Supporting Information

Identifications of putative PKA substrates with quantitative phosphoproteomics and primary-sequence based scoring.

Haruna Imamura[1, 2], Omar Wagih[2], Tomoya Niinae[1], Naoyuki Sugiyama[1], Pedro Beltrao[2] and Yasushi Ishihama[1*]

[1]*Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan*
[2]*European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD*

Contents

Supplementary Methods

Experimental materials

DMEM, kanamycin, phosphate-buffered saline (PBS), DMSO, HEPES, sucrose, $MgCl_2$, KCl, Ammonium bicarbonate, sodium deoxycholate (SDC), sodium N-lauroylsarcosinate (SLS), Tris, DTT, iodoacetamide (IAA), and Lys-C were obtained from Wako (Osaka, Japan). Forskolin, H-89, protease inhibitor and phosphatase inhibitor were obtained from Sigma-Aldrich (St Louis, MO). Fetal bovine serum (FBS) was obtained from Gibco, and BCA protein assay kit was obtained from Thermo Fisher Scientific (Waltham, MA). Trypsin was obtained from Promega (Madison, WI).

Cell culture

HeLa S3 cells were cultured in DMEM containing 10% FBS and 100 µg/ml kanamycin. For SILAC labelling, cells were grown in the presence of either [$^{12}C_6$, $^{14}N_2$]-Lys and [$^{12}C_6$, $^{14}N_4$]-Arg (light labelling) or [$^{13}C_6$, $^{15}N_2$]-Lys and [$^{13}C_6$, $^{15}N_4$]-Arg (heavy labeling). The amount of each added Arg and Lys used for supplementation followed the defined concentrations in DMEM.

Drug stimulation to the cells

Forskolin and H89 in DMSO was diluted with culture medium to a final concentration of 50 µM and 100 µM, respectively. For controls, the same concentration of DMSO (0.01%) was added to drug-treated cells. Drug treatment was performed for 60 min in 37°C. After the treatments, cells were washed with ice-cold PBS, harvested on ice with ice-cold PBS, and frozen at -80°C until use.

Sample preparation for LC-MS/MS analysis

For quantitation with LC-MS/MS, activator experiments were done with SILAC labelling while TMT labelling was applied for the inhibitor experiments.

For activator experiments, the equal amount of the cell pellets from the drug-

treated and control samples (either light/heavy labelled) were mixed at this stage. The proteins were extracted by following the method described previously: for activator experiments, 'supernatant fraction' in [1] with slight modification on lysis buffer (20 mM HEPES (pH 7.5), 250 mM sucrose, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5% Nonidet P-40, containing 1% protease inhibitor and phosphatase inhibitor); for inhibitor experiments, the phase transfer surfactant protocol [2] with modified lysis buffer (12 mM SDC, 12 mM SLS, 100 mM Tris-HCl (pH 9.0), containing 1% protease inhibitor and phosphatase inhibitor).

Protein amount was measured with BCA protein assay kit. After protein reduction/alkylation with DTT/IAA, Lys-C/trypsin digestion was performed as described previously [3]. Then, the peptides were desalted with StageTip [4]. For inhibitor experiments, the equal amount of the peptides from either drug-treated or control samples of three biological replicates was independently labelled with 6-plex TMT reagents by following the previously described method [5], and they were mixed into one sample.

Phosphopeptides were enriched by HAMMOC [6], and desalted using StageTip for subsequent LC-MS/MS analyses.

LC-MS/MS analysis

For activator experiments, a self-pulled analytical column (250 mm length x 100 µm i.d.) was prepared with ReproSil-Pur C18-AQ materials (3 µm, Dr. Maisch, Ammerbuch, Germany). The injection volume was set to 5 µL and the flow rate was set to 500 nL/min. The mobile phases consisted of (A) 0.5% acetic acid and (B) 0.5% acetic acid in 80% acetonitrile. A gradient condition was employed, i.e., 5-40% B in 180 min, 40–100% B in 5 min, and 5% B for 30 min. NanoLC-MS/MS analyses were conducted using a TripleTOF 5600 System (AB Sciex, Foster City, CA) equipped with an Ultimate 3000 pump (Thermo Fisher Scientific) and a HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland). The analysis was repeated twice for each of two biological replicates, switching the heavy and light labelling for either sample.

For inhibitor experiments, monolithic silica columns (100 µm i.d., 2 m long) were prepared as described previously [7]. The mobile phases consisted the same with the activator experiments, but longer gradient with 5-40% B in 8 hours. LC-MS/MS analyses were conducted using a Q Exactive (Thermo Fisher Scientific) equipped with the same LC and autosampler system described above. The analysis was repeated seven times

for a sample, which contained three biological replicates.

The MS raw data and analysis files have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the jPOST partner repository (http://jpostdb.org) with the data set identifier PXD005922 & PXD005925 [8].

Identification and quantitation

Peak lists were created from the raw data files based on the recorded fragmentation spectra.

Peptides from activator experiment (SILAC labelled samples with TripleTOF 5600 system) were identified by Mascot v. 2.4 (Matrix Sciences, London. U.K.) against human SwissProt Database (version 2013-11) with cysteine carbamidomethylation as a fixed modification, and [$^{13}C_6$, $^{15}N_2$]-Lys and [$^{13}C_6$, $^{15}N_4$]-Arg and methionine oxidation as well as phosphorylation of serine, threonine, and tyrosine as variable modifications. Precursor mass tolerance was set to 10 ppm, and fragment ion mass tolerance was 0.1 Da, allowing for up to 2 missed cleavages.

Peptides from inhibitor experiment (TMT labelled samples with Q Exactive system) were identified by Mascot v. 2.4 against human SwissProt Database (version 2016-04) with cysteine carbamidomethylation as a fixed modification, and TMT 6-plex modification on N-terminal and lysine, methionine oxidation as well as phosphorylation of serine, threonine, and tyrosine as variable modifications. Precursor mass tolerance was set to 5 ppm, and fragment ion mass tolerance was 0.02 Da, allowing for up to 2 missed cleavages.

Peptides were considered identified if the Mascot score was over the 95% confidence limit based on the ‘identity’ score of each peptide. We also used the additional criterion that at least three successive y- or b-ions with a further two or more y-, b- and/or precursor-origin neutral loss ions were observed, based on the error-tolerant peptide sequence tag concept [9]. False discovery rates (FDR) were estimated with these criteria by searching against a randomized decoy database (<1%). In addition, phosphosite localization was confirmed with a site-determining ion combination (SIDIC) method, as described before [10]. Briefly, this method is based on the presence of site-determining y- or b-ions in the peak lists of the fragment ions, which unambiguously identify the phosphosites. Note that SIDIC has been evaluated to be equivalent to other phosphosite

localization scorings (*i.e*., PTM score [11] and Mascot delta score [12])[13].

Quantitation for each peptide was calculated by integrating the peak area in MS1 scan using Mass Navigator v1.2 (Mitsui Knowledge Industry, Tokyo, Japan), and only those had QuanScore > 0.8 were accepted from SILAC labelled sample. For TMT labelled samples, the peak intensities of reporter ions were obtained from MS2 scan. The ratio of drug-treated to control were logged (base 2), and centralized to zero among each technical replicate. The mean was taken if they had identical sequences and modifications. Only the quantitation of localization-confirmed sites from singly phosphorylated peptide was accepted for the following analysis.

Dataset for informatics analysis

We collected experimentally validated human phosphosites from three online databases (PhosphoSitePlus[14], PhosphoELM[15], HPRD[16]) as previously described [17] . After excluding duplicates and sites without annotated literature reference, phosphosites sequences (defined as the site and ±7 flanking residues) were matched against the longest isoforms of CCDS proteins. Nonmatching phosphosites sequences were discarded. We limited the sites used to only those that have at least one experimentally annotated kinase. This resulted in 9,595 pairs of kinase-substrate relationships including PKA (pS=389, pT=62), ERK1 (pS=183, pT=66), and AKT1 (pS=144, pT=42).

PWM construction and scoring

The probability of observing residue *x* in position *i* is computed as follows:

$$p(x,i) = \frac{f_{x,i} + c(x)}{N},$$

$$c(x) = p(x) \cdot \epsilon$$

where $f_{\mathrm{x,i}}$ is the frequency of observing residue x at position *i* and *N* is the total number of sequences. *c(x)* is a pseudo count function which is computed as the probability of observing residue b in the proteome, *p(x)*, multiplied by ε, defined as the square root of the total number of sequences used to train the PWM. This avoids infinite values when

computing logarithms.

Probabilities are then converted to weights as follows:

$$w_{x,i} = \log_2 \frac{p(x,i)}{p(x)}$$

where *p(x)* = background probability of amino acid *x*; *p(x,i)* = corrected probability of amino acid *x* in position *i*; $W_{x,i}$ = PWM value of amino acid *x* in position *i*.

Given a sequence *q* of length *l*, a score λ is then computed by summing $\log_2$ weights:

$$\lambda = \sum_{i=1}^{l} w_{q_i,i}$$

where q_i is the i^{th} residue of *q*. In this study, the score was computed using the flanking ±7 residues surrounding each phosphosite.

Motif extraction and the fold increase score

Motifs were obtained using the rmotifx package (https://github.com/omarwagih/rmotifx) [18], which is an implementation of the motif-x algorithm [19]. Default parameters (minimum sequences ≥ 20, and p-value ≤ $1x10^{-6}$) were used. Each motif *m*, of length *l* is reported with an associated fold increase score, t_m. Given a sequence *q*, matching a set of motifs $m_1,m_2,\ldots m_n$ the fold increase score γ is computed as follows:

$$\Upsilon = max(t_{m_1}, t_{m_2}, \ldots, t_{m_n})$$

Performance

We conducted 5-fold cross validation to estimate the prediction power of PWMs, motif fold increase (FINC), and PSP scores. For a given kinase, the positive test set was

defined as all sites known to be phosphorylated by that kinase, as defined in the public databases. The negative test set consisted of all kinase-associated sites for all kinases except the one in question. Each dataset was split into 5, and the one fifth of the sequences were used as the test sites and the remaining was regarded as training sites. Using the training sites, PWM and motifs were re-generated with the procedure described above. The trained models are then scored on the test sites for evaluating performance. This is repeated five times such that each fifth segment of data is used as a test set once.

To evaluate the performance of our predictors, we generated ROC curves using the true positive rates (TPR) and false positive rates (FPR) as follows:

$$FPR = \frac{FP}{FP + TN} \qquad TPR = \frac{TP}{TP + FN}$$

All ROC analysis were carried out using the ROCR package in R [20]. The averaged AUC were calculated from the five trials.

References

[1] H. Imamura, N. Sugiyama, M. Wakabayashi, Y. Ishihama, Large-scale identification of phosphorylation sites for profiling protein kinase selectivity, J Proteome Res. 13 (2014) 3410–3419. doi:10.1021/pr500319y.

[2] T. Masuda, N. Sugiyama, M. Tomita, Y. Ishihama, Microscale Phosphoproteome Analysis of 10 000 Cells from Human Cancer Cell Lines, Anal Chem. 83 (2011) 7698–7703. doi:10.1021/ac201093g.

[3] K. Imami, N. Sugiyama, H. Imamura, M. Wakabayashi, M. Tomita, M. Taniguchi, et al., Temporal Profiling of Lapatinib-suppressed Phosphorylation Signals in EGFR/HER2 Pathways, Mol Cell Proteomics. 11 (2012) 1741–1757. doi:10.1074/mcp.M112.019919.

[4] J. Rappsilber, M. Mann, Y. Ishihama, Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips, Nat Protoc. 2 (2007) 1896–1906. doi:10.1038/nprot.2007.261.

[5] G. Böhm, P. Prefot, S. Jung, S. Selzer, V. Mitra, D. Britton, et al., Low-pH Solid-Phase Amino Labeling of Complex Peptide Digests with TMTs Improves Peptide Identification Rates for Multiplexed Global Phosphopeptide Analysis, J Proteome Res. 14 (2015) 2500–2510. doi:10.1021/acs.jproteome.5b00072.

[6] N. Sugiyama, T. Masuda, K. Shinoda, A. Nakamura, M. Tomita, Y. Ishihama, Phosphopeptide Enrichment by Aliphatic Hydroxy Acid-modified Metal Oxide Chromatography for Nano-LC-MS/MS in Proteomics Applications, Mol Cell Proteomics. 6 (2007) 1103–1109. doi:10.1074/mcp.T600060-MCP200.

[7] M. Iwasaki, S. Miwa, T. Ikegami, M. Tomita, N. Tanaka, Y. Ishihama, One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the Escherichia coli proteome on a microarray scale, Anal Chem. 82 (2010) 2616–2620. doi:10.1021/ac100343q.

[8] S. Okuda, Y. Watanabe, Y. Moriya, S. Kawano, T. Yamamoto, M. Matsumoto, T. Takami, D. Kobayashi, N. Araki, A. C. Yoshizawa, T. Tabata, N. Sugiyama, S. Goto, Y. Ishihama, jPOSTrepo: an international standard data repository for proteomes, Nucleic Acids Research 45 (2017), D1107-D1111.

[9] M. Mann, M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags, Anal Chem. 66 (1994) 4390–4399.

[10] H. Nakagami, N. Sugiyama, K. Mochida, A. Daudi, Y. Yoshida, T. Toyoda, et al., Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants, Plant Physiol. 153 (2010) 1161–1174.

doi:10.1104/pp.110.157347.

[11] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, et al., Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks, Cell. 127 (2006) 635–648. doi:10.1016/j.cell.2006.09.026.

[12] M.M. Savitski, S. Lemeer, M. Boesche, M. Lang, T. Mathieson, M. Bantscheff, et al., Confident Phosphorylation Site Localization Using the Mascot Delta Score, Mol Cell Proteomics. 10 (2011) M110.003830–M110.003830. doi:10.1074/mcp.M110.003830.

[13] M.-H. Lin, N. Sugiyama, Y. Ishihama, Systematic profiling of the bacterial phosphoproteome reveals bacterium-specific features of phosphorylation, Science Signaling. 8 (2015) rs10. doi:10.1126/scisignal.aaa3117.

[14] P.V. Hornbeck, B. Zhang, B. Murray, J.M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, Nucleic Acids Res. 43 (2015) D512–20. doi:10.1093/nar/gku1267.

[15] H. Dinkel, C. Chica, A. Via, C.M. Gould, L.J. Jensen, T.J. Gibson, et al., Phospho.ELM: a database of phosphorylation sites--update 2011, Nucleic Acids Res. 39 (2011) D261–7. doi:10.1093/nar/gkq1104.

[16] T.S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, et al., Human Protein Reference Database--2009 update, Nucleic Acids Res. 37 (2009) D767–72. doi:10.1093/nar/gkn892.

[17] O. Wagih, J. Reimand, G.D. Bader, MIMP: predicting the impact of mutations on kinase-substrate phosphorylation, Nat Methods. (2015). doi:10.1038/nmeth.3396.

[18] O. Wagih, N. Sugiyama, Y. Ishihama, P. Beltrao, Uncovering Phosphorylation-Based Specificities through Functional Interaction Networks, Mol Cell Proteomics. 15 (2016) 236–245. doi:10.1074/mcp.M115.052357.

[19] D. Schwartz, S.P. Gygi, An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets, Nat Biotechnol. 23 (2005) 1391–1398. doi:10.1038/nbt1146.

[20] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics. 21 (2005) 3940–3941. doi:10.1093/bioinformatics/bti623.

Figure S1

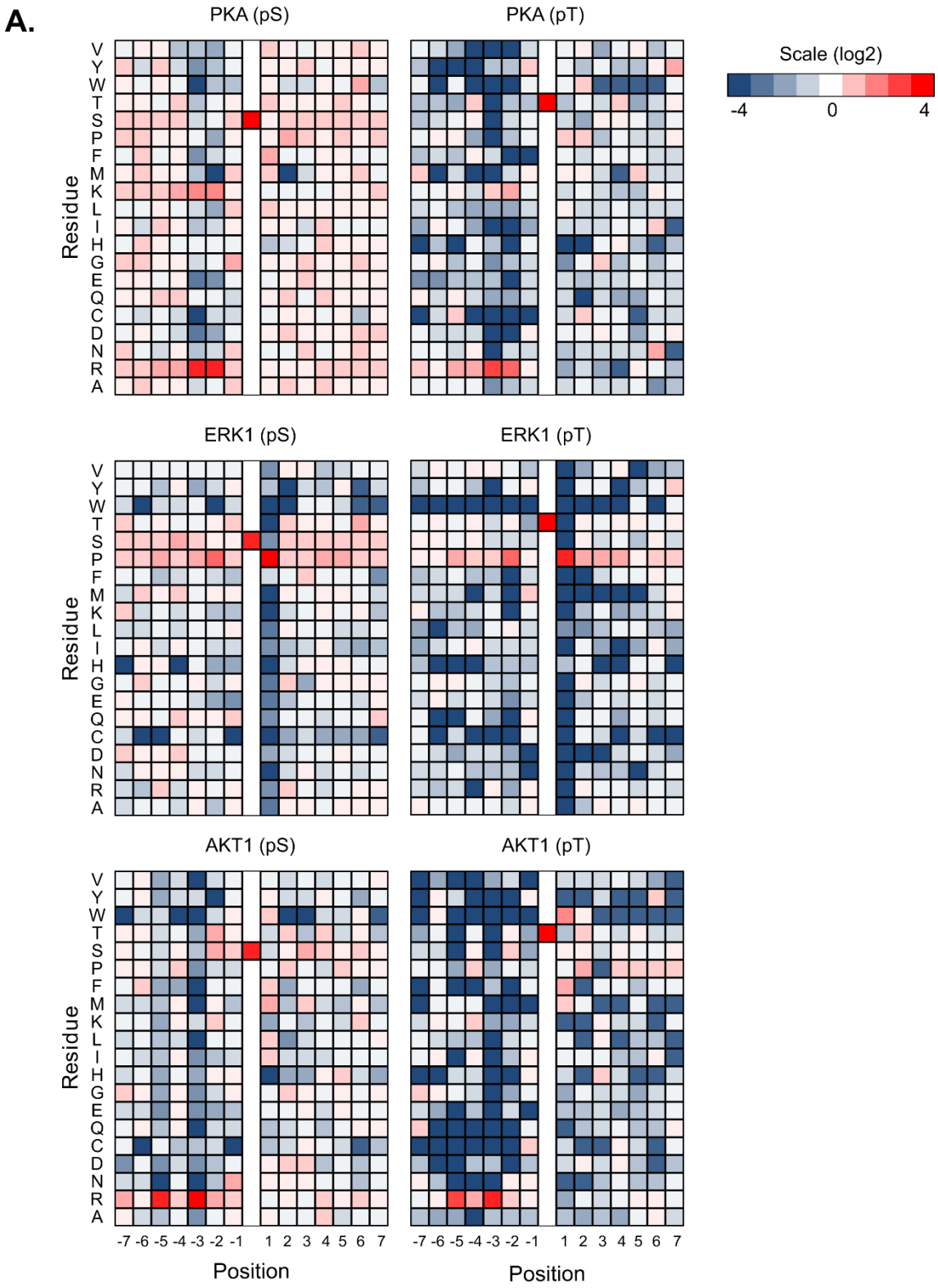

A.
PKA (pS)
PKA (pT)
Scale (log2)
-4
0
4
ERK1 (pS)
ERK1 (pT)
AKT1 (pS)
AKT1 (pT)
Residue
V Y W T S P F M K L I H G E Q C D N R A
-7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7
Position

Figure S1

B.

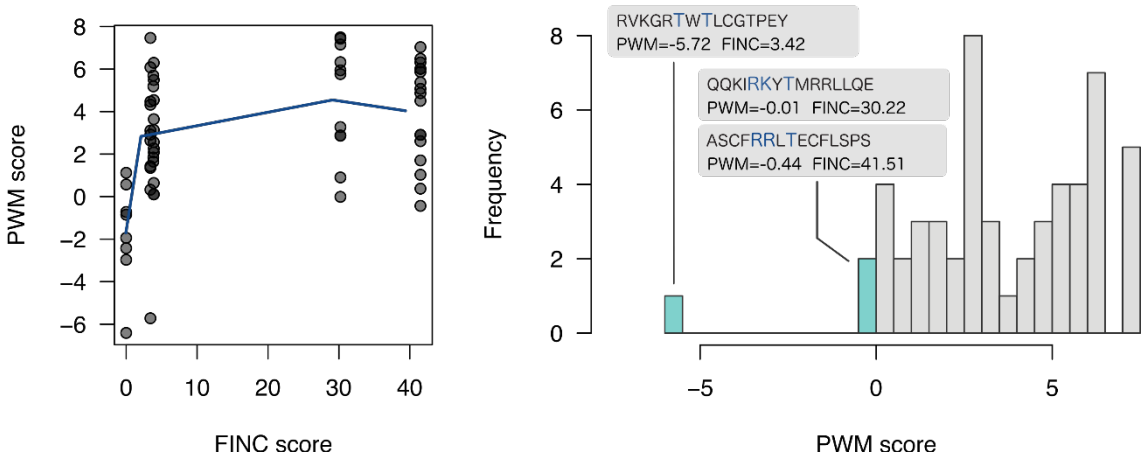
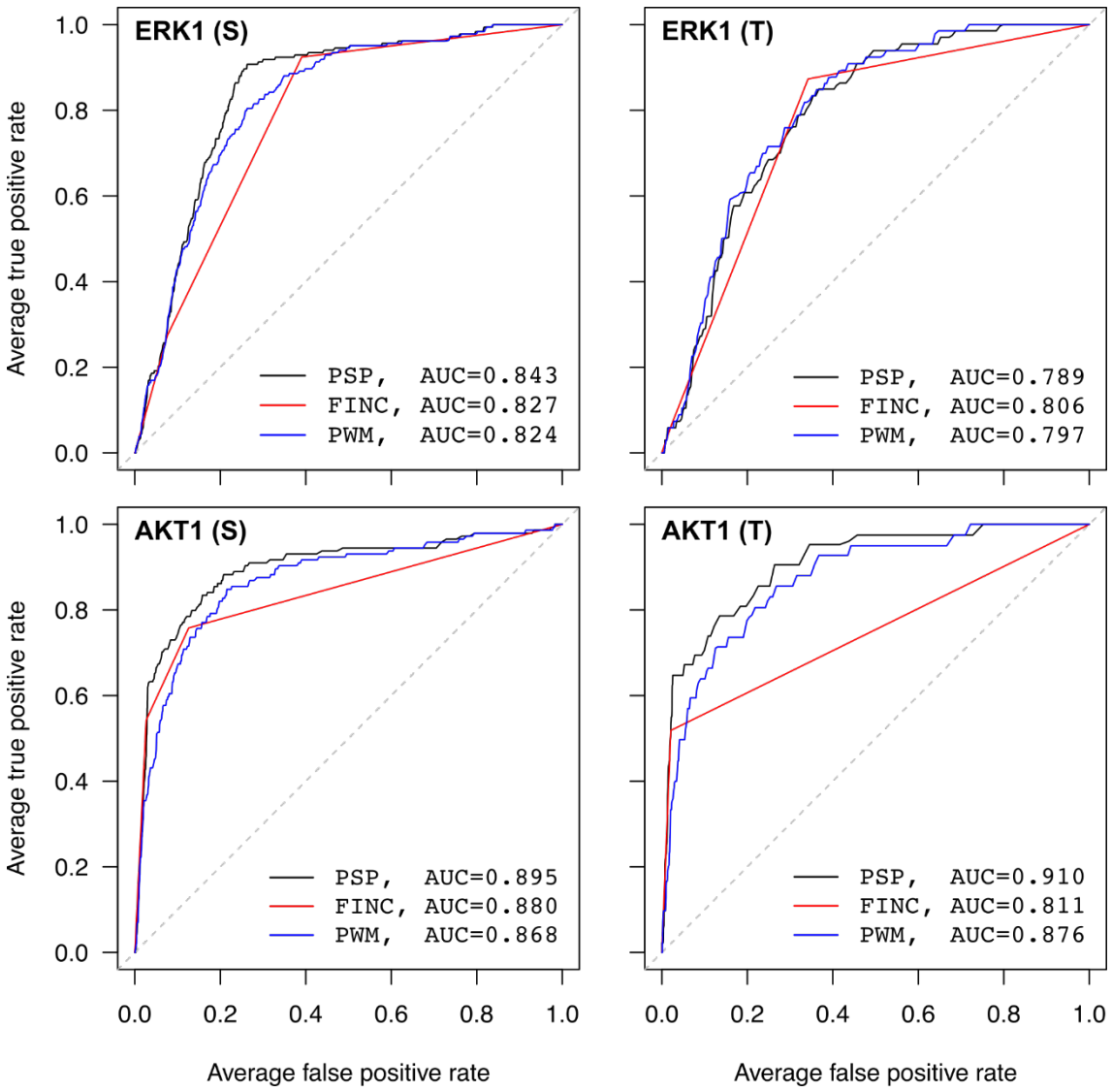

PWM score
FINC score
Frequency
PWM score
RVKGRTWTLCGTPEY
PWM=-5.72 FINC=3.42
QQKIRKYTMRRLLQE
PWM=-0.01 FINC=30.22
ASCFRRLTECFLSPS
PWM=-0.44 FINC=41.51

Figure S2


ERK1 (S)
PSP, AUC=0.843
FINC, AUC=0.827
PWM, AUC=0.824
ERK1 (T)
PSP, AUC=0.789
FINC, AUC=0.806
PWM, AUC=0.797
AKT1 (S)
PSP, AUC=0.895
FINC, AUC=0.880
PWM, AUC=0.868
AKT1 (T)
PSP, AUC=0.910
FINC, AUC=0.811
PWM, AUC=0.876
Average true positive rate
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0