
Osteopontin
KI67R
Histone
%
log2(FE)
Phosphosites enriched for domain (%)
High or Medium conservation in primates only (n=273)
log2(Fold enrichment)
OGFr_III
AKAP2_C
Utp14
zf-RanBP
ARHGEF5_35
BEX
Treacle
GRIP
GAGE
KI67R
%
log2(FE)
Phosphosites enriched for domain (%)
Medium conservation in all vertebrates except birds (n=326)
log2(Fold enrichment)
Saf4_Yju2
PTRF_SDPR
PPI_Ypi1
Treslin_N
DUF3498
Spectrin
SART-1
TFIIF_alpha
MARCKS
Filament
%
log2(FE)
Phosphosites enriched for domain (%)
High conservation in all vertebrates except amphibians (n=288)
log2(Fold enrichment)
Spectrin
FAM117
UCH
FAM219A
RhoGAP_pG1_pG2
Vinculin
SPAR_C
PDZ
eIF-3c_N
THRAP3_BCLAF1
%
log2(FE)
Phosphosites enriched for domain (%)
High conservation in mammals and amphibians only (n=77)
log2(Fold enrichment)
AKAP2_C
C2
CENP_C_N
Filament_head
DUF4695
KI67R
RSRP
Granin
Filamin
%
log2(FE)
Phosphosites enriched for domain (%)
Medium conservation in all vertebrates except fish (n=539)
log2(Fold enrichment)
Atrophin-1
TMEM131_like
BMP2K_C
Tropomyosin
AbLIM_anchor
HMG14_17
Histone_H2A_C
THRAP3_BCLAF1
Keratin_2_head
Filament
%
log2(FE)
Phosphosites enriched for domain (%)

B

pY sites

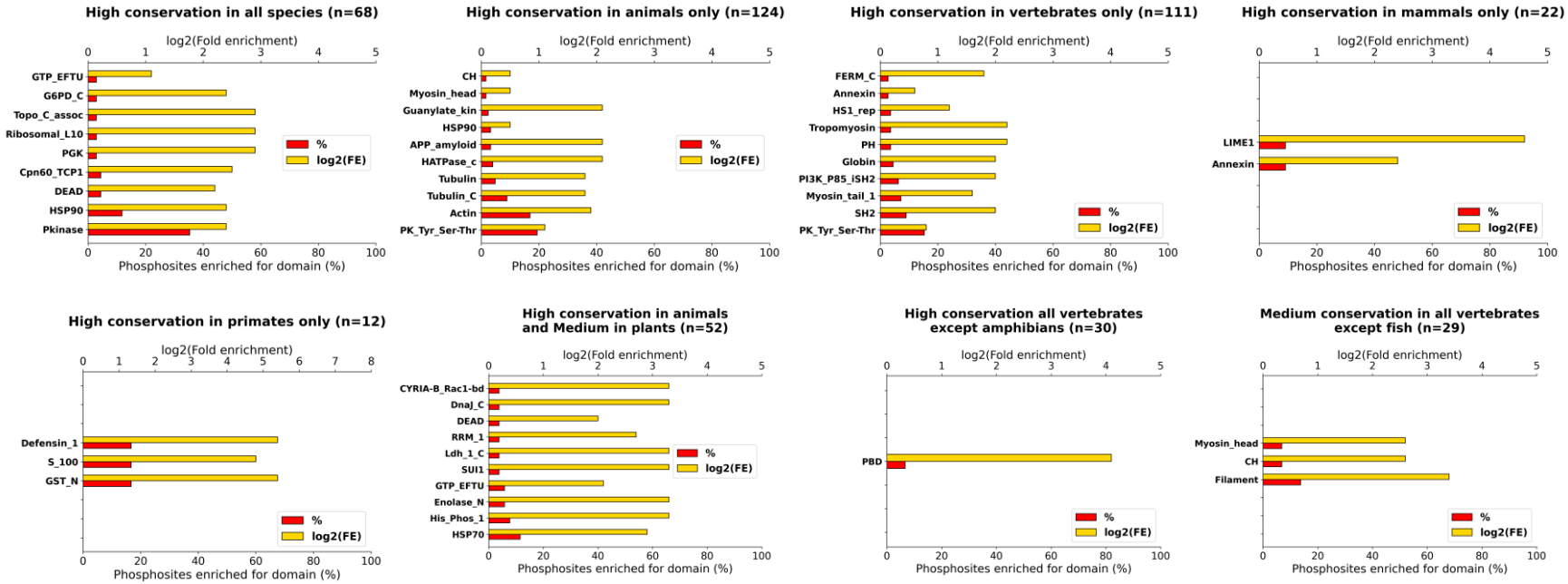


Figure 18. Protein domains from Pfam database (y-axis) for which **(A)** Ser/Thr and **(B)** Tyr phosphosites with specific conservation patterns across eukaryotes were enriched against a control background of all phosphosites mapped to any Pfam domains. For each conservation pattern, the percentage (%) of phosphosites with that pattern mapped to a specific domain is given, as well as the log2(fold enrichment). The total number of phosphosites with a particular conservation pattern mapped to any domain is presented by *n*.

Finally, we identified a small number of tyrosine phosphosites highly conserved in animals and plants which were enriched for a heat shock protein domain "*HSP70*" and domain "*His_Phos_1*" which belongs to a histidine phosphatase superfamily of proteins (Fig. 18). Upon further investigation, we found that those Tyr phosphosites were associated with enzymes called phosphoglycerate mutases. Those enzymes typically have a conserved histidine residue in their active site which is phosphorylated to regulate glucose metabolism during glycolysis[288]. By visualising an example alignment of protein P15259 (Phosphoglycerate mutase 2; PGAM2), we first found the phosphorylated target Tyr92 site which was highly conserved in all animals and plants, along with some fungi and protists (Fig. 19). According to UniProt and an associated study[289], the phosphorylated Tyr92 and its neighbouring His91 form a 4-amino acid binding site motif Glu-Arg-His-Tyr in PGAM enzymes. Interestingly, however, the neighbouring His91 site was absent in plants (and some protists) where it was mutated to methionine but conserved in all other species (Fig. 19). Previous studies of PGAM enzymes in plants reported its involvement in plant growth and photosynthesis, suggesting different molecular pathways compared to other eukaryotic species which may not involve histidine in its binding site[290, 291]. In terms of the active site of PGAM with phosphorylated His, previous reports indicated that it is typically found and conserved at position 11 of the PGAM sequence[292, 293]. In fact, we also found conserved His11 in PGAM2 (Fig. 19) isoform which confirms the functional relevance of that site in the characterised "*His_Phos_1*" protein domain that spans between positions 6 and 134 in the sequence. This result highlights the importance of non-canonical histidine phosphorylation across eukaryotes and provides evidence of its involvement in phosphorylation motifs alongside conserved functional tyrosine sites.

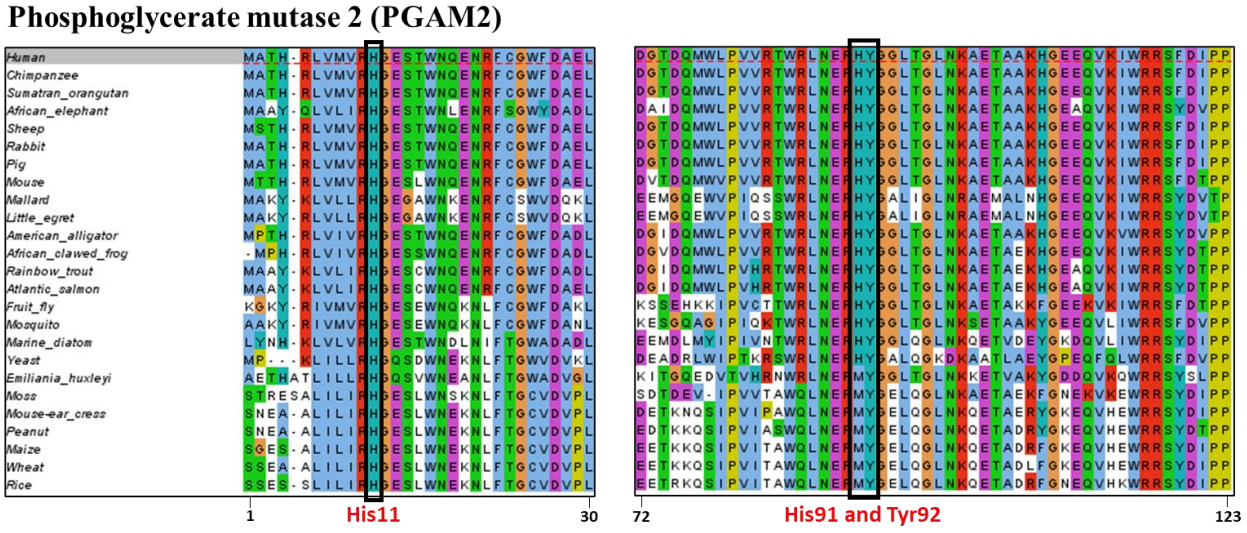


Figure 19. Sections of a multiple sequence alignment of human Phosphoglycerate mutase 2 (PGAM2) with its potential eukaryotic homologues. The target human sequence is located at the top. Conserved functional sites His11, His91 and Tyr92 are marked by black rectangle boxes and the flanking positions in the human sequence are given at the bottom of each alignment section. Alignment annotation was done in Jalview.

4.4.3 Predicting phosphorylation sites in eukaryotes

In our conservation analysis we only used high-quality human phosphorylation sites with plenty of identification evidence in databases PhosphoSitePlus and PeptideAtlas (Table S10) in order to limit the potential use of false positive phosphosite identifications. As a result, we used those phosphosites and the proteins in which they were found as a reference set to predict phosphorylation sites in other eukaryotes. In particular, we analysed multiple sequence alignments between the target human proteins and their top protein matches in BLAST from other eukaryotes to identify any sites which were aligned with the human phosphosites. For each aligned amino acid site from a certain species, we assumed that it was likely to be a true phosphosite if it was the same as the human phosphosite (considering Ser/Thr substitutions) and if its adjacent site at the +1 position in the sequence was also conserved in the human protein (i.e., a site which is likely to also be involved in a typical phosphorylation motif[90]). Using this assumption, we predicted a total of 830,078 Ser, 148,818 Thr and 56,480 Tyr potential phosphosites in the analysed species (Table 8, Table S19). The majority of phosphosites were predicted for primates, likely due to their closest evolutionary relationship with humans compared to the other analysed species, leading to similarities in protein sequences and common functional relevance. In addition, by using confident human phosphosites as a reference, we predicted hundreds of potential phosphosites in lower eukaryotes such as plants, fungi and protists based on their sequence alignment with human sites and likely involvement in common functions regulated by phosphorylation (Table 8, Table S19).

To validate our phosphosite predictions, we investigated how many of our predicted sites in species such as mouse and Arabidopsis had any actual experimental identification evidence in phosphorylation databases PSP[57] and Plant PTM Viewer[109], respectively. We found that 82% and 75% of our predicted Ser/Thr and Tyr phosphosites in mouse had reported experimental evidence in PSP, respectively (Table S19). Out of those sites, 61% had at least 5 pieces of phosphorylation evidence, indicating a confident set of phosphosites with a low false discovery rate (Table S19) as discussed in Chapter 3. By taking into account the total number of Ser, Thr and Tyr phosphosites in the mouse proteome and the overall number of phosphosites reported in PSP for mouse, we estimated enrichment factors of 14 and 43 for identifying confident phosphorylation sites that had any experimental evidence in PSP and at least 5 pieces of evidence, respectively, against a probability of identifying those sites by random chance. In other words, our method is much more likely to identify confident phosphosites in mouse than if they were selected at random. Furthermore, we found that 35% of our predicted Ser, Thr and Tyr phosphosites in Arabidopsis were supported by experimental evidence in Plant PTM Viewer resource (Table S19). The proportion of predicted Arabidopsis phosphosites mapped to experimental evidence was lower than for mouse predictions, likely because

many of the target Arabidopsis protein sequences were missing phosphosite annotations or because many sequences have not yet been analysed experimentally and reported in Plant PTM Viewer. Nevertheless, our results suggest that our method of analysing multiple sequence alignments between human proteins with confident phosphosites and their top sequence matches in BLAST from other eukaryotic species can successfully predict confident phosphosites in those species. In our analysis, the resulting phosphosite predictions in eukaryotes were based on the conservation of target sites and their adjacent +1 sites against the aligned human phosphosites. However, it might be possible to improve our phosphosite predictions by making sure that the whole phosphorylation motifs encompassing target phosphosites in the human protein sequences were conserved, which would further increase the likelihood of the aligned sites from other species being real phosphosites.

To summarise, our phosphosite predictions and results from the functional enrichment analysis (Fig. 17, Fig. S7) identified areas of potential functional relevance in protein sequences from species which do not have much experimental evidence or comprehensive protein sequence annotations. Researchers can therefore analyse our predicted phosphosites to ultimately improve protein sequence annotations in different eukaryotic species and direct further research involving the use of these species as biological models to study conserved cell signalling and disease development pathways in which phosphorylation plays an important role.

Table 8. The counts of predicted Ser, Thr and Tyr phosphosites across 100 eukaryotic species. The species are ranked from high to low in terms of their total counts of predicted phosphosites.

Species	pSer count	pThr count	pTyr count	Total predicted phosphosites
Chimpanzee	16179	2625	938	19742
Western lowland gorilla	15642	2547	909	19098
Olive baboon	15569	2540	913	19022
Crab-eating macaque	15574	2525	909	19008
Rhesus macaque	15549	2531	894	18974
Pygmy chimpanzee	15482	2527	900	18909
Sooty mangabey	15413	2493	897	18803
Green monkey	15075	2470	887	18432
White-tufted-ear marmoset	14998	2415	878	18291
Sumatran orangutan	14945	2439	894	18278
Northern white-cheeked gibbon	14881	2455	876	18212
Ma's night monkey	14879	2393	867	18139
Black snub-nosed monkey	14670	2427	869	17966
Drill	14601	2412	888	17901
Bolivian squirrel monkey	14404	2347	851	17602
Pacific walrus	13812	2256	870	16938
Horse	13817	2252	853	16922
Cat	13571	2251	833	16655
Bovine	13502	2239	868	16609
Yangtze river dolphin	13534	2223	821	16578
Goat	13445	2241	843	16529

Beluga whale	13458	2196	840	16494
Sperm whale	13360	2223	817	16400
Coquerel's sifaka	13424	2163	808	16395
Caribbean manatee	13285	2203	833	16321
Small-eared galago	13229	2171	827	16227
Giant panda	13101	2211	840	16152
Mouse	13079	2196	875	16150
Dog	13045	2200	824	16069
Thirteen-lined ground squirrel	13074	2153	842	16069
Pig	12993	2225	834	16052
European domestic ferret	13024	2175	827	16026
Philippine tarsier	12801	2124	800	15725
Golden hamster	12699	2123	825	15647
Sheep	12661	2119	809	15589
Polar bear	12555	2168	817	15540
African elephant	12554	2083	829	15466
Ord's kangaroo rat	12464	2110	793	15367
Western European hedgehog	12339	2072	816	15227
Guinea pig	12270	1995	823	15088
Little brown bat	12115	2054	769	14938
Black flying fox	11981	2021	735	14737
Rabbit	11903	1992	781	14676
Naked mole rat	11678	1984	787	14449
Atlantic bottle-nosed dolphin	11405	1897	690	13992
Damaraland mole rat	11127	1926	783	13836
Common wombat	10735	1887	758	13380
Chinese hamster	9710	1687	645	12042
Chinese alligator	8886	1666	669	11221
American alligator	8225	1616	632	10473
Weddell seal	8172	1467	590	10229
Chinese softshell turtle	7951	1537	604	10092
Green anole	7884	1533	617	10034
African clawed frog	7540	1569	623	9732
Zebra finch	7360	1375	596	9331
Rock dove	7207	1402	586	9195
Wild turkey	7234	1394	564	9192
Blue-fronted Amazon parrot	7206	1395	553	9154
Scaled quail	6831	1311	539	8681
Western clawed frog	6648	1387	563	8598
Duckbill platypus	6726	1248	515	8489
Emperor penguin	6482	1257	520	8259
Atlantic salmon	6216	1463	553	8232
Greater amberjack	5948	1416	546	7910
Anna's hummingbird	6131	1219	484	7834
Little egret	6036	1201	480	7717
Downy woodpecker	5961	1165	474	7600
Mallard	5801	1186	483	7470
Three-spined stickleback	5392	1308	497	7197
Hoatzin	5504	1134	443	7081
White-throated tinamou	5483	1061	469	7013

Japanese pufferfish	5203	1286	496	6985
Rainbow trout	4812	1218	429	6459
Red flour beetle	1341	567	193	2101
Dampwood termite	1347	509	195	2051
Clonal raider ant	1300	512	188	2000
Fruit fly	1221	563	165	1949
Jerdon's jumping ant	1247	478	168	1893
Water flea	1145	454	190	1789
Asian swallowtail butterfly	1133	482	159	1774
Green bottle fly	1086	527	160	1773
Mosquito	1064	501	163	1728
Silk moth	984	427	153	1564
Neoptera	661	293	116	1070
Moss	445	224	82	751
Wheat	424	205	82	711
Slime mold	389	230	63	682
Maize	393	204	77	674
Wild banana	387	178	77	642
Mouse-ear cress	392	177	65	634
Rice	360	186	70	616
Peanut	360	171	75	606
Emericella nidulans	326	201	55	582
Neurospora crassa	318	192	67	577
Yarrowia lipolytica	264	160	57	481
Chlamydomonas reinhardtii	274	141	51	466
Baker's yeast	247	125	53	425
Emiliania huxleyi	191	128	52	371
Marine diatom	184	119	50	353
Plasmodium falciparum	145	84	44	273
Total	**830,078**	**148,818**	**56,480**	**1,035,376**

4.5 Conclusion

In this Chapter, we analysed a set of confident human Ser, Thr and Tyr phosphosites in terms of their conservation across specific groups of eukaryotic species. We established various evolutionary trends for the target human phosphosites and linked them to key conserved protein functions, highlighting the importance of phosphorylation within different groups of eukaryotes. In particular, we established the relevance of phosphosites in several ancient functions involved in cell signalling and metabolism, as well as functions which were only conserved in species closely related to humans, such as regulation of brain and muscle development, motility and immune response. We explored the identified protein functions further and provided an insight into the evolution of protein domains containing the target phosphosites. Our results emphasised the importance of conservation analysis in predicting the functional significance of phosphosites and identifying organisms which can be used as biological models to study conserved signalling pathways relevant to human biology and disease.

By providing a comprehensive dataset describing the conservation and functional relevance of human phosphosites (Table S18), we suggest directions for further phosphoproteomics research. For example, phosphosite conservation can be applied in the investigation of human kinase enzymes and their phosphosite pairings. Despite the presence of a large number of identified human phosphosites[57, 100, 248], less than 5% of the characterised human phosphoproteome has been linked to specific kinase enzymes responsible for the corresponding phosphorylation events[57, 102]. Similarly, comprehensive kinase-substrate pairings have only been reported for a small number of well-researched kinase enzymes usually involved in cancer or neurodegenerative diseases, with several human kinases only having a few associated phosphosites or no known substrates at all[294, 295]. Numerous publicly available bioinformatics resources including NetPhorest[255] and NetworKIN[296] offer algorithms which analyse protein sequence data to predict kinase-substrate pairings. In addition, a recent large-scale proteomics study by Johnson *et al.* (2023) profiled the human kinome to identify likely kinase matches for many reported human Ser/Thr phosphosites and link the predictions to various cell signalling pathways[102]. Our analysis of confident human phosphosites can therefore be expanded to the motif-level and be used to validate the available kinase-substrate predictions (i.e., if a kinase phosphorylation motif or site is conserved then, it is more likely to be true), identify potential kinase substrates and enhance the understanding of kinase evolution.

Furthermore, we successfully applied conservation analysis to predict over 1,000,000 potential Ser, Thr and Tyr phosphosites across various eukaryotic species ranging from primates and other mammals to plants, fungi and protists. Our predictions can ultimately be used to improve proteome annotations of species which do not have many experimentally confirmed phosphosites and which lack comprehensive curation. In conclusion, our analysis clearly demonstrates the diversity of

evolutionary patterns in protein functions regulated by phosphorylation and our results can also be used to infer phosphorylation sites in other species and direct further research surrounding the evolution and functional relevance of protein phosphorylation.

Chapter 5

Thesis Conclusion and Future Research Directions

The research described in this Thesis profiled the human phosphoproteome to estimate the true extent of protein phosphorylation, highlight the issue of high phosphosite false discovery rate in large datasets, analyse the evolutionary conservation trends of human phosphosites, explore their key functions and predict large numbers of phosphosites in other eukaryotic species.

First, in Chapter 2, we developed and validated an accessible Python pipeline which can determine the conservation of specific amino acid sites such as PTMs and perform several steps of a typical conservation analysis in a single step. In particular, for each query protein sequence, the pipeline identifies its likely homologous sequences from the selected species using the BLAST algorithm, generates multiple sequence alignments and calculates the conservation of target amino acid sites. We demonstrated that the pipeline is robust, easy to use and can be readily applied to analyse the conservation of thousands of protein targets at once from several selected species. The pipeline also generates multiple useful outputs which can be used for an in-depth downstream analysis of target sequences, such as BLAST results, FASTA sequences of top hits from each species, MSAs and percentage conservation scores of the target and adjacent sites. Currently, it is possible to calculate the conservation of -1 and +1 sites around each target site in the protein sequence. However, in the future, it may be possible to extend this motif sequence to analyse the conservation of selected phosphorylation motifs rather than a set motif of three amino acids, which can be useful in linking conservation results to substrate-kinase relationships. Nevertheless, the current pipeline is ideal for studying the evolutionary conservation of any specific biological sites of interest such as PTMs across multiple species. The pipeline is supported by a guide containing detailed installation and running instructions, explanations of any inputs and outputs, a troubleshooting guide and links to example inputs and results. Finally, no prior knowledge of Python programming is required to run the pipeline and it can work on multiple systems including Windows, MacOS and Linux. In the future, it may also be possible to incorporate the pipeline into a convenient online tool which would make it even more accessible.

In Chapter 3, we addressed the issue of the likely overestimation of true phosphosites in public databases by developing a method for independent FDR estimation and predicting a more realistic count of true identifications. In particular, we ranked all reported human Ser, Thr and Tyr phosphosites into sets according to the amount of identification evidence they had in public databases such as PSP and PA, and analysed the sets in terms of their orthogonal properties such as conservation across 100 species, sequence properties and functional annotations. We demonstrated significant

differences between the sets and estimated that around 62,000 Ser, 8,000 Thr and 12,000 Tyr phosphosites in the human proteome were likely to be true, which is lower than most published estimates. Remarkably, we reported an estimated FDR of 84%, 98% and 82% within sets of phosphorylated Ser, Thr and Tyr sites, respectively, that were supported by only a single piece of identification evidence - the majority of sites in PSP. In fact, our analysis estimated that a total of 86,000 Ser, 50,000 Thr and 26,000 Tyr phosphosites were likely false-positive identifications, highlighting the significant potential of false positive data to be present in phosphorylation databases. As a general rule, the results in Chapter 3 suggested that phosphorylation sites with <5 independent observations should be treated with caution, and those with only one observation in a database are likely to be false positives. Overall, Chapter 3 provides a methodological framework for estimating global FDR in large-scale phosphorylation data sets, which does not rely on native scores from search engines or site localisation software. Methods for estimating global FDR in meta-analyses of phosphosites are not yet robust, and thus we recommend that other groups profile orthogonal properties of ranked sets to estimate the real distribution of true and false phosphosites in their data. The analysis in Chapter 3 investigated phosphosites from the human phosphoproteome which had any evidence in PSP and PA databases. However, it could be useful to expand the analysis and include all phosphosites from selected phosphorylation resources to obtain clearer estimates of database-level false discovery rates. Phosphosites from other species can also be considered. Finally, our methodology developed for the analysis of orthogonal phosphosite properties can be applied to study the emerging non-canonical phosphorylation in proteins.

In Chapter 4, we analysed the evolutionary conservation of human phosphosites across different groups of eukaryotic species ranging from mammals and other eukaryotes to plants, fungi and protists. We also linked the conservation patterns of phosphosites to their functional relevance in their associated species to further understand the evolution of protein functions regulated by phosphorylation. In particular, we established the relevance of phosphosites in several ancient functions involved in cell signalling and metabolism, as well as functions which were only conserved in species closely related to humans, such as regulation of brain and muscle development, motility and immune response. The functions were analysed further by investigating protein domains containing the target phosphosites with varying conservation patterns. Our results emphasised the importance of conservation analysis in predicting the functional significance of phosphosites and identifying organisms which can be used as biological models to study conserved signalling pathways relevant to human biology and disease. We also proposed that the conservation analysis could be expanded to the motif-level and be used to validate the available kinase-substrate predictions (i.e., if a kinase phosphorylation motif or site is conserved, then it is more likely to be true), identify potential

kinase substrates and enhance the understanding of kinase evolution. In addition, since our analysis was limited to a specific set of around 20,000 human phosphosites which had strong phosphorylation evidence, it may be possible to include additional human phosphosites to obtain a bigger profile of phosphosite conservation in eukaryotes, although this would increase the risk of including potential false positive identifications as discussed in Chapter 3.

Finally, we applied the conservation analysis in Chapter 4 to predict over 1,000,000 potential phosphosites in eukaryotes by using confident human phosphosites as a reference. Our predictions were validated by matching a significant proportion of the predicted phosphosites in mouse and Arabidopsis to actual experimental evidence reported in relevant databases. However, additional experimental validation can be carried out as part of further research to test the predictions. The predicted phosphosites could also be explored in relation to their known kinase enzymes or protein interactions in associated species to support their functional significance and their likelihood of being real identifications. Moreover, the conservation analysis of protein motifs encompassing the predicted phosphosites would also increase the confidence of predicted phosphorylation sites. Nevertheless, our resulting phosphosite predictions are a good starting point for further analysis and can ultimately be used to improve proteome annotations of species which do not have many experimentally confirmed phosphosites and which lack comprehensive curation. Overall, the analysis in Chapter 4 clearly demonstrates the diversity of evolutionary patterns of phosphosites and the associated protein functions regulated by phosphorylation, highlights how several model organisms can be identified to study human functions and predicts phosphorylation sites in other species.

In conclusion, this Thesis provides a comprehensive profiling analysis of human phosphosites in relation to their identification evidence in large datasets, conservation in eukaryotes and functional annotations, all of which are crucial aspects of proteomics research surrounding phosphorylation.

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., *Molecular Biology of the Cell, Fifth Edition*. 2008, Garland Science. p. 329-399.
2. Petsko, G. A., Ringe, D., *Protein Structure and Function*. 2003, Oxford University Press. p. 2-82.
3. Pauling, L. and Corey, R. B., *Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets.* Proc Natl Acad Sci U S A, 1951. **37**(11): p. 729-40.
4. Voet, D., Voet, J. G., Pratt, C. W., *Fundamentals of Biochemistry: Life at the Molecular Level, 5th Edition*. 2016, Wiley. p. 131-165.
5. Perutz, M. F., *Hemoglobin Structure and Respiratory Transport.* Scientific American, 1978. **239**(6): p. 92-125.
6. Crick, F. H. C., *The packing of α-helices: simple coiled-coils.* Acta Crystallographica, 1953. **6**(8-9): p. 689-697.
7. Weng, Z., Rickles, R. J., Feng, S., Richard, S., Shaw, A. S., Schreiber, S. L., Brugge, J. S., *Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions.* Mol Cell Biol, 1995. **15**(10): p. 5627-34.
8. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., *Molecular Biology of the Cell, Fifth Edition*. 2008, Garland Science. p. 152-191.
9. Parker, M. W., *Protein structure from x-ray diffraction.* J Biol Phys, 2003. **29**(4): p. 341-62.
10. Hu, Y., Cheng, K., He, L., Zhang, X., Jiang, B., Jiang, L., Li, C., Wang, G., Yang, Y., Liu, M., *NMR-Based Methods for Protein Analysis.* Anal Chem, 2021. **93**(4): p. 1866-1879.
11. Artigues, A., Nadeau, O. W., Rimmer, M. A., Villar, M. T., Du, X., Fenton, A. W., Carlson, G. M., *Protein Structural Analysis via Mass Spectrometry-Based Proteomics.* Adv Exp Med Biol, 2016. **919**: p. 397-431.
12. Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y. P., Voigt, M., Westbrook, J. D., Young, J. Y., Zardecki, C., Zhuravleva, M., *RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences.* Nucleic Acids Res, 2021. **49**(D1): p. D437-d451.
13. The UniProt Consortium, *UniProt: the Universal Protein Knowledgebase in 2023.* Nucleic Acids Res, 2023. **51**(D1): p. D523-d531.
14. Dominguez, R. and Holmes, K. C., *Actin structure and function.* Annu Rev Biophys, 2011. **40**: p. 169-86.
15. Lorenz, M. and Holmes, K. C., *The actin-myosin interface.* Proc Natl Acad Sci U S A, 2010. **107**(28): p. 12529-34.
16. Herrmann, H., Bär, H., Kreplak, L., Strelkov, S. V., Aebi, U., *Intermediate filaments: from cell architecture to nanomechanics.* Nature Reviews Molecular Cell Biology, 2007. **8**(7): p. 562-573.
17. Shapiro, F., Cahill, C., Malatantis, G., Nayak, R. C., *Transmission electron microscopic demonstration of vimentin in rat osteoblast and osteocyte cell bodies and processes using the immunogold technique.* Anat Rec, 1995. **241**(1): p. 39-48.
18. Locher, K. P., *Review. Structure and mechanism of ATP-binding cassette transporters.* Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1514): p. 239-45.
19. Azad, A. K., Raihan, T., Ahmed, J., Hakim, A., Emon, T. H., Chowdhury, P. A., *Human Aquaporins: Functional Diversity and Potential Roles in Infectious and Non-infectious Diseases.* Front Genet, 2021. **12**: p. 654865.
20. Hiller-Sturmhöfel, S. and Bartke, A., *The endocrine system: an overview.* Alcohol Health Res World, 1998. **22**(3): p. 153-64.
21. Robinson, P. K., *Enzymes: principles and biotechnological applications.* Essays in Biochemistry, 2015. **59**: p. 1-41.
22. Boyer, P. D., *The ATP synthase--a splendid molecular machine.* Annu Rev Biochem, 1997. **66**: p. 717-49.
23. Freeman, H. J., Kim, Y. S., and Sleisenger, M. H., *Protein digestion and absorption in man. Normal mechanisms and protein-energy malnutrition.* Am J Med, 1979. **67**(6): p. 1030-6.

24. Chondrogianni, N., Petropoulos, I., Grimm, S., Georgila, K., Catalgol, B., Friguet, B., Grune, T., Gonos, E. S., *Protein damage, repair and proteolysis.* Mol Aspects Med, 2014. **35**: p. 1-71.
25. Cohen, P., *The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture.* European journal of biochemistry / FEBS, 2001. **268**: p. 5001-10.
26. Cohen, P., *The origins of protein phosphorylation.* Nature Cell Biology, 2002. **4**(5): p. E127-E130.
27. Cohen, P., *The structure and regulation of protein phosphatases.* Annu Rev Biochem, 1989. **58**: p. 453-508.
28. Milanesi, L., Petrillo, M., Sepe, L., Boccia, A., D'Agostino, N., Passamano, M., Di Nardo, S., Tasco, G., Casadio, R., and Paolella, G., *Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity.* BMC Bioinformatics, 2005. **6**(4): p. S20.
29. Amanchy, R., Kalume, D. E., Iwahori, A., Zhong, J., and Pandey, A., *Phosphoproteome Analysis of HeLa Cells Using Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC).* Journal of Proteome Research, 2005. **4**(5): p. 1661-1671.
30. Nousiainen, M., Silljé, H. H. W., Sauer, G., Nigg, E. A., and Körner, R., *Phosphoproteome analysis of the human mitotic spindle.* Proceedings of the National Academy of Sciences, 2006. **103**(14): p. 5391.
31. Olsen, J., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M., *Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks.* Cell, 2006. **127**: p. 635-48.
32. Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M., *Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis.* Science Signaling, 2010. **3**(104): p. ra3.
33. Sharma, K., D'Souza, R. C., Tyanova, S., Schaab, C., Wiśniewski, J. R., Cox, J., and Mann, M., *Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling.* Cell Rep, 2014. **8**(5): p. 1583-94.
34. Ding, L., Cao, J., Lin, W., Chen, H., Xiong, X., Ao, H., Yu, M., Lin, J., Cui, Q., *The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer.* Int J Mol Sci, 2020. **21**(6).
35. Otto, T. and Sicinski, P., *Cell cycle proteins as promising targets in cancer therapy.* Nat Rev Cancer, 2017. **17**(2): p. 93-115.
36. Zhang, W. and Liu, H. T., *MAPK signal pathways in the regulation of cell proliferation in mammalian cells.* Cell Research, 2002. **12**(1): p. 9-18.
37. Cohen, P., *The regulation of protein function by multisite phosphorylation--a 25 year update.* Trends Biochem Sci, 2000. **25**(12): p. 596-601.
38. Ortiz, M. A., Mikhailova, Tatiana, Li, Xiang, Porter, Baylee A., Bah, Alaji, Kotula, Leszek, *Src family kinases, adaptor proteins and the actin cytoskeleton in epithelial-to-mesenchymal transition.* Cell Communication and Signaling, 2021. **19**(1): p. 67.
39. Mansueto, M. S., Reens, A., Rakhilina, L., Chi, A., Pan, B. S., and Miller, J. R., *A reevaluation of the spleen tyrosine kinase (SYK) activation mechanism.* J Biol Chem, 2019. **294**(19): p. 7658-7668.
40. Hunter, T., *Why nature chose phosphate to modify proteins.* Philos Trans R Soc Lond B Biol Sci, 2012. **367**(1602): p. 2513-6.
41. Hardman, G., Perkins, S., Brownridge, P. J., Clarke, C. J., Byrne, D. P., Campbell, A. E., Kalyuzhnyy, A., Myall, A., Eyers, P. A., Jones, A. R., and Eyers, C. E., *Strong anion exchange-mediated phosphoproteomics reveals extensive human non-canonical phosphorylation.* The EMBO journal, 2019. **38**(21): p. e100847-e100847.
42. Duan, G. and Walther, D., *The roles of post-translational modifications in the context of protein interaction networks.* PLoS Comput Biol, 2015. **11**(2): p. e1004049.
43. Wang, Y.-C., Peterson, S. E., and Loring, J. F., *Protein post-translational modifications and regulation of pluripotency in human stem cells.* Cell Research, 2014. **24**(2): p. 143-160.
44. Hunter, T., *The age of crosstalk: phosphorylation, ubiquitination, and beyond.* Mol Cell, 2007. **28**(5): p. 730-8.
45. Sakaguchi, K., Herrera, J. E., Saito, S., Miki, T., Bustin, M., Vassilev, A., Anderson, C. W., and Appella, E., *DNA damage activates p53 through a phosphorylation-acetylation cascade.* Genes Dev, 1998. **12**(18): p. 2831-41.
46. Bannister, A. J. and Kouzarides, T., *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.

47. Khoury, G. A., Baliban, R. C., and Floudas, C. A., *Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.* Scientific Reports, 2011. **1**(1): p. 90.
48. Hebert, T. E., Loisel, T. P., Adam, L., Ethier, N., Onge, S. S., and Bouvier, M., *Functional rescue of a constitutively desensitized beta2AR through receptor dimerization.* Biochem J, 1998. **330**(Pt 1): p. 287-293.
49. Samuel, C. E., *Procedures for measurement of phosphorylation of ribosome-associated proteins in interferon-treated cells*, in *Methods in Enzymology*, S. Petska, Editor. 1981, Academic Press. p. 168-178.
50. Wagner, P. D. and Vu, N.-D., *Phosphorylation of ATP-Citrate Lyase by Nucleoside Diphosphate Kinase.* Journal of Biological Chemistry, 1995. **270**(37): p. 21758-21764.
51. Wettenhall, R. E., Aebersold, R. H., and Hood, L. E., *Solid-phase sequencing of 32P-labeled phosphopeptides at picomole and subpicomole levels.* Methods Enzymol, 1991. **201**: p. 186-99.
52. Wettenhall, R. E., Erikson, E., and Maller, J. L., *Ordered multisite phosphorylation of Xenopus ribosomal protein S6 by S6 kinase II.* J Biol Chem, 1992. **267**(13): p. 9021-7.
53. Kaufmann, H., Bailey, J. E., and Fussenegger, M., *Use of antibodies for detection of phosphorylated proteins separated by two-dimensional gel electrophoresis.* Proteomics, 2001. **1**(2): p. 194-9.
54. Mandell, J. W., *Phosphorylation state-specific antibodies: applications in investigative and diagnostic pathology.* Am J Pathol, 2003. **163**(5): p. 1687-98.
55. White, C. D. and Toker, A., *Using phospho-motif antibodies to determine kinase substrates.* Curr Protoc Mol Biol, 2013. **Chapter 18**: p. Unit 18.20.
56. Honda, T., Obara, Y., Yamauchi, A., Couvillon, A. D., Mason, J. J., Ishii, K., and Nakahata, N., *Phosphorylation of ERK5 on Thr732 is associated with ERK5 nuclear localization and ERK5-dependent transcription.* PLoS One, 2015. **10**(2): p. e0117914.
57. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E., *PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.* Nucleic acids research, 2015. **43**(Database issue): p. D512-D520.
58. Rutherford, N. J., Brooks, M., and Giasson, B. I., *Novel antibodies to phosphorylated α-synuclein serine 129 and NFL serine 473 demonstrate the close molecular homology of these epitopes.* Acta Neuropathologica Communications, 2016. **4**(1): p. 80.
59. Kanner, S. B., Reynolds, A. B., Vines, R. R., and Parsons, J. T., *Monoclonal antibodies to individual tyrosine-phosphorylated protein substrates of oncogene-encoded tyrosine kinases.* Proc Natl Acad Sci U S A, 1990. **87**(9): p. 3328-32.
60. Lemeer, S. and Heck, A. J. R., *The phosphoproteomics data explosion.* Current Opinion in Chemical Biology, 2009. **13**(4): p. 414-420.
61. Thingholm, T. E., Jensen, O. N., and Larsen, M. R., *Analytical strategies for phosphoproteomics.* Proteomics, 2009. **9**(6): p. 1451-68.
62. Glish, G. L. and Vachet, R. W., *The basics of mass spectrometry in the twenty-first century.* Nature Reviews Drug Discovery, 2003. **2**(2): p. 140-150.
63. Nita-Lazar, A., Saito-Benz, H., and White, F. M., *Quantitative phosphoproteomics by mass spectrometry: Past, present, and future.* PROTEOMICS, 2008. **8**(21): p. 4433-4443.
64. Palumbo, A. M., Smith, S. A., Kalcic, C. L., Dantus, M., Stemmer, P. M., and Reid, G. E., *Tandem mass spectrometry strategies for phosphoproteome analysis.* Mass Spectrometry Reviews, 2011. **30**(4): p. 600-625.
65. Wojtkiewicz, M., Berg Luecke, L., Kelly, M. I., and Gundry, R. L., *Facile Preparation of Peptides for Mass Spectrometry Analysis in Bottom-Up Proteomics Workflows.* Current Protocols, 2021. **1**(3): p. e85.
66. Gundry, R. L., White, M. Y., Murray, C. I., Kane, L. A., Fu, Q., Stanley, B. A., and Van Eyk, J. E., *Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow.* Curr Protoc Mol Biol, 2009. **Chapter 10**: p. Unit10.25.
67. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., and Yates, J. R., III, *Protein Analysis by Shotgun/Bottom-up Proteomics.* Chemical Reviews, 2013. **113**(4): p. 2343-2394.
68. Porath, J., *Immobilized metal ion affinity chromatography.* Protein Expr Purif, 1992. **3**(4): p. 263-81.
69. Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., and Jørgensen, T. J., *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns.* Mol Cell Proteomics, 2005. **4**(7): p. 873-86.

70. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D., and Comb, M. J., *Immunoaffinity profiling of tyrosine phosphorylation in cancer cells.* Nat Biotechnol, 2005. **23**(1): p. 94-101.
71. Zhang, Y., Wolf-Yadlin, A., Ross, P. L., Pappin, D. J., Rush, J., Lauffenburger, D. A., and White, F. M., *Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules.* Mol Cell Proteomics, 2005. **4**(9): p. 1240-50.
72. Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R., *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.
73. Orsburn, B. C., *Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics.* Proteomes, 2021. **9**(1): p. 15.
74. Tyanova, S., Temu, T., and Cox, J., *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nature Protocols, 2016. **11**(12): p. 2301-2319.
75. Johnson, A. R. and Carlson, E. E., *Collision-Induced Dissociation Mass Spectrometry: A Powerful Tool for Natural Product Structure Elucidation.* Analytical Chemistry, 2015. **87**(21): p. 10668-10678.
76. Riley, N. M. and Coon, J. J., *The Role of Electron Transfer Dissociation in Modern Proteomics.* Analytical Chemistry, 2018. **90**(1): p. 40-64.
77. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
78. Verheggen, K., Raeder, H., Berven, F. S., Martens, L., Barsnes, H., and Vaudel, M., *Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows.* Mass Spectrom Rev, 2020. **39**(3): p. 292-306.
79. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
80. Eng, J. K., Hoopmann, M. R., Jahan, T. A., Egertson, J. D., Noble, W. S., and MacCoss, M. J., *A deeper look into Comet--implementation and features.* J Am Soc Mass Spectrom, 2015. **26**(11): p. 1865-74.
81. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
82. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., and Gygi, S. P., *A probability-based approach for high-throughput protein phosphorylation analysis and site localization.* Nature Biotechnology, 2006. **24**(10): p. 1285-1292.
83. Eyrich, B., Sickmann, A., and Zahedi, R. P., *Catch me if you can: Mass spectrometry-based phosphoproteomics and quantification strategies.* PROTEOMICS, 2011. **11**(4): p. 554-570.
84. Timm, T., Lenz, C., Merkel, D., Sadiffo, C., Grabitzki, J., Klein, J., and Lochnit, G., *Detection and Site Localization of Phosphorylcholine-Modified Peptides by NanoLC-ESI-MS/MS Using Precursor Ion Scanning and Multiple Reaction Monitoring Experiments.* Journal of the American Society for Mass Spectrometry, 2015. **26**(3): p. 460-471.
85. Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K., *Universal and confident phosphorylation site localization using phosphoRS.* J Proteome Res, 2011. **10**(12): p. 5354-62.
86. Fermin, D., Walmsley, S. J., Gingras, A. C., Choi, H., and Nesvizhskii, A. I., *LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach.* Mol Cell Proteomics, 2013. **12**(11): p. 3409-19.
87. Shteynberg, D. D., Deutsch, E. W., Campbell, D. S., Hoopmann, M. R., Kusebauch, U., Lee, D., Mendoza, L., Midha, M. K., Sun, Z., Whetton, A. D., and Moritz, R. L., *PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline.* Journal of proteome research, 2019. **18**(12): p. 4262-4272.

88. Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D., and Reiter, L., *Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results.* Mol Cell Proteomics, 2017. **16**(12): p. 2296-2309.
89. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R., *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): p. O111.016717.
90. Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A., *A curated compendium of phosphorylation motifs.* Nature Biotechnology, 2007. **25**(3): p. 285-286.
91. Dephoure, N., Gould, K. L., Gygi, S. P., and Kellogg, D. R., *Mapping and analysis of phosphorylation sites: a quick guide for cell biologists.* Molecular biology of the cell, 2013. **24**(5): p. 535-542.
92. Elias, J. E. and Gygi, S. P., *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nature Methods, 2007. **4**(3): p. 207-214.
93. Lee, D. C. H., Jones, A. R., and Hubbard, S. J., *Computational phosphoproteomics: from identification to localization.* Proteomics, 2015. **15**(5-6): p. 950-963.
94. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R., *Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.* Analytical Chemistry, 2002. **74**(20): p. 5383-5392.
95. Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S., *Posterior error probabilities and false discovery rates: two sides of the same coin.* J Proteome Res, 2008. **7**(1): p. 40-4.
96. Käll, L., Storey, J. D., and Noble, W. S., *QVALITY: non-parametric estimation of q-values and posterior error probabilities.* Bioinformatics (Oxford, England), 2009. **25**(7): p. 964-966.
97. Ferries, S., Perkins, S., Brownridge, P. J., Campbell, A., Eyers, P. A., Jones, A. R., and Eyers, C. E., *Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer.* J Proteome Res, 2017. **16**(9): p. 3448-3459.
98. Baker, P. R., Trinidad, J. C., and Chalkley, R. J., *Modification site localization scoring integrated into a search engine.* Mol Cell Proteomics, 2011. **10**(7): p. M111.008078.
99. Ramsbottom, K. A., Prakash, A., Riverol, Y. P., Camacho, O. M., Martin, M.-J., Vizcaíno, J. A., Deutsch, E. W., and Jones, A. R., *Method for Independent Estimation of the False Localization Rate for Phosphoproteomics.* Journal of Proteome Research, 2022. **21**(7): p. 1603-1615.
100. Ochoa, D., Jarnuczak, A. F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A. A., Hill, A., Garcia-Alonso, L., Stein, F., Krogan, N. J., Savitski, M. M., Swaney, D. L., Vizcaíno, J. A., Noh, K.-M., and Beltrao, P., *The functional landscape of the human phosphoproteome.* Nature Biotechnology, 2020. **38**(3): p. 365-373.
101. Zhao, M.-X., Chen, Q., Li, F., Fu, S., Huang, B., and Zhao, Y., *Protein phosphorylation database and prediction tools.* Briefings in Bioinformatics, 2023: p. bbad090.
102. Johnson, J. L., Yaron, T. M., Huntsman, E. M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D. M., Cohen, B. M., Vasan, N., Ma, Y., Krismer, K., Robles, J. T., van de Kooij, B., van Vlimmeren, A. E., Andrée-Busch, N., Käufer, N. F., Dorovkov, M. V., Ryazanov, A. G., Takagi, Y., Kastenhuber, E. R., Goncalves, M. D., Hopkins, B. D., Elemento, O., Taatjes, D. J., Maucuer, A., Yamashita, A., Degterev, A., Uduman, M., Lu, J., Landry, S. D., Zhang, B., Cossentino, I., Linding, R., Blenis, J., Hornbeck, P. V., Turk, B. E., Yaffe, M. B., and Cantley, L. C., *An atlas of substrate specificities for the human serine/threonine kinome.* Nature, 2023. **613**(7945): p. 759-766.
103. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R., *The PeptideAtlas project.* Nucleic acids research, 2006. **34**(Database issue): p. D655-D658.
104. Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R., *A guided tour of the Trans-Proteomic Pipeline.* Proteomics, 2010. **10**(6): p. 1150-1159.
105. Ma, K., Vitek, O., and Nesvizhskii, A. I., *A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet.* BMC Bioinformatics, 2012. **13**(16): p. S1.
106. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I., *iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates.* Mol Cell Proteomics, 2011. **10**(12): p. M111.007690.

107. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.
108. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A., *Human Protein Reference Database--2009 update.* Nucleic acids research, 2009. **37**(Database issue): p. D767-D772.
109. Willems, P., Horne, A., Van Parys, T., Goormachtig, S., De Smet, I., Botzki, A., Van Breusegem, F., and Gevaert, K., *The Plant PTM Viewer, a central resource for exploring plant protein modifications.* Plant J, 2019. **99**(4): p. 752-762.
110. Hu, Y., Sopko, R., Chung, V., Foos, M., Studer, R. A., Landry, S. D., Liu, D., Rabinow, L., Gnad, F., Beltrao, P., and Perrimon, N., *iProteinDB: An Integrative Database of Drosophila Post-translational Modifications.* G3 (Bethesda), 2019. **9**(1): p. 1-11.
111. Trost, B., Kusalik, A., and Napper, S., *Computational Analysis of the Predicted Evolutionary Conservation of Human Phosphorylation Sites.* PLoS One, 2016. **11**(4): p. e0152809.
112. Farris, S. M., *The rise to dominance of genetic model organisms and the decline of curiosity-driven organismal research.* PLoS One, 2020. **15**(12): p. e0243088.
113. Strumillo, M. J., Oplová, M., Viéitez, C., Ochoa, D., Shahraz, M., Busby, B. P., Sopko, R., Studer, R. A., Perrimon, N., Panse, V. G., and Beltrao, P., *Conserved phosphorylation hotspots in eukaryotic protein domain families.* Nat Commun, 2019. **10**(1): p. 1977.
114. Landry, C. R., Levy, E. D., and Michnick, S. W., *Weak functional constraints on phosphoproteomes.* Trends in Genetics, 2009. **25**(5): p. 193-197.
115. Lienhard, G. E., *Non-functional phosphorylations?* Trends in Biochemical Sciences, 2008. **33**(8): p. 351-352.
116. Capra, J. A. and Singh, M., *Predicting functionally important residues from sequence conservation.* Bioinformatics, 2007. **23**(15): p. 1875-1882.
117. Devos, D. and Valencia, A., *Practical limits of function prediction.* Proteins, 2000. **41**(1): p. 98-107.
118. Panchenko, A. R., Kondrashov, F., and Bryant, S., *Prediction of functional sites by analysis of sequence and structure conservation.* Protein Sci, 2004. **13**(4): p. 884-92.
119. Todd, A. E., Orengo, C. A., and Thornton, J. M., *Evolution of function in protein superfamilies, from a structural perspective.* J Mol Biol, 2001. **307**(4): p. 1113-43.
120. Wangler, M. F., Yamamoto, S., Chao, H. T., Posey, J. E., Westerfield, M., Postlethwait, J., Hieter, P., Boycott, K. M., Campeau, P. M., and Bellen, H. J., *Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research.* Genetics, 2017. **207**(1): p. 9-27.
121. Saucedo, L. J. and Edgar, B. A., *Filling out the Hippo pathway.* Nat Rev Mol Cell Biol, 2007. **8**(8): p. 613-21.
122. McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M., *Systematic discovery of nonobvious human disease models through orthologous phenotypes.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6544-9.
123. Mizushima, N., Noda, T., Yoshimori, T., Tanaka, Y., Ishii, T., George, M. D., Klionsky, D. J., Ohsumi, M., and Ohsumi, Y., *A protein conjugation system essential for autophagy.* Nature, 1998. **395**(6700): p. 395-398.
124. Huang, L., Kuo, Y.-M., and Gitschier, J., *The pallid gene encodes a novel, syntaxin 13-interacting protein involved in platelet storage pool deficiency.* Nature Genetics, 1999. **23**: p. 329-332.
125. Budovskaya, Y. V., Stephan, J. S., Deminoff, S. J., and Herman, P. K., *An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase.* Proceedings of the National Academy of Sciences, 2005. **102**(39): p. 13933-13938.
126. Malik, R., Nigg, E. A., and Körner, R., *Comparative conservation analysis of the human mitotic phosphoproteome.* Bioinformatics, 2008. **24**(12): p. 1426-1432.
127. Pearson, W. R., *An introduction to sequence similarity ("homology") searching.* Curr Protoc Bioinformatics, 2013. **Chapter 3**: p. Unit3.1.
128. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
129. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

130. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
131. Wootton, J. C., *Non-globular domains in protein sequences: Automated segmentation using complexity measures.* Computers & Chemistry, 1994. **18**(3): p. 269-285.
132. Henikoff, S. and Henikoff, J. G., *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
133. Pertsemlidis, A. and Fondon, J. W., 3rd, *Having a BLAST with bioinformatics (and avoiding BLASTphemy).* Genome Biol, 2001. **2**(10): p. Reviews2002.
134. Kerfeld, C. A. and Scott, K. M., *Using BLAST to teach "E-value-tionary" concepts.* PLoS Biol, 2011. **9**(2): p. e1001014.
135. Doolittle, R. F. and Feng, D.-F., *Nearest neighbor procedure for relating progressively aligned amino acid sequences.* Methods in Enzymology, 1990. **183**: p. 659-669.
136. Jankun-Kelly, T. J., Lindeman, A. D., and Bridges, S. M., *Exploratory visual analysis of conserved domains on multiple sequence alignments.* BMC Bioinformatics, 2009. **10**(11): p. S7.
137. Johansson, F. and Toh, H., *A comparative study of conservation and variation scores.* BMC Bioinformatics, 2010. **11**(1): p. 388.
138. Tamura, K., Stecher, G., and Kumar, S., *MEGA11: Molecular Evolutionary Genetics Analysis Version 11.* Mol Biol Evol, 2021. **38**(7): p. 3022-3027.
139. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J., *Jalview Version 2—a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-1191.
140. Edgar, R. C. and Batzoglou, S., *Multiple sequence alignment.* Curr Opin Struct Biol, 2006. **16**(3): p. 368-73.
141. Notredame, C., *Recent Evolutions of Multiple Sequence Alignment Algorithms.* PLOS Computational Biology, 2007. **3**(8): p. e123.
142. Thompson, J. D., Higgins, D. G., and Gibson, T. J., *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
143. Saitou, N. and Nei, M., *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
144. Katoh, K., Misawa, K., Kuma, K., and Miyata, T., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.
145. Edgar, R. C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic acids research, 2004. **32**(5): p. 1792-1797.
146. Sievers, F. and Higgins, D. G., *Clustal Omega for making accurate alignments of many protein sequences.* Protein Sci, 2018. **27**(1): p. 135-145.
147. Notredame, C., Higgins, D. G., and Heringa, J., *T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton.* Journal of Molecular Biology, 2000. **302**(1): p. 205-217.
148. Jacques, F., Bolivar, P., Pietras, K., and Hammarlund, E. U., *Roadmap to the study of gene and protein phylogeny and evolution-A practical guide.* PLoS One, 2023. **18**(2): p. e0279597.
149. Wang, L., Vervoort, V., Wallez, Y., Coré, N., Cremer, H., and Pasquale, E. B., *The SRC homology 2 domain protein Shep1 plays an important role in the penetration of olfactory sensory axons into the forebrain.* J Neurosci, 2010. **30**(39): p. 13201-10.
150. Kaneko, T., Joshi, R., Feller, S. M., and Li, S. S. C., *Phosphotyrosine recognition domains: the typical, the atypical and the versatile.* Cell Communication and Signaling, 2012. **10**(1): p. 32.
151. Salomon, A. R., Ficarro, S. B., Brill, L. M., Brinker, A., Phung, Q. T., Ericson, C., Sauer, K., Brock, A., Horn, D. M., Schultz, P. G., and Peters, E. C., *Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry.* Proc Natl Acad Sci U S A, 2003. **100**(2): p. 443-8.
152. Conceicao, C., Thakur, N., Human, S., Kelly, J. T., Logan, L., Bialy, D., Bhat, S., Stevenson-Leggett, P., Zagrajek, A. K., Hollinghurst, P., Varga, M., Tsirigoti, C., Tully, M., Chiu, C., Moffat, K., Silesian, A. P., Hammond, J. A., Maier, H. J., Bickerton, E., Shelton, H., Dietrich, I., Graham, S. C., and Bailey, D., *The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins.* PLOS Biology, 2020. **18**(12): p. e3001016.
153. Lei, K. C. and Zhang, X. D., *Conservation analysis of SARS-CoV-2 spike suggests complicated viral adaptation history from bat to human.* Evol Med Public Health, 2020. **2020**(1): p. 290-303.

154. Holub, A. S., Bouley, R. A., Petreaca, R. C., and Husbands, A. Y., *Identifying Cancer-Relevant Mutations in the DLC START Domain Using Evolutionary and Structure-Function Analyses.* Int J Mol Sci, 2020. **21**(21): p. 8175.
155. Takeda, T., Takahashi, M., Shimizu, M., Sugihara, Y., Yamashita, T., Saitoh, H., Fujisaki, K., Ishikawa, K., Utsushi, H., Kanzaki, E., Sakamoto, Y., Abe, A., and Terauchi, R., *Rice apoplastic CBM1-interacting protein counters blast pathogen invasion by binding conserved carbohydrate binding module 1 motif of fungal proteins.* PLOS Pathogens, 2022. **18**(9): p. e1010792.
156. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., and Ben-Tal, N., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(W1): p. W344-50.
157. Padhi, E. M., Ng, J. K., Mehinovic, E., Sams, E. I., and Turner, T. N., *ACES: Analysis of Conservation with an Extensive list of Species.* Bioinformatics, 2021. **37**(21): p. 3920-3922.
158. Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D., *Large-scale gene function analysis with the PANTHER classification system.* Nature Protocols, 2013. **8**(8): p. 1551-1566.
159. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P., *Ensembl comparative genomics resources.* Database (Oxford), 2016. **2016**: p. 1-17.
160. Saxena, R., Bishnoi, R., and Singla, D., *Chapter 9 - Gene Ontology: application and importance in functional annotation of the genomic data*, in *Bioinformatics*, D.B. Singh and R.K. Pathak, Editors. 2022, Academic Press. p. 145-157.
161. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
162. Huang, D. W., Sherman, B. T., and Lempicki, R. A., *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Research, 2009. **37**(1): p. 1-13.
163. Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., and Chang, W., *DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update).* Nucleic Acids Res, 2022. **50**(W1): p. W216-21.
164. Fisher, R. A., *Statistical Methods for Research Workers*, in *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N.L. Johnson, Editors. 1992, Springer New York: New York, NY. p. 66-70.
165. Kanehisa, M. and Goto, S., *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
166. Letunic, I., Khedkar, S., and Bork, P., *SMART: recent updates, new developments and status in 2020.* Nucleic Acids Research, 2021. **49**(D1): p. D458-D460.
167. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, Gustavo A., Bileschi, Maxwell L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, Daniel H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, Darren A., Orengo, Christine A., Pandurangan, Arun P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, Cathy H., and Bateman, A., *InterPro in 2022.* Nucleic Acids Research, 2023. **51**(D1): p. D418-D427.
168. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., and Yu, G., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.* Innovation (Camb), 2021. **2**(3): p. 100141.
169. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.
170. Mintseris, J. and Weng, Z., *Structure, function, and evolution of transient and obligate protein–protein interactions.* Proceedings of the National Academy of Sciences, 2005. **102**(31): p. 10930-10935.
171. Teppa, E., Zea, D. J., and Marino-Buslje, C., *Protein-protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces.* Protein Sci, 2017. **26**(12): p. 2438-2444.
172. Cross, F. R. and Jacobson, M. D., *Conservation and function of a potential substrate-binding domain in the yeast Clb5 B-type cyclin.* Mol Cell Biol, 2000. **20**(13): p. 4782-90.
173. Magliery, T. J. and Regan, L., *Sequence variation in ligand binding sites in proteins.* BMC Bioinformatics, 2005. **6**: p. 240.

174. Soccio, R. E., Tuteja, G., Everett, L. J., Li, Z., Lazar, M. A., and Kaestner, K. H., *Species-specific strategies underlying conserved functions of metabolic transcription factors.* Mol Endocrinol, 2011. **25**(4): p. 694-706.
175. Arslan, A. and van Noort, V., *Evolutionary conservation of Ebola virus proteins predicts important functions at residue level.* Bioinformatics, 2017. **33**(2): p. 151-154.
176. Chaikuad, A., Koschade, S. E., Stolz, A., Zivkovic, K., Pohl, C., Shaid, S., Ren, H., Lambert, L. J., Cosford, N. D. P., Brandts, C. H., and Knapp, S., *Conservation of structure, function and inhibitor binding in UNC-51-like kinase 1 and 2 (ULK1/2).* Biochem J, 2019. **476**(5): p. 875-887.
177. Liban, T. J., Medina, E. M., Tripathi, S., Sengupta, S., Henry, R. W., Buchler, N. E., and Rubin, S. M., *Conservation and divergence of C-terminal domain structure in the retinoblastoma protein family.* Proc Natl Acad Sci U S A, 2017. **114**(19): p. 4942-4947.
178. Petrova, N. V. and Wu, C. H., *Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties.* BMC Bioinformatics, 2006. **7**: p. 312.
179. Johnson, L. S., Eddy, S. R., and Portugaly, E., *Hidden Markov model speed heuristic and iterative HMM search procedure.* BMC Bioinformatics, 2010. **11**: p. 431.
180. Pearson, W. R., *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.* Genomics, 1991. **11**(3): p. 635-50.
181. Smith, T. F. and Waterman, M. S., *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
182. Nichio, B. T. L., Marchaukoski, J. N., and Raittz, R. T., *New Tools in Orthology Analysis: A Brief Review of Promising Perspectives.* Front Genet, 2017. **8**: p. 165.
183. Ullah, I., Sjöstrand, J., Andersson, P., Sennblad, B., and Lagergren, J., *Integrating Sequence Evolution into Probabilistic Orthology Analysis.* Syst Biol, 2015. **64**(6): p. 969-82.
184. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.
185. Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A., *Genomic evidence for two functionally distinct gene classes.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 6239-44.
186. Remm, M., Storm, C. E., and Sonnhammer, E. L., *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.* J Mol Biol, 2001. **314**(5): p. 1041-52.
187. Li, L., Stoeckert, C. J., Jr., and Roos, D. S., *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.
188. Ekseth, O. K., Kuiper, M., and Mironov, V., *orthAgogue: an agile tool for the rapid prediction of orthology relations.* Bioinformatics, 2014. **30**(5): p. 734-6.
189. Hu, X. and Friedberg, I., *SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier.* GigaScience, 2019. **8**(10): p. giz118.
190. Wang, K. and Samudrala, R., *Incorporating background frequency improves entropy-based residue conservation measures.* BMC Bioinformatics, 2006. **7**: p. 385.
191. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
192. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. **18 Suppl 1**: p. S71-7.
193. Toporik, A., Borukhov, I., Apatoff, A., Gerber, D., and Kliger, Y., *Computational identification of natural peptides based on analysis of molecular evolution.* Bioinformatics, 2014. **30**(15): p. 2137-41.
194. Sanner, M. F., *Python: a programming language for software integration and development.* J Mol Graph Model, 1999. **17**(1): p. 57-61.
195. Oliphant, T. E., *Python for Scientific Computing.* Computing in Science & Engineering, 2007. **9**(3): p. 10-20.
196. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold,

G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y. and SciPy, C., *SciPy 1.0: fundamental algorithms for scientific computing in Python.* Nature Methods, 2020. **17**(3): p. 261-272.

197. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
198. Bi, C., Huang, X., Tang, D., Shi, Y., Zhou, L., Hu, Y., Chen, X., Qi, S., and Lin, S., *A python script to design site-directed mutagenesis primers.* Protein Sci, 2020. **29**(4): p. 1054-1059.
199. Szabó, T. G., Palotai, R., Antal, P., Tokatly, I., Tóthfalusi, L., Lund, O., Nagy, G., Falus, A., and Buzás, E. I., *Critical role of glycosylation in determining the length and structure of T cell epitopes.* Immunome Res, 2009. **5**: p. 4.
200. Weil, P., Hoffgaard, F., and Hamacher, K., *Estimating sufficient statistics in co-evolutionary analysis by mutual information.* Computational Biology and Chemistry, 2009. **33**(6): p. 440-444.
201. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
202. Holt, L. J., Tuch, B. B., Villén, J., Johnson, A. D., Gygi, S. P., and Morgan, D. O., *Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution.* Science, 2009. **325**(5948): p. 1682-1686.
203. Mithoe, S. C., Boersema, P. J., Berke, L., Snel, B., Heck, A. J., and Menke, F. L., *Targeted quantitative phosphoproteomics approach for the detection of phospho-tyrosine signaling in plants.* J Proteome Res, 2012. **11**(1): p. 438-48.
204. Andoh, T., Hirata, Y., and Kikuchi, A., *Yeast glycogen synthase kinase 3 is involved in protein degradation in cooperation with Bul1, Bul2, and Rsp5.* Mol Cell Biol, 2000. **20**(18): p. 6712-20.
205. Jonak, C. and Hirt, H., *Glycogen synthase kinase 3/SHAGGY-like kinases in plants: an emerging family with novel functions.* Trends Plant Sci, 2002. **7**(10): p. 457-61.
206. Goedert, M., Spillantini, M. G., Cairns, N. J., and Crowther, R. A., *Tau proteins of alzheimer paired helical filaments: Abnormal phosphorylation of all six brain isoforms.* Neuron, 1992. **8**(1): p. 159-168.
207. Alessi, D. R., Andjelkovic, M., Caudwell, B., Cron, P., Morrice, N., Cohen, P., and Hemmings, B. A., *Mechanism of activation of protein kinase B by insulin and IGF-1.* Embo j, 1996. **15**(23): p. 6541-51.
208. Söderholm, S., Hintsanen, P., Öhman, T., Aittokallio, T., and Nyman, T. A., *PhosFox: a bioinformatics tool for peptide-level processing of LC-MS/MS-based phosphoproteomic data.* Proteome science, 2014. **12**: p. 36-36.
209. Wang, Y., Tian, Y., Liu, X., Dong, J., Wang, L., and Ye, M., *A New Workflow for the Analysis of Phosphosite Occupancy in Paired Samples by Integration of Proteomics and Phosphoproteomics Data Sets.* Journal of Proteome Research, 2020. **19**(9): p. 3807-3816.
210. Chalkley, R. J. and Clauser, K. R., *Modification site localization scoring: strategies and performance.* Mol Cell Proteomics, 2012. **11**(5): p. 3-14.
211. Hoopmann, M. R., Kusebauch, U., Palmblad, M., Bandeira, N., Shteynberg, D. D., He, L., Xia, B., Stoychev, S. H., Omenn, G. S., Weintraub, S. T., and Moritz, R. L., *Insights from the First Phosphopeptide Challenge of the MS Resource Pillar of the HUPO Human Proteome Project.* J Proteome Res, 2020. **19**(12): p. 4754-4765.
212. Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F., *Phospho.ELM: a database of phosphorylation sites--update 2011.* Nucleic acids research, 2011. **39**(Database issue): p. D261-D267.
213. Gnad, F., Gunawardena, J., and Mann, M., *PHOSIDA 2011: the posttranslational modification database.* Nucleic acids research, 2011. **39**(Database issue): p. D253-D260.
214. Huang, K.-Y., Lee, T.-Y., Kao, H.-J., Ma, C.-T., Lee, C.-C., Lin, T.-H., Chang, W.-C., and Huang, H.-D., *dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications.* Nucleic Acids Research, 2019. **47**(D1): p. D298-D308.

215. Giansanti, P., Aye, T. T., van den Toorn, H., Peng, M., van Breukelen, B., and Heck, A. J., *An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas.* Cell Rep, 2015. **11**(11): p. 1834-43.
216. Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., Hu, Y., Tan, Z., Stokes, M., Sullivan, L., Mitchell, J., Wetzel, R., Macneill, J., Ren, J. M., Yuan, J., Bakalarski, C. E., Villen, J., Kornhauser, J. M., Smith, B., Li, D., Zhou, X., Gygi, S. P., Gu, T. L., Polakiewicz, R. D., Rush, J., and Comb, M. J., *Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer.* Cell, 2007. **131**(6): p. 1190-203.
217. Boekhorst, J., van Breukelen, B., Heck, A., Jr., and Snel, B., *Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes.* Genome biology, 2008. **9**(10): p. R144-R144.
218. Studer, R. A., Rodriguez-Mias, R. A., Haas, K. M., Hsu, J. I., Viéitez, C., Solé, C., Swaney, D. L., Stanford, L. B., Liachko, I., Böttcher, R., Dunham, M. J., de Nadal, E., Posas, F., Beltrao, P., and Villén, J., *Evolution of protein phosphorylation across 18 fungal species.* Science, 2016. **354**(6309): p. 229-232.
219. Chen, S. C.-C., Chen, F.-C., and Li, W.-H., *Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals.* Molecular biology and evolution, 2010. **27**(11): p. 2548-2554.
220. Byrne, D. P., Clarke, C. J., Brownridge, P. J., Kalyuzhnyy, A., Perkins, S., Campbell, A., Mason, D., Jones, A. R., Eyers, P. A., and Eyers, C. E., *Use of the Polo-like kinase 4 (PLK4) inhibitor centrinone to investigate intracellular signalling networks using SILAC-based phosphoproteomics.* Biochem J, 2020. **477**(13): p. 2451-2475.
221. Hutti, J. E., Jarrell, E. T., Chang, J. D., Abbott, D. W., Storz, P., Toker, A., Cantley, L. C., and Turk, B. E., *A rapid method for determining protein kinase phosphorylation specificity.* Nat Methods, 2004. **1**(1): p. 27-9.
222. Kettenbach, A. N., Wang, T., Faherty, B. K., Madden, D. R., Knapp, S., Bailey-Kellogg, C., and Gerber, S. A., *Rapid determination of multiple linear kinase substrate motifs by mass spectrometry.* Chem Biol, 2012. **19**(5): p. 608-18.
223. Hall, F. L. and Vulliet, P. R., *Proline-directed protein phosphorylation and cell cycle regulation.* Current Opinion in Cell Biology, 1991. **3**(2): p. 176-184.
224. Johnson, L. N., Lowe, E. D., Noble, M. E., and Owen, D. J., *The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases.* FEBS Lett, 1998. **430**(1-2): p. 1-11.
225. Keshwani, M. M., Aubol, B. E., Fattet, L., Ma, C.-T., Qiu, J., Jennings, P. A., Fu, X.-D., and Adams, J. A., *Conserved proline-directed phosphorylation regulates SR protein conformation and splicing function.* The Biochemical journal, 2015. **466**(2): p. 311-322.
226. Lu, K. P., Liou, Y.-C., and Zhou, X. Z., *Pinning down proline-directed phosphorylation signaling.* Trends in Cell Biology, 2002. **12**(4): p. 164-172.
227. Pietrangelo, A. and Ridgway, N. D., *Phosphorylation of a serine/proline-rich motif in oxysterol binding protein-related protein 4L (ORP4L) regulates cholesterol and vimentin binding.* PloS one, 2019. **14**(3): p. e0214768-e0214768.
228. Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H., and Cantley, L. C., *Use of an oriented peptide library to determine the optimal substrates of protein kinases.* Curr Biol, 1994. **4**(11): p. 973-82.
229. Sugiyama, N., Imamura, H., and Ishihama, Y., *Large-scale Discovery of Substrates of the Human Kinome.* Scientific Reports, 2019. **9**(1): p. 10503.
230. Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J., DeMaggio, A. J., Hoekstra, M. F., Blenis, J., Hunter, T., and Cantley, L. C., *A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1.* Molecular and cellular biology, 1996. **16**(11): p. 6486-6493.
231. Wälchli, S., Espanel, X., Harrenga, A., Rossi, M., Cesareni, G., and Hooft van Huijsduijnen, R., *Probing protein-tyrosine phosphatase substrate specificity using a phosphotyrosine-containing phage library.* J Biol Chem, 2004. **279**(1): p. 311-8.
232. Espinos, E., Le Van Thaï, A., Pomiès, C., and Weber, M. J., *Cooperation between phosphorylation and acetylation processes in transcriptional control.* Mol Cell Biol, 1999. **19**(5): p. 3474-84.
233. Habibian, J. and Ferguson, B. S., *The Crosstalk between Acetylation and Phosphorylation: Emerging New Roles for HDAC Inhibitors in the Heart.* Int J Mol Sci, 2018. **20**(1).

234. Leney, A. C., El Atmioui, D., Wu, W., Ovaa, H., and Heck, A. J. R., *Elucidating crosstalk mechanisms between phosphorylation and O-GlcNAcylation.* Proc Natl Acad Sci U S A, 2017. **114**(35): p. E7255-e7261.
235. Naro, C. and Sette, C., *Phosphorylation-mediated regulation of alternative splicing in cancer.* Int J Cell Biol, 2013. **2013**: p. 151839.
236. Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J., *Systematic functional prioritization of protein posttranslational modifications.* Cell, 2012. **150**(2): p. 413-25.
237. Barbara, K. E., Willis, K. A., Haley, T. M., Deminoff, S. J., and Santangelo, G. M., *Coiled coil structures and transcription: an analysis of the S. cerevisiae coilome.* Molecular Genetics and Genomics, 2007. **278**(2): p. 135-147.
238. Baxevanis, A. D. and Vinson, C. R., *Interactions of coiled coils in transcription factors: where is the specificity?* Current Opinion in Genetics & Development, 1993. **3**(2): p. 278-285.
239. Pogenberg, V., Ballesteros-Álvarez, J., Schober, R., Sigvaldadóttir, I., Obarska-Kosinska, A., Milewski, M., Schindl, R., Ögmundsdóttir, M. H., Steingrímsson, E., and Wilmanns, M., *Mechanism of conditional partner selectivity in MITF/TFE family transcription factors with a conserved coiled coil stammer motif.* Nucleic Acids Research, 2020. **48**(2): p. 934-948.
240. Gu, G. M. and Wang, J. K., *[DNA-binding profiles of mammalian transcription factors].* Yi Chuan, 2012. **34**(8): p. 950-68.
241. Keegan, S., Cortens, J. P., Beavis, R. C., and Fenyö, D., *g2pDB: A Database Mapping Protein Post-Translational Modifications to Genomic Coordinates.* Journal of Proteome Research, 2016. **15**(3): p. 983-990.
242. Safaei, J., Maňuch, J., Gupta, A., Stacho, L., and Pelech, S., *Prediction of 492 human protein kinase substrate specificities.* Proteome Science, 2011. **9**(1): p. S6.
243. Campbell, A. E., Ferraz Franco, C., Su, L.-I., Corbin, E. K., Perkins, S., Kalyuzhnyy, A., Jones, A. R., Brownridge, P. J., Perkins, N. D., and Eyers, C. E., *Temporal modulation of the NF-κB RelA network in response to different types of DNA damage.* Biochemical Journal, 2021. **478**(3): p. 533-551.
244. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
245. Arabidopsis Genome Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.
246. Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S., *Evolution of protein kinase signaling from yeast to man.* Trends Biochem Sci, 2002. **27**(10): p. 514-20.
247. Hunter, T. and Plowman, G. D., *The protein kinases of budding yeast: six score and more.* Trends Biochem Sci, 1997. **22**(1): p. 18-22.
248. Kalyuzhnyy, A., Eyers, P. A., Eyers, C. E., Bowler-Barnett, E., Martin, M. J., Sun, Z., Deutsch, E. W., and Jones, A. R., *Profiling the Human Phosphoproteome to Estimate the True Extent of Protein Phosphorylation.* J Proteome Res, 2022. **21**(6): p. 1510-1524.
249. Ullah, S., Lin, S., Xu, Y., Deng, W., Ma, L., Zhang, Y., Liu, Z., and Xue, Y., *dbPAF: an integrative database of protein phosphorylation in animals and fungi.* Scientific Reports, 2016. **6**(1): p. 23534.
250. Sadowski, I., Breitkreutz, B. J., Stark, C., Su, T. C., Dahabieh, M., Raithatha, S., Bernhard, W., Oughtred, R., Dolinski, K., Barreto, K., and Tyers, M., *The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update.* Database (Oxford), 2013. **2013**: p. bat026.
251. Rinschen, M. M., Pahmeyer, C., Pisitkun, T., Schnell, N., Wu, X., Maaß, M., Bartram, M. P., Lamkemeyer, T., Schermer, B., Benzing, T., and Brinkkoetter, P. T., *Comparative phosphoproteomic analysis of mammalian glomeruli reveals conserved podocin C-terminal phosphorylation as a determinant of slit diaphragm complex architecture.* Proteomics, 2015. **15**(7): p. 1326-31.
252. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A., *Pfam: The protein families database in 2021.* Nucleic Acids Research, 2021. **49**(D1): p. D412-D419.
253. Meyer, C., Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D., *Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes.* BMC Bioinformatics, 2020. **21**(1): p. 513.
254. Reeves, G. A., Talavera, D., and Thornton, J. M., *Genome and proteome annotation: organization, interpretation and integration.* Journal of The Royal Society Interface, 2008. **6**(31): p. 129-147.
255. Blom, N., Gammeltoft, S., and Brunak, S., *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.* J Mol Biol, 1999. **294**(5): p. 1351-62.

256. Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I., *NCBI Taxonomy: a comprehensive update on curation, resources and tools.* Database (Oxford), 2020. **2020**: p. baaa062.
257. Letunic, I. and Bork, P., *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Res, 2021. **49**(W1): p. W293-w296.
258. Kolde, R. *pheatmap: Pretty Heatmaps. R package version 1.0.12.* 2019; Available from: https://CRAN.R-project.org/package=pheatmap.
259. R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.* 2021; Available from: https://www.R-project.org.
260. Fossati, M., Pizzarelli, R., Schmidt, E. R., Kupferman, J. V., Stroebel, D., Polleux, F., and Charrier, C., *SRGAP2 and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses.* Neuron, 2016. **91**(2): p. 356-69.
261. Sporny, M., Guez-Haddad, J., Kreusch, A., Shakartzi, S., Neznansky, A., Cross, A., Isupov, M. N., Qualmann, B., Kessels, M. M., and Opatowsky, Y., *Structural History of Human SRGAP2 Proteins.* Mol Biol Evol, 2017. **34**(6): p. 1463-1478.
262. Guen, V. J., Gamble, C., Lees, J. A., and Colas, P., *The awakening of the CDK10/Cyclin M protein kinase.* Oncotarget, 2017. **8**(30): p. 50174-50186.
263. Malumbres, M. and Barbacid, M., *Cell cycle, CDKs and cancer: a changing paradigm.* Nat Rev Cancer, 2009. **9**(3): p. 153-66.
264. Friedrich, J. K., Panov, K. I., Cabart, P., Russell, J., and Zomerdijk, J. C., *TBP-TAF complex SL1 directs RNA polymerase I pre-initiation complex formation and stabilizes upstream binding factor at the rDNA promoter.* J Biol Chem, 2005. **280**(33): p. 29551-8.
265. Nowick, K., Hamilton, A. T., Zhang, H., and Stubbs, L., *Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes.* Mol Biol Evol, 2010. **27**(11): p. 2606-17.
266. Turelli, P., Playfoot, C., Grun, D., Raclot, C., Pontis, J., Coudray, A., Thorball, C., Duc, J., Pankevich, E. V., Deplancke, B., Busskamp, V., and Trono, D., *Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons.* Sci Adv, 2020. **6**(35): p. eaba3200.
267. Gjerstorff, M. F. and Ditzel, H. J., *An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members.* Tissue Antigens, 2008. **71**(3): p. 187-92.
268. Herrmann, H., Bär, H., Kreplak, L., Strelkov, S. V., and Aebi, U., *Intermediate filaments: from cell architecture to nanomechanics.* Nature Reviews Molecular Cell Biology, 2007. **8**(7): p. 562-573.
269. Ho, M., Thompson, B., Fisk, J. N., Nebert, D. W., Bruford, E. A., Vasiliou, V., and Bunick, C. G., *Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders.* Human Genomics, 2022. **16**(1): p. 1.
270. Vandebergh, W. and Bossuyt, F., *Radiation and functional diversification of alpha keratins during early vertebrate evolution.* Mol Biol Evol, 2012. **29**(3): p. 995-1004.
271. Cuny, G. D., Robin, M., Ulyanova, N. P., Patnaik, D., Pique, V., Casano, G., Liu, J. F., Lin, X., Xian, J., Glicksman, M. A., Stein, R. L., and Higgins, J. M., *Structure-activity relationship study of acridine analogs as haspin and DYRK2 kinase inhibitors.* Bioorg Med Chem Lett, 2010. **20**(12): p. 3491-4.
272. Ikeda, D., Ono, Y., Snell, P., Edwards, Y. J., Elgar, G., and Watabe, S., *Divergent evolution of the myosin heavy chain gene family in fish and tetrapods: evidence from comparative genomic analysis.* Physiol Genomics, 2007. **32**(1): p. 1-15.
273. Aguilar-Cuenca, R., Llorente-González, C., Chapman, J. R., Talayero, V. C., Garrido-Casado, M., Delgado-Arévalo, C., Millán-Salanova, M., Shabanowitz, J., Hunt, D. F., Sellers, J. R., Heissler, S. M., and Vicente-Manzanares, M., *Tyrosine Phosphorylation of the Myosin Regulatory Light Chain Controls Non-muscle Myosin II Assembly and Function in Migrating Cells.* Curr Biol, 2020. **30**(13): p. 2446-2458.
274. Vicente-Manzanares, M., Ma, X., Adelstein, R. S., and Horwitz, A. R., *Non-muscle myosin II takes centre stage in cell adhesion and migration.* Nature Reviews Molecular Cell Biology, 2009. **10**(11): p. 778-790.
275. Kim, B. J., Kim, A. R., Han, J. H., Lee, C., Oh, D. Y., and Choi, B. Y., *Discovery of MYH14 as an important and unique deafness gene causing prelingually severe autosomal dominant nonsyndromic hearing loss.* J Gene Med, 2017. **19**(4): p. e2950.

276. Babuke, T., Ruonala, M., Meister, M., Amaddii, M., Genzler, C., Esposito, A., and Tikkanen, R., *Hetero-oligomerization of reggie-1/flotillin-2 and reggie-2/flotillin-1 is required for their endocytosis.* Cell Signal, 2009. **21**(8): p. 1287-97.
277. Banning, A., Kurrle, N., Meister, M., and Tikkanen, R., *Flotillins in receptor tyrosine kinase signaling and cancer.* Cells, 2014. **3**(1): p. 129-49.
278. Francois, C. M., Durand, F., Figuet, E., and Galtier, N., *Prevalence and Implications of Contamination in Public Genomic Resources: A Case Study of 43 Reference Arthropod Assemblies.* G3 (Bethesda), 2020. **10**(2): p. 721-730.
279. Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., and Hein, J., *Uncertainty in homology inferences: assessing and improving genomic sequence alignment.* Genome Res, 2008. **18**(2): p. 298-309.
280. Zhang, C., Zhao, Y., Braun, E. L., and Mirarab, S., *TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution.* Methods in Ecology and Evolution, 2021. **12**(11): p. 2145-2158.
281. de la Cruz, J., Kressler, D., and Linder, P., *Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families.* Trends Biochem Sci, 1999. **24**(5): p. 192-8.
282. Garwain, O., Sun, X., Iyer, D. R., Li, R., Zhu, L. J., and Kaufman, P. D., *The chromatin-binding domain of Ki-67 together with p53 protects human chromosomes from mitotic damage.* Proc Natl Acad Sci U S A, 2021. **118**(32): p. e2021998118.
283. Hur, E. M., Son, M., Lee, O. H., Choi, Y. B., Park, C., Lee, H., and Yun, Y., *LIME, a novel transmembrane adaptor protein, associates with p56lck and mediates T cell activation.* J Exp Med, 2003. **198**(10): p. 1463-73.
284. Park, I., Son, M., Ahn, E., Kim, Y. W., Kong, Y. Y., and Yun, Y., *The Transmembrane Adaptor Protein LIME Is Essential for Chemokine-Mediated Migration of Effector T Cells to Inflammatiory Sites.* Mol Cells, 2020. **43**(11): p. 921-934.
285. Araújo, T. G., Mota, S. T. S., Ferreira, H. S. V., Ribeiro, M. A., Goulart, L. R., and Vecchi, L., *Annexin A1 as a Regulator of Immune Response in Cancer.* Cells, 2021. **10**(9): p. 2245.
286. Varkey, J. and Nagaraj, R., *Antibacterial activity of human neutrophil defensin HNP-1 analogs without cysteines.* Antimicrob Agents Chemother, 2005. **49**(11): p. 4561-6.
287. Quinlan, R., Hutchison, C., and Lane, B., *Intermediate filament proteins.* Protein profile, 1995. **2**(8): p. 795-952.
288. Rigden, D. J., *The histidine phosphatase superfamily: structure and function.* Biochem J, 2008. **409**(2): p. 333-48.
289. Rigden, D. J., Walter, R. A., Phillips, S. E. V., and Fothergill-Gilmore, L. A., *Polyanionic Inhibitors of Phosphoglycerate Mutase: Combined Structural and Biochemical Analysis.* Journal of Molecular Biology, 1999. **289**(4): p. 691-699.
290. Westram, A., Lloyd, J. R., Roessner, U., Riesmeier, J. W., and Kossmann, J., *Increases of 3-phosphoglyceric acid in potato plants through antisense reduction of cytoplasmic phosphoglycerate mutase impairs photosynthesis and growth, but does not increase starch contents.* Plant, Cell & Environment, 2002. **25**(9): p. 1133-1143.
291. Zhao, Z. and Assmann, S. M., *The glycolytic enzyme, phosphoglycerate mutase, has critical roles in stomatal movement, vegetative growth, and pollen production in Arabidopsis thaliana.* J Exp Bot, 2011. **62**(14): p. 5179-89.
292. Hitosugi, T., Zhou, L., Fan, J., Elf, S., Zhang, L., Xie, J., Wang, Y., Gu, T. L., Alečković, M., LeRoy, G., Kang, Y., Kang, H. B., Seo, J. H., Shan, C., Jin, P., Gong, W., Lonial, S., Arellano, M. L., Khoury, H. J., Chen, G. Z., Shin, D. M., Khuri, F. R., Boggon, T. J., Kang, S., He, C., and Chen, J., *Tyr26 phosphorylation of PGAM1 provides a metabolic advantage to tumours by stabilizing the active conformation.* Nat Commun, 2013. **4**: p. 1790.
293. Wang, Y., Wei, Z., Bian, Q., Cheng, Z., Wan, M., Liu, L., and Gong, W., *Crystal Structure of Human Bisphosphoglycerate Mutase.* Journal of Biological Chemistry, 2004. **279**(37): p. 39132-39138.
294. Berginski, M. E., Moret, N., Liu, C., Goldfarb, D., Sorger, Peter K., and Gomez, S. M., *The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases.* Nucleic Acids Research, 2021. **49**(D1): p. D529-D535.
295. Needham, E. J., Parker, B. L., Burykin, T., James, D. E., and Humphrey, S. J., *Illuminating the dark phosphoproteome.* Science Signaling, 2019. **12**(565): p. eaau8645.
296. Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. T. M., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G.,

Samson, L. D., Woodgett, J. R., Russell, Robert B., Bork, P., Yaffe, M. B., and Pawson, T., *Systematic Discovery of In Vivo Phosphorylation Networks.* Cell, 2007. **129**(7): p. 1415-1426.