Published in final edited form as:

Proteomics. 2016 July ; 16(13): 1868–1871. doi:10.1002/pmic.201600068.

KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis

Pengyi Yang^{1,2,3}, Ellis Patrick⁴, Sean J. Humphrey^{2,5}, Shila Ghazanfar¹, David E. James², Raja Jothi³, and Jean Yee Hwa Yang¹

¹School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia

²Charles Perkins Centre, School of Molecular Biosciences, University of Sydney, Sydney, NSW, Australia

³Systems Biology Section, Epigenetics & Stem Cell Biology Laboratory, National Institute of Environmental, Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

⁴Brigham and Women's Hospital, Harvard Medical School, Broad Institute, Boston, MA, USA

⁵Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, Germany

Abstract

Mass spectrometry (MS)-based quantitative phosphoproteomics has become a key approach for proteome-wide profiling of phosphorylation in tissues and cells. Traditional experimental design often compares a single treatment with a control, whereas increasingly more experiments are designed to compare multiple treatments with respect to a control. To this end, the development of bioinformatic tools that can integrate multiple treatments and visualise kinases and substrates under combinatorial perturbations is vital for dissecting concordant and/or independent effects of each treatment. Here, we propose a hypothesis driven kinase perturbation analysis (KinasePA) to annotate and visualise kinases and their substrates that are perturbed by various combinatorial effects of treatments in phosphoproteomics experiments. We demonstrate the utility of KinasePA through its application to two large-scale phosphoproteomics datasets and show its effectiveness in dissecting kinases and substrates within signalling pathways driven by unique combinations of cellular stimuli and inhibitors. We implemented and incorporated KinasePA as part of the “directPA” R package available from the comprehensive R archive network (CRAN). Furthermore, KinasePA also has an interactive web interface that can be readily applied to annotate user provided phosphoproteomics data (<http://kinasepa.pengyiyang.org>).

Keywords

Bioinformatics; Hypothesis testing; Kinase; Perturbation; Phosphoproteomics; Signalling

Correspondence: Professor Jean Yee Hwa Yang School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia, jean.yang@sydney.edu.au, **Fax:** +61-93514534.

The authors have declared no conflict of interest.

Protein phosphorylation is a post-translational modification that plays a critical role in signal transduction. The most widely used approach for proteome-wide phosphorylation profiling relies on the use of high-throughput tandem mass spectrometry (LC-MS/MS). Coupled with isotopic/isobaric labelling or increasingly label-free quantitation, thousands of phosphopeptides can be accurately identified and quantified [1].

A key goal in the majority of phosphoproteomics studies is to identify kinases that are the key nodes in controlling signalling cascades from perturbation experiments [2]. Traditional experimental design consists of measuring phosphorylation changes between a single treatment such as a stimuli or an inhibitor and the corresponding basal condition. Here, methodologies developed for gene expression analysis can be adapted and utilised to identify phosphorylation sites that are perturbed by the treatment. Analogous to gene expression studies, it is often hard to interpret a list of perturbed phosphorylation sites, and pathway analyses are often applied. This enables scientists to determine differentially regulated kinases based on the coordinated change of their substrates (i.e. targeted phosphorylation sites) in treatment compared to control [3, 4].

Increasingly, phosphoproteomics studies comprising multiple treatments are conducted, for example, to compare more than two individual inhibitors with the basal conditions to discover concordant and/or independent effects of treatments [5–7]. In such studies, it is nontrivial to directly apply methods from pathway analysis to identify kinases that are perturbed by a combination of treatments. To this end, we propose to adapt a directional hypothesis testing framework that we developed for pathway analysis [8] to annotate and visualise kinases and their substrates that are perturbed by multiple treatments. We have extended the original approach presented in [8] by introducing a kinase perturbation plot where the degree of kinase activity regulated by the experimental treatments are visualised using Stouffer's statistics. We implemented the proposed approach as an interactive web tool, called “kinase perturbation analysis” or KinasePA. This web tool is intuitive and can be easily applied to annotate user provided phosphoproteomics data. We also implemented and incorporated KinasePA as part of the “directPA” R package. This provides more flexibility to advanced users.

Firstly, we extracted kinase-substrate relationships from public databases [9, 10] to annotate experimentally identified phosphorylation sites for mouse and human. Briefly, a phosphorylation site i ($i = 1, 2, \dots, n$) is said to belong to a kinase κ if the annotation is found in a database as κ phosphorylates i . In this way, a set of phosphorylation sites is assigned to a kinase κ as its substrate set based on the database annotation. The procedure is repeated for all kinases annotated in the database for a given organism (e.g. *Mus musculus*, *Homo sapiens*). Then, for a given phosphoproteomics dataset the kinase-substrate pairs are filtered to select those that are quantified in the phosphoproteomics experiments. This allows only experimentally relevant kinases and substrates to be considered in the enrichment analysis.

We then adapted the directional hypothesis testing framework we designed for pathway analysis [8] for annotating phosphoproteomics data. Let t_{i1} and t_{i2} denote the observed test statistics for the i phosphorylation site in the comparison of two treatments and controls,

respectively. Briefly, we test if any of the phosphorylation sites within a kinase κ is differentially regulated in the following steps:

Step 1. Convert the observed test statistics into z -scores where $z_{ij} = -\Phi^{-1}(P(t_{ij}))$, ($j = 1, 2$) and $\Phi(\cdot)$ is the cumulative distribution.

Step 2. Based on spherical coordinates in a 2D Euclidean space, the test statistics from the original direction can be aligned toward the direction of interest δ by calculating θ' where $\theta' = \theta + \delta$. That is, for the i^{th} phosphorylation site, find the angle θ between the x -axis and the vector from

the origin to (z_{i1}, z_{i2}) by calculating $\tan^{-1}\left(\frac{z_{i1}}{z_{i2}}\right)$ and rotate z -scores as $z'_{i1} = \sqrt{z_{i1}^2 + z_{i2}^2} \times \cos(\theta')$ and $z'_{i2} = \sqrt{z_{i1}^2 + z_{i2}^2} \times \sin(\theta')$, respectively.

Step 3. Combine the rotated z -scores across treatments for each phosphorylation site using a one-sided Pearson's method

$$p_i^S = P\left(\chi_4^2 < -2 \times \sum_{j=1,2} \ln\left(1 - \Phi(-z'_{ij})\right)\right)$$

Step 4. Combine the p -values of phosphorylation sites that belong to a kinase κ

$$\text{as } p^\kappa = \Phi\left(\frac{\sum_{i=1}^n \Phi^{-1}(\mathbf{p}_i^S) \mathbf{I}\{i \in \kappa\}}{\sqrt{\sum_{i=1}^n \mathbf{I}\{i \in \kappa\}}}\right).$$

To visualise the combinatorial effect of treatments, we introduce a kinase perturbation plot where for each experimental treatment j ($j = 1, 2$) we calculate Stouffer's statistics

$$z_j^\kappa = \frac{\sum_{i=1}^n z_{ij} \mathbf{I}\{i \in \kappa\}}{\sqrt{\sum_{i=1}^n \mathbf{I}\{i \in \kappa\}}}. \text{ These statistics calculated for each treatment are plotted against each other to show the degree of kinase activity regulated by the experimental treatments.}$$

To allow broad accessibility for bioinformaticians and non-bioinformaticians alike, we implemented and incorporated KinasePA as part of the “directPA” R package and also implemented it as a “Shiny” application [11]. While the R package provides more flexibility to experienced R users, the Shiny application allows the users to upload their own phosphoproteomics data in the form of a csv file and interactively visualise and query the data. Human and mouse annotation databases extracted from PhosphositePlus are preloaded into the application. The interactive graph of the kinases allows the interrogation of phosphorylation sites that are driving the signal in kinase perturbation. The Shiny application is available from: <http://kinasepa.pengyiyang.org>

We used two large-scale phosphoproteomics datasets to illustrate KinasePA. The first dataset was generated from phosphoproteomics profiling of insulin signalling pathways upon MK2206 and LY294002 inhibitions, respectively, in insulin stimulated adipocytes [5]. The second dataset was generated from a screen of mTOR substrates in HEK-293E cells using insulin stimulation and inhibition by Torin1 or Rapamycin [6].

Several interesting hypotheses can be tested from these two datasets using KinasePA, including which kinases are inhibited by the treatment of both inhibitors and which kinases

are more potently inhibited by the treatment of an individual inhibitor. Specifically, both Akt1 and mTOR are inhibited by the treatment of MK2206 and LY294002 (Fig. 1A) in adipocytes. This is consistent with the current knowledge that both LY294002 and MK2206 target the canonical Akt/mTOR pathway [5]. Interestingly, Akt1 is more potently inhibited by MK2206 (Fig. 1B) whereas mTOR is more potently inhibited by LY294002 (Fig. 1C). This suggests that while MK2206 specifically targets Akt1 and its substrates, LY294002 preferentially targets mTOR and its substrates. The independent effects of these two kinases are further illustrated in Fig. 2A using kinase perturbation plot. Specifically, mTOR falls into the upper region of the diagonal line whereas Akt1 falls into the lower region, indicating a stronger inhibition of LY294002 on mTOR and a stronger inhibition of MK2206 on Akt1. These combinatorial observations become more transparent with KinasePA whereas approaches based on pathway analysis for testing up, down, or mixed regulations are unable to elucidate the combinatorial relationships between Akt1 and mTOR.

In the second application to HEK-293E dataset, the results from KinasePA suggest that both Torin1 and Rapamycin inhibit mTOR (Fig. 1D). However, Torin1 is more potent in inhibiting mTOR compared to rapamycin, as evidenced by the statistical significance of mTOR in Fig. 1F but not in Fig. 1E. A similar conclusion could also be drawn from the kinase perturbation plot (Fig. 2B) where mTOR shows significantly stronger down-regulation on the x -axis (i.e. Torin1/Insulin) than on the y -axis (i.e. Rapamycin/Insulin). It has been reported that Rapamycin only affects a subset of mTORC1 (mTOR complex 1) substrates, and has no effect on the activity of mTORC2 (mTOR complex 2) whereas Torin1 inhibits both mTOR complexes [6]. Using KinasePA, we clearly dissect such combinatorial relationships of Rapamycin and Torin1 in inhibiting mTOR signalling pathways.

Together, these applications demonstrate that KinasePA can be used to investigate coordinated effects and independent effects of inhibitors on kinases and their substrates using phosphoproteomics data. Direct adaptation of conventional pathway analysis cannot disentangle these combinatorial effects. The integration of multiple experimental treatments is the unique strength in KinasePA for dissecting such combinatorial effects. KinasePA is flexible and can be applied to analyse gene and protein networks using databases such as KEGG and GO. It allows complex biological questions to be formulated seamlessly into hypotheses and tested in an intuitive manner. Kinases identified to be perturbed by a combination of treatments can be tested by targeted knockdown and/or gene mutation experiments. This will help biologists to establish causality between signalling cascades and the observed phenotypes. The implementation of KinasePA as an interactive web tool makes the proposed approach easily accessible to all scientists in general.

References

- [1]. Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* 2010; 11:427–439. [PubMed: 20461098]
- [2]. Hoffman NJ, Parker BL, Chaudhuri R, Fisher-Wellman KH, Kleinert M, Humphrey SJ, Yang P, et al. Global phosphoproteomic analysis of human skeletal muscle reveals a network of exercise-regulated kinases and AMPK substrates. *Cell. Metab.* 2015; 5:922–935. [PubMed: 26437602]
- [3]. Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. *Bioinformatics.* 2009; 25:684–686. [PubMed: 19176546]

- [4]. Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, et al. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* 2013; 6:rs6–rs6. [PubMed: 23532336]
- [5]. Humphrey SJ, Yang G, Yang P, Fazakerley DJ, et al. Dynamic adipocyte phosphoproteome reveals that akt directly regulates mTORC2. *Cell. Metab.* 2013; 17:1009–1020. [PubMed: 23684622]
- [6]. Hsu PP, Kang SA, Rameseder J, Zhang Y, et al. The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling. *Science.* 2011; 332:1317–1322. [PubMed: 21659604]
- [7]. Kirkpatrick DS, Bustos DJ, Dogan T, Chan J, et al. Phosphoproteomic characterization of DNA damage response in melanoma cells following MEK/PI3K dual inhibition. *Proc. Natl. Acad. Sci. U. S. A.* 2013; 110:19426–19431. [PubMed: 24218548]
- [8]. Yang P, Patrick E, Tan S-X, Fazakerley DJ, et al. Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. *Bioinformatics.* 2014; 30:808–14. [PubMed: 24167158]
- [9]. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, 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–D270. [PubMed: 22135298]
- [10]. Dinkel H, Chica C, Via A, Gould CM, et al. Phospho.ELM: a database of phosphorylation sites-update 2011. *Nucleic Acids Res.* 2011; 39:261–267.
- [11]. Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. Shiny: web application framework for R. 2015. <http://CRAN.R-project.org/package=shiny>. R package version 0.11

Figure 1. Scatter plots of (A, B, and C) the adipocyte dataset and (D, E, and F) the HEK-293E dataset on three tested directions with phosphorylation sites ranked from enriched in red to depleted in purple based on p_i^S ($i=1, 2, \dots, n$). The arrow indicates the tested direction. Known substrates of Akt1 and mTOR are highlighted in (A, B, and C) and known mTOR substrates are highlighted in (D, E, and F). The tables beneath each panel are the top-3 enriched kinases corresponding to the tested directions in each dataset.

Figure 2.

Kinase perturbation plot. Kinases are plotted based on z_j^K ($j=1, 2$), the integrated z-scores calculated from the coordinated change of their substrates in (A) LY294002 and MK2206 compared to insulin in Adipocytes dataset, and (B) Torin1 and Rapamycin compared to insulin in HEK-293E dataset. The larger the distance to the origin the more perturbed a given kinase. The quadrant and the crosshairs provide the perspective of the perturbation to each treatment.