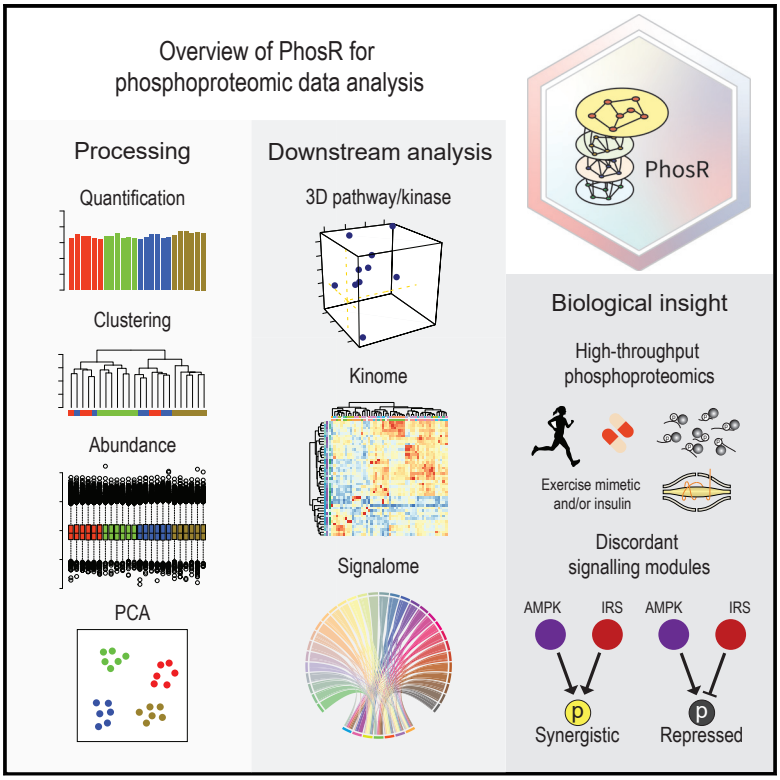
Cell Reports

PhosR enables processing and functional analysis of phosphoproteomic data

Graphical Abstract



Authors

Hani Jieun Kim, Taiyun Kim, Nolan J. Hoffman, Di Xiao, David E. James, Sean J. Humphrey, Pengyi Yang

Correspondence

pengyi.yang@sydney.edu.au

In brief

Protein phosphorylation regulates all aspects of cell biology. Although phosphoproteomics enables profiling of global phosphorylation, extracting knowledge from such data requires specialized computational methods. Kim et al. present PhosR for comprehensive analysis of phosphoproteomic data. Their analysis of muscle phosphoproteome uncovers unappreciated interactions between AMPK and insulin signaling pathways.

Highlights

- PhosR implements a suite of methods for comprehensive phosphoproteomic data analysis
- Stably phosphorylated sites are useful for phospho-data normalization and integration
- Signalomes constructed from PhosR enable interpretation of global signal transduction
- PhosR reveals unappreciated interactions between the AMPK and insulin signaling

Kim et al., 2021, Cell Reports *34*, 108771
February 23, 2021
https://doi.org/10.1016/j.celrep.2021.108771

CellPress

Cell Reports

CellPress
OPEN ACCESS

Report

PhosR enables processing and functional analysis of phosphoproteomic data

Hani Jieun Kim,[1,2,3,6] Taiyun Kim,[1,2,3,6] Nolan J. Hoffman,[3,4,7] Di Xiao,[2] David E. James,[3,4,5] Sean J. Humphrey,[3,4] and Pengyi Yang[1,2,3,8,*]

[1]School of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia
[2]Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia
[3]Charles Perkins Centre, The University of Sydney, Sydney, NSW, Australia
[4]School of Environmental and Life Sciences, The University of Sydney, Sydney, NSW, Australia
[5]Sydney Medical School, The University of Sydney, Sydney, NSW, Australia
[6]These authors contributed equally
[7]Present address: Exercise and Nutrition Research Program, Mary MacKillop Institute for Health Research, Australian Catholic University, Melbourne, VIC, Australia
[8]Lead contact
*Correspondence: pengyi.yang@sydney.edu.au
https://doi.org/10.1016/j.celrep.2021.108771

SUMMARY

Mass spectrometry (MS)-based phosphoproteomics has revolutionized our ability to profile phosphorylation-based signaling in cells and tissues on a global scale. To infer the action of kinases and signaling pathways in phosphoproteomic experiments, we present PhosR, a set of tools and methodologies implemented in a suite of R packages facilitating comprehensive analysis of phosphoproteomic data. By applying PhosR to both published and new phosphoproteomic datasets, we demonstrate capabilities in data imputation and normalization by using a set of "stably phosphorylated sites" and in functional analysis for inferring active kinases and signaling pathways. In particular, we introduce a "signalome" construction method for identifying a collection of signaling modules to summarize and visualize the interaction of kinases and their collective actions on signal transduction. Together, our data and findings demonstrate the utility of PhosR in processing and generating biological knowledge from MS-based phosphoproteomic data.

INTRODUCTION

Protein phosphorylation is an essential regulatory mechanism in cellular signal transduction. Elucidating changes in phosphorylation is crucial for understanding how cells sense and respond to environmental cues and perturbations (Humphrey et al., 2015a). Advances in mass spectrometry (MS)-based technologies have enabled us to quantify changes in phosphorylation of tens of thousands of phosphorylation sites in the phosphoproteome of cells (Sharma et al., 2014). Although these technological advances have enabled the generation of large-scale phosphoproteomic data (Macek et al., 2009), computational methods for phosphoproteomic data analysis remain in relative infancy. Upstream challenges in the analysis workflow include phosphosite filtering, handling missing values, and batch effect correction (Tyanova et al., 2016). Beside challenges in data processing, a major obstacle in phosphoproteomics is the lack of annotated phosphosites (Needham et al., 2019). Without knowledge of cognate kinase(s) for the majority of phosphosite sites, the identification of regulated phosphosites by themselves provides an incomplete view of signaling network function. Moreover, most phosphoproteomic studies still rely on an analysis framework in which phosphorylation is evaluated site specifically, although studies have revealed that many proteins are phosphorylated at multiple sites, of which some are targeted by orthogonal kinases. Adopting a phosphosite-centric analysis would therefore ignore any interactions and relationships between phosphosites from the same protein and any co-regulation of proteins at multiple sites.

Currently, only a handful of computational tools are suited to processing and downstream analysis of phosphoproteomic data. For example, although a large number of imputation algorithms have been developed for proteomic data (Webb-Robertson et al., 2015), significantly fewer methods are available for phosphoproteomic data (Tyanova et al., 2016). Similarly, a variety of methods developed for normalizing and batch-correcting genomic and transcriptomic data (Johnson et al., 2007; Risso et al., 2014) have been used for phosphoproteomic data normalization, but very few are specifically tailored for this task. For the downstream analysis of phosphoproteomic data, a number of tools (Beekhof et al., 2019; Casado et al., 2013; Mischnik et al., 2016) use kinase-substrate annotations to infer the activity of a kinase by evaluating the phosphorylation status of its substrates; however, these tools rely on a limited number of kinase-substrate relationships predicted or curated in databases (Dinkel

et al., 2011; Hornbeck et al., 2012) and may therefore restrict insight that could be obtained from unannotated sites. Although these methods can be used in conjuncture with methods that predict kinase-substrate relationships in a phosphoproteomic data agnostic manner (Horn et al., 2014; Wong et al., 2007), the use of motifs or a protein-protein interaction network overlooks the dynamic and context-specific nature of phosphorylation. Compared to these approaches, more recent methods such as PHOTON (Rudolph et al., 2016), CoPhosK (Ayati et al., 2019), CoPPNet (Ayati et al., 2020), and PHONEMeS (Terfve et al., 2015) use a data-driven approach to infer phosphorylation-based networks by using phosphosite-level interactions. These methods enable us to begin addressing issues related to the specificity of kinase-substrate relationships and context-specific regulation but do not currently take into account how a set of phosphosites may be co-regulated within and across proteins.

Here, we developed a phosphoproteomic analysis pipeline called PhosR (Figure 1A) to address key issues in processing and downstream analysis of large-scale phosphoproteomic data and applied the components of PhosR to a panel of published and new skeletal muscle cell phosphoproteomic datasets. We demonstrate the impact of imputation on downstream analysis, introduce "stably phosphorylated sites" (SPSs) and highlight their utility in phosphoproteomic data normalization and integration; and develop a kinase-substrate scoring method that leverages the dynamic profiles of canonical substrates and through which the global relationships of kinases and substrates can be annotated. We then use these annotations (1) to identify cognate proteins with phosphosites of similar regulatory profiles by interpreting phosphorylation sites in the context of their protein of origin; and (2) to construct "signalomes" of large-scale kinase and substrate relationships on the basis of these protein modules and, by doing, so reconstruct the interactions between kinases and their collective action on signal transduction pathways. Using our approach, we demonstrate distinct modules of cognate proteins that characterize the response of rat L6 myotubes to treatment with the 5′ AMP-activated protein kinase (AMPK) agonist 5-aminoimidazole-4-carboxamide-1-β-D-ribofuranoside (AICAR) and/or insulin stimulation. In particular, our approach revealed a module that is predominantly regulated by AMPK and is characterized by AMPK-activated phosphosites whose phosphorylation is attenuated with concurrent insulin stimulation. This finding reveals previously unappreciated interactions between the AMPK and insulin signaling pathways involved in coordinating skeletal muscle glucose transport and insulin sensitivity. Together, our data and findings demonstrate the utility of PhosR from various aspects of phosphoproteomic data processing toward generating biological insight from MS-based phosphoproteomic data and facilitating deeper understanding of existing and future large-scale phosphoproteomic resources.

RESULTS

Pre-processing phosphoproteomic data with PhosR strengthens biological signal for downstream analysis

To first demonstrate the handling of missing data in PhosR, we used a phosphoproteome profiling dataset from FL83B and Hepa 1-6 cells stimulated with insulin (Humphrey et al., 2015b). We conducted a stepwise imputation approach, comprising site- and condition-specific imputation (*scImpute*) and paired-tail imputation (*ptImpute*) that takes into account the value of quantified phosphosites as well as the experimental design (see STAR methods). We demonstrate that with the PhosR imputation, the percentage of quantified phosphosites rises to 80% across all samples (Figure S1A) and the biological replicates from each of the four conditions are now correctly clustered (Figure 1B). In comparison, both the background imputation and the k-nearest neighbor imputation did not lead to correct clustering of biological replicates under all conditions (Figure S1B). Next, to evaluate the impact of imputation on downstream analysis, we compared the number of differentially phosphorylated sites before and after imputation (Figure S1C). Interestingly, we found that although the number of differentially up- and downregulated sites nearly doubled in FL84B cells, imputation also led to an almost 3-fold decrease in the number of upregulated sites in Hep 1-6 cells, bringing the up- and downregulated sites to a relatively comparable range (Figure S1C). Using pathway enrichment analysis on the phosphoproteome summarized to protein level with the *phosCollapse* function in PhosR (see STAR methods), we further show that the differentially phosphorylated sites from the imputed datasets are more enriched for the key pathways related to insulin signaling (Figure 1C). Together, these findings suggest PhosR imputation strengthens biological signals and facilitates downstream pathway analysis.

Identification of a set of SPSs from phosphoproteomic data

Several commonly used data normalization approaches such as the "removal of unwanted variation" (RUV) (Gagnon-Bartsch and Speed, 2012) require a set of internal standards that are known to be unchanged biologically in the samples measured. This is a challenge for phosphoproteomics because phosphorylation is highly dynamic, with diverse regulation across different cell types and experimental conditions. To explore whether we could identify a set of phosphorylation sites that might meet the criteria of being "stably phosphorylated" across multiple phosphoproteomics datasets, we used four high-quality datasets generated from different cell types and experimental conditions (see STAR methods). We performed a four-way overlap of the four datasets and found 1,207 phosphosites common to all four datasets (Figure S1D). To identify SPSs, we ranked the overlapping phosphosites on the basis of their absolute log2 fold change (Figure S1E) and generated a consensus ranking by using a statistical framework (see STAR methods) for which phosphosites with consistently small fold changes are highly ranked and those with large fold changes are lowly ranked (Figures S1F and S1G). The top 100 phosphosites from the consensus list are referred hereafter as SPSs.

Normalization using SPSs removes unwanted variation in phosphoproteomic data

To evaluate the utility of SPSs in phosphoproteomic data normalization, we applied RUV-III (Molania et al., 2019) with SPSs (denoted as "*RUVphospho*") to normalize the phosphoproteomic data of rat L6 myotubes treated with AICAR, an

CellPress
OPEN ACCESS

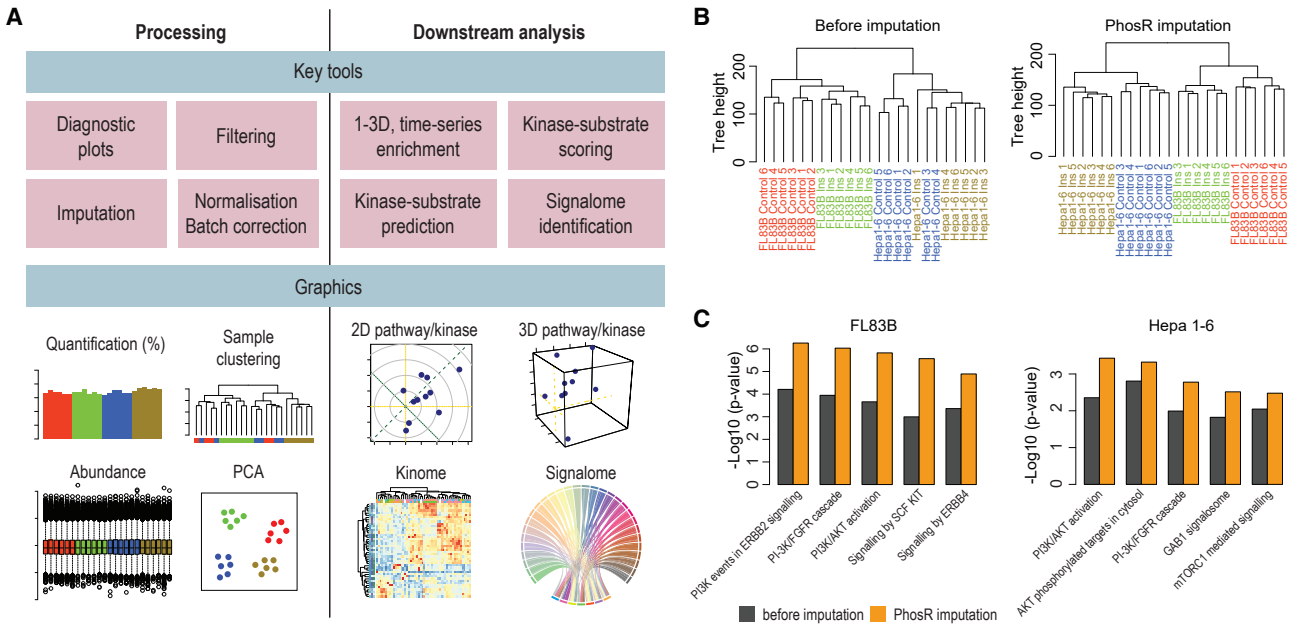


Figure 1. Overview of the main components of PhosR and impact on downstream analysis

(A) The key modules in PhosR are categorized into two broad steps of data analytics—processing and downstream analysis.

(B) Hierarchical clustering of biological replicates from phosphoproteomic experiments profiling FL83B and Hepa1-6 liver cells under basal or insulin-stimulated conditions.

(C) Enrichment of various signaling pathways known to be associated with insulin signaling before (black) and after (orange) imputation.

analog of adenosine monophosphate (AMP) that stimulates AMPK activity, and insulin either individually or in combination (Figure 2A; see STAR methods). Before normalization, hierarchical clustering and principal component analysis (PCA) of the myotube phosphoproteomic data revealed a batch effect that is driven by experiment runs for samples treated with insulin (Figures 2B and 2C, left panels), whereas normalization with *RUVphospho* effectively corrects this (Figures 2B and 2C, right panels).

Integrating multiple phosphoproteomics datasets from independent studies is typically challenging because signal derived from technical sources such as high-performance liquid chromatography (HPLC) and mass spectrometer performance characteristics often dominate biological signals. To illustrate the ability of *RUVphospho* to enable the integration of phosphoproteomics data from independent studies, we used two time-course datasets from early and intermediate insulin signaling from mouse liver (Humphrey et al., 2015b). This study was not included among the four used to select the SPSs. It contains two overlapping time points in each time series (0 s and 30 s), thus providing the opportunity to integrate the time series into a single comprehensive dataset. Prior to normalization, hierarchical clustering of the combined datasets reveals separation of the independent time series. Applying *RUVphospho* effectively integrates the two datasets, as demonstrated by the clustering of 0- and 30-s time point samples from the two datasets (Figure 2D). Closer inspection of the temporal phosphorylation change of phosphosite AKT1 T309, one of the most important markers of AKT activity in response to insulin stimulation (Humphrey and James, 2012), reveals a smoother temporal profile following normalization with *RUVphospho*. Importantly, normalization using data scaling and quantile normalization did not result in the correction of batch effect found in these datasets (Figures S1H–S1J). Collectively, these results demonstrate the normalization procedure in PhosR facilitates effective batch correction and integration of phosphoproteomic data.

Dual-centric analyses to detect regulated pathways and kinases in phosphoproteomic data

Most phosphoproteomic studies have adopted a phosphosite-level analysis of the data. To enable phosphoproteomic data analysis on the protein level, PhosR implements both site- and protein-centric analyses for detecting changes in kinase activities and signaling pathways through traditional enrichment analyses (over-representation or rank-based gene set test, together referred to as "one-dimensional enrichment analysis") as well as two-dimensional (2D) and three-dimensional (3D) analyses (Figure 1A). To test which signaling pathways are activated upon insulin stimulation in myotubes, we performed protein-centric enrichment analyses on the normalized myotube phosphoproteomic dataset by using both over-representation and rank-based gene set tests (Figure S2A). We found several expected pathways including those associated with mTORC1, AKT, and ERK signaling. Although these highly enriched pathways were largely in agreement between the two types of enrichment analyses, the rank-based gene set test had much greater statistical power in detecting these pathways (Figure S2B).

The two- and three-D analyses implemented in PhosR use direction-based statistics (Yang et al., 2014, 2016) that enables the investigation of kinases regulated by different combinations of

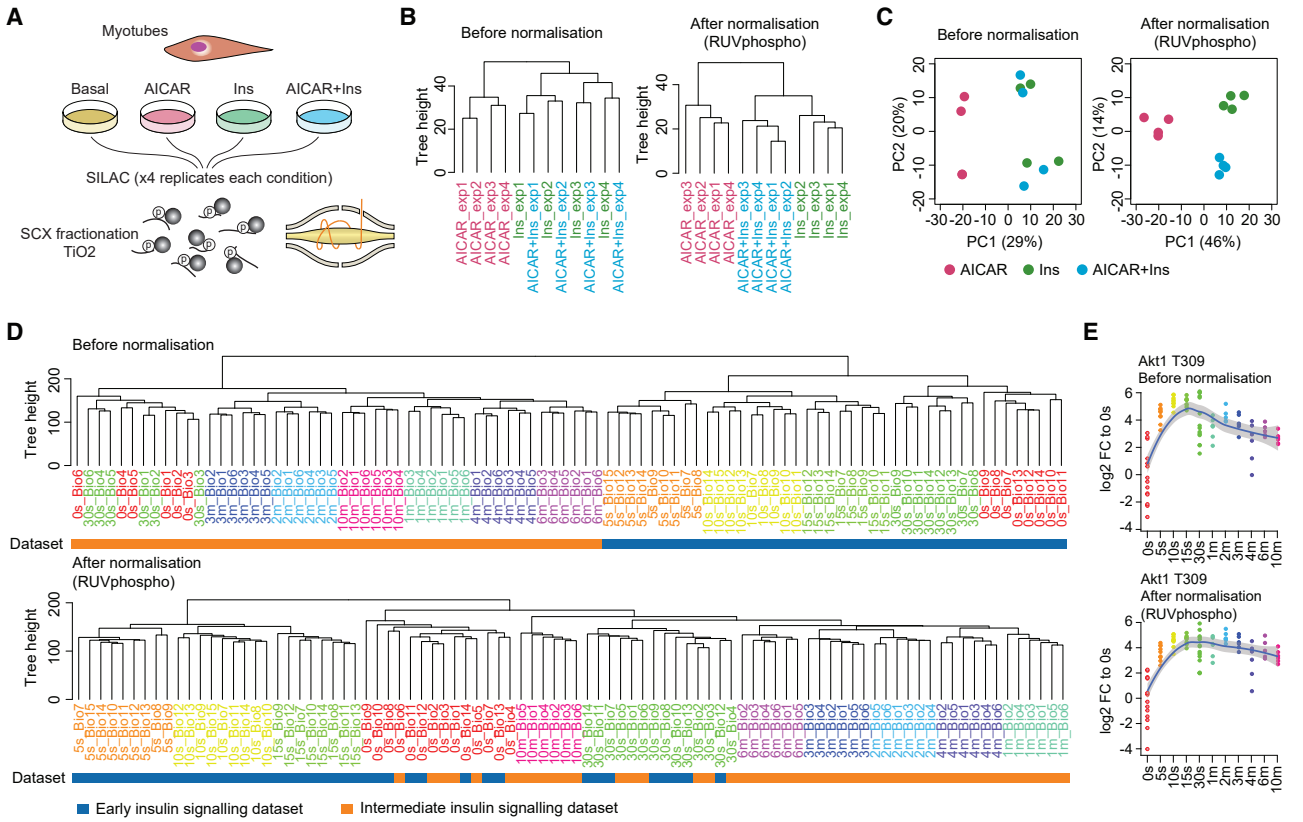


Figure 2. Normalization and batch correction using RUV and SPSs in PhosR

(A) Experimental setup of the phosphoproteomic profiling experiment in L6 myotubes in which phosphoproteomic analysis of cells were performed under the basal conditions or following treatment with the AMPK agonist AICAR, insulin (Ins), or in combination (AICAR+Ins).

(B) Dendrogram of all biological replicates before and after *RUVphospho* normalization of the myotube phosphoproteomic data. Samples are colored by experimental condition.

(C) PCA of the myotube phosphoproteomes before and after *RUVphospho* normalization. Each point represents a sample and is colored by experimental condition.

(D) Hierarchical clustering of phosphoproteomic datasets from early and intermediate *in situ* insulin stimulation of mouse liver before and after *RUVphospho* normalization. Samples are colored by time point and dataset.

(E) Log2 fold change in phosphorylation of AKT1 T309 upon insulin stimulation before and after implementation of PhosR normalization.

treatments. Applying this to the myotube phosphoproteome datasets, we found that, as expected, the activity of AMPK, marked by PRKAA1 (the catalytic alpha-1 subunit of AMPK), is upregulated following both AICAR and AICAR+Ins treatments but remains unchanged by insulin treatment alone (Figure S2C, top two panels). Strikingly, we found that the AICAR-induced upregulation of AMPK catalytic activity is attenuated by the addition of insulin as is observed from the kinase activity plot of AICAR versus AICAR+Ins (Figure S2C, bottom left panel). These pairwise comparisons can be summarized using the 3D analysis for which the three comparisons are integrated into a single statistical analysis to highlight the combinatorial effect of different treatments on PRKAA1 activity (Figure S2D).

Global kinase-substrate relationship scoring of phosphosites using PhosR

A key challenge in analyzing phosphoproteomics data is in identifying kinases responsible for the phosphorylation of specific sites. although various computational tools can be applied to annotate potential kinases of particular phosphosites on the basis of their amino acid sequences or structural information (Trost and Kusalik, 2011), most methods do not directly consider cell type and/or treatment/condition specificity of phosphorylation. To this end, PhosR implements a multi-step kinase-substrate scoring method in which first the likelihood of a kinase to regulate a phosphosite is scored by combining both kinase recognition motifs and the dynamic phosphorylation profiles of sites. The combined scores across all kinases are then integrated using an adaptive-sampling-based positive-unlabeled learning method (Yang et al., 2019a) to prioritize the kinase most likely to regulate a phosphosite (see STAR methods). The application of the proposed scoring method to the myotube phosphoproteome uncovers potential kinase-substrate pairs (Figure 3A, row dendrogram) and global relationships between kinases (Figure 3A, column dendrogram). A Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway overrepresentation analysis of the kinase-substrate pairs highlights over-represented pathways known to be associated with each kinase

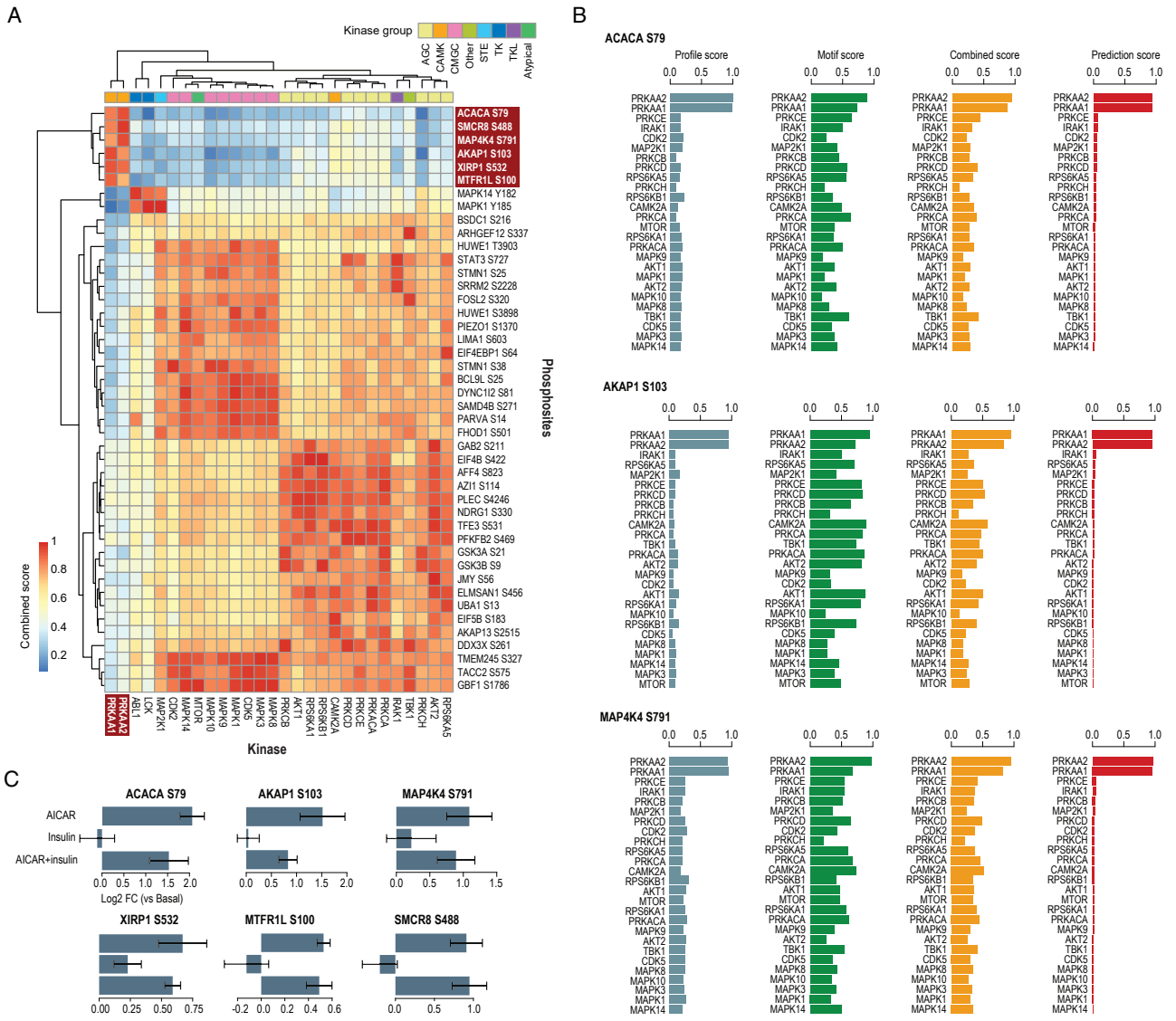


Figure 3. Global kinase-substrate relationship scoring of the myotube phosphoproteome

(A) A clustered heatmap of the combined kinase-substrate score for the top three phosphosites of all evaluated kinases. A higher combined score denotes a better fit to a kinase motif and kinase-substrate phosphorylation profile of a phosphosite.

(B) Bar plots showing profile, motif, and combined scores and positive-unlabeled ensemble learning prediction score of the top-ranked AMPK substrates.

(C) Bar plots showing the log2 fold change in phosphorylation level of the top-ranked AMPK substrates after treatment with AICAR, insulin, and combined treatment. Error bars denote standard deviation from the four biological replicates.

(Figure S3A). The kinase dendrogram reveals three major kinase groups governing the myotube phosphoproteome (AGC kinases [e.g., RPS6K and AKT isoforms], CMGC kinases [e.g., MAPKs], and CAMK kinases [e.g., AMPK catalytic subunits PRKAA1 and PRKAA2]). In particular, our kinase-substrate scoring method confirms several well-established substrates of AMPK such as ACACA S79, AKAP1 S103, SMCR8 S488 (Hoffman et al., 2015), and MTFR1L S100 (Schaffer et al., 2015) while also finding new candidate AMPK substrates such as XIRP1 S532 and MAP4K4 S791 (Figures 3A, 3B, and S3B). In agreement with our 2D kinase enrichment analysis (Figure S2C, bottom left panel), we demonstrate that the phosphorylation profiles of several of these AMPK substrates show a strong upregulation of phosphorylation upon AICAR stimulation that is attenuated when myotubes are co-stimulated with insulin (Figure 3C).

Construction of signalomes from discrete modules of co-regulated proteins

Proteins are frequently phosphorylated at multiple sites and often by orthogonal kinases. Site- and protein-centric analyses of phosphoproteomics data lie at opposite ends of the spectrum, with the former treating phosphosites on the same protein independently and ignoring the host protein information and the latter focusing on a specific site, losing information from other sites on

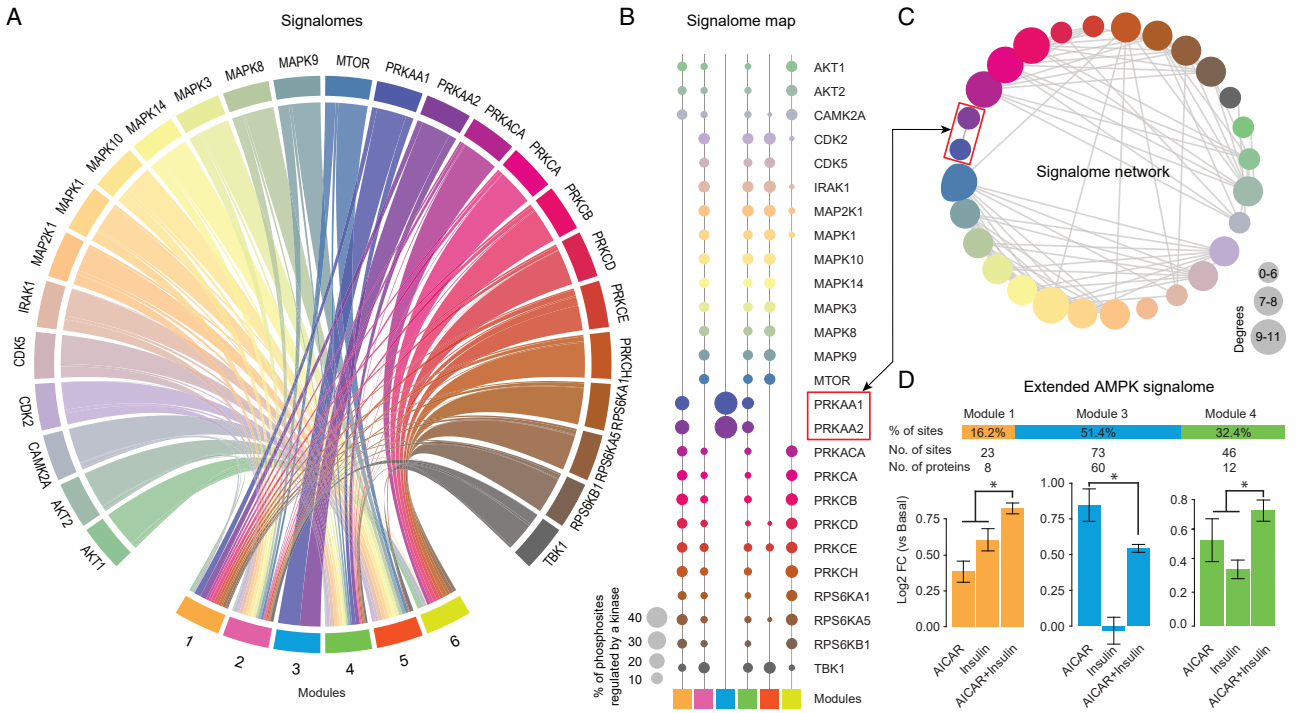


Figure 4. Construction of signalomes in the myotube phosphoproteome

(A) Signalomes identified from the myotube phosphoproteome. The branching nodes consist of 26 kinases, and the stem nodes consist of 6 protein modules each with a distinct phosphorylation and regulatory profile. Edges between nodes connect kinases to the protein modules they regulate.

(B) The signalome map demonstrating the proportion of phosphosites regulated by kinases in each protein module. The size of the balloon denotes the percentage of phosphosites the kinase regulates within a module.

(C) An interaction network of kinases. The higher the number of interactions with other kinases (degrees), the larger the circle.

(D) Summary of the three protein modules in the extended AMPK signalomes. Bar plots of log2 fold change of the phosphosites in each module are summarized for each condition against basal. The error bars denote standard deviation and * indicates a $p < 0.05$ using Wilcoxon rank-sum test.

the same protein. Because of the lack of appropriate methods, the question of whether proteins are co-regulated across multiple phosphosites remains poorly investigated. Leveraging our global kinase-substrate scoring of phosphosites, we set out to generate signalomes wherein dynamic changes in phosphorylation within and across proteins are conjointly analyzed.

We developed an approach to generate signalomes comprising discrete protein modules with phosphosites sharing similar dynamic phosphorylation profiles and kinase regulation (Figure 4A; see STAR methods). Using this approach, we show that the myotube phosphoproteome stimulated by AMPK activation and/or insulin stimulation contains six discrete protein modules. The resulting map of signalomes demonstrates that the modules are regulated by different kinases and at various proportions (Figure 4B). Notably, the signalome map highlights a module (blue, module 3) entirely regulated by AMPK catalytic activity (PRKAA1 and PRKAA2) and others (orange, module 1; and green, module 4) that are co-regulated by AMPK with other kinases, suggesting potential signaling crosstalk (Figure 4B). We then zoomed in to the extended AMPK signalome (see STAR methods) from the signalome network (Figure 4C) and found distinct phosphorylation profiles between the three protein modules (Figures 4D and S3D). Consistent with previous reports (Kjøbsted et al., 2015), the phosphorylation of sites in modules 1 and 4 show synergistic effects upon AICAR and insulin stimulation. Yet, in agreement with our kinase activity analysis, module 3—predominately regulated by AMPK alone—displays activity that is enhanced by AICAR and attenuated by insulin (Figure 4D).

DISCUSSION

Here, we present PhosR, a complete set of methods and tools for phosphoproteomic data processing and downstream analysis. Using PhosR, we have at once addressed many current challenges facing phosphoproteomic data analysis. We have addressed issues of data imputation, normalization, and integration through *RUVphospho*, supported by defining a set of SPSs, which we include in the PhosR package as a resource to the community. Processing of phosphoproteomics data with PhosR facilitated the extraction of differentially phosphorylated proteins with greater biological relevance, demonstrated by the strengthened signal of known pathways. *RUVphospho* normalization enabled datasets from independent studies to be integrated and eliminated batch effects without affecting biological signal.

Biochemically assigning phosphosites to their cognate kinase is an experimentally labor-intensive process and may be affected by the experimental system used. Moreover, because

of the great complexity within phosphoproteomes, many kinase-substrate relationships are likely to be context- and cell-type specific, further complicating efforts to elucidate them. Our global kinase-substrate scoring method enables computational inference of kinase activities specific to experimental conditions and cellular systems and the construction of signalomes wherein both dynamic and differential phosphorylation changes in phosphosites within and across proteins are taken into account. Using this approach, we could identify proteins co-regulated at the following three levels: (1) across experimental conditions, (2) between multiple phosphosites, and (3) by similar kinase regulation. In doing so, we can begin investigating these layers of complexities in signal transduction networks. Although the lack of annotated phosphosites remains a major challenge in phosphoproteomics studies (Needham et al., 2019), much progress has being made with the systematic mapping of kinases acting upstream of a large number of phosphorylation sites (Hijazi et al., 2020). We anticipate that as the number of experimentally validated kinase-substrate annotations increases the prediction accuracy and our capacity to recapitulate the underlying signaling network will also increase.

Previous research has demonstrated that skeletal muscle AMPK activation, following AICAR treatment or exercise, influences muscle glucose transport and insulin sensitivity. In particular, prior stimulation of skeletal muscle with AICAR to stimulate AMPK activity has been shown to enhance the sensitivity with which insulin stimulates glucose uptake (Kjøbsted et al., 2015). Our approach to generate modules of co-regulated proteins enabled the discovery of three sets of proteins with phosphosites that are regulated by AMPK in the stimulated myotube phosphoproteome (Figures 4A and 4B). Consistent with previous knowledge, we found two modules that exhibited enhanced phosphorylation upon insulin treatment if they were first stimulated by AICAR, demonstrating a synergistic effect between insulin and AMPK signaling pathways (Figure 4D). Indeed, we observed TBC1D1 among the proteins, which has been implicated in the AMPK-dependent increase of muscle glucose uptake and insulin sensitivity (Dokas et al., 2013; Kjøbsted et al., 2015; Taylor et al., 2008). Intriguingly, our approach also revealed a module entirely regulated by AMPK, and unlike the other two, the phosphosites found here demonstrated strong activation by AICAR treatment and no sensitivity to insulin stimulation alone. Strikingly, the AICAR-induced activation of phosphorylation on these sites was attenuated by the addition of insulin, suggesting a negative regulatory effect of insulin on the phosphorylation of AMPK substrates (Figure 4D). Because the key differences between these modules are differential kinase regulation of phosphosites and the presence of insulin-sensitive sites, we postulate that the interplay of AMPK with other kinases such as MAPKs and S6K may occur to stimulate diverse actions on different signaling pathways. In conclusion, our signalome construction method is applicable to diverse datasets that profile dynamic changes in the phosphoproteomes, enables inference of kinase activities through visualization of kinase interactions and their collective action on signal transduction pathways, and supports the interpretation of phosphoproteomic data at a level beyond the analysis of phosphosites in isolation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Phosphosite Filtering and Imputation
  - Identification of Stably Phosphorylated Sites and Data Normalization
  - Protein- and Phosphosite-centric Enrichment Analyses
  - Kinase-substrate Prioritisation of Phosphosites
  - Signalome Construction
  - Additional Functions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Differentially Phosphorylated Phosphosites
  - KEGG Pathway Over-representation Analysis

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2021.108771.

ACKNOWLEDGMENTS

The authors thank the colleagues at the School of Mathematics and Statistics, The University of Sydney, and Sydney Precision Bioinformatics Alliance for their intellectual engagement. This work is supported by an Australian Research Council (ARC)/Discovery Early Career Researcher Award (DE170100759) and a National Health and Medical Research Council (NHMRC) Investigator Grant (1173469) to P.Y., an Australian Research Council (ARC) Postgraduate Research Scholarship and Children's Medical Research Institute Postgraduate Scholarship to H.J.K., and the Judith and David Coffey Life Lab Scholarship to T.K.

AUTHOR CONTRIBUTIONS

P.Y. designed the computational methods with input from H.J.K. and S.J.H. H.J.K., T.K., and P.Y. performed the computational analysis. T.K., H.J.K., and P.Y. implemented the R package. D.X. assisted with the computational analysis and package implementation. N.J.H. and S.J.H. performed myotube sample preparation and MS analysis under the supervision of D.E.J. All authors wrote and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: June 1, 2020
Revised: December 7, 2020
Accepted: January 28, 2021
Published: February 23, 2021

REFERENCES

Ayati, M., Wiredja, D., Schlatzer, D., Maxwell, S., Li, M., Koyutürk, M., and Chance, M.R. (2019). CoPhosK: A method for comprehensive kinase substrate

annotation using co-phosphorylation analysis. PLoS Comput. Biol. *15*, e1006678.

Ayati, M., Chance, M.R., and Koyutürk, M. (2020). Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer. Bioinformatics, btaa678. https://doi.org/10.1093/bioinformatics/btaa678.

Beekhof, R., van Alphen, C., Henneman, A.A., Knol, J.C., Pham, T.V., Rolfs, F., Labots, M., Henneberry, E., Le Large, T.Y., de Haas, R.R., et al. (2019). INKA, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. Mol. Syst. Biol. *15*, e8250.

Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Casado, P., Rodriguez-Prados, J.-C., Cosulich, S.C., Guichard, S., Vanhaesebroeck, B., Joel, S., and Cutillas, P.R. (2013). Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. Sci. Signal. *6*, rs6.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res. *39*, D261–D267.

Dokas, J., Chadt, A., Nolden, T., Himmelbauer, H., Zierath, J.R., Joost, H.G., and Al-Hasani, H. (2013). Conventional knockout of Tbc1d1 in mice impairs insulin- and AICAR-stimulated glucose uptake in skeletal muscle. Endocrinology *154*, 3502–3514.

Gagnon-Bartsch, J.A., and Speed, T.P. (2012). Using control genes to correct for unwanted variation in microarray data. Biostatistics *13*, 539–552.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812.

Hijazi, M., Smith, R., Rajeeve, V., Bessant, C., and Cutillas, P.R. (2020). Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. Nat. Biotechnol. *38*, 493–502.

Hoffman, N.J., Parker, B.L., Chaudhuri, R., Fisher-Wellman, K.H., Kleinert, M., Humphrey, S.J., Yang, P., Holliday, M., Trefely, S., Fazakerley, D.J., et al. (2015). Global Phosphoproteomic Analysis of Human Skeletal Muscle Reveals a Network of Exercise-Regulated Kinases and AMPK Substrates. Cell Metab. *22*, 922–935.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods *11*, 603–604.

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40*, D261–D270.

Humphrey, S.J., and James, D.E. (2012). Uncaging akt. Sci. Signal. *5*, pe20.

Humphrey, S.J., Yang, G., Yang, P., Fazakerley, D.J., Stöckli, J., Yang, J.Y., and James, D.E. (2013). Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. Cell Metab. *17*, 1009–1020.

Humphrey, S.J., James, D.E., and Mann, M. (2015a). Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. Trends Endocrinol. Metab. *26*, 676–687.

Humphrey, S.J., Azimifar, S.B., and Mann, M. (2015b). High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. Nat. Biotechnol. *33*, 990–995.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Kjøbsted, R., Treebak, J.T., Fentz, J., Lantier, L., Viollet, B., Birk, J.B., Schjerling, P., Björnholm, M., Zierath, J.R., and Wojtaszewski, J.F.P. (2015). Prior AICAR stimulation increases insulin sensitivity in mouse skeletal muscle in an AMPK-dependent manner. Diabetes *64*, 2042–2055.

Macek, B., Mann, M., and Olsen, J.V. (2009). Global and site-specific quantitative phosphoproteomics: principles and applications. Annu. Rev. Pharmacol. Toxicol. *49*, 199–221.

Minard, A.Y., Tan, S.-X., Yang, P., Fazakerley, D.J., Domanova, W., Parker, B.L., Humphrey, S.J., Jothi, R., Stöckli, J., and James, D.E. (2016). mTORC1 Is a Major Regulatory Node in the FGF21 Signaling Network in Adipocytes. Cell Rep. *17*, 29–36.

Mischnik, M., Sacco, F., Cox, J., Schneider, H.C., Schäfer, M., Hendlich, M., Crowther, D., Mann, M., and Klabunde, T. (2016). IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. Bioinformatics *32*, 424–431.

Molania, R., Gagnon-Bartsch, J.A., Dobrovic, A., and Speed, T.P. (2019). A new normalization for Nanostring nCounter gene expression data. Nucleic Acids Res. *47*, 6073–6083.

Needham, E.J., Parker, B.L., Burykin, T., James, D.E., and Humphrey, S.J. (2019). Illuminating the dark phosphoproteome. Sci. Signal. *12*, eaau8645.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics *1*, 376–386.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. *47*, D442–D450.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rudolph, J.D., de Graauw, M., van de Water, B., Geiger, T., and Sharan, R. (2016). Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Syst. *3*, 585–593.e3.

Schaffer, B.E., Levin, R.S., Hertz, N.T., Maures, T.J., Schoof, M.L., Hollstein, P.E., Benayoun, B.A., Banko, M.R., Shaw, R.J., Shokat, K.M., and Brunet, A. (2015). Identification of AMPK Phosphorylation Sites Reveals a Network of Proteins Involved in Cell Invasion and Facilitates Large-Scale Substrate Prediction. Cell Metab. *22*, 907–921.

Sharma, K., D'Souza, R.C.J., Tyanova, S., Schaab, C., Wiśniewski, J.R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep. *8*, 1583–1594.

Taylor, E.B., An, D., Kramer, H.F., Yu, H., Fujii, N.L., Roeckl, K.S.C., Bowles, N., Hirshman, M.F., Xie, J., Feener, E.P., and Goodyear, L.J. (2008). Discovery of TBC1D1 as an insulin-, AICAR-, and contraction-stimulated signaling nexus in mouse skeletal muscle. J. Biol. Chem. *283*, 9787–9796.

Terfve, C.D.A., Wilkes, E.H., Casado, P., Cutillas, P.R., and Saez-Rodriguez, J. (2015). Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. Nat. Commun. *6*, 8033.

Trost, B., and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. Bioinformatics *27*, 2927–2935.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520–525.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Välikangas, T., Suomi, T., and Elo, L.L. (2018). A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief. Bioinform. *19*, 1344–1355.

Webb-Robertson, B.J.M., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G., and Waters, K.M. (2015). Review, evaluation, and discussion of the challenges of

missing value imputation for mass spectrometry-based label-free global proteomics. J. Proteome Res. *14*, 1993–2001.

Wong, Y.H., Lee, T.Y., Liang, H.K., Huang, C.M., Wang, T.Y., Yang, Y.H., Chu, C.H., Huang, H.D., Ko, M.T., and Hwang, J.K. (2007). KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res. *35*, W588–W594.

Yaffe, D. (1968). Retention of differentiation potentialities during prolonged cultivation of myogenic cells. Proc. Natl. Acad. Sci. USA *61*, 477–483.

Yang, P., Patrick, E., Tan, S.-X., Fazakerley, D.J., Burchfield, J., Gribben, C., Prior, M.J., James, D.E., and Hwa Yang, Y. (2014). Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. Bioinformatics *30*, 808–814.

Yang, P., Patrick, E., Humphrey, S.J., Ghazanfar, S., James, D.E., Jothi, R., and Yang, J.Y.H. (2016). KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. Proteomics *16*, 1868–1871.

Yang, P., Ormerod, J.T., Liu, W., Ma, C., Zomaya, A.Y., and Yang, J.Y.H. (2019a). AdaSampling for Positive-Unlabeled and Label Noise Learning With Bioinformatics Applications. IEEE Trans. Cybern. *49*, 1932–1943.

Yang, P., Humphrey, S.J., Cinghu, S., Pathania, R., Oldfield, A.J., Kumar, D., Perera, D., Yang, J.Y.H., James, D.E., Mann, M., and Jothi, R. (2019b). Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. Cell Syst. *8*, 427–445.e10.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

STAR★METHODS

KEY RESOURCES TABLE

RESOURCE	SOURCE	IDENTIFIER
Experimental models: cell lines		
Rat L6 myoblasts	Dr David Yaffe, Weizmann Institute of Science	RRID: CVCL_0385
Deposited data		
L6 myotube phosphoproteome	This paper	PRIDE: PXD019127
Hepa 1-6 & FL83 phosphoproteome	Humphrey et al., 2015b	PRIDE: PXD001792
Adipocyte insulin, LY, & MK phosphoproteome	Humphrey et al., 2013	NA
Adipocyte FGF2 phosphoproteome	Minard et al., 2016	PRIDE: PXD003631
Mouse liver insulin phosphoproteome	Humphrey et al., 2015b	PRIDE: PXD001792
ESC phosphoproteome	(Yang et al., 2019b)	PRIDE: PXD010621
Software and algorithms		
R version 3.6.1		https://www.R-project.org/
MaxQuant 1.5	Cox and Mann, 2008	https://www.biochem.mpg.de/5111795/maxquant
DirectPA 1.4	Yang et al., 2014	https://cran.r-project.org/web/packages/directPA/index.html
Limma 3.32.2	Ritchie et al., 2015	https://bioconductor.org/packages/release/bioc/html/limma.html
clusterProfiler 3.14.4	Yu et al., 2012	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
circlize 0.4.9	(Gu et al., 2014)	https://cran.r-project.org/web/packages/circlize/index.html
PhosR 1.0	This paper	https://bioconductor.org/packages/release/bioc/html/PhosR.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for reagent and resource may be directed to and will be fulfilled by the Lead Contact, Dr. Pengyi Yang (pengyi.yang@sydney.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The myotube phosphoproteomic data described in this study are deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/cgi/GetDataset) via the PRIDE (Perez-Riverol et al., 2019) partner repository. The accession number for the data reported in this paper is PRIDE: PXD019127.

The PhosR package is available as a Bioconductor package (https://bioconductor.org/packages/release/bioc/html/PhosR.html) and the latest development version and associated vignette are available from Github repository (https://pyanglab.github.io/PhosR/). All source code is published under the open-source license of GPL-3.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study utilized a collection of published phosphoproteomic datasets and profiled the phosphoproteome of myotubes in response to AICAR and insulin stimulation (new data). The published datasets used in this study include (i) the 'ESC differentiation' dataset

where a cocktail of treatments were applied to differentiate mouse embryonic stem cells (ESCs) to epiblast-like cells (Yang et al., 2019b) (PRIDE: PXD010621); (ii) the 'Adipocyte FGF21' dataset (Minard et al., 2016) (PXD003631) where the phosphoproteome of 3T3-L1 adipocytes treated with either insulin or FGF21 were profiled using SILAC quantification; (iii) the 'Adipocyte insulin, LY, and MK' dataset (Humphrey et al., 2013) where the phosphoproteome of 3T3-L1 adipocytes treated with either insulin in a time-course or LY and MK prior to insulin were profiled using SILAC quantification; (iiii) the 'FL83B and Hepa 1-6 Insulin' dataset (Humphrey et al., 2015b) (PXD001792) where the phosphoproteome of FL83B and Hepa 1-6 cells in basal and treated with insulin were profiled using label-free quantification; and (iv) the 'Mouse liver insulin' datasets (Humphrey et al., 2015b) (PXD001792) where the phosphoproteomes of mouse livers treated with insulin were profiled in an early time-course and an intermediate time-course using label-free quantification.

Rat L6 myoblasts (RRID: CVCL_0385), which have no reported sex (Yaffe, 1968), were grown and maintained in α-Minimum essential medium (α-MEM; GIBCO by Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS; Hyclone Laboratories) in a 10% CO_2 humidified incubator at 37°C. Cells were routinely checked for the absence of mycoplasma using the MycoAlert PLUS Mycoplasma Detection Kit (Lonza).

Stable isotope labeling by amino acids (SILAC) labeling (Ong et al., 2002) of L6 myoblasts was performed by supplementing SILAC DMEM (deficient in Lysine, Arginine and Leucine; Thermo Fisher Scientific) with 10% FBS (Hyclone Laboratories), 'light' Leucine and either 'light' or 'heavy' Lysine ($^{13}C_8$) and Arginine ($^{13}C_{10}$) (Silantes) to generate two different isotopically labeled cell populations. L6 myoblasts were cultured for at least five passages to allow sufficient SILAC amino acid incorporation (i.e., > 98%). SILAC labels were switched between the basal and AICAR and/or insulin-treated groups in two out of four biological replicates to account for any effects of isotopically labeled amino acids. SILAC-labeled L6 myoblasts at ~90% confluence and between passage number 14 and 16 were differentiated into myotubes in SILAC DMEM containing 2% FBS. L6 myotubes were treated and harvested between 6 and 8 days post-initiation of differentiation. Prior to treatments, myotubes were washed twice with PBS and twice with SILAC DMEM with 0.2% BSA (Bovogen Biologicals) prior to serum starvation in SILAC DMEM with 0.2% BSA. All cells remained in serum starvation medium for 1.5 h and either left in the basal condition or stimulated for the final 30 min with 2 mM AICAR (Toronto Research Chemicals) and/or 20 min with 100 nM bovine insulin (Sigma).

Following cell harvesting and mixing of equal protein content from light and heavy SILAC cell populations, proteins were trypsinised and fractionated using strong cation exchange chromatography and phosphopetides were enriched and analyzed by LC-MS/MS as described previously (Hoffman et al., 2015). Raw MS data were processed using MaxQuant (Cox and Mann, 2008) (version 1.5) by searching with the following variable modifications: methionine oxidation; and serine, threonine and tyrosine phosphorylation. First search and main search peptide tolerances were set to 20 ppm and 4.5 ppm, respectively, in MaxQuant (default settings) and product-ion mass tolerance set to 0.02 Da. An FDR cutoff of 0.01 was used at the peptide level for selecting high confidence peptide identifications. Phosphosites with a localization score of 0.75 or higher were retained for analysis.

METHOD DETAILS

Phosphosite Filtering and Imputation

MS-based phosphoproteomic data commonly contain a large amount of missing values due to biological and technical reasons. PhosR implements a collection of data filtering and imputation methods for dealing with missing values in a phosphoproteomic dataset. For filtering, PhosR allows users to specify an overall quantification rate (i.e., the percentage of quantification) of a phosphosite across all biological replicates of all conditions (or time points in a time-course experiment) from which phosphosites with lower quantification rate would be removed from further analysis. While this overall quantification rate filtering is straightforward to implement, more flexible filtering procedures are needed in many scenarios. One common scenario is that a phosphosite is only phosphorylated in a specific condition (or treatment) but not in other conditions. Let us denote the quantification rate of a phosphosite in biological replicates of a condition as q^t $(t = 1 \ldots T)$ where T is the number of conditions (or time points in the case of time-course data). PhosR implements the function *selectGrps* which allows phosphosites with a q^t value equal to or greater than a predefined threshold in one or more conditions to be retained. A schematic example is shown in Figure S4A. In addition, for time-course data, PhosR implements the function *selectTimes* which allows phosphosites with a q^t value equal or greater than a predefined threshold in two or more consecutive time points to be retained.

For data imputation, PhosR implements multiple methods to take advantage of data structure and experimental design. These include site- and condition-specific imputation (*scImpute*) where, in a condition, the missing values of a phosphosite with a q^t value equal or greater than a predefined threshold will be imputed by sampling from the empirical normal distribution constructed from the quantified values of that phosphosite in that condition (Figure S4B); tail-based imputation (*tImpute*), similar to those described in Tyanova et al. (2016), where the missing values were imputed from the tail of the empirical normal distribution with a default setting of $\mathcal{N}(\mu - \sigma \times 1.6,\ \sigma \times 0.6)$ constructed from the quantified values across all sites in a sample (Figure S4C); and paired tail-based imputation (*ptImpute*) where for a phosphosite that have missing values in all replicates in a condition (e.g., 'basal') and a q^t value equal or greater than a predefined threshold in another condition (e.g., 'stimulation'), the tail-based imputation is applied to impute for the missing values in the first condition (Figure S4D).

For comparison, we also applied generic imputation methods including background imputation and k-nearest neighbor imputation. For background imputation, all missing values were replaced with the lowest detected value in a phosphoproteomics data (Välikangas et al., 2018). For k-nearest neighbor imputation, the algorithm identifies k (default of 10) most similar phosphosites for a site that contains missing value(s) (Troyanskaya et al., 2001).

Identification of Stably Phosphorylated Sites and Data Normalization

To identify a set of stably phosphorylated sites (SPSs) for subsequent data normalization and batch effect correction, we utilized four independent datasets including 'ESC differentiation', 'Adipocyte FGF21', 'Adipocyte insulin, LY, and MK' and 'Hepa 1-6 and FL83B insulin' (see **Experimental Data**). Specifically, these include the phosphoproteome data from the time-course of mouse embryonic stem cell differentiation (Yang et al., 2019b) ('ESC differentiation'), phosphoproteomic data of control and FGF21 treated mouse 3T3-L1 adipocytes (Minard et al., 2016) ('Adipocyte FGF21'), phosphoproteomic data of FL83B and Hep 1-6 cells ('FL83B & Hep 1-6 insulin'; processed as described in the previous section), and phosphoproteomes of control and insulin stimulated, and kinase inhibitor treated mouse 3T3-L1 adipocytes (Humphrey et al., 2013) ('Adipocyte insulin, LY, & MK'). We selected phosphosites that were identified in all four datasets and then applied multiple steps to rank the selected phosphosites. For each dataset, let us denote the log2 quantification of a selected phosphosite compared to a control condition as $s_i^t\ (t = 1 \ldots T)$ where T is the number of conditions or time points in that dataset. We first calculated the rank of each phosphosite by $\max(\mathrm{abs}(s_i^t))$ in each dataset. This captures the maximum magnitude of changes, either up- or downregulation, of each phosphosite in each of the four datasets. We next converted the ranks of phosphosites in each dataset into z-scores from which we calculated the *p-values* $p_i\ (i = 1,\ 2,\ 3,\ 4)$ for each phosphosite in each of the four datasets. The *p-value*s of each phosphosite were then integrated into a single combined *p-value* using Fisher's methods:

$$p_{combined} = p\left(\chi_d^2 > -2\sum_{i=1}^{4}\log(p_i)\right)$$

The $p_{combined}$ was used to generate the final consensus ranking of phosphosites identified in all four dataset and the top-100 sites that show the overall minimum phosphorylation level changes were selected as SPSs.

To perform data normalization and batch effect correction, we implemented a wrapper function *RUVphospho* which makes use of SPSs identified above as 'negative controls' in the RUV method using the version RUV-III (Molania et al., 2019). When the input data contains missing values, tail-based imputation will be applied to impute for the missing values since RUV-III requires a complete data matrix (Figure S4E). The imputed values are removed by default after normalization but can be retained for downstream analysis. For comparison, data scaling, where each sample was first centered and then divided by its standard deviation, and quantile normalization (Bolstad et al., 2003) were also applied for normalizing and correcting batch effects.

Protein- and Phosphosite-centric Enrichment Analyses

To enable enrichment analyses on both gene and phosphosite levels, PhosR implements a simple method called *phosCollapse* which reduces phosphosite level of information to the protein level by selecting the sites with either the maximum (by default) or minimum $\mathrm{abs}(s_i^t)\ (t = 1 \ldots T)$ values as the representative of phosphorylation changes of their respective proteins. Phosphosite-centric analyses are performed using kinase-substrate annotation information from PhosphoSitePlus and protein-centric analyses are performed using Reactome and KEGG databases while other pathway annotation databases such as Gene Ontology can also be used as well. For testing enrichment, PhosR implements two typical methods including over-representation test (using Fisher's Exact test) and rank-based gene set test (using Wilcoxon rank-sum test), and together refer to as 1-dimensional enrichment analyses. PhosR also provide a single interface to unify several methods developed previously for analyzing multiple experimental conditions simultaneously (refered to as 2- and 3-dimensional enrichment analyses) (Yang et al., 2014, 2016).

Kinase-substrate Prioritisation of Phosphosites

To identify potential kinases that could be responsible for the phosphorylation change of a phosphorylation site, we implement a multi-step framework that contains two major components including (i) a *kinaseSubstrateScore* function which scores a given phosphosite using kinase recognition motif and phosphoproteomic dynamics, and (ii) a *kinaseSubstratePred* function which synthesize the scores generated from (i) for predicting kinase-substrate relationships using an adaptive sampling-based positive-unlabelled learning method (Yang et al., 2019a). The kinase-substrate scoring function combines both kinase recognition motif (i.e., motif matching score) and experimental perturbation (i.e., profile matching score) for prioritising kinases that may be regulating the phosphorylation level of each site quantified in the dataset. To calculate the motif matching score for each kinase, all kinases and their substrate peptide sequences from PhosphoSitePlus database were used to compile position-specific scoring matrices (PSSMs) as follows:

$$P_j^k = \frac{1}{N_M^k}\sum_{N_M^k} I(x_j = a)$$

where $k\ (k = 1 \ldots K_M)$ is the index of kinases, N_M^k is the number of substrate sequences included for calculating the PSSM for the kth kinase, j is the index to a position in sequence x (with a window size of 13 surrounding the sites of phosphorylation), and a is the set of

characters corresponding to the 22 amino acids. Then, a motif matching score is calculated for each of all phosphorylation sites s_i by scoring their surrounding amino acid sequences x_i against each of all PSSMs for quantifying the phosphorylation preference of kinases to each phosphosite:

$$M^k_{s_i} = \sum x_{i,j} \times P^k_j$$

For calculating profile matching score, the phospho-quantification of each site in the phosphoproteomic data is first z-score transformed. Then, for each of all kinases, PhosR searches in the phosphoproteomic data for any known substrates of each of all kinases. For each kinase that have one or more known substrates quantified in the phosphoproteomic data (N^k_D), the z-score transformed dynamic phosphorylation profiles of its known substrates are median averaged (denoted as d_k $(k = 1 \ldots K_D)$, where K_d is the total number of kinases that have a quantified substrate profile). Next, the profile matching scores of each phosphosite quantified in the dataset are calculated by using Pearson's correlation with respect to the averaged profiles of known substrates of each of all kinases:

$$D^k_{s_i} = \frac{\sum \left(s_i - \overline{s_i}\right)\left(d_k - \overline{d_k}\right)}{\sqrt{\sum \left(s_i - \overline{s_i}\right)^2 \sum \left(d_k - \overline{d_k}\right)^2}}$$

The final combined score of a phosphosite s_i with respect to a kinase k is the weighted average of the motif matching score and profile matching score by taking into the number of sequences and substrates used for calculating the motif and profile of the kinase (Figure S3C). Specifically, the weights for the two parts of a kinase are calculated as $w^k_M = \log2(r(N^k_M) + 1))$ and $w^k_D = \log2(r(N^k_D) + 1))$ and the combined score is calculated as:

$$C^k_{s_i} = \left(\frac{w^k_M}{w^k_M + w^k_D}\right) M^k_{s_i} + \left(\frac{w^k_D}{w^k_M + w^k_D}\right) D^k_{s_i}$$

While the combined score calculated above takes into account both motif and phosphorylation profile of a phosphosite in prioritising kinases that may be responsible for their phosphorylation changes, these scores for each kinase are calculated independently from each other. To maximize the information in determining kinase-substrate relationships, PhosR implements a machine learning method where the combined scores across all kinases are used as learning features to predict for kinases $P(k \mid s_i,\ C^1_{s_i} \ldots C^K_{s_i})$ for a given phosphosite s_i. One of the key issues in training machine learning models for predicting kinase substrates is the need to curate a set of training examples for each kinase. This is difficult for most kinases because the numbers of known substrates are prohibitively small for training predictive models. To this end, we implemented in PhosR the AdaSampling-based positive-unlabelled ensemble of support vector machines (SVMs) as described previously (Yang et al., 2019a). For each kinase the top 30 highly ranked phosphosites (based on the combined scores) are used initially as positive examples for training SVMs for predicting substrates of that kinase and the AdaSampling procedure is used to subsequently update the training examples based on the model confidence on each phosphosite.

Signalome Construction

To construct signalomes wherein kinase regulation of protein modules can be identified, we developed an approach where we make direct use of the kinase-substrate prioritisation scores from the functions *kinaseSubstrateScore* and *kinaseSubstratePred*.

A similarity matrix of phosphosites is generated from the combined score from the *kinaseSubstrateScore* function by using Pearson's correlation as the similarity metric. The resulting matrix provides a correlation of the kinase-substrate scoring of phosphosites against all other phosphosites. The similarity matrix is then used to hierarchically cluster the phosphosites into groups with distinct profiles. Because the kinase-substrate scoring is a combined score of both kinase recognition motif (i.e., motif matching score) and experimental perturbation (i.e., profile matching score) for a phosphosite against all kinases, the phosphosites are partitioned into clusters on the basis of all these components, while taking into account the global relationships between kinases. The total number of phosphosite clusters is determined as the number of clusters wherein the mean correlation is equal to or above 0.5 for all clusters. When there are multiple scenarios where all clusters have an average correlation equal to or above 0.5, the set of clusters with the highest average correlation is chosen.

Given that many proteins are found to have multiple differentially regulated phosphosites, many of which were predicted to be regulated by different kinases, we devised a method to evaluate phosphoproteomic data whereby the regulation of multiple phosphosites can be analyzed at the protein-level (therefore allowing both a protein- *and* site-centric analysis). To this end, we constructed a phosphosite co-assignment matrix based on the phosphosite clusters and the proteins they reside on. The co-assignment matrix essentially provides a way to assign phosphosites of each protein across the clusters, generating a profile of assignment. As proteins will show different profiles in terms of their overall phosphosite membership across the clusters, we are able to create multiple combinations of protein assignment. The assignment is a binary score, meaning that the frequency with which a protein is assigned is not considered, ensuring that the co-assignment matrix does not bias toward proteins with many phosphosites. The final co-assignments, herein referred to as "protein modules," consist of exclusive sets of proteins with similar phosphorylation profiles and kinase regulation.

The *Signalomes* function generates a visualization of the signalomes present in the phosphoproteomic data. For the visualization of signalomes, it does so by using the protein modules identified from above and the kinase-substrate predictions from the *kinaseSubstratePred* function. A cut-off of 0.5 is used as default (*signalomeCutoff* = 0.5) to capture kinase-substrate relationships (Figure S3B). Then an adjacency matrix depicting the regulation of proteins by kinases is used to generate a chord diagram from the *circlize* package. This method of visualization provides a summary of the kinase regulation of each protein module. The *Signalomes* function also outputs signalomes associated to any kinase of interest (referred to as extended signalome of a kinase). To facilitate assessment of proteins and phosphosites that are under similar regulation, the extended signalome of a kinase combines cognate signalomes from other kinases that share a high degree of similarity in substrate regulation.

Additional Functions

PhosR also provide a set of helper functions that enable various tasks related to phosphoproteomics data processing and analysis. These include (but not limited to) data standardization, centering and scaling normalization, calculation of amino acid position-specific frequency matrix, ANOVA analysis, and filtering for phosphosite localization probability. Details of these functions are provided in the PhosR Bioconductor package (https://pyanglab.github.io/PhosR/) and the associated vignette.

QUANTIFICATION AND STATISTICAL ANALYSIS

Differentially Phosphorylated Phosphosites

Differentially phosphorylated sites were identified using the two-sided moderated *t*-test implemented in the *limma* R package (Ritchie et al., 2015). Analyses were done on log (base 2)-transformed data and *p-value*s were adjusted for multiple testing using Benjamini-Hochberg FDR correction at $\alpha = 0.05$.

KEGG Pathway Over-representation Analysis

Pathway over-representation analysis was performed on protein sets identified from our kinase-substrate scoring analysis (kinase-substrate pairs with score > 0.5 were selected) using the over-representation analysis implemented in the *clusterProfiler* R package (Yu et al., 2012). The KEGG pathway database was used and *p-value*s were adjusted for multiple testing using Benjamini-Hochberg FDR correction at $\alpha = 0.05$.