Briefings in Bioinformatics, 2021, 1–11
https://doi.org/10.1093/bib/bbab510
Problem Solving Protocol

PhosPiR: an automated phosphoproteomic pipeline in R

Ye Hong, Dani Flinkman †, Tomi Suomi †, Sami Pietilä †, Peter James, Eleanor Coffey and Laura L. Elo

Corresponding author: Eleanor Coffey, Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland. Tel: +358-401822424; E-mail: eleanor.coffey@bioscience.fi

†These authors contributed equally to this work.

Abstract

Large-scale phosphoproteome profiling using mass spectrometry (MS) provides functional insight that is crucial for disease biology and drug discovery. However, extracting biological understanding from these data is an arduous task requiring multiple analysis platforms that are not adapted for automated high-dimensional data analysis. Here, we introduce an integrated pipeline that combines several R packages to extract high-level biological understanding from large-scale phosphoproteomic data by seamless integration with existing databases and knowledge resources. In a single run, PhosPiR provides data clean-up, fast data overview, multiple statistical testing, differential expression analysis, phosphosite annotation and translation across species, multilevel enrichment analyses, proteome-wide kinase activity and substrate mapping and network hub analysis. Data output includes graphical formats such as heatmap, box-, volcano- and circos-plots. This resource is designed to assist proteome-wide data mining of pathophysiological mechanism without a need for programming knowledge.

Keywords: phosphoproteomics, proteomics, pipeline, bioinformatics, statistics, data visualization

Introduction

Protein phosphorylation is a reversible post-translational modification (PTM) catalyzed by protein kinases of the transferase class [1]. Since its discovery in 1932, numerous studies have highlighted the importance of phosphorylation as a central regulatory process in cells [2]. High numbers of often tightly interconnected phosphoproteins participate in cell signaling and all aspects of cellular function from proliferation and differentiation to metabolism and neurotransmission, to name a few [3–5]. Phosphorylation is the most abundant 'signaling' PTM exceeding ubiquitination, methylation and acetylation [6]. To understand how complex phosphorylation changes, especially shifts introduced by pathophysiological states coordinate function, systems-level phosphoproteomics study becomes necessary [6]. Advanced mass spectrometry methods enable high-throughput measurement of phosphoproteomes [7], however traditional downstream analysis does little beyond phosphopeptide identification and quantification. Recent developments in R packages have taken advantage of protein phosphorylation databases and annotation advances, thereby supporting the creation of an analysis tool that can better exploit phosphopeptide data.

Here we introduce PhosPiR, a pipeline which takes advantage of available open-source tools for a complete downstream analysis of mass spectrometry-derived data after phosphopeptide identification. No programing knowledge is required to run the pipeline. Our workflow consists of peptide quality control, data overviewing with histogram, boxplot, heatmap, and principal component analysis (PCA) plots, data annotation utilizing UniProt and Ensembl database, differential expression analysis including four statistical test options and post hoc testing, phosphosite translation across species, four enrichment analyses for phosphoproteins,

Ye Hong is a PhD student at the University of Turku and Åbo Akademi University specializing on bioinformatics solutions for proteomics and phosphoproteomics data with a particular focus on analysis of data related to neurological and psychological disorders.
Dani Flinkman is a PhD student at Lund University and Åbo Akademi University researching mass spectrometry analysis methods.
Tomi Suomi is a postdoc from the Medical Bioinformatics Center of Turku Bioscience. He works with proteomics data and develops statistical tools.
Sami Pietilä is a PhD student at the University of Turku majoring in bioinformatics.
Peter James is a professor at Lund University and visiting professor at Åbo Akademi University and the University of Turku. He is a pioneer of proteomics and works with neuronal and immunology disfunctions.
Eleanor Coffey is a research director at Turku Bioscience Centre. She leads the Kinase Function in Brain group which studies kinase regulation of brain disorders and biomarker discovery, using a range of methods including mass spectrometry and bioinformatics.
Laura Elo is professor of computational medicine at University of Turku and a research director at Turku Bioscience Centre. She leads the Medical Bioinformatics Center. Her expertise is in developing and benchmarking bioinformatics tools related to medical bioinformatics.
Received: September 8, 2021. **Revised:** October 25, 2021. **Accepted:** November 4, 2021

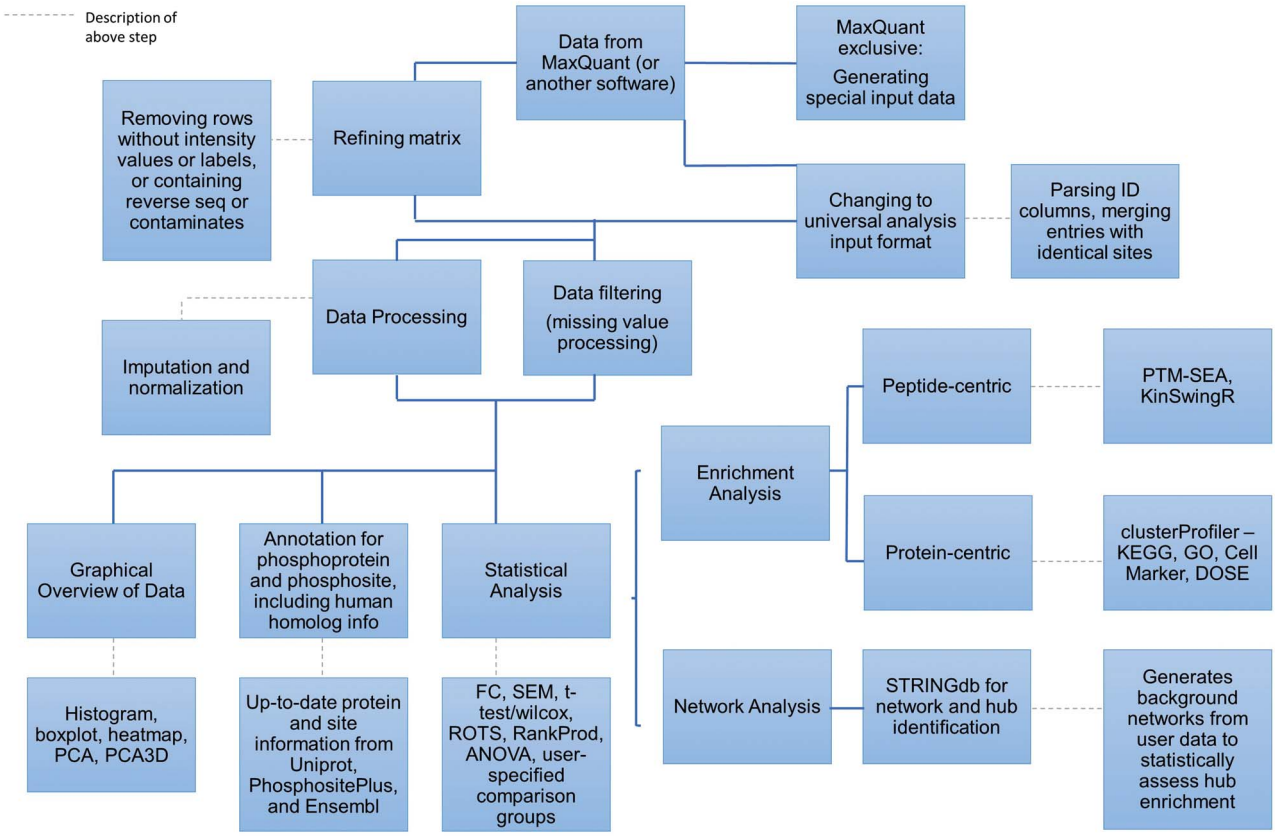

Figure 1. Flowchart overview of pipeline architecture. The main pipeline steps are outlined. Information on software packages utilized and/or information on the approach used is provided in adjoining boxes.

post-translational modification set enrichment analysis (PTM-SEA) for phosphopeptide, kinase analysis, network analysis and hub analysis (Figure 1). The pipeline is accompanied by video tutorials and we exemplify the functionality of the tool using previously published large-scale phosphoproteomic study of circadian clock changes in synaptoneuroscomes [8].

Methods

The current pipeline marks version 1.1.

Input data

PhosPiR accepts output files directly from MaxQuant or preprocessed files from similar MS data processing platforms that provide peptide sequence identification and intensity data on PTMs [9]. The input file from MaxQuant for PhosPiR is the 'Phospho (STY)Sites.txt' file from the 'combined/txt' folder within MaxQuant. In case another spectra analysis tool is preferred instead of MaxQuant, such as Progenesis, Spectronaut, openMS or PEAKS, the user should select 'Other' for data input. The input format for this option is explained in Supplementary Table 1A, and demonstrated with an example input file in Supplementary Table 1B. The steps for user data input are also outlined in the user support videos found at https://youtu.be/c7n7yE0DMsA (short setup introduction video), and https://youtu.be/n4EagNoxusI (long pipeline run demonstration and result file introduction video). To start the workflow, simply run the 'run.R' script in R program with version 4.0.3 or above. R can be downloaded from https://www.r-project.org/

Special input generation

If the input file is 'Phospho (STY)Sites.txt' directly from MaxQuant, the pipeline will automatically remove MaxQuant-marked reverse sequences and potential contaminants and sum the intensities for each phosphosite entry. However, if the user wishes to keep track of possibly falsely labeled potential contaminant sequences, or have the analysis done on individual phosphosite intensities separated by sequence phosphocount number for each site (similar to the 'expand site' option in Perseus), a 'Special Input Generation' step can be selected, where details can be found in Supplementary Table 1C. This is specific for MaxQuant output and is not included as part of the analysis in the 'All steps' choice of the analysis run option.

Filtering and normalization setup

Filtering and normalization steps are implemented as data preprocessing. It is important to understand the data and make educated choices here to bring forth the most reliable result. Rows and/or columns with excessive missing values can be filtered out here. The user can

also choose not to filter. For normalization, there are three choices, normalize and impute, only normalize, or neither. proBatch package [10] and MSImpute package [11] are utilized for normalization and imputation, respectively. Median normalization is performed when 'only normalize' is chosen. Quantile normalization and low-rank approximation imputation is performed when 'normalize and impute' is chosen. These normalization methods were chosen as they have proven successful with phosphoproteomics [8, 10, 12]. Imputation requires at least four nonmissing values per row, if the input data do not satisfy this requirement, the user will be forced to choose one of the other two options.

Other information setup

Organism information, sample group information and comparison information needed to be setup by the user. PhosPiR provides a prompt window accompanied by a guide for the setup process. For organism selection, the user should highlight an organism from the organism list, or if the organism is not available from the list, select 'Other' from the list, then input the scientific name of the source organism. A few analysis steps are only available for human, mouse or rat data.

For sample group information, the user could set it up either by inputting an information file or by setting up groups within the pipeline. For the first option, a template of the information file is included in Supplementary Table 2. Both .csv and .xlsx formats are accepted by PhosPiR. For the second option, the user will be asked how many group classifications are found in the data, for each classification, a brief description is entered, e.g. treatment or genotype, followed by group order. Group order is recorded from the user's selection of sample names in a list of all sample names within the dataset for each group. After the user has selected samples for group 1, the pipeline asks if the current group classification contains another group, upon pressing 'yes', the list of all sample names will be displayed again, this time with all previously selected samples marked with their respective group numbers. The selection process repeats until all groups from the current group classification are setup. Multiple group comparisons can also be setup by providing an information file (Supplementary Table 2) or going through the pipeline process. In case of setting up in PhosPiR, the user chooses the group classification first, and then enters, e.g. '1,2' to specify the groups that should be compared against each other. The process can be repeated for as many group comparisons are wished. Lastly, the user can select whether to run all analysis steps, or only special input generation/annotation/overview figure/differential analysis step. After setup, PhosPiR will run the analyses automatically, however, between analyses, step-specific choices will be given to the user, and the pipeline will not continue until the user has responded.

Overview figures

Several figures will be plotted automatically in order to provide an overview of the data distribution. Boxplot and histograms are suitable visuals for comparing sample distributions. Heatmaps, PCA with k-means clustering, and 3D PCA plots will automatically display the results from unsupervised clustering of the data, providing informative biological patterns from the data. Heatmap clustering utilizes Euclidean distance and complete-linkage clustering. Two heatmaps are generated, one with only column (sample) clustering, and one with column and row clustering. K-means clustering on PCA plot sets the number of clusters to the reference group count number plus 1. Log 2 intensity values are automatically generated and used for overview figures. Missing values are assigned intensity of 1. A few packages are utilized for this step. Boxplot and histogram are plotted with 'ggplot2' [13]; heatmap is plotted with 'pheatmap' [14]; PCAs and 3D PCA are plotted with 'fingerprint' [15], 'vegan' [16], 'rgl' [17], 'FactoMineR' [18], 'factoextra' [19], 'plot3D' [20] and 'magick' [21].

Data annotation

A data annotation step utilizes information from all organisms found in the Ensembl database to identify for example reviewed accession and phosphorylation site position, Entrez ID, genome position of the protein, human ortholog genome position, accession, identify score and sequence alignment, protein pathology, expression, PTMs, subcellular location and links to publications containing information on the protein in question. For each unique protein i.d., this information is extracted from both the Ensembl and Uniprot database. For nonhuman organisms, the human ortholog information is also included for comparison. Due to the long run time, the user has the choice to opt out of UniProt and human ortholog information mining.

Nonhuman organism data usually have many unreviewed accessions within the dataset. Some databases such as PhosphoSitePlus, host site information based on Swiss-Prot accessions, and does not include unreviewed accessions. This results in difficulties matching the input phosphosite identity to the database's information. PhosPiR solves this issue by identifying the Swiss-Prot accession for the protein and aligning the sequences to generate the equivalent reviewed phosphosite position. This reviewed site information can be used for database searches by data annotation tools, thereby maximizing the identification of associated biological information. Human ortholog information allows for direct comparison of model organism data to human information. Pairwise alignment to human ortholog protein sequences allows the user to identify orthologous phosphorylation sites in human for any site of interest identified in their organism. Sequence alignment should only serve as a reference. Although it should be accurate for alignments with high identity scores, its practicality decreases as sequence identify score decreases. The

following packages are utilized for this step: 'biomaRt' [22], 'Biostrings' [23], 'GenomicAlighments' [24], 'protr' [25] and 'UniprotR' [26].

Differential expression analysis

Statistical tests are performed based on the group comparison setup. The user selects whether or not the data are paired in the given comparison and is offered a choice of statistical tests. For two group comparisons, fold change is automatically calculated. The fold change direction is determined by the group number, where the group with the larger number is the numerator. Group numbering should therefore take this into account. Four statistical tests *T*-test, Wilcoxon signed-rank test, reproducibility-optimized test statistic (ROTS) and rank product test can be chosen. *T*-test should not be chosen for nonparametric data. All tests can be selected if desired. Each test will yield a *P*-value and an FDR value for each data row. For a multiple group comparison, ANOVA with post hoc Tukey HSD Test will be performed if the groups are not paired, and linear mixed effect (LME) modeling is performed if groups are paired. Next, the user can set thresholds based on *P*-value or FDR for example. The significant lists will then be used as input for enrichment and network analyses. The user can choose volcano plots to visualize the statistical results. Multiple comparisons (maximum 4) can be plotted together. Based on the selected statistical cutoff, significant entries in the plot will be labeled if the number of significant changes is less than 60 in total. ROTS analysis utilizes the 'ROTS' package [27], 'RankProd' [28] performs the rank product analysis, 'multcompView' [29], 'lsmeans' [30] and 'nlme' [31] are ultilized for LME modeling. 'ggrepel' [32] and 'gridExtra' [33] are utilized in addition to 'ggplot2' for volcano plots.

Enrichment analysis

Enrichment analysis is performed on both phosphoprotein intensity data and phosphosite data. Protein level enrichment utilizes the 'clusterProfiler' package [34]. This powerful analysis tool enables gene ontology (GO) enrichment, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment and cell marker enrichment, and for human data, disease-association enrichment. For KEGG analysis, it automatically utilizes the current latest online version of KEGG database which for example includes the COVID-19 pathway from the newest release (v.98). All the listed enrichment analyses are performed for each significant list created earlier. Both universal background and dataset background are applied for separate analyses. Only enrichments with significant entries are recorded as a result. Phosphosite enrichment utilizes the 'PTM-SEA' (PTM-signature enrichment analysis) tool and its library PTMsigDB [35]. PTMsigDB curates detailed PTM information based on perturbation-induced site-specific changes, such as the direction of phosphorylation change upon a signaling event, and the affected signature sets that are collectively regulated. PTM-SEA analysis is performed on the entire dataset for all group comparisons. It is available for human, mouse and rat.

Kinase–substrate analysis

A kinase–substrate relationship analysis is performed with the 'KinSwingR' package [36], on each group comparison. It predicts a kinase–substrate interactome based on a library of motifs, and then integrates the fold change direction and *P*-value from the statistical results, to calculate a normalized swing score. The score resembles a *z*-score which predicts kinases' activity changes. The *P*-value is calculated to determine the significance level based on a permutation test. PhosPiR utilizes the latest kinase information from the PhosphoSitePlus human, mouse and rat data for customized kinase library [37] instead of the kinase dataset included in KinSwingR, which has become outdated and is only available for human.

A kinase–substrate network Circos plot is automatically created with the 'circlize' package [38] which shows the top 250 significant substrates. Kinases are connected with edges to their specific substrate sites with phosphorylation sites grouped by phosphoprotein.

Phosphoprotein/protein network analysis

The phosphoprotein network is built using the 'STRINGdb' package [39]. The STRING tool uses its own protein–protein interaction database which is used in PhosPiR to build an interaction network from each significant data list. Only interactions with greater than 0.4 confidence score (ranging from 0 to 1) are included in the STRING network figure (e.g. Supplementary Figure 1). The user can choose to identify hubs from within each network, and a hub interaction-enrichment score is also calculated. Hub phosphoproteins represent highly connected proteins within the network and are therefore likely to be functionally informative. The user can also choose from two cutoffs for defining hubs: the top 10% highest interactions, or an interaction count of 1 SD above the mean. The hub interaction enrichment is calculated by generating 1000 control networks for each hub and comparing hub interactions in control networks to the query network. Both *P*-value and FDR are presented in the result.

Results

Description of example setup and output

To demonstrate the functionality of PhosPiR, we analyzed synaptoneurosome phosphoproteome data from Brüning *et al.* [8]. In the original study, the authors studied the phosphorylation changes over time in synaptic terminals (otherwise known as synaptoneurosomes) from sleep-deprived mice and control mice. Here, we compared the overall differences in synaptoneurosome protein phosphorylation from mice while awake or asleep under control or sleep-deprived conditions.

The original study processed the synaptoneurosome phosphopeptide data with the EasyPhos platform. The sleep/wake cycles were controlled as follows: mice were kept in a light:dark 12 hours cycle, synaptoneurosomes were taken every 4 hours ($n = 4$) in a single day for sleep-deprived mice and baseline mice, totaling to 48 samples [8]. For the reanalysis of this data, we downloaded the raw MS files from the PRIDE database with identifier PXD010697. Taking the original data preparation as a reference [8], we preprocessed the raw data in a similar fashion using MaxQuant 1.6.17.0 [9] and Perseus 1.6.7.0 [40] (Supplementary Table 1). When inputting the pre-processed data into PhosPiR, 'Neither' was selected for normalization and imputation, as it is done outside of the pipeline.

The folder organization of the output files from PhosPiR is shown schematically in Supplementary Figure 2. The *Group Information* folder describes the group comparisons as setup by the user. Examples of group information can be found in Supplementary Tables 3 and 4, respectively. In the Overview Figure folder, the overall data distribution is visualized in several ways including histograms, heat-maps, PCA and boxplots plots. An example is shown in Supplementary Figure 3. The Statistical Analysis folder presents the statistical data including significance threshold for several tests as well as volcano plot output of significantly changing phosphopeptides or proteins, for each comparison and for every statistical test selected. Examples of statistical analysis output are shown in Figure 2A and Supplementary Table 5. The Enrichment folder stores the protein-centric result on cell marker enrichment, GO enrichment and KEGG enrichment (Figure 2B and Supplementary Table 6), and peptide-centric PTM-SEA enrichment result. PTM-SEA data are stored in the PhosphoSite enrichment subfolder, which includes converted. GCT input files and PTM-SEA analysis results, one per comparison, each in its own folder. Example rank plots of phosphorylation signatures can be found in Figure 2C. Kinase Analysis folder includes predicted significant kinases analyzed from KinSwingR, and resembled by swing scores, similar to z-scores for kinase activity change, and *P*-values, which indicate the significance of this change. Kinase activities are evaluated from the entire set rather than from the significant list. Examples of motif diagrams and kinase swing score output can be found in Figure 3 and Supplementary Table 7, respectively. The Network folder provides interaction figures of kinases to substrates spinning off from kinase analysis (Figure 4), also provides output from the STRING database network analysis and hub significance analysis. Examples are shown in Supplementary Figure 1 and Figure 5. The Annotation folder includes important ID information, UniProt database information, human homolog information and sequence alignment for all proteins as well as phosphorylation sites from the dataset. An example of human homolog ID information and UniProt database information can be found in Supplementary Tables 8 and 9, respectively.

Description of example results

The MS data used here [8], incorporated intensity data for 13 634 phosphosites from which, 8386 remained after filtering. Among these, PhosPiR identified 61 known disease associated phosphorylation sites, and 256 known regulatory sites using the automatic detection annotation tools (Supplementary Tables 10 and 11). The group comparisons (control versus sleep-deprived, and wake period versus sleep period with or without sleep deprivation), identified 367 significantly changing phosphorylation sites with fold change ≥ 2, and Rank product FDR of <0.05. These results can be seen from volcano plot and csv file output (Figure 2A, Supplementary Table 12). Interestingly, the proteins with significantly altered phosphorylation between wake and sleep time were enriched for changes on the dopaminergic synapse pathway, as shown in Figure 2B. Thus, significant phosphopeptide changes were identified for voltage gated ion channels VGCC, VSSC and Cav2.1/2.2, and for signaling proteins PLC, PKC and CamKII (Figure 2).

In the phosphosite-centric enrichment analysis, the signature set 'rapamycin' was 40% downregulated and the 'mTOR' signature set was 14% upregulated, in sleep-deprived synaptoneurosomes (Figure 2C), consistent with known negative regulation of mTOR by Rapamycin [41]. This demonstrates the utility of the PhosPiR pipeline to make functionally accurate predictions as the mTOR pathway is known to regulate sleep-deprivation induced responses [42, 43]. Moreover, as the PhosPiR identifies specific phosphorylation sites from these signature sets, as well as their regulatory function, where documented (Supplementary Table 13), far more detailed mechanistic insight can be gained using PhosPiR's integrated approach than would be possible with a stand-alone phosphosite analysis. This is further supported by the PhosPiR kinase activity analysis, which uses the KinSwingR package to predict kinase activity changes (Figure 3, Supplementary Table 14).

Examples of identified kinase substrates that undergo altered phosphosphorylation at synapses during wake versus sleep hours, or following sleep deprivation, are shown in the Circos plots in Figure 4A and B. In Figure 4A, neurofilament M (NEFM) which shows the greatest fold change in phosphorylation (−115×) can be seen to be regulated by SRC, ADRBK1 (GRK2), CSNK1D and PRKCD in synaptoneurosomes when comparing wake hours to sleep hours. Examination of the corresponding statistics file reveals that RPS6KA1 has the most increased activity among kinases based on motif phosphorylation, and PRKCZ shows the largest decrease in activity during wake hours in sleep-deprived mice based on observed phosphorylation changes on known kinase motifs (Figure 4B). This figure also highlights mTOR, this time as one of the kinases with the largest number of substrates that undergo altered phosphorylation in synaptoneurosomes

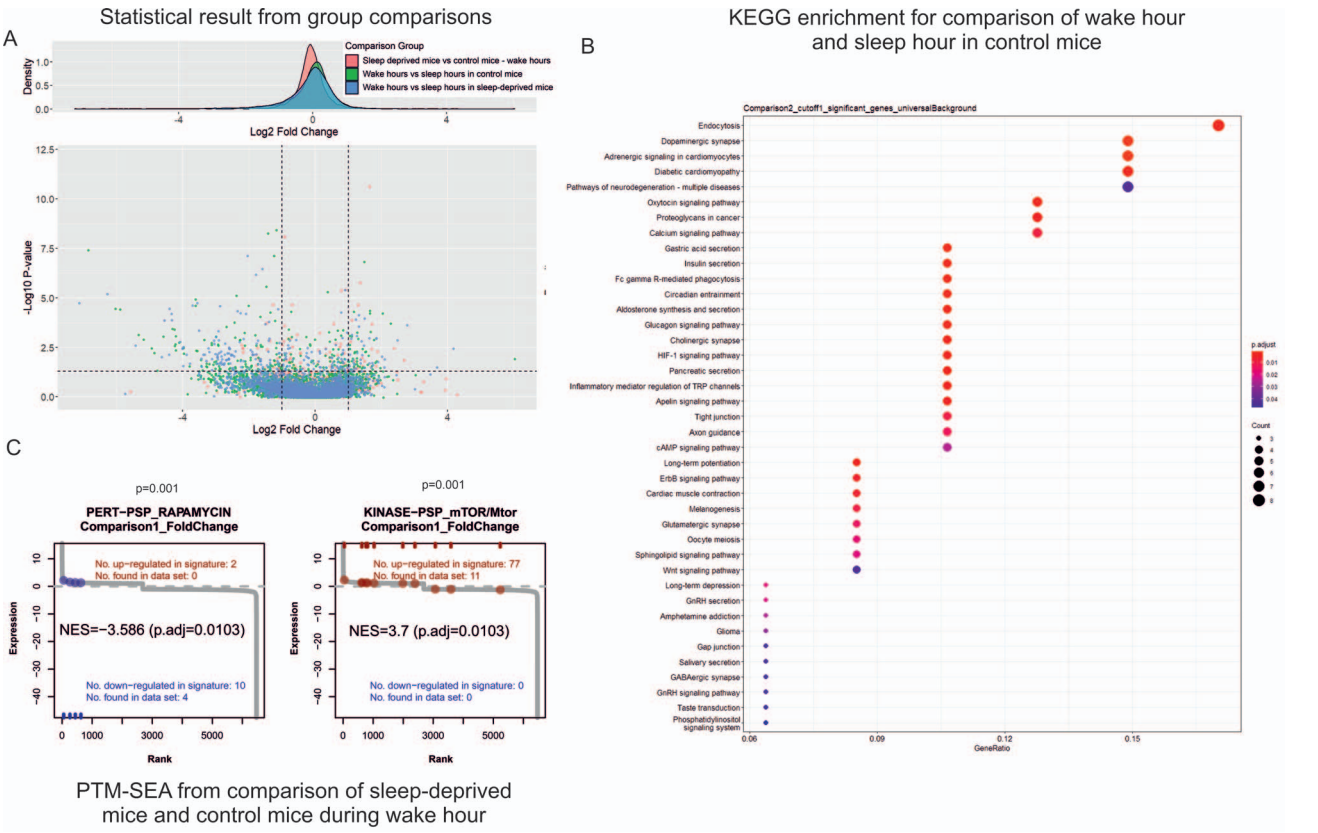

Figure 2. Sample output from the statistical and enrichment analysis feature of PhosPiR. Phosphoproteome changes in synaptoneurosomes of sleep-deprived versus normal sleep cycle mice. (A) Normalized phosphoproteomic data were loaded to PhosPiR. Automated statistical analysis was done on user-defined group comparisons for up to four statistical tests and visualized as volcano plots and csv files. A representative volcano plot is shown for the rank product FDR statistical analysis output. Significant proteins are labeled in the volcano plots only when there are ≤60 significant datapoints, otherwise the labels overlap. In this example the number of significant hits are >60. Every volcano plot is accompanied by a csv file providing detailed numerical output from all statistical tests, including UniProt and gene accession identifiers. (B) PhosPiR performs several enrichment analyses on the data, e.g. GO, cell marker and KEGG enrichment analyses. The KEGG analysis output is shown from the comparison between wake and sleep time synaptoneurosomes phosphopoteome, as an example. (C) Phosphosite enrichment using the post-translational modification set enrichment analysis (PTM-SEA) compares synaptoneurosomes from sleep-deprived mice and control mice during wake hours. Enrichment *P*-values and FDR (adjusted *P*-value) are indicated. This analysis highlights synaptic upregulation of mTOR pathway phosphorylation in sleep-deprived mice. Information on specific proteins and regulated sites are found in the accompanying csv file in the Enrichment\PhosphoSite enrichment folder.

from sleep-deprived mice. The precise substrates; adhesion G-protein coupled receptor L1 (LPHN1), cell cycle exit and neural differentiation protein 1 (CEND1), piccolo (PCLO), F-actin mono-oxygenase (MICAL3), spectrin beta chain (SPTBN1), MARCKS, GJA1 and MAP1B, and their altered phosphorylation sites, can be read directly from the plot (Figure 4B).

The hub analysis of protein phosphorylation in synaptoneurosomes during wake time versus sleep time has identified that the NMDA receptor subunit GRIN2B was a highly significant signaling hub (Figure 5A). This is consistent with several reports pointing to NMDAR in sleep regulation especially in the context of autoimmune encephalitis-induced sleep disturbance [44, 45]. Similarly, SHANK3 was highly connected to the changing phosphoproteins, consistent with its reported action in the control of circadian rhythm [46]. Synapsin I (SYN1), a neurotransmitter release regulatory protein, was also highly networked with the wake cycle phosphoproteins. SYN1 has previously been associated with synaptic changes following sleep deprivation [47]. Conversely, in sleep-deprived mice, there were fewer hubs overall consistent with the finding of Bruning *et al.*, which showed that phosphorylation cycling was reduced upon sleep deprivation [8]. Nonetheless, synapsin I, neurofilament (NEFM) and MAPT showed increased connectivity to the regulated phosphoproteins (Figure 5B). MAPT phosphorylation has been shown to increase upon sleep-deprivation stress [48]. Thus, PhosPiR automated analysis identified known regulators of sleep/wake cycle and sleep-deprivation stress in synaptic terminal preparations from mouse brain. Moreover, PhosPiR provides site-specific and network information that can assist detailed parsing of mechanism.

PhosPiR also enables kinase–substrate predictions and links to kinase activity directional changes. Moreover, as all identified sites are matched to their human homolog with pairwise alignment (Supplementary Table 15). Analysis of pathological implications could be further confirmed from database search of aligned homolog sites. These are some of the highlights of PhosPiR.

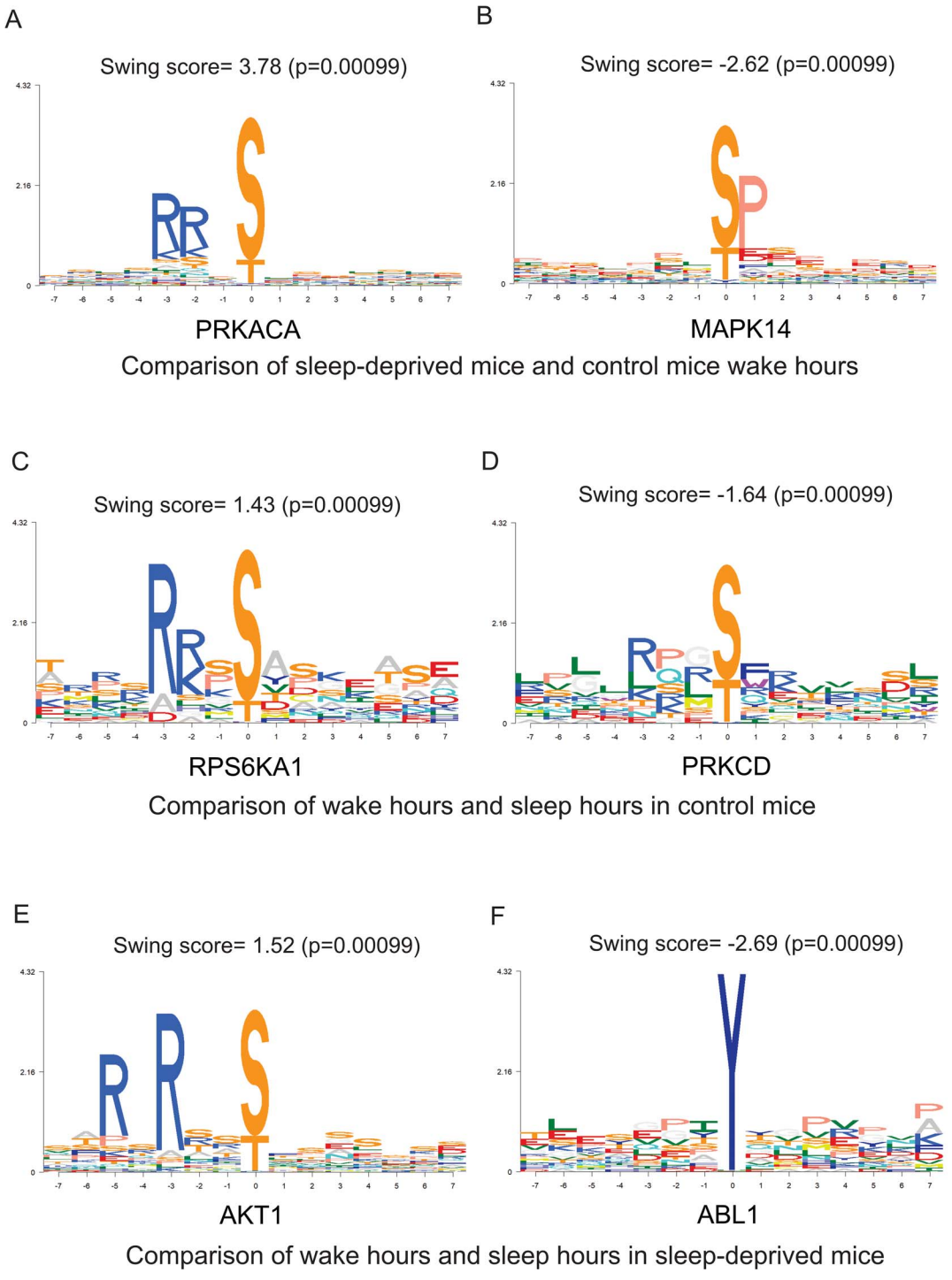

Figure 3. PhosPiR utilizes the KinSwingR tool to predict increases or decreases in kinase activities for defined group comparisons. PhosPiR integrates the KinSwingR tool which assesses local connectivity (swing) of kinase–substrate networks. Automated output from PhosphoPiR kinase analysis predicts regulated kinase activity based on identified substrate motifs. The final swing score is a normalized and weighted score of predicted kinase activity. Swing scores, positive and negative represent the direction of kinase activity change. Representative output tiffs are shown and accompanying csv file (ComparisonX_swingscore) is found in the Kinase analysis output folder.

Discussion

PhosPiR is an automated pipeline that does not require any coding knowledge from its user. It integrates several new phosphoproteomic analysis tools such as PTM-SEA and KinSwingR into a single pipeline while it simultaneously translates phosphoproteomic data from model organisms to human in order to exploit a range of customized databases that facilitate identification of functionally relevant information.

Although all analysis steps are automatic, the pipeline retains flexibility through its setup options. The user is free to fully customize group comparisons. For example, in addition to the examples shown here, a time series

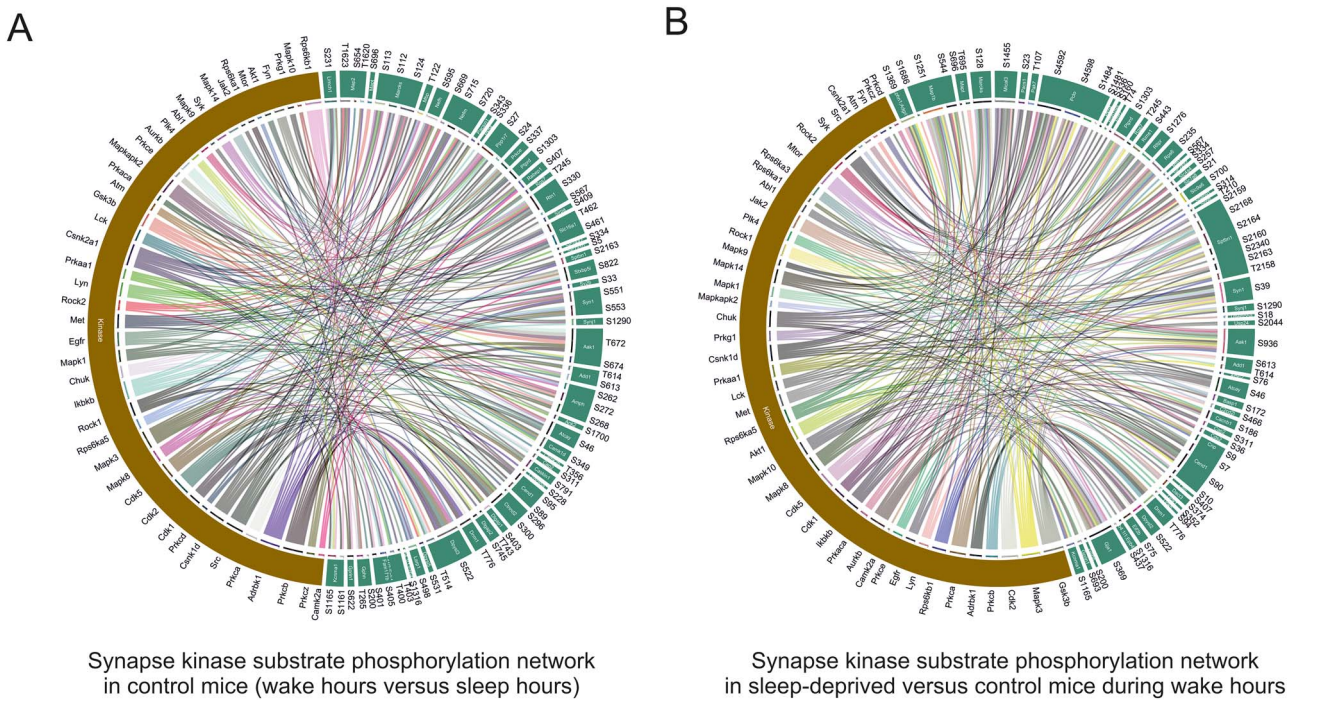

Figure 4. Kinase substrate network prediction tool. PhosPiR performs a proteome-wide kinase analysis using the KinSwingR package as shown (Figure 3). The PhosPiR Network Analysis tool finds the top kinase–substrate relations and presents them in a Circos plot. (A and B) Predicted kinase–substrate connections from the significantly changing data for group comparisons (A) wake hours versus sleep hours from control mice and (B) sleep deprived versus control mice during wake hours are shown. Colored ribbons link the kinase of interest with the substrate phosphorylation site that is significantly changed in the comparison. Predictions rely on known kinase–substrate phosphorylation sites. Only the top 250 most significant kinase–substrate relationships are plotted, to facilitate labeling. All output data are available in the accompanying csv file ComparisonX_significant_kinaseNetwork.

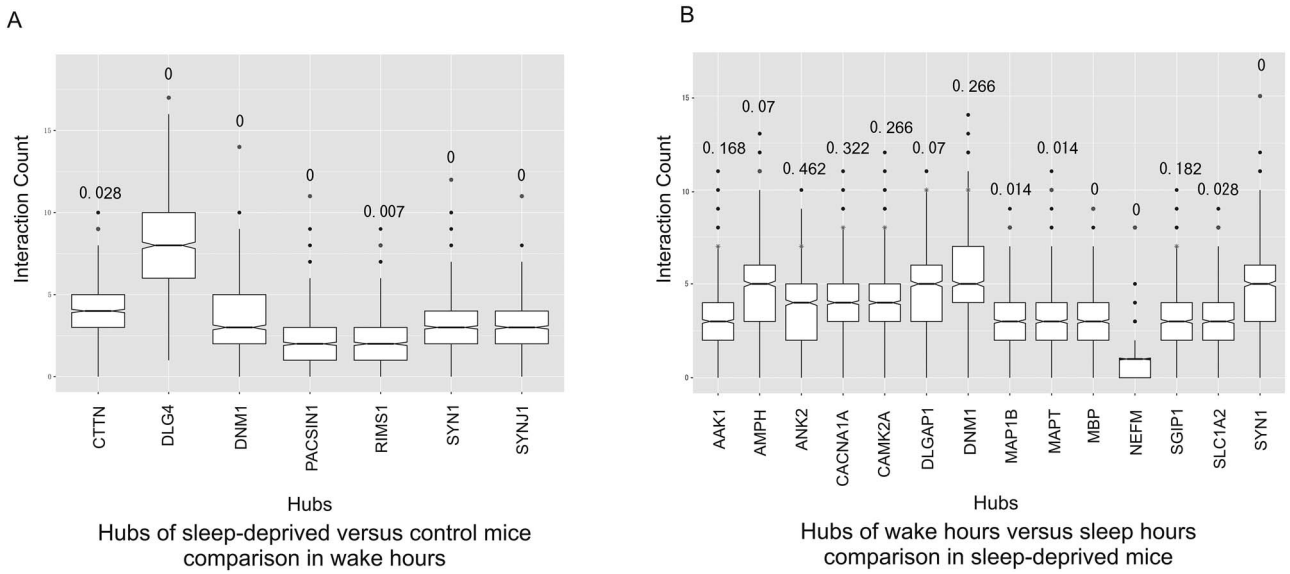

Figure 5. PhosPiR identifies network hubs based on protein:protein interactions. Sample output from the Network Analysis tool hub analysis is shown in (A) and (B). Hubs are defined as proteins with interaction number > 1 SD from the mean. Hub significance is calculated from the number of interactions within the data set compared to 1000 equal sized background datasets randomly generated from the total data. The hub interaction count in the background dataset is shown as a boxplot, and interaction count (hubness) in the target network is indicated by a red star. FDR values calculated from the permutation test are indicated above the boxplots. Hubs from comparisons of sleep deprived versus control mice during wake hours, and wake hours versus sleep hours from sleep-deprived mice are shown in (A) and (B), respectively.

analysis can also be included, by choosing the pairwise multigroup comparison option. All user options are presented with textboxes created using the 'svDialogs' package [49] creating a straightforward and seamless experience for the user. Although the pipeline can be applied also to nonphosphoproteomic data, several of the functions are specific to phosphoproteomics. For example, the ortholog alignment function, PTM-SEA enrichment analysis, kinase–substrate analysis and the kinase network figure generation all rely on phosphorylation-site

information. Among the unique features is the ortholog alignment tool that provides a human reference site for every single phosphorylation-site from mouse or rat. An important benefit of PhosPiR is that it integrates several packages, such as PTM-SEA that utilize customized, curated libraries that contain up to date information. The integrated 'kinase analysis' function in PhosPiR not only predicts which kinases are linked to the input data, but also predicts activity changes based on the generated statistics. The 'kinase network' function clearly labels substrate phosphosite position and organizes them by protein, thereby providing a clear visual summary for significant kinase substrate changes.

Many important functions of the pipeline come from recently developed, powerful R-packages, which the pipeline unifies and provides important customizations. For example, most of the included analysis packages require strict input formats and generate diverse output formats. PhosPiR utilizes packages such as 'reshape2' [50], 'vroom' [51], 'openxlsx' [52], 'textreadr' [53], 'plyr' [54] and 'cmapR' [55] to transform data between analysis steps, so that input requirements are satisfied, and output results are unified. Output figures are expanded from the originals, modified with packages such as 'gplots' [56], 'gridExtra' [33] and 'RColorBrewer' [57] to be informative at a glance. Many more packages are utilized and listed within Method section, we wanted to include all of them to offer their functionalities and customizations to a wider audience of nonbioinformaticians. PhosPiR also supports a wide range of organisms. With the exclusion of the disease ontology semantic enrichment analysis, kinase analysis and PTM-SEA which search only human data, whereas all other analysis steps support up to 18 organisms.

Our pipeline offers unique functionalities compared to even the most recent analysis packages such as PhosR [58], which provides a kinase analyses toolkit for bioinformaticians working in R coding language. In contrast, the PhosPiR workflow does not require coding knowledge and can be performed by nonbioinformaticians. Furthermore, PhosPiR provides automated protein and site annotation from UniProt, Ensembl and PhosphoSitePlus. The annotation files provide information on functionality and associated pathologies. Scientific references for identified functions are included in the output. Another unique feature of the pipeline is the protein-centric network and hub analysis, which provides aligned sequence information on human homolog when for example the input data are from a model organism. Finally, the annotations accompanying the kinase enrichment function (using the PTM-SEA database) takes into account directionality of phosphorylation change when identifying pathology and regulatory signatures. These many exclusive features enable users to study their data from multiple angles and distinguishes PhosPiR from existing phosphoproteomic data analysis software.

Conclusion

In summary, PhosPiR is an automated pipeline that integrates a full range of analysis tools to maximize the discovery potential for phosphoproteomics datasets. PhosPiR is compatible with datasets from all MS spectral interpretation tools. In terms of compatibility, most of the analysis tools will support up to 18 different organisms for either phosphoproteomics or proteomics datasets. In terms of functionality, PhosPiR performs data preprocessing, normalization and imputation, figure overviews, statistical analysis, enrichment analysis, PTM-SEA, kinase analysis, network and hub analysis, while also carrying out annotation mining and homolog alignment. Altogether it produces over 100 result files and figures from an average dataset, without counting annotation files. In this paper, we use PhosPiR automated analysis to identify regulators of sleep/wake cycle and sleep-deprivation stress in synaptic terminal preparations from a mouse brain. This identifies known and novel regulators, thereby validating the pipeline's utility while at the same time contributing to discovery. We hope that PhosPiR will provide an opportunity for users with limited programming knowledge to experience great R packages for comprehensive functional prediction analysis with statistical support, from their phosphoproteomics data.

Key Points

- PhosPiR is an automatic phosphoproteomics pipeline which does not require any programing knowledge from the users.
- In a single run, PhosPiR provides peptide quality control, data overviewing with histogram, boxplot, heatmap and PCAs, data annotation utilizing UniProt and Ensembl database, differential expression analysis including four statistical test options and post hoc testing, phosphosite translation across species, four enrichment analyses for phosphoproteins, PTM-SEA (post-translational modification set enrichment analysis) for phosphopeptide, kinase analysis, network analysis and hub analysis.
- We included a variety of recently updated R package in the pipeline to offer their functionalities and customizations to a wider audience of nonbioinformaticians.
- Our code and tutorial videos can be found at https://github.com/TCB-yehong/PhosPiR.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

Example data are part of the study F. Brüning, S.B. Noya, T. Bange, S. Koutsouli, J.D. Rudolph, S.K. Tyagarajan, J. Cox, M. Mann, S.A. Brown and M.S. Robles, 'Sleep-wake cycles drive daily dynamics of synaptic phosphorylation,' Science, vol. 366, pp. eaav3617, 10/11. 2019. It is available on PRIDE database, at https://www.ebi.ac.uk/pride/archive/projects/PXD010697. The data files generated with PhosPiR are available at 10.6084/m9.figshare.16583390. This manuscript is accompanied by a short video tutorial which can be viewed here: https://youtu.be/c7n7yE0DMsA PhosPiR code can be accessed at https://github.com/TCB-yehong/PhosPiR.

Software availability statement

The PhosPiR software is available at the GitHub repository at the following site: https://github.com/TCB-yehong/PhosPiR. Any issues or questions during the run can be submitted to the 'Issue' tab on our GitHub page. The tutorial videos can be accessed via these links: https://youtu.be/c7n7yE0DMsA (short setup introduction video), https://youtu.be/n4EagNoxusI (long pipeline run demonstration and result file introduction video).

Funding

Business Finland grants (1817/31/2015 and 1545/31/2019), Michael J Fox Foundation grant 008489 and Academy of Finland grant 26080953 to ETC, the MATTI graduate school at the University of Turku to YH, and Academy of Finland (310561 and 335611) to LLE. Our research is also supported by Biocenter Finland and ELIXIR Finland.

References

1. Fischer EH. Cellular regulation by protein phosphorylation. *Biochem Biophys Res Commun* 2013;**430**:865–7.
2. Cohen P. The origins of protein phosphorylation. *Nat Cell Biol* 2002;**4**:E127–30.
3. Jouy F, Müller SA, Wagner J, *et al.* Integration of conventional quantitative and phospho-proteomics reveals new elements in activated Jurkat T-cell receptor pathway maintenance. *Proteomics* 2015;**15**:25–33.
4. Francavilla C, Papetti M, Rigbolt KTG, *et al.* Multilayered proteomics reveals molecular switches dictating ligand-dependent EGFR trafficking. *Nat Struct Mol Biol* 2016;**23**:608–18.
5. Robles MS, Humphrey SJ, Mann M. Phosphorylation is a central mechanism for circadian control of metabolism and physiology. *Cell Metab* 2017;**25**:118–27.
6. Derouiche A, Cousin C, Mijakovic I. Protein phosphorylation from the perspective of systems biology. *Curr Opin Biotechnol* 2012;**23**:585–90.
7. Bekker-Jensen DB, Bernhardt OM, Hogrebe A, *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun* 2020;**11**:787.
8. Brüning F, Noya SB, Bange T, *et al.* Sleep-wake cycles drive daily dynamics of synaptic phosphorylation. *Science* 2019;**366**:eaav3617.
9. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**:1367–72.
10. Cuklina J, Lee CH, Willams EG, *et al.* Computational challenges in biomarker discovery from high-throughput proteomic data. Ph.D. thesis, ETH Zurich, 2018. https://doi.org/10.3929/ethz-b-000307772.
11. Hediyeh-zadeh S. *msImpute: Imputation of Label-Free Mass Spectrometry Peptides*, 2021. R package version 1.2.0.
12. Ressa A, Fitzpatrick M, van den Toorn H, *et al.* PaDuA: a python library for high-throughput (Phospho)proteomics data analysis. *J Proteome Res* 2019;**18**:576–84.
13. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
14. Kolde R. *pheatmap: Pretty Heatmaps*, 2019. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.
15. Guha R. *fingerprint: Functions to Operate on Binary Fingerprint Data*, 2018. R package version 3.5.7. https://CRAN.R-project.org/package=fingerprint.
16. Oksanen J, Blanchet FG, Friendly M, *et al. vegan: Community Ecology Package*, 2020. R package version 2.5-7. https://CRAN.R-project.org/package=vegan.
17. Adler D, Murdoch D. *rgl: 3D Visualization Using OpenGL*, 2021. R package version 0.107.14. https://CRAN.R-project.org/package=rgl.
18. Lê S, Josse J, Husson F. FactoMineR: AnRPackage for multivariate analysis. *J Stat Softw* 2008;**25**:1–18.
19. Kassambara A, Mundt F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.
20. Soetaert K. *plot3D: Plotting Multi-Dimensional Data*, 2021. R package version 1.4. https://CRAN.R-project.org/package=plot3D.
21. Ooms J. *magick: Advanced Graphics and Image-Processing in R*, 2021. R package version 2.7.3. https://CRAN.R-project.org/package=magick.
22. Durinck S, Spellman PT, Birney E, *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91.
23. Pagès H, Aboyoun P, Gentleman R, *et al. Biostrings: Efficient manipulation of biological strings*, 2021. R package version 2.60.2. https://bioconductor.org/packages/Biostrings.
24. Lawrence M, Huber W, Pagès H, *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;**9**:e1003118.
25. Xiao N, Cao D, Zhu M, *et al.* Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 2015;**31**:1857–9.
26. Soudy M, Anwar AM, Ahmed EA, *et al.* UniprotR: retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase). *J Proteomics* 2020;**213**:103613.
27. Suomi T, Seyednasrollah F, Jaakkola M, *et al.* ROTS: an R package for reproducibility-optimized statistical testing. *PLoS Comput Biol* 2017;**13**:e1005562.
28. del Carratore F, Jankevics A, Eisinga R, *et al.* RankProd 2.0: a refactored Bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics* 2017;**33**:2774–5.
29. Graves S, Piepho H, Luciano Selzer with help from Sundar Dorai-Raj. *multcompView: Visualizations of Paired Comparisons*, 2019. R package version 0.1-8. https://CRAN.R-project.org/package=multcompView.

30. Lenth RV. Least-squares means: the R package lsmeans. *J Stat Softw* 2016;**69**(1):1–33. doi: 10.18637/jss.v069.i01.
31. Pinheiro J, Bates D, DebRoy S, *et al. nlme: Linear and Nonlinear Mixed Effects Models*, 2021. R package version 3.1-152. https://CRAN.R-project.org/package=nlme.
32. Slowikowski K. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*, 2021. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel.
33. Auguie B. *gridExtra: Miscellaneous Functions for Grid Graphics*, 2017.
34. Yu G, Wang LG, Han Y, *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 2012;**16**:284–7.
35. Krug K, Mertins P, Zhang B, *et al.* A curated resource for phosphosite-specific signature analysis. *Mol Cell Proteomics* 2019;**18**:576–93.
36. Waardenberg AJ. *KinSwingR: KinSwingR: Network-Based Kinase Activity Prediction*, 2021. R package version 1.10.0.
37. Hornbeck PV, Kornhauser JM, Tkachev S, *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012;**40**:D261–70.
38. Gu Z, Gu L, Eils R, *et al.* Circlize implements and enhances circular visualization in R. *Bioinformatics* 2014;**30**:2811–2.
39. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2018;**47**:D607–13.
40. Tyanova S, Cox J. Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research. In: von Stechow L (ed). *Cancer Systems Biology: Methods and Protocols*. New York, NY: Springer New York, 2018, 133–48.
41. Dutcher JP. Mammalian target of rapamycin inhibition. *Clin Cancer Res* 2004;**10**:6382S–7.
42. Maiese K. Moving to the rhythm with clock (circadian) genes, autophagy, mTOR, and SIRT1 in degenerative disease and cancer. *Curr Neurovasc Res* 2017;**14**:299–304.
43. Merrow M, Roenneberg T. Cellular clocks: coupled circadian and cell division cycles. *Curr Biol* 2004;**14**:R25–6.
44. Burgdorf JS, Vitaterna MH, Olker CJ, *et al.* NMDAR activation regulates the daily rhythms of sleep and mood. *Sleep* 2019;**42**:zsz135. https://doi.org/10.1093/sleep/zsz135.
45. Muñoz-Lopetegi A, Graus F, Dalmau J, *et al.* Sleep disorders in autoimmune encephalitis. *Lancet Neurol* 2020;**19**:1010–22.
46. Ingiosi AM, Schoch H, Wintler T, *et al.* Shank3 modulates sleep and expression of circadian transcription factors. *Elife* 2019;**8**:10.7554/eLife.42819.
47. Gilestro GF, Tononi G, Cirelli C. Widespread changes in synaptic markers as a function of sleep and wakefulness in Drosophila. *Science* 2009;**324**:109–12.
48. Barthélemy NR, Liu H, Lu W, *et al.* Sleep deprivation affects tau phosphorylation in human cerebrospinal fluid. *Ann Neurol* 2020;**87**:700–9.
49. Grosjean P. *SciViews-R*, 2019. Belgium: UMONS, MONS. http://www.sciviews.org/SciViews-R.
50. Wickham H. Reshaping data with the reshape package. *J Stat Softw* 2007;**21**:1–20.
51. Hester J, Wickham H. *vroom: Read and Write Rectangular Text Data Quickly*, 2020. R package version 1.5.5. https://CRAN.R-project.org/package=vroom.
52. Schauberger P, Walker A. *openxlsx: Read, Write and Edit xlsx Files*, 2021. R package version 4.2.4. https://CRAN.R-project.org/package=openxlsx.
53. Rinker TW. *textreadr: Read Text Documents into R*, 2018. R package version 0.9.1. Buffalo, New York. http://github.com/trinker/textreadr.
54. Wickham H. The split-apply-combine strategy for data analysis. *J Stat Softw* 2011;**40**:1–29.
55. Natoli T. *cmapR: CMap Tools in R*, 2020.
56. Warnes GR, Bolker B, Bonebakker L, *et al. gplots: Various R Programming Tools for Plotting Data*, 2020.
57. Neuwirth E. *RColorBrewer: ColorBrewer Palettes*, 2014.
58. Yang P, Kim T, Kim JH. *PhosR: A Set of Methods and Tools for Comprehensive Analysis of Phosphoproteomics Data*, 2020.