Monitoring Functional Post-Translational Modifications Using a Data-Driven Proteome Informatic Pipeline Based on PEIMAN2

Payman Nickchi[1], Mehdi Mirzaie[2,3], Marc Baumann[3], Amir Ata Saei*§[4], Mohieddin Jafari*[3]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, British Columbia

[2]Department of Pharmacology, Faculty of Medicine & Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland

[3]Medicum, Department of Biochemistry and Developmental Biology, Meilahti Clinical Proteomics Core Facility, University of Helsinki, Helsinki, Finland

[4]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 65 Stockholm, Sweden

§Current address: Biozentrum, University of Basel, 4056 Basel, Switzerland

Abstract

Post-translational modifications (PTMs) are under significant focus in molecular biomedicine due to their importance in signal transduction in most cellular and organismal processes. Characterization of PTMs, discrimination between functional and inert PTMs, quantification of their occupancies and PTM crosstalk are demanding tasks in each biosystem. On top of that, the study of each PTM often necessitates a particular laborious experimental design. Here, we present a PTM-centric proteome informatic pipeline for prediction of most probable and relevant PTMs in mass spectrometry-based proteomics data. Upon prediction, such PTMs can be incorporated in a refined database search. To demonstrate the applicability of our approach, using expression profiling, we identified cellular proteins that are differentially regulated in response to multikinase inhibitors dasatinib and staurosporine. Computational enrichment analysis was employed to determine the potential PTMs of protein targets for both drugs. Finally, we conducted an additional round of database search with the predicted probable PTMs. Our pipeline helped to analyze the enriched PTMs and even the detected proteins that were not identified in the initial search. Our findings support the idea of PTM-centric searching of MS data in proteomics based on computational enrichment analysis and we believe this strategy should be incorporated in future proteomics search engines.

Main

Proteins are the primary functional units of cellular systems, but they often gain activity when modified post-translationally. In addition to regulation of protein activity, their function, stability/solubility, interactions with other biomolecules and their cellular localization are governed by transient modulation of post-translational modifications

(PTMs)[1,2]. By regulating such diverse characteristics, PTMs can modulate the involvement of proteins in biochemical reactions, signaling, transport, structural remodeling, gene regulation, cell motility and cell death[3,4]. Due to the importance of PTMs in signal transduction in health and disease, the mechanisms and kinetics of PTMs have turned into an active research area[5-9].

Analysis of PTMs would provide valuable information regarding the status and function of proteins upon diverse perturbations[10]. Therefore, understanding the nature, quantity, and temporal progression of PTMs has arguably been one of the most substantial contributions of MS-based proteomics to modern biology[1]. However, the sub-stoichiometric nature and dynamic regulation of PTMs makes it challenging to capture and detect PTMs[11]. Thus, unique enrichment techniques and sample-processing workflows are often required for enriching PTMs before analysis by mass spectrometry.

Experimental techniques exploit the unique chemical properties of a given PTM for their enrichment. For example, at both protein and peptide levels, PTM-directed antibodies can be used to enrich a specific chemical group within a given proteome[12]. Another routinely used strategy involves the enrichment of phosphorylated peptides (and/or proteins) using metal oxide resins, such as titanium and zirconium[13]. These enrichment strategies are very prominent when modulation of a certain PTM is expected; for example, when investigating the function of a kinase, modulation of phosphorylation levels is an expected outcome.

Furthermore, the inclusion of more PTMs in database searches can dramatically enlarge the search space, imposing time constraints and heavily straining the search engine[14]. It should also be noted that including any extra PTMs increases the chance of false identifications in a given database search and thus increasing the burden of proof for PTM identification. Therefore, only most common PTMs, such as asparagine deamidation[15], methionine oxidation and cysteine carbamidomethylation are usually included in routine database searches.

Altogether, due to a lack of prior knowledge on the most important PTMs in a particular study condition, many PTMs are usually not monitored. Although no practical issues exist in the biochemical characterization of stable and common PTMs such as phosphorylation or acetylation, researchers do not monitor them in the lack of presumption. Despite the presence of proteome-wide PTM approaches such as ModifiComb[16,17], the analysis of less-common or unstable PTMs still remains challenging and needs a more complex study design, especially for PTMs without highly specific antibodies or reagents.

Previously, we introduced a computational enrichment analysis for PTMs by a standalone software called “Post-translational modification Enrichment Integration and Matching Analysis” or PEIMAN, to explore most probable and enriched PTMs on protein lists[18]. The enriched PTMs are the PTM terms (extracted from UniProt for all the proteins) that are occurring more frequently than by pure chance for the analyzed proteins. The software identifies these terms by applying a hypergeometric test.

Here, we present in detail how our new PEIMAN2 R package boosts PTM-centric discovery proteomics by providing two case studies involving multikinase inhibitors, i.e., dasatinib and staurosporine. To discover the differentially regulated mechanistic proteins for the

above drugs, we perform deep expression profiling of the model cell line A549 (representing lung cancer) treated by above drugs in a series of concentrations. Then, PEIMAN2 (Fig. 1) is used to identify the most probable and enriched PTMs among the differentially-regulated proteins. Finally, the most probable PTMs are incorporated in a refined database search to assess the applicability of PEIMAN2 pipeline.

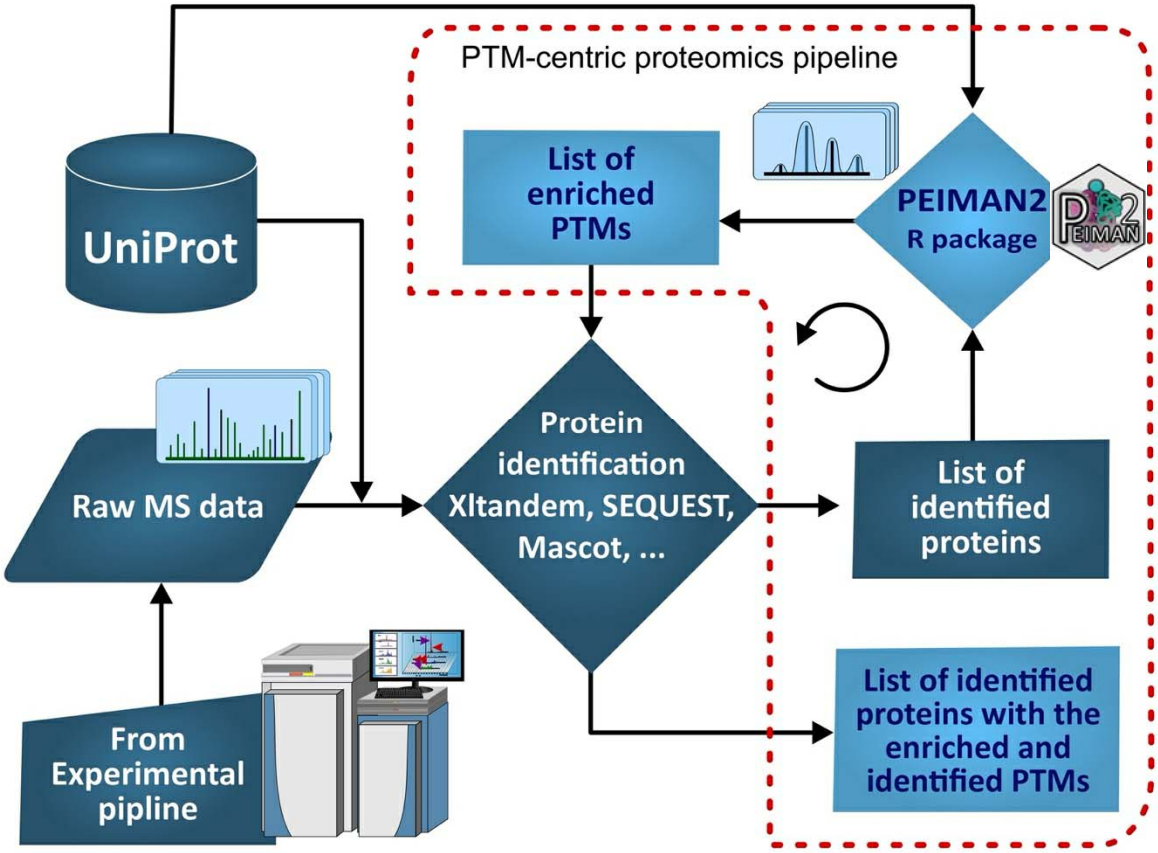


Figure 1: *An informatic workflow for PTM-centric proteomics using PEIMAN2 R package.*

Results

PTM data extraction for PEIMAN2 R package

The PEIMAN2 R package contains as of August 2022 a database that encompasses 223,292 proteins from 9,480 organisms, covering 525 PTM types. We compared the distribution of the 10 most prevalent PTMs across all known organisms to gain insight into the prevalence of PTMs in the current version of UniProt vs 2015 version. Figure 2 panel (a) shows the distribution of the ten most common PTMs in the current version of UniProt. The PTM "Nucleotide-binding" had a frequency of 101,625 occurrences which is about 10 orders of magnitude more frequent than "Methylation" (which is the 10th most frequent PTM) with a frequency of 8,532 occurrences (however, unlike covalent intercalation, non-covalent binding of proteins to nucleotides may not be considered a PTM). Panel (b) in the figure compares the changes in the frequencies of these ten most common PTMs in the current and previous versions of UniProt database. Note that the database version is denoted on the x axis and the y axis represents the frequency of PTMs. To better highlight the changes, the frequency of PTMs is shown in log-10 base. Over the course of the past seven years, we expected to discover a consistent pattern of growth among the top 10 selected PTMs. However, the rates of growth are not consistent among all terms. For example, "Phosphoserine", "Phosphothreonine", and "Methylation" have a higher rate of growth compared to the other PTM-terms. This difference in the rate of growth might be related to two reasons. First, the identification of some PTMs is subject to experimental limitations, therefore we cannot expect a consistent rate of growth. Second, the assigned biological activities of some PTMs such as phosphorylation has attracted more research attention than some other PTMs.

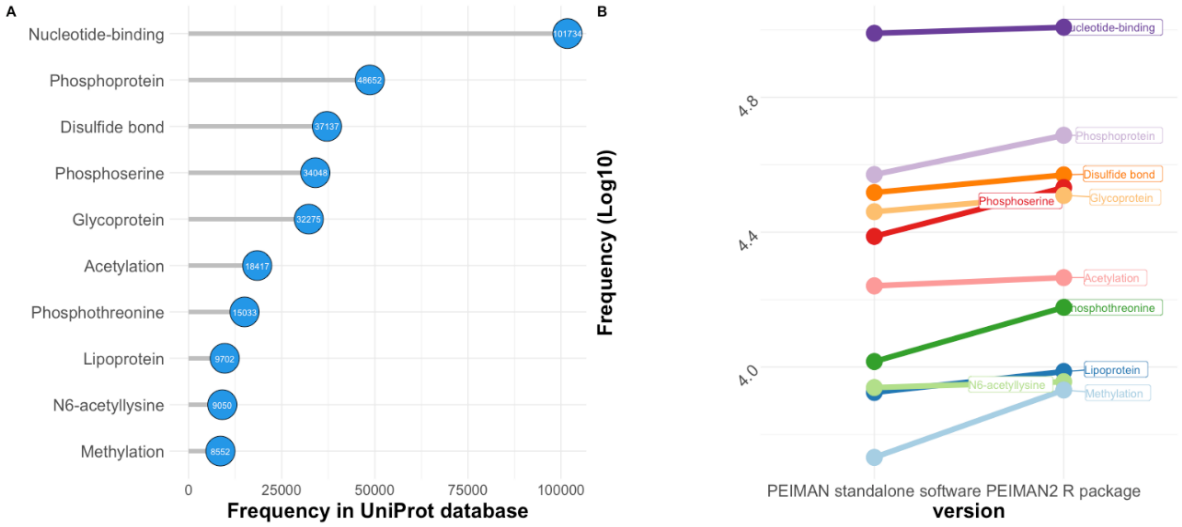


Figure 2: *An update on PTM statistics based on UniProt database. (A) The top ten most common PTMs in UniProt/SwissProt across all available species. (B) The frequency changes of the ten most common PTMs in PEIMAN standalone software (2015) vs PEIMAN2 R package (2022).*

In addition, it was interesting to investigate the distribution of all PTMs among different species in UniProt/SwissProt. We plotted the tree maps for 8 model organisms from diverse taxonomic branches of life (Supplementary Figure 1). This distribution of PTMs among the distinct model organisms provides a hint that PTMs can be utilized for taxonomic discrimination within the tree of life. A t-SNE plot was also built; it clearly showed how four major super kingdoms can be classified depending on the PTM profiles of their species (Supplementary Figure 2).

PEIMAN2 R package functionality

We introduced PEIMAN standalone software in 2015 to facilitate single enrichment analysis based on PTMs in proteomics studies[18]. Single enrichment analysis is a popular method providing insight into biological pathways altered in disease or under various perturbations. The idea of single enrichment analysis is to check whether the genes/proteins with a specific biological feature in a given list are occurring more frequently than by pure chance. As simple and powerful as this approach is, there are some known drawbacks to it, including the difficulty of identifying significant signals from noise, subjective interpretations among biologists, and for the same data getting different final list of significant genes/proteins among different laboratories. Gene/Protein set enrichment is an alternative way to resolve these problems, therefore we provided a new enrichment method for PTM study called protein set enrichment analysis (PSEA) in an R package, making this tool accessible for a broader community of researchers.

The PEIMAN2 R package offers a wider range of features and functionalities compared with the PEIMAN standalone software. First note that single enrichment analysis is still included in the PEIMAN2 R package and can be utilized by calling `runEnrichment()` function. This function requires the user to pass a list of protein accession codes (UniProt/SwissProt) for a specific organism along with an adjustment method for multiple testing. The output is a table of enriched PTMs sorted based on their adjusted p-values. To visualize the single enrichment analysis of a list of proteins or integrating the results of two single enrichment analyses, one can use the `plotEnrichment()` function.

As a new feature, PEIMAN2 package implements PSEA as an additional tool for proteomics studies based on PTMs. `runPSEA()` function in the package allows the user to perform a PSEA analysis on a given list of proteins for a specific organism. This function requires a list of protein accession codes along with taxonomy name of the organism. Some additional parameters of the function include enrichment weighting (refer to the methods section for more details), number of permutations to estimate the false discovery rate, FDR (default number of permutations is 1000), choice of the method to adjust p-values, and a controlling cut off to include specific PTMs with a certain occurrence rate in the analysis. A table of enriched proteins along with their enrichment and normalized enriched score, adjusted p-value, FDR, and proteins in the leading-edge will be produced. For each PTM, the leading-edge proteins are the proteins that show up in the ranked list at or before the point where the enrichment score (ES) reaches its maximum deviation from zero.

There are two functions available in the package to visualize the results of PSEA, `plotPSEA()` and `plotRunningScore()`. The first function plots the results of one PSEA

analysis or matches and integrates the results of two PSEA analyses. The resulting plot shows the normalized enrichment score for each PTM. The second function presents the running enrichment score plot for each PTM in the table generated by `runPSEA()` function. In each plot, x-axis is the sorted protein list by proteins score and y-axis is the enrichment score. The leading-edge proteins are shown with a rug plot on the x-axis (for example see Supplementary Figure 3 and 4).

In PTM-centric proteomics, we suggest including PEIMAN2 results within the workflow of mass spectrometry data analysis. In other words, one can perform SEA or PSEA to obtain a list of enriched terms. These are PTM terms of which the counts in a given list of protein is statistically significant. The SEA and PSEA are among two methods to obtain the list of enriched PTM terms (see methods section for more details). The enriched terms can then be used to extract the subset of protein modifications. These modifications can be used to search in a proteomics search engine software to gain more insight into designing the experiment and investigating the effect of a treatment on PTMs. For this purpose, we included functions to prepare results for such a re-search in MaxQuant software. The results of SEA or PSEA can be passed to `sea2mass()` or `psea2mass()` functions, respectively, to extract a subset of protein modifications. This subset of chemical modifications can be used to parametrize the search engine for mass spectrometry data, such as MaxQuant. For more information, we have provided a detailed vignette manual along with the package and a Readme page on PEIMAN2's GitHub directory (https://github.com/jafarilab/PEIMAN2).

Multikinase inhibitors case study with PEIMAN2

In order to investigate the ability of PEIMAN2 to computationally enrich and predict probable PTMs, we performed deep expression profiling of two drugs, i.e., multikinase inhibitors dasatinib and staurosporine, to identify the mechanistic proteins exhibiting differential expression upon treatment. Since both drugs are multikinase inhibitors, the treatment would be expected to modulate the number of phosphorylation events and/or modification occupancies, as shown before for AKT1/2 inhibitor and ipatasertib[19]. Therefore, we expected to see an enrichment of phosphorylation PTM terms after applying PEIMAN2 on the differentially regulated proteins in response to both drugs. To compute the differentially expressed proteins, we tested four increasing concentrations of both drugs in an A549 cell line model of lung cancer taking advantage of TMTpro 16 multiplexing[20]. As expected, a higher number of differentially expressed proteins were identified at higher concentrations of drug. Then, we tested PEIMAN2 on all concentrations of drugs and checked the number of differentially changing proteins with annotated PTM term changes at different drug concentration levels based on enrichment analysis. Panels (A) and (B) in Figure 3, present fold changes of proteins between control and conc.4 (highest concentration of drugs) measured in log2 scale versus their p-value derived from a two-sided t-test with equal variance assumption for dasatinib and staurosporine drugs, respectively. In the plot, each data point represents one protein, where rectangles denote the corresponding enriched PTM terms. The presence of various PTMs on the differentially abundant proteins can be observed on the plots. Figure 3 panels (C) and (D) present the number of differentially expressed proteins with annotated PTM terms for each drug at each concentration, grouped by PTM. Figure 3 panels (E) and (F) show the number of differentially expressed proteins for each drug. The percentage on the bar indicates the

proportion of proteins that have phosphorylation related PTM terms. One can note that as the concentration of drug increases from Conc.1 to conc.4, the number of potential protein targets with specific PTM terms increases. Therefore, we considered conc.4 for the downstream analyses including a refined MaxQuant database search. For emphasis, we tried to detect actual PTMs in mass spectrometry data based on the PTMs that were enriched in the identified proteins.

More specifically, for obtaining the most probable modifications changing under the treatment, at the first step of analysis, we implemented PSEA method by calling `runPSEA()` function on the samples treated with the highest concentration of drugs, to identify enriched modification terms. We considered the default value for weight exponent of one. As for permutation, we considered randomly permuting scores of proteins 1000 times to adjust for FDR. A significance level of 5 percent along with a reasonable cut-off for PTM frequency in UniProt/Swissprot was applied to each drug list, separately. The exact modification of each enriched PTM was obtained by calling `psea2mass()` function. The top five modifications for dasatinib were: 'O-phospho-L-threonine', 'N6-acetyl-L-lysine', 'N-acetyl-L-alanine', 'O4-phospho-L-tyrosine', and 'N-acetyl-L-methionine'. The corresponding PTMs are: 'Phosphothreonine', 'N6-acetyllysine', 'N-acetylalanine', 'Phosphotyrosine', 'N-acetylmethionine'. On the other hand, the top five modifications for staurosporine were: 'O-phospho-L-serine', 'O-phospho-L-threonine', 'O4-phospho-L-tyrosine', 'N-acetyl-L-alanine', and 'N-acetyl-L-methionine'. The corresponding PTMs are: 'Phosphoserine', 'Phosphoprotein', 'Acetylation', 'Phosphothreonine', and 'N6-acetyllysine'. In Supplementary Figure 3 and 4, we show the running score plot of the top five modifications shown in the integrated normalized enrichment score plot for dasatinib and staurosporine, respectively. In summary, PSEA shows enriched phosphorylation for both drug treatments (P-value < 0.05), however, the top PTMs for these kinase inhibitors were distinct and specific. In the next step, we performed a refined search using MaxQuant, where the top 5 modifications suggested by PEIMAN2 were added as further variable modifications (other parameters were kept constant as with the initial search). Figure 3G presents the integrated results of PSEA for two drugs. The x-axis in Figure 3G shows the normalized enrichment score obtained from 1000 permutations and y-axis shows the PTM type.

Number of proteins, peptides, and sequence coverage before/after using PEIMAN2

First, we performed a quality control to ensure that the number of unique peptides, sequence coverage, and q-values are not significantly changed after including additional enriched modifications suggested by PEIMAN2 on the proteomics search engine software, i.e., MaxQuant. Note that the re-searching parameters (except the selected modifications) as well as filtering criteria of selecting the identified proteins remained the same both before/after including PEIMAN2 suggestions. Figure 4 shows the box plot of sequence coverage of proteins for each unique number of peptides colored by their q-value before/after including additional enriched modifications suggested by PEIMAN2. The results of analysis before/after applying PEIMAN2 for dasatinib and staurosporine are presented in panels (A-B) and (C-D), respectively. For both drugs, the median of sequence

coverage at each unique number of peptides was similar before and after applying PEIMAN2 suggested modifications.

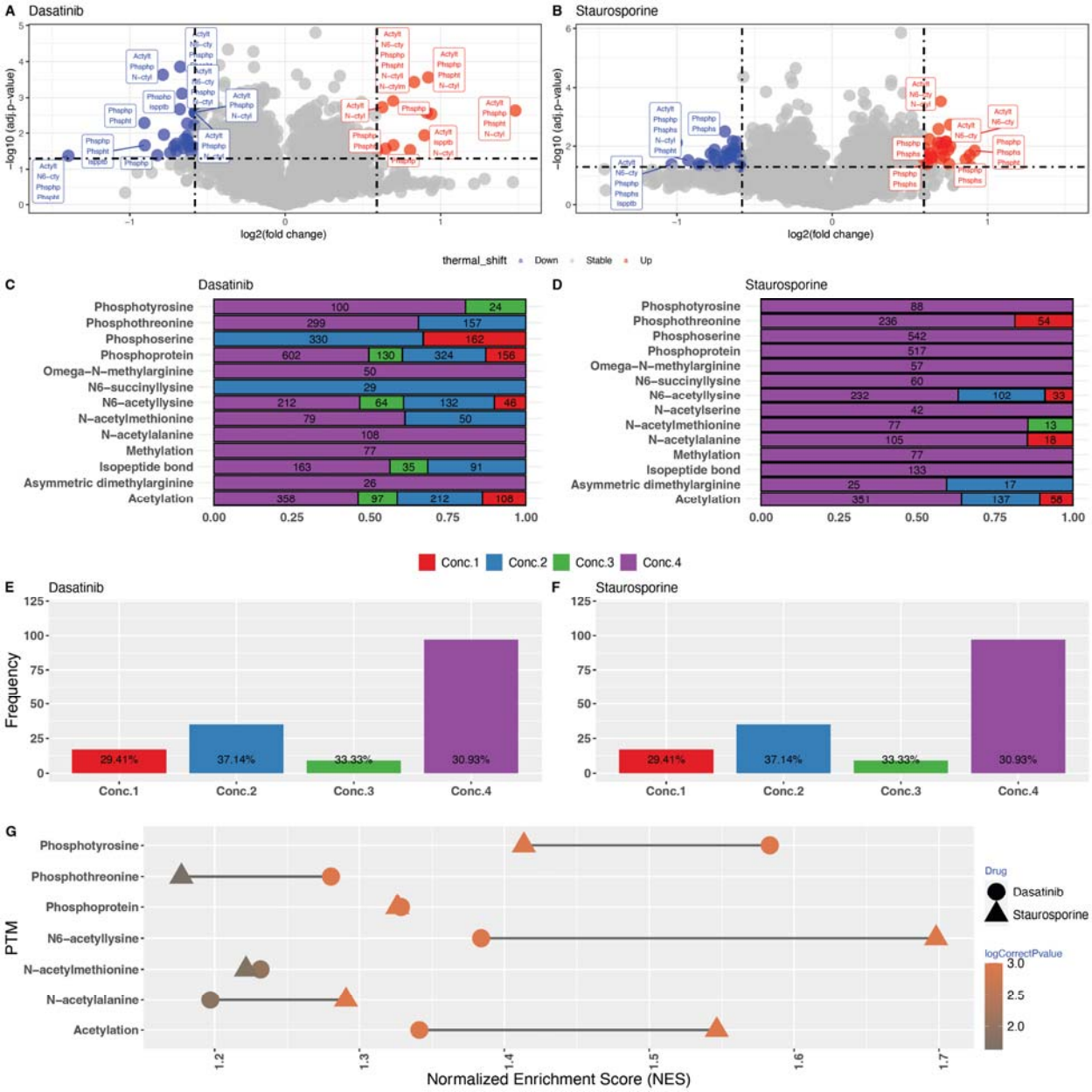


Figure 3: *Volcano plot of fold change versus p-value colored by the sign of regulation for both drugs; the PTMs annotated in UniProt for each protein are shown in rectangles; (A) Dasatinib, (B) Staurosporine. The results are shown for concentration 4 vs. vehicle treated cells. Bar plot of number of differentially regulated proteins with PTM at four concentration levels of both drugs (see Supplementary data 1 and 2 for more details); (C) Dasatinib, (D) Staurosporine. The value of absolute numbers of differentially regulated proteins carrying different PTMs at each concentration level is labeled in the bars of the plot; (E) Dasatinib, (F) Staurosporine. The percentage of proteins with phosphorylation related PTMs are labeled on the bars. (G) Integrated normalized enrichment score (NES) plot for both drugs colored by corrected p-value. The data points for dasatinib and staurosporine are plotted with a filled circle and triangle, respectively. The points are colored with their corrected p-value presented in log10 scale.*

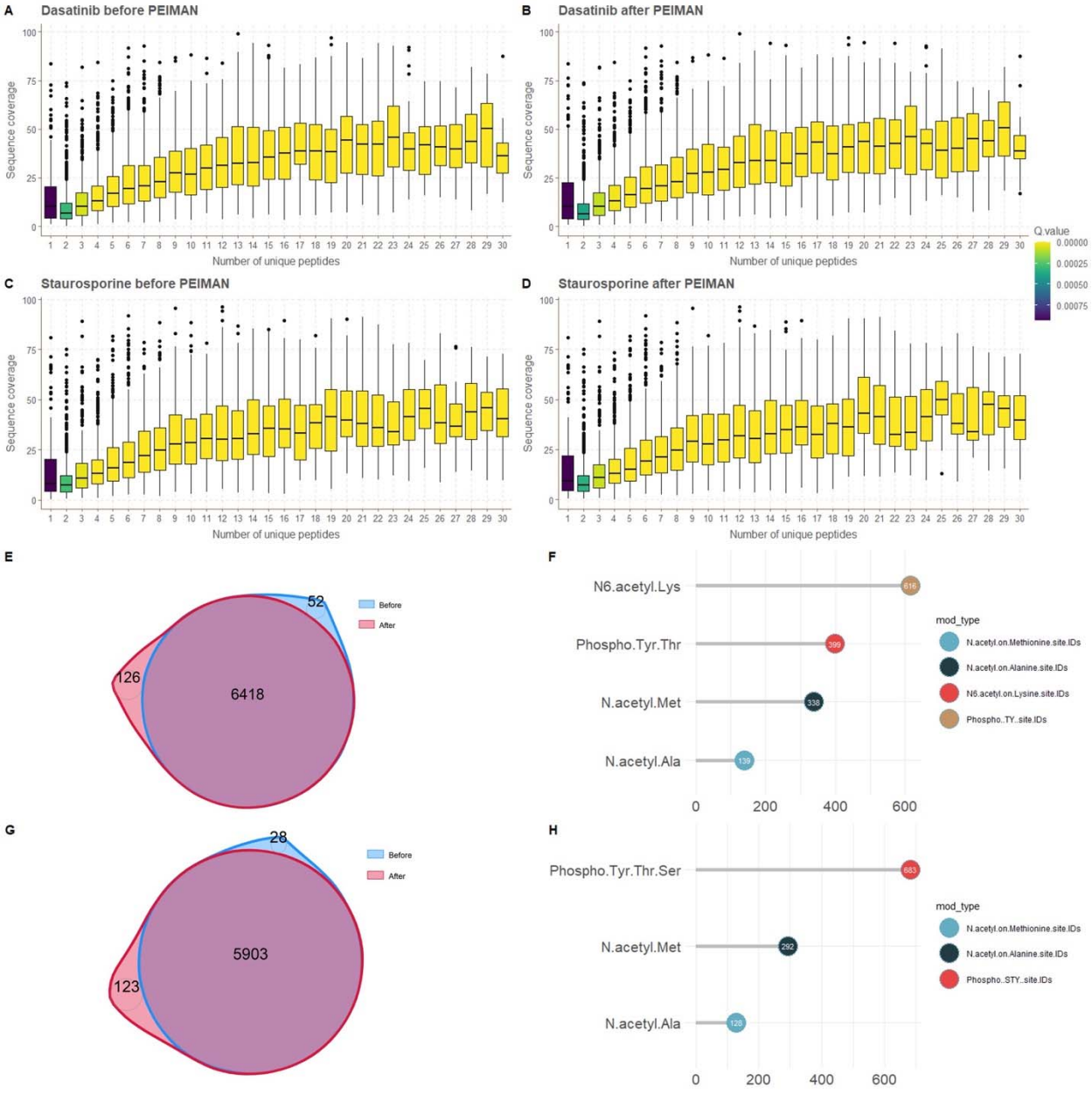


Figure 4: *PEIMAN2 analysis-based before/after box plots for dasatinib and staurosporine. In these two pair plots (A:B and C:D), the effect of including additional enriched modifications suggested by PEIMAN2 on the searching database by MaxQuant is depicted based on sequence coverage versus the number of unique peptides for both drugs separately. Figure 5EFGH: PEIMAN2 analysis-based before-and-after venn diagrams and PTM frequency bar plots for dasatinib and staurosporine. The venn diagrams (A) and (C), respectively, depict the number of proteins identified by MaxQuant for dasatinib and staurosporine before/after PEIMAN2 analysis. Based on PEIMAN2 analysis and re-searching the database by MaxQuant, the frequency plot of identified modifications of proteins is depicted for each drug (B and D).*

Then, we performed another quality control to check the number of proteins that are identified before/after using PEIMAN2 and checked whether the newly identified proteins carry any PTMs. Figure 4 panels (E) and (G) present the Venn diagrams of the number of newly identified (or lost) proteins by MaxQuant before/after using PEIMAN2 for dasatinib and staurosporine drugs, respectively. In Figure 4, panels (F) and (H) depict the frequency bar plot of identified modifications in newly identified proteins based on re-searching the database by MaxQuant considering PEIMAN2 suggestions. For dasatinib 126 new proteins were identified by including the modifications suggested by PEIMAN2. On the other hand, 52 proteins that were previously identified were lost in the refined search, perhaps indicating that their corresponding peptides did not pass the 1% FDR threshold set in MaxQuant search. The majority of disappeared proteins were identified with two peptides in the initial search, with a median score value of 2.861 and small Q values. Note that 86% of proteins after including the suggested modification in the search had more than two peptides. There were 6,418 proteins in common before/after using PEIMAN2. For staurosporine, refined searched yielded 5,903 proteins that were also found in the initial database search. Furthermore, 123 proteins were identified that were not included in the initial search results, and 28 proteins were not found in the refined search after including PEIMAN2 suggested modifications. When comparing the results of the experiment before/after including PEIMAN2 suggestions, this result implies that more than 97 percent of the proteins that were quantified before/after PEIMAN2 implementation in the refined database search are the same. However, as a result of employing PEIMAN2, we now have information regarding the PTM status of previously identified proteins in addition to previously undiscovered proteins. It was interesting to investigate the types and frequency of modifications in newly identified proteins after searching MaxQuant by considering PEIMAN2 suggestions. In Figure 4, panels (F) and (H) show the bar plot frequency of PTMs of newly identified proteins for two drugs dasatinib and staurosporine, respectively.

PTMs on protein targets of drugs

We investigated the differential protein expression analysis before/after considering PEIMAN2's suggestions in the database search. Figure 5 presents the fold change of proteins for each drug compared to control group, before/after considering PEIMAN2's suggestions separately. The fold change is calculated as the log2 base of proportion of two sample means (conc.4 vs control). Each data point in both plots corresponds to a protein. The data points are colored depending on the results of a two-sample t-test (with equal variance assumption) before/after incorporating PEIMAN2's suggestions in the database search. For example, purple color indicates that a protein was significantly different in conc.4 group vs control group, both before/after using modification suggestions. These are drug targets that were found repeatedly by MaxQuant and whose PTM status was clarified by PEIMAN2. The panels in Figure 5 shows the distinct modifications of proteins, if any.

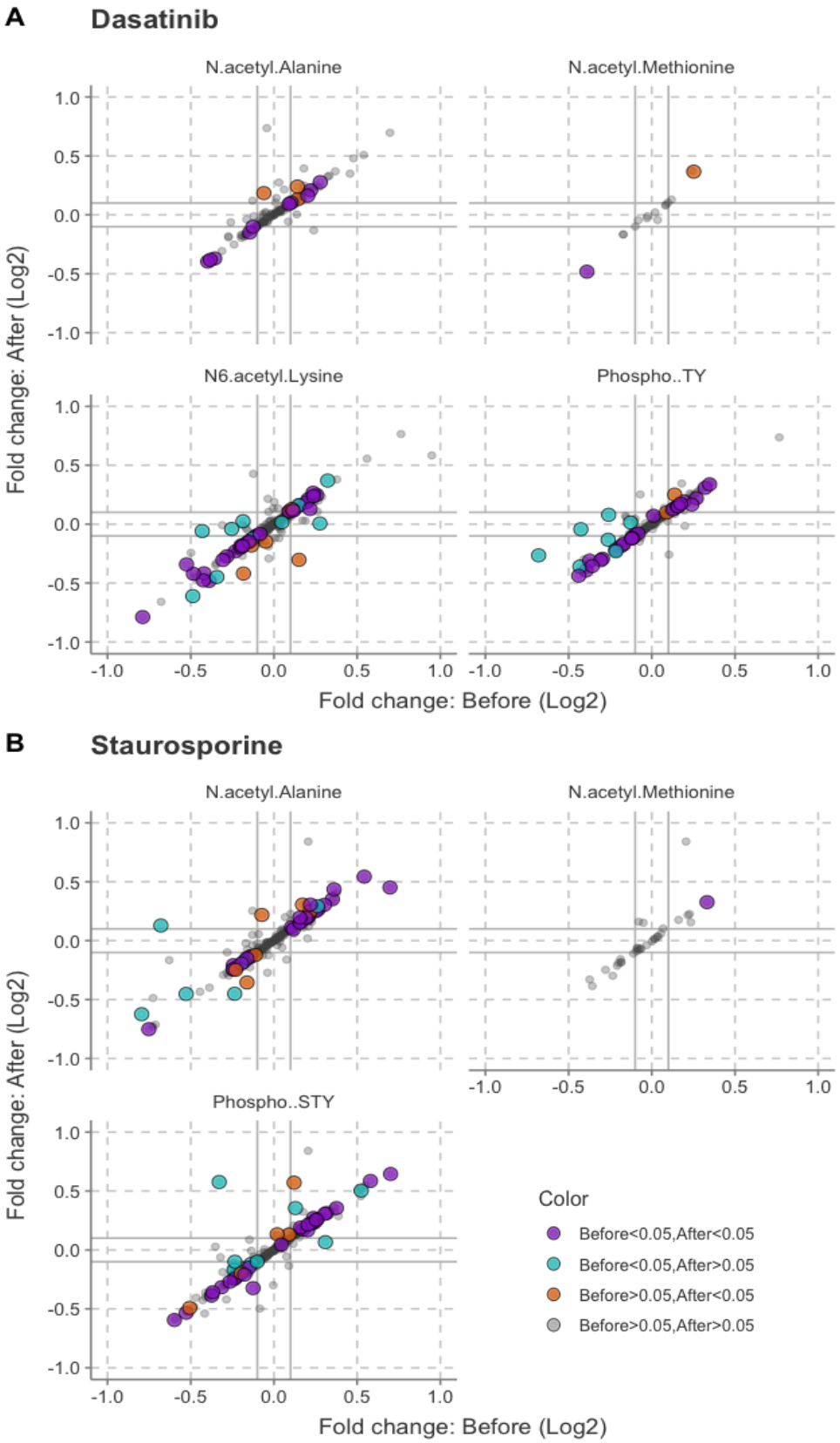


Figure 5: *Differential analysis of proteins in response to the treatments before/after applying PEIMAN2. The fold changes of proteins with respect to untreated cells for each drug compared is displayed before/after considering PEIMAN2's recommendations for incorporating enriched modifications in the database search. Four distinct colors were also used to depict the p-values of the t-statistics before/after PEIMAN2 analysis. Using shape icons, differential proteins with distinct modifications are also indicated.*

It is interesting to note that proteins that were significant before incorporating PEIMAN2's suggestions are still significant after considering the suggestions (purple color dots) with an enriched modification. In addition, the proteins that were not significantly changed before and became significant after including modification in MaxQuant search, are modified too. In comparison to other PTMs, the level of phosphorylated proteins is also more decreased or perturbed, necessitating further research to determine whether this is because of the primary or secondary effects of inhibitors on protein targets. The above findings are consistent for both drugs. These results suggest that studies on drug targets and mechanism of action considering PTMs are helpful for identifying new proteins that are involved in drug mechanism of action. Basically, by including relevant PTMs in database search, the results of studies on different perturbations can be brought closer to reality using PEIMAN2.

Monitoring of PTM changes after PEIMAN2 implementation

Finally, we investigated the perturbation of the quantified PTMs upon treatment with dasatinib and staurosporine at the peptide level. First, the intensity of a given peptide was normalized by total intensity for each TMT channel. Then this normalized intensity was divided by the sum normalized intensity of the other peptides not carrying any PTMs for the same protein. The latter normalization would cancel the effect of abundance changes in the protein level upon treatment with the drug. Finally, we provided the trend of abundance changes for each modified peptides across different concentrations vs. respective controls and the results for all the PTM-carrying peptides for both dasatinib and staurosporine are shown in Supplementary Figure 5. While some modified peptides show a trend of decreasing or increasing abundance in a concentration-dependent manner, some other PTM-carrying peptides are unchanged upon different treatments, as expected. For example, the levels of two phosphopeptides (with three phosphothreonines) belonging to myristoylated alanine-rich C-kinase substrate (MARCKS) decreased by 1.5-fold upon treatment with the highest concentration of dasatinib. The PTM-carrying peptides following a concentration-dependent trend could be involved in drug mechanism of action.

Discussion

In the present work, we introduced an informatic pipeline called PTM-centric proteomics for prediction of most probable and relevant PTMs using PEIMAN2 R package. We show this tool to predict the relevant PTMs for the top regulated proteins upon perturbations or interventions and demonstrate that including the predicted modifications in a refined database search can lead to identification of more proteins as well as PTMs on the top regulated proteins. When the answer to the research question lies beyond expression proteomics (usually does), in order to explain a particular phenotype, these findings are highly relevant and informative.

Any alterations in PTM processing in response to perturbations can potentially impact various biochemical and biophysical aspects of proteins, subsequently affecting the cellular and even organismal phenotype. For example, a simple hypusination event on eIF5A drives

protein synthesis and cell proliferation[21]. Therefore, PTM studies are one of the major forefronts or proteomics research and their widespread use is not limited only to mammals or eukaryota. PTM types and sites in bacteria, archaea, and even viral proteins have been characterized and reported in an extensive number of studies[22-24]. For example, PTMs can modulate protein turnover[25], stability/solubility[2,5,26,27], folding and localization of proteins, the interactions between proteins[28], the direct and indirect effects on genome function[29], the trafficking of molecules[30], and the activation of receptors[31]. In addition, since many proteins include numerous PTMs and one PTM can change the prevalence or occupancy of another, a phenomenon known as PTM crosstalk, it is difficult to understand these complex control mechanisms without characterization and analysis of PTMs[32].

There have been many efforts to focus on investigating the identity and effect of a single PTM or multiple PTMs on protein function[33-35]. Although mass spectrometry-based proteomics is the golden standard for PTM analyses, the high-throughput experimental procedures used to identify PTMs are labor intensive and time-consuming[36]. Suppose one has a presumption about the role of protein phosphorylation. In that case, we need to design a phosphoproteomics study using titanium/zirconium dioxide-based beads to enrich phosphopeptides with high specificity. Otherwise, the study design is independent of adding extra experimental steps and routine proteomics database searches are applied. In spite of these advances, PTM identification is not the focal point of any proteomics investigation that lacks a prior PTM-specific hypothesis. Therefore, there is an immediate demand for computational methodologies and effective tools that can predict PTMs that are most probably found in a given biological sample or are occurring upon a specific perturbation [9].

Savitski *et al.* provided a computational method called ModifiComb based on the difference between the molecular masses and the retention time of the modified and unmodified peptides[37]. The authors provided a method that is independent of PTM-related *priori* assumptions. Compared to searching all possible modifications, this method succeeded to reduce search space and, as a result, the propensity of false positive PTM identification. To improve the understanding of this dark matter of proteomics, Kong *et al.* also presented a fragment-ion indexing method and implemented it into MSFragger tool to computationally speed up searching proteomics database with PTMs [38,39]. However, both methods rely on the identification of unmodified peptides to compute differences in mass and retention time. It should also be noted that not all detectable chemical modifications in mass spectra have a biological significance and cannot be inferred functional PTMs whereas PEIMAN2 provide enrichment analysis to avoid the detection of inert stochastically modified peptides.

We believe that PTM-centric proteomics based on enrichment analysis is a successful attempt to bring the results of perturbational studies closer to reality. Such predictions present opportunities for developing myriad PTM-related hypotheses and a particular follow-up experimental design in biological studies. To carry out PTM-centric proteomics, it is recommended to incorporate PEIMAN2 after the initial round of mass spectrometry database search and analysis in order to carry out a second round of mass spectrometry database search and downstream analysis with a given set of PTMs. PEIMAN2 is not dependent on data from any MS instrument and can be easily integrated into the majority

of existing data analysis pipelines. Accordingly, PEIMAN2 has the potential to become a valuable option for routine analysis of the most probable PTMs in shotgun proteomics data. Altogether, the application of this package will help unravel PTM crosstalk in homeostasis and disease.

Methods

Cell culture

Human lung carcinoma A549 were grown in McCoy's 5A medium, supplemented with 10% FBS superior (Biochrom, Berlin, Germany), 2 mM L-glutamine (Lonza, Wakersville, MD, USA) and 100 units/mL penicillin/streptomycin (Gibco, Invitrogen) and incubated at 37 °C in 5% CO2. Cells were routinely checked for mycoplasma contamination by PCR and low passage number cells from ATCC were used in the experiments.

Cell viability assay

Cell viability upon compound treatment was measured using CellTiter-Blue assay (Promega) according to manufacturer protocol and the LC50s were determined as the concentration of compound causing 50% cytotoxicity.

Multikinase inhibitor treatment

After seeding 250,000 A549 cells in triplicates in 6 well plates, cells were allowed to grow for 24 h, after which they were treated with the compounds for 24 h in triplicates. Dasatinib was profiled at 100 nM, 1 μM, 5 μM and 25 μM. Staurosporine was profiled at 8 nM, 40 nM, 200 nM and 1 μM. Cells treated with DMSO were used as controls.

Table 1: TMT labeling scheme for the experiments

Compounds	Dasatinib	Staurosporine
TMT126	Control replicate 1	Control replicate 1
TMT127N	Control replicate 2	Control replicate 2
TMT127C	Control replicate 3	Control replicate 3
TMT128N	100 nM replicate 1	8 nM replicate 1
TMT128C	100 nM replicate 2	8 nM replicate 2
TMT129N	100 nM replicate 3	8 nM replicate 3
TMT129C	1 μM replicate 1	40 nM replicate 1
TMT130N	1 μM replicate 2	40 nM replicate 2
TMT130C	1 μM replicate 3	40 nM replicate 3
TMT131N	5 μM replicate 1	200 nM replicate 1
TMT131C	5 μM replicate 2	200 nM replicate 2
TMT132N	5 μM replicate 3	200 nM replicate 3
TMT132C	25 μM replicate 1	1 μM replicate 1
TMT133N	25 μM replicate 2	1 μM replicate 2

Compounds	Dasatinib	Staurosporine
TMT133C	25 μM replicate 3	1 μM replicate 3

LC-MS/MS sample preparation

Sample preparation was done according to our previous protocol[40]. After treatment, cells were trypsinized, washed with PBS and lysed with the lysis buffer (8 M urea, 1% SDS, 50 mM Tris pH 8.5). Protein concentration was measured using Pierce BCA Protein Assay Kit (Thermo), and the volumes corresponding to 25 µg of protein was transferred from each sample to new low-bind Eppendorf tubes. DTT was added to a final concentration of 10 mM and samples were incubated for 1 h at room temperature. Subsequently, iodoacetamide (IAA) was added to a final concentration of 50 mM and samples were incubated at room temperature for 1 h in the dark. The reaction was quenched by adding an additional 10 mM of DTT. After precipitation of proteins using methanol/chloroform, the semi-dry protein pellets were dissolved in 25 µL of 8 M urea in 20 mM EPPS (pH 8.5) and were then diluted with EPPS buffer to reduce urea concentration to 4 M. Lysyl Endopeptidase (Wako) was added at a 1:75 w/w ratio to protein and incubated at room temperature overnight. After diluting urea to 1 M, trypsin (Promega) was added at the ratio of 1:75 w/w and the samples were incubated for 6 h at room temperature. TMT reagents were added 4x by weight to each sample, followed by incubation for 2 h at room temperature. The reaction was quenched by addition of 0.5% hydroxylamine. Samples were combined, acidified by TFA, cleaned using Sep-Pak (Waters) and dried using a DNA 120 SpeedVac concentrator (Thermo). Samples were resuspended in 20 mM ammonium hydroxide and separated into 96 fractions on an XBrigde BEH C18 2.1x150 mm column (Waters; Cat#186003023), using a Dionex Ultimate 3000 2DLC system (Thermo Scientific) over a 48 min gradient of 1-63% B (B=20 mM ammonium hydroxide in acetonitrile) in three steps (1-23.5% B in 42 min, 23.5-54% B in 4 min and then 54-63%B in 2 min) at 200 μL/min flow. Fractions were then concatenated into 24 samples in sequential order (e.g. A1, C1, E1 and G1).

Proteomics

After resuspension in 0.1% FA (Fluka), fractions (1 μg) were analyzed by LC-MS/MS. Samples were loaded onto a 50 cm column (EASY-Spray, 75 μm internal diameter (ID), PepMap C18, 2 μm beads, 100 Å pore size) connected to a nanoflow Dionex UltiMate 3000 UHPLC system (Thermo) and eluted in an organic solvent gradient increasing from 4% to 26% (B: 98% ACN, 0.1% FA, 2% H2O) at a flow rate of 300 nL/min over a total 110 min method time. The eluent was ionized by electrospray and mass spectra of the molecular ions were acquired with an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) in data-dependent mode at MS1 resolution of 120,000 and MS2 resolution of 60,000, in the m/z range from 400 to 1600. Peptide fragmentation was performed via higher-energy collision dissociation (HCD) with energy set at 35 NCE and MS2 isolation width at 1.6 Th.

Proteomic Data, Bioinformatic and statistical Analysis

The raw LC-MS data were analyzed MaxQuant version 1.6.2.3 [41]. The Andromeda search engine [42] was run against the International Protein Index (human version UP000005640_9606, 92957 entries). Methionine oxidation was selected as variable

modifications, while cysteine carbamidomethylation was set as a fixed modification. No more than two missed cleavages were allowed, and a 1% FDR was used as a filter at both protein and peptide levels. All the contaminants were removed in the first step and only proteins with at least two peptides and a Qvalue less than of 5% were considered in all cases. After PEIMAN2 analysis, the mentioned modifications were added to the raw LC-MS search. All the experiments were performed in triplicates. Two-tailed Student t-test was applied to calculate p-value. The data were normalized by the total intensity of each TMT channel and subsequently, the expression ratio for each protein was calculated relative to the DMSO-treated controls.

Preparing PEIMAN2 R package

At the first step, we downloaded 566,996 "Reviewed (UniProt) - Manually annotated" proteins (as of August 2022 from UniProt online repository (available at https://www.uniprot.org). The database records various useful functional information about proteins. We reduced the size of file by narrowing each protein record to include unique UniProt accession code (AC), organism taxonomy name (OS), keywords (KW) and features (FT). We were particularly interested in KW and FT as any manually curated information regarding PTM are available in these fields. At the next step, we used R statistical software to prepare a database to obtain PTM profile of proteins for all species. In our search for PTMs in proteins, we used a list of controlled PTM vocabulary provided by UniProt (available at https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/ptmlist.txt). To find proteins with any PTM modification, we searched the downloaded file against controlled PTM vocabulary by looking into KW and FT entry of each protein. The presence of any PTM in CROSSLNK, LIPID, or MOD_RES feature of any protein was of interest. At the time of preparing this manuscript, we obtained the PTM profile of 220,389 proteins. To keep up with monthly changes in UniProt, we automated the preparation process and will update the database each month accordingly. The latest PEIMAN2 scripts and PEIMAN2 database are available at JafariLab GitHub repository (https://github.com/jafarilab/PEIMAN2).

Enrichment Analysis

Single Enrichment Analysis (SEA)

The enrichment analysis is a powerful strategy which facilitates the identification of biological processes for a list of genes or proteins. The single enrichment analysis (SEA) is known as one of the traditional methods to infer the biological functions in a given list of genes. The analysis in SEA starts with a list of differentially expressed genes provided by researcher (selected with some criteria: p-value or fold-change). The idea behind SEA is to test if the number of genes in the list with a certain biological function (for example PTM) is significantly different from occurrence through random chance. In a general sense, enrichment analysis investigates whether a group of genes or proteins are over/under-presented for a specific biological pathway in a large set of genes/proteins. Different statistical methods are introduced to measure this discrepancy such as Chi-Square, Fisher's exact test, and hypergeometric test. We previously implemented a standalone software to

run SEA in a list of proteins and infer any enriched modification by applying a hypergeometric test[18]. The idea in PEIMAN standalone software and PEIMAN2 is to investigate if a subset of proteins is over/under presented for any particular PTM, in a large set of proteins. We here briefly describe the idea of hypergeometric test in this context. Assume there are N proteins in the database and $K(\leq N)$ of these proteins have one of the known modifications, for example "Acetylation". We pass a list of n proteins. We can apply hypergeometric test to check if "Acetylation" is over-under/represented in the sample list of n proteins using a hypergeometric test. The p-value of such a test is calculated as:

$$P - value = \sum_{x=m}^{minX(K,n)} \frac{\binom{K}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

This simple idea is very helpful in inferring biological meaning from a large list of genes. In the next section, we discuss some of the weaknesses associated with SEA and review an alternative powerful approach to SEA.

Protein Set Enrichment Analysis (PSEA)

SEA takes a list of differentially expressed genes/proteins and identifies if the number of genes/proteins with a certain biological feature is significantly enriched in the list. However, there are some known disadvantageous to SEA as follows[43]. First, the p-values obtained after correcting for multiple testing may result in no differentially significant gene because the real differences are not large compared to the inevitable noise. Second, the final list may contain many declared significant genes/proteins leading to difficulty in interpretation and subjective interpretations among biologists with different expertise. Third, single gene/protein enrichment analysis potentially misses the vital effects on pathways. Finally, it is common that different research lab groups report several lists of significant genes/proteins for the same perturbation or biological process.

Gene set enrichment analysis or GSEA has been introduced by Subramanian *et al.* [43] to overcome these drawbacks. We briefly highlight the key points of GSEA method here. GSEA is applied on profiles of genome-wide expression data that belongs to two experimental groups (control vs treatment). Genes are then sorted based on a score, for example correlation between their expression profile and the class they belong to. A set of genes, S, is defined as genes that belong to a certain set with a distinct biological annotation (e.g., metabolic pathway, GO category, PTM). The idea behind GSEA is to identify if the members of set S tend to show more often toward the top (or bottom) of the gene list or are randomly distributed throughout the list.

The GSEA method can be summarized in three steps as follows. First, an enrichment score (ES) is calculated to measure if the gene set S is over-presented at the top or bottom of the ranked gene list. This is achieved by calculating a signed version of the Kolmogorov-Smirnov statistic while running from the top to the bottom of the ranked list. Whenever we encounter a gene in the S set, the value of the statistic is increased proportional to an exponent power of the gene's score. Likewise, when the gene is not in the S set, the value of statistic is decreased. The enrichment score for gene set S is defined as the maximum

observed deviation from zero in the running score profile. In the second step, the significance of calculated ES score for a given list of genes is evaluated by randomly permuting the score of each gene for certain number of times (usually 1000 times) and calculating ES for each random profile to generate a null distribution for the ES. A nominal p-value is then calculated according to the null distribution. Finally, GSEA accounts for the effect of multiple testing by calculating FDR. This is achieved by normalizing ES score of each gene set relative to gene set size. For more details refer to supplementary material of [31].

Inspired by the idea and usefulness of GSEA in elucidating biological inferences in a given list of genes, we implement protein set enrichment analysis or PSEA in PEIMAN2 package to infer biological meaning from a list of proteins. In our work, the gene sets are replaced by a set of proteins that belong to a certain modification group, for example "Acetylation". For any list of protein given by researcher, the set of proteins with a certain modification are identified. For each set of proteins, we calculate enrichment score and assess the significance of ES by the methods described in [43]. Finally, we provide a list of modification that are most probably enriched in a given list. All these functionalities are implemented in an R package to serve a broader community of researchers. For more details on functionality of package, please read the Vignette and Readme page at PEIMAN2 GitHub page.

Data availability

The LC-MS/MS raw data files and extracted peptides and protein abundances are deposited in the jPOST repository of the ProteomeXchange Consortium [44] under the dataset identifiers PXD037679 and PXD037681

Code availability

All analyses reported in this study used the statistical software R (v.4.0.0). The R package can be found on GitHub (https://github.com/jafarilab/PEIMAN2).

Acknowledgements

This study was financially supported by the Academy of Finland [Grant 332454 to M.J.], Swedish Research Council [grant 2020-00687 to A.A.S.], and the Swedish Society of Medicine [grant SLS-961262, 1086 Stiftelsen Albert Nilssons forskningsfond to A.A.S.]. We would like to thank Prof. Roman A. Zubarev for his valuable comments and input in the manuscript. Meilahti Clinical Proteomics Core Unit is supported by Biocenter Finland and Helsinki Institute of Life Sciences (HiLIFE).

Author information

Contributions

M.J. conceived of the study and supervised the project. P.N. and M.J. developed the computational analysis as well as the PEIMAN2 R package. A.A.S designed, developed, and led the experimental methods for deep expression profiling. M.M., A.A.S, and M.J.

contributed to the interpretation of the findings, and M.B. advised on the work. All authors contributed to the final manuscript by discussing the findings and reviewing and modifying it.

Corresponding authors

Correspondence to Amir Ata Saei (amirata.saei.dibavar@ki.se) or Mohieddin Jafari (mohieddin.jafari@helsinki.fi).

Ethics declarations

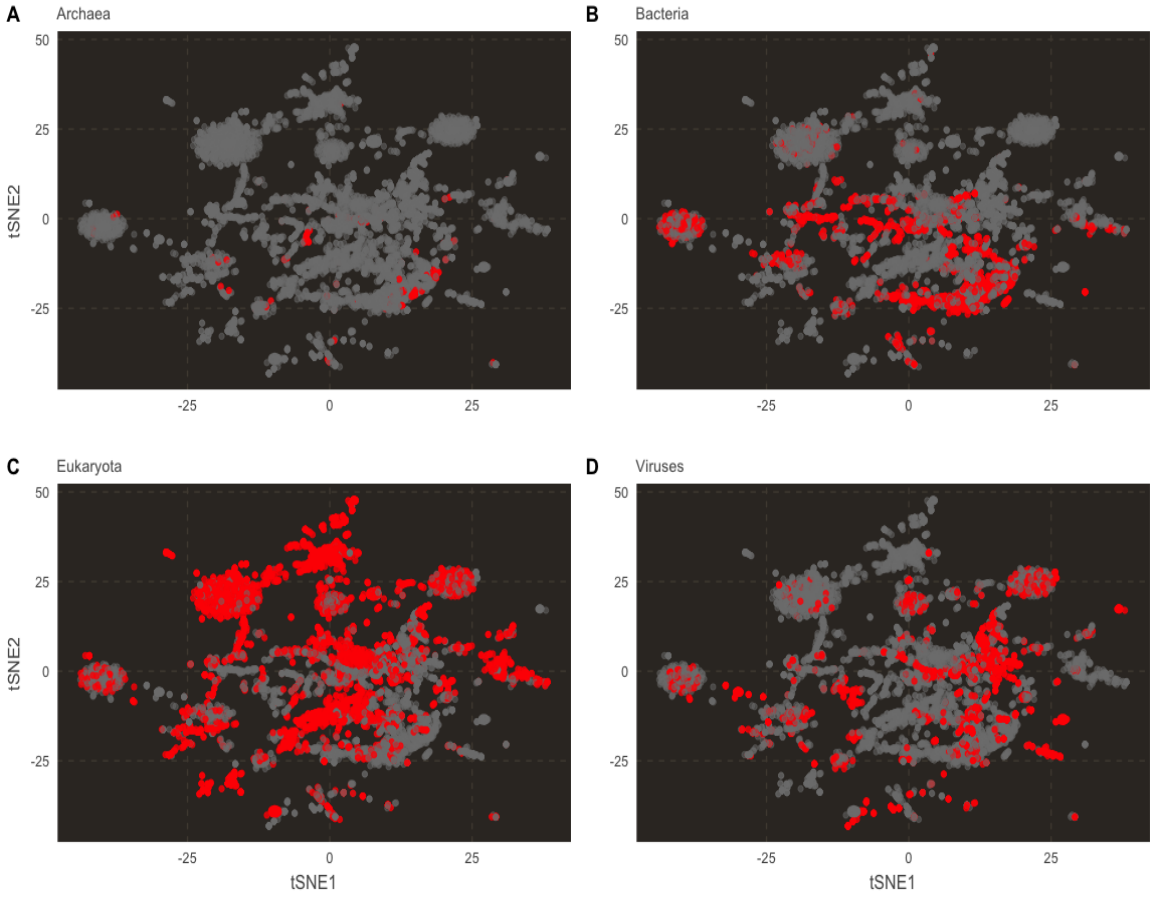
Competing interests

The authors declare no competing interests.

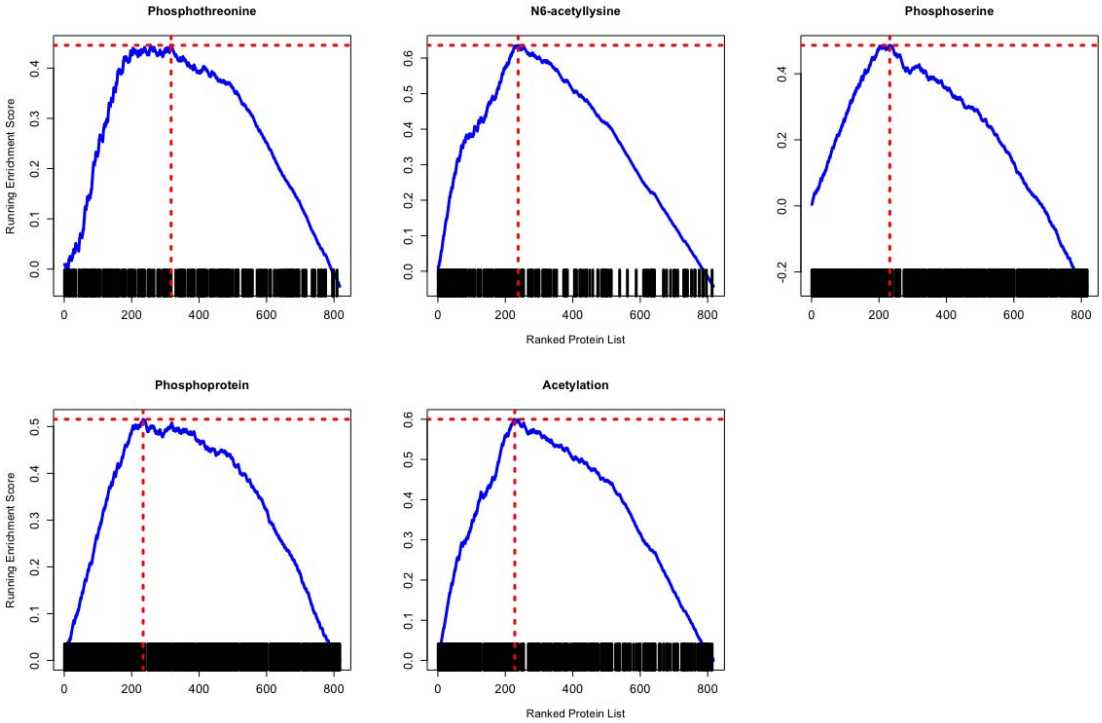
Supplementary information

Supplementary Figures

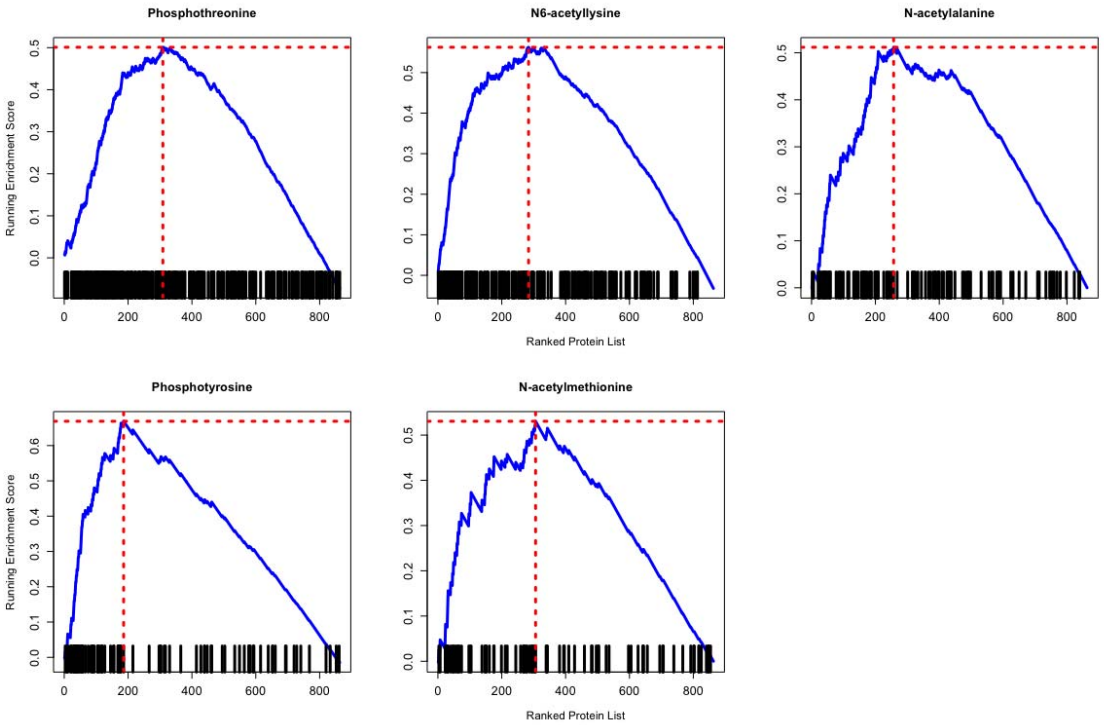
Supplementary Figure 1: *PTM frequency treemaps for eight popular organisms from diverse taxonomic branches of life.*



Supplementary Figure 2: *The t-SNE plots based on PTM profiles in four super kingdoms of life, i.e., Archaea, Bacteria, Eukaryota and Viruses (panels A-D). Each dot in the plots presents one organism. The red and grey colors indicates if the point (organism) belongs to corresponding super kingdom of life or not. Note that we anticipate a more uniform distribution of viruses across all three of the other phyla of life (Panel D).*



Supplementary Figure 3: *The running score plot of the first top five PTMs identified on differentially expressed proteins upon dasatinib treatment. The x-axis is the ranked protein based on their score and the y-axis is their enrichment score. The rug in the x-axis indicates the proteins with the corresponding PTM. The position of maximum running enrichment score is denoted by a red dashed line.*



Supplementary Figure 4: *The running score plot of the first top five PTMS identified on differentially regulated proteins upon staurosporine treatment. The x-axis is the ranked protein based on their score and the y-axis is their enrichment score. The rug in the x-axis indicates the proteins with the corresponding PTM. The position of maximum running enrichment score is denoted by a red dashed line.*

ATTACHED PDF FILE

Supplementary Figure 5: *The modified peptides with probabilities. The modified peptides with all above-mentioned PTMs are listed separately upon dasatinib and staurosporine treatment. The x-axis is the four drug concentrations and the control, and the y-axis is the proportional abundance of the corresponding peptide compared to unmodified peptide (see Supplementary data 3 and 4 for more details).*

Supplementary data

Supplementary data 1*. A table of all identified proteins for dasatinib following PTM-centric proteome informatic pipeline.*

Supplementary data 2*. A table of all identified proteins for staurosporine following PTM-centric proteome informatic pipeline.*

Supplementary data 3*. A compiled table of all PTM-carrying peptides for dasatinib and calculated final fold changes that are plotted in Supp figure 5.*

Supplementary data 4*. A compiled table of all PTM-carrying peptides for staurosporine and calculated final fold changes that are plotted in Supp figure 5.*

References

1 Mann, M. & Jensen, O. N. *Nature biotechnology* **21**, 255-261, (**2003**).
2 Saei, A. A. *et al. Nature communications* **12**, 1-13, (**2021**).
3 Khoury, G. a., Baliban, R. C. & Floudas, C. a. *Scientific reports* **1**, 1-5, (**2011**).
4 Nussinov, R., Tsai, C.-J., Xin, F. & Radivojac, P. *Trends in biochemical sciences* **37**, 447-455, (**2012**).
5 Saei, A. A. *et al. bioRxiv*, (**2022**).
6 Ochoa, D. *et al. Nature biotechnology* **38**, 365-373, (**2020**).
7 Weinert, B. T. *et al. Cell* **174**, 231-244. e212, (**2018**).
8 Brüning, F. *et al. Science* **366**, eaav3617, (**2019**).
9 Kirkpatrick, D. S., Gerber, S. A. & Gygi, S. P. *Methods* **35**, 265-273, (**2005**).
10 Mnatsakanyan, R. *et al. Expert Review of Proteomics* **15**, 515-535, (**2018**).
11 Darie, C. C. *Modern Chemistry & Applications*, (**2013**).
12 Tyanova, S. *et al. Nature Methods* **13**, 731-740, (**2016**).
13 Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P. & Jørgensen, T. J. *Molecular & cellular proteomics* **4**, 873-886, (**2005**).
14 Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. *J Proteome Res* **17**, 1844-1851, (**2018**).
15 Yang, H. & Zubarev, R. A. *Electrophoresis* **31**, 1764-1772, (**2010**).
16 Savitski, M., Nielsen, M. & Zubarev, R. *Mol Cell Proteomics* **5**, 935-948, (**2006**).
17 Chick, J. M. *et al. Nature biotechnology* **33**, 743-749, (**2015**).
18 Nickchi, P., Jafari, M. & Kalantari, S. *Database : the journal of biological databases and curation* **2015**, bav037, (**2015**).
19 Saei, A. A. *et al. Nature Communications* **12**, 1296, (**2021**).
20 Li, J. *et al. Nature methods* **17**, 399-404, (**2020**).
21 Park, M., Nishimura, K., Zanelli, C. F. & Valentini, S. R. *Amino acids* **38**, 491-500, (**2010**).
22 Carabetta, V. J. & Hardouin, J. *Frontiers in Microbiology* **13**, 874602, (**2022**).
23 Cohen, M. S. & Chang, P. *Nature chemical biology* **14**, 236-243, (**2018**).
24 Stevens, K. M. & Warnecke, T. in *Seminars in Cell & Developmental Biology* (2022).
25 Zecha, J. *et al. Nature communications* **13**, 1-14, (**2022**).
26 Potel, C. M. *et al. Nature methods* **18**, 757-759, (**2021**).
27 Huang, J. X. *et al. Nature methods* **16**, 894-901, (**2019**).
28 Wang, S., Osgood, A. O. & Chatterjee, A. *Current Opinion in Structural Biology* **74**, 102352, (**2022**).
29 Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. *Nature Reviews Genetics*, 1-18, (**2022**).
30 May, E. A., Sroka, T. J. & Mick, D. U. *Frontiers in Cell and Developmental Biology* **9**, 664279, (**2021**).
31 Ramazi, S. & Zahiri, J. *Database* **2021**, (**2021**).
32 Adoni, K. R., Cunningham, D. L., Heath, J. K. & Leney, A. C. *Journal of proteome research* **21**, 930-939, (**2022**).
33 Bradley, D. *Current Opinion in Genetics & Development* **76**, 101956, (**2022**).
34 Brandi, J., Noberini, R., Bonaldi, T. & Cecconi, D. *Journal of Chromatography A*, 463352, (**2022**).

35 Brodbelt, J. S. *Current Opinion in Chemical Biology* **70**, 102180, (**2022**).
36 Hermann, J., Schurgers, L. & Jankowski, V. *Molecular Aspects of Medicine*, 101066, (**2022**).
37 Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. *Molecular & Cellular Proteomics* **5**, 935-948, (**2006**).
38 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. *Nature Methods* **14**, 513-520, (**2017**).
39 *PROTEOMICS* **n/a**, 2100369, (**2022**).
40 Saei, A. A. *et al. Nature communications* **10**, 1-13, (**2019**).
41 Cox, J. & Mann, M. *Nature biotechnology* **26**, 1367-1372, (**2008**).
42 Cox, J. *et al. Journal of proteome research* **10**, 1794-1805, (**2011**).
43 Subramanian, A. *et al. Proceedings of the National Academy of Sciences* **102**, 15545-15550, (**2005**).
44 Vizcaino, J. A. *et al. Nature biotechnology* **32**, 223-226, (**2014**).