Published in final edited form as:
Nat Methods. 2019 August ; 16(8): 703–706. doi:10.1038/s41592-019-0498-4.

Thesaurus: quantifying phosphopeptide positional isomers

Brian C. Searle, **Robert T. Lawrence**, **Michael J. MacCoss**, **Judit Villén**[*]
Department of Genome Sciences, University of Washington, Seattle, WA, USA

Abstract

Proteins can be phosphorylated at neighboring sites resulting in different functional states, and studying the regulation of these sites has been challenging. Here we present Thesaurus, a search engine that detects and quantifies phosphopeptide positional isomers from parallel reaction monitoring and data independent acquisition mass spectrometry experiments. We apply Thesaurus to analyze phosphorylation events in the PI3K/AKT signaling pathway and show neighboring sites with distinct regulation.

Hundreds of thousands of amino acids in thousands of proteins are estimated to be actively phosphorylated in every human cell (1). Many proteins are phosphorylated at neighboring sites (2) and over half of sites in multi-phosphorylated proteins are within four amino acids of each other (3). Several well-studied proteins make use of neighboring phosphorylation sites to act as switches (MAPK (4), CDC4 (5)), timers (PER (6)) or as negative inhibition toggles (IRS1 (7)) but global analysis of these phosphorylation clusters has remained impractical. Tandem mass spectrometry (MS/MS) of tryptic peptides is a key tool in discovering and quantifying sites of protein phosphorylation. Typical phosphoproteomic workflows use data dependent acquisition (DDA) to collect MS/MS spectra based on precursor m/z as peptides chromatographically elute. Site localization software tools such as Ascore (8) assign the most likely phosphorylation position for each peptide using site-specific fragment ions. To increase the number of distinct peptides that are sampled, DDA

[*]Corresponding author, jvillen@uw.edu.

AUTHOR CONTRIBUTIONS
R.T.L. and J.V. conceived the study. B.C.S., R.T.L., and J.V. designed the experiments. B.C.S. and R.T.L. performed the experiments. B.C.S. designed and wrote the software, and analyzed the data. R.T.L. generated the spectrum library. M.J.M. and J.V. supervised the work. B.C.S., R.T.L., M.J.M. and J.V. wrote the paper.

SOFTWARE AVAILABILITY:
Thesaurus is an open-source, cross-platform Java application available at http://bitbucket.org/searleb/thesaurus/. While Thesaurus can read and produce reports to enable Skyline analysis, it is a stand-alone tool and does not require any other software to run. Thesaurus is fully multi-threaded and designed to work on typical desktop computers where searches of individual mzML files in this experiment took on average of 10–12 minutes to complete with commodity PC hardware.

DATA AVAILABILITY:
All mass spectrometry files (Supplementary Dataset 3) presented here have been deposited to the Chorus Project (https://chorusproject.org/) with the project identifier 1374 and at the MassIVE proteomics repository (https://massive.ucsd.edu/) with project identifier MSV000082956. The high-resolution human phosphopeptide spectrum library used in this study is available through Phosphopedia (https://phosphopedia.gs.washington.edu/PhosphoproteomicsAssay/sections/elib.xhtml).

COMPETING FINANCIAL INTERESTS STATEMENT
The MacCoss Lab at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific, the manufacturer of the instrumentation used in this research. Additionally, M.J.M. is a paid consultant for Thermo Fisher Scientific and B.C.S. is a shareholder and paid consultant of Proteome Software.

dynamically excludes peptides of the same m/z from being sampled repeatedly within a narrow elution time. However, phosphopeptides that exist as multiple positional isomers are difficult to sample and assign using DDA because they have the same mass, similar retention times, and share many fragment ions.

Parallel reaction monitoring (PRM) and data independent acquisition (DIA) are alternative approaches that systematically collect MS/MS spectra across the chromatographic elution profile of peptides, improving quantitative reproducibility. While PRM methods target specific peptide precursors (9), DIA methods acquire MS/MS spectra systematically across the m/z space (10). These methods are free of both intensity biases during data collection and active exclusion of previously sequenced precursors, making it possible to detect closely eluting positional isomers. Despite the strengths of these methods, assigning phosphorylation to a specific amino acid remains difficult. Recently, Rosenberger et al (11) reported on IPF, a peptide-centric tool that uses OpenSwath (12) to determine the most likely positional isomer from fragment ions in a peak. An alternate spectrum-centric approach, PIQED (13), deconvolves DIA data with DIA-Umpire (14) to enable site localization tools originally designed for DDA. Finally, Specter (15) deconvolves DIA signals using linear combinations of spectra in libraries, and in some instances can resolve positional isomers. A limitation of IPF and PIQED is that they compete potential positional isomers with similar retention times against each other and only the best scoring isomer is reported. On the other hand, Specter was not designed for phosphopeptide localization and lacks site localization statistics. Here we extend these approaches and present a new DIA and PRM search engine named Thesaurus, which is designed to specifically look for positional isomers.

Thesaurus detects phosphopeptides with EncyclopeDIA and a spectrum library (16), and using the detections as retention time anchors, iteratively finds new positional isomers that share many of the same fragment ions but differ in their phosphorylation site-specific ions (Supplementary Figure 1). Thesaurus can detect multiple co-eluting positional isomers because it calculates localization probabilities directly using an interference distribution, rather than by competing isomers against each other. For each phosphopeptide, Thesaurus determines every possible positional isomer and extracts corresponding site-specific fragment ion signals. Each ion has a unique frequency of interference across the experiment, and this frequency is highest with low m/z ions (Supplementary Figure 2). Thesaurus uses this frequency to calculate a background distribution for each run and precursor isolation window, since these distributions depend on peptide mass and various acquisition settings. Localization p-values are calculated as the probability that all site-specific ions were observable by chance in this background distribution and FDR corrected using the Benjamini-Hochberg method. Thesaurus detects positional isomers absent from the spectrum library by generating synthetic spectra with shifted fragment ions. Thesaurus quantifies positional isomers even if their precursor signals are convolved, using site-specific ions to determine peak boundaries and including additional fragment ions that fit that shape.

We validated Thesaurus using a synthetic phosphopeptide DIA dataset described previously (11) (Supplementary Figure 3), and found that it produced both more detections and more accurate error estimates than IPF and PIQED. In addition to correctly localizing 240 synthetic phosphopeptides, Thesaurus was also able to identify and flag 11 products of a gas-phase phosphate rearrangement (Supplementary Figure 4). We further demonstrated Thesaurus' performance with phosphopeptides derived from serum-stimulated HeLa cells. Previously we reported a human phosphopeptide library based on nearly a thousand DDA experiments (17). Here we used a subset of this library containing 82,029 phosphopeptides, where 44% of phosphopeptides are phosphorylated at multiple positions (Supplementary Figure 5). Thesaurus was able to detect an average of 10,780 phosphopeptides across four technical replicates (Supplementary Dataset 1), corresponding to an average of 6,288 confidently localized positional isomers (Supplementary Figure 6a). We found that within phosphopeptides containing multiple acceptor sites, approximately 13% were phosphorylated at multiple positions (Supplementary Figure 6b). While overall Thesaurus performed comparably to DDA/Ascore (Figure 1a), Thesaurus found four times more phosphopeptides with multiple positional isomers per run (Figure 1b), predominantly when the retention time difference was less than 60 seconds (Supplementary Figure 6c). The detection of positional isomers was also more reproducible, as demonstrated for phosphopeptides with only two acceptor sites (Figure 1c). For example, Thesaurus consistently detected two isomers of the peptide AITGASLADIMAK from the 60S ribosomal protein RPL24 with phosphorylation at either T83 or S86 (Figure 1d). These isomers elute within 25 seconds of each other, and while the precursor signal represents a mixture of both isomers (Figure 1e), Thesaurus confidently assigned them using site-specific ions (Figure 1f and 1g) to calculate localization scores (Figure 1h). Although the site-specific fragment ions observed in DIA were confirmed with DDA (Supplementary Figure 7), DDA reliably triggered MS/MS on the early-eluting isomer (pS86) and excluded the more intense late-eluting isomer (pT83) in 3 of the 4 replicates (Figure 1d). Here, precursor quantification was unreliable because the total signal was only assigned to the lower abundance isomer. In contrast, IPF and PIQED assigned the higher intensity pT83 isomer in every DIA replicate but never detected the lower intensity pS86 isomer.

We designed a DIA quantitative experiment to resolve positional isomers in the PI3K/AKT signaling network in MCF-7 cells after stimulation with insulin or IGF-1. We found that 2,273 of the 7,434 localized phosphopeptides that were measured consistently in six cell culture replicates changed abundance at an FDR-corrected p-value <0.05 (Supplementary Dataset 2), including several known AKT substrate sites (Supplementary Figure 8) and 48 of 759 positional isomer pairs (Supplementary Figure 9 and 10). For example, the peptide KGSGDYMPMSPK from the insulin receptor scaffold protein IRS1 (Figure 2a) contains three residues that are putatively phosphorylated by three different kinases: Y632 by INSR (upstream of AKT), S636 by S6K1 (downstream of AKT), and S629 by either PKA (18) or AKT (19). While our spectrum library did not contain KGSGDpYMPMSPK, Thesaurus was able to independently detect, localize, and quantify all three singly phosphorylated isomers (Figure 2b–e), which we confirmed with targeted PRM (Supplementary Figure 11). As expected from the model, phosphorylation of IRS1 Y632 increased by >10-fold after both insulin and IGF-1 stimulation (Figure 2e). Similarly, S636 phosphorylation also increased, but that effect was lower and significantly diminished after treatment with the AKT inhibitor MK-2206. We saw a decrease in phosphorylation of S629 after insulin and IGF-1 stimulation, likely associated with an increase in the doubly phosphorylated (S629 and

S636) peptide (Supplementary Figure 12). This result suggests that S629 is phosphorylated by PKA and not AKT. While Thesaurus consistently detected all three forms, both IPF and PIQED make assumptions that can complicate positional isomer detection (Figure 2b, Supplementary Note).

Some phosphopeptide positional isomers were indistinguishable by retention time, yet could be localized and quantified using site-specific ions. For example, MARK3 positional isomers at S469 and S476 co-eluted under our chromatographic conditions. Using Thesaurus, we were able to detect that the S469 isomer responded to insulin/IGF-1 and AKT inhibition while the S476 isomer remained constant (Supplementary Figure 13). However, one should be cautious about interpreting co-eluting positional isomers. Potentially a third to a half of these isomers may actually be the result of gas-phase phosphate rearrangement, and only additional evidence such as measured differential regulation can confirm that both isomers exist biologically (Supplementary Note).

Positional isomers represent an important concept for understanding signaling biology. Thesaurus provides a new avenue to study positional isomers even if their precursor signals cannot be resolved. With a search engine specifically designed to analyze neighboring sites of phosphorylation it is possible to determine whether they have distinct functional implications, are redundant mechanisms for regulation, or are simply representative of a background phosphorylation state. Additionally, Thesaurus could be used to estimate phosphate rearrangements occurring in the mass spectrometer and how these interfere with phosphoproteome analysis. Other types of PTM studies will also benefit from this approach, so we have extended Thesaurus to support other modifications. Our results indicate that PRM and DIA strategies will be crucial in assessing the complex regulatory nature of the human phosphoproteome.

METHODS

Methods, including statements of software and data availability and any associated accession codes and references, are available in the online version of the paper.

ONLINE METHODS

Cell Culture:

HeLa cervical cancer cells were cultured at 37°C and 5% CO_2 in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with L-glutamine, 10% FBS, and 0.5% streptomycin/penicillin. Cells were grown to an estimated 90% confluence in 10-cm plates, where one plate was used for each replicate/condition. Prior to harvest, cells were incubated for 4 hours under serum starvation conditions and then serum stimulated for 30 minutes. MCF-7 breast cancer cells were similarly cultured and starved, followed by stimulation with insulin (100 ng/ml) or IGF-1 (100 ng/ml) in phosphate-buffered saline (PBS) or unstimulated (control, added same volume of PBS) for 20 minutes. Some MCF-7 cells were additionally treated with DMSO or the pan-AKT inhibitor MK-2206 for 40 minutes before stimulation. After stimulation cells were quickly washed three times with refrigerated PBS and immediately flash frozen with liquid nitrogen. With the MCF-7 experiment, six cell culture replicates were performed for each of the six conditions: control/DMSO, insulin/DMSO, IGF-1/DMSO, control/MK-2206, insulin/MK-2206, and IGF-1/MK-2206. The six replicates were performed in three cell culture batches to simplify sample handling and ensure precise timing.

Sample Preparation:

Frozen cells were lysed in a buffer of 9 M urea, 50 mM Tris (pH 8), and 75 mM NaCl, with a cocktail of protease inhibitors (Roche Complete-mini EDTA-free) and phosphatase inhibitors (50 mM NaF, 50 mM β-glycerophosphate, 10 mM pyrophosphate, and 1 mM orthovanadate). After scraping, cells were subjected to 2 cycles of 25 seconds of probe sonication each followed by 10 minutes of incubation on ice. Lysates were centrifuged for 10 minutes at 21,000 × *g* and 4°C to eliminate cell debris. The protein content of the supernatant was estimated using BCA. For every condition/replicate, an estimated 850 μg of protein was reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature, and the alkylation was quenched with an additional 5 mM dithiothreitol for 15 minutes at room temperature. The proteins were diluted to 1.8 M urea and then digested with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 4 hours at 37°C. The digestion was quenched by adding 10% trifluoroacetic acid (TFA) to achieve pH ~ 2. Resulting peptides were desalted with 100 mg tC18 SepPak cartridges (Waters) using vendor-provided protocols and dried with vacuum centrifugation. Phosphopeptides were enriched using immobilized metal affinity chromatography (IMAC) using Fe-NTA magnetic agarose beads (Cube Biotech). Enrichment was performed with a KingFisher Flex magnetic particle processor (Thermo Scientific), which incubated peptides with 150 μl 5% bead slurry in 0.1% TFA 80% acetonitrile for 30 minutes, washed them three times with the same solution, and eluted them with 60 μl 25% acetonitrile 0.5% NH_4OH. Phosphopeptides were then acidified with 10% formic acid and dried. Phosphopeptides were brought to 1 μg / 3 μl in 0.1% formic acid assuming a 1:100 reduction in peptide abundance from the IMAC enrichment.

Liquid Chromatography Mass Spectrometry:

Phosphopeptides were separated with a Waters NanoAcquity UPLC and emitted into a Thermo Q-Exactive HF or a Thermo Fusion tribrid mass spectrometer. Pulled tip columns were created from 75 μm inner diameter fused silica capillary in-house using a laser pulling device and packed with 3 μm ReproSil-Pur C18 beads (Dr. Maisch) to 300 mm. Trap columns were created from 150 μm inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 25 mm. Solvent A was 0.1% formic acid in water, while solvent B was 0.1% formic acid in 98% acetonitrile. For each injection, 3 μl (approximately 1 μg) was loaded and eluted using a 90-minute gradient from 5 to 25% B, followed by a 40-minute washing gradient. Data were acquired using data-dependent acquisition (DDA), data-independent acquisition (DIA), or parallel reaction monitoring (PRM). Four DDA and DIA HeLa technical replicates were acquired in alternating mode to avoid bias. Single injections for 36 MCF-7 samples (six cell culture replicates of six conditions) was randomized within blocks to enable downstream statistical analysis.

DDA Acquisition and Processing:

The Thermo Q-Exactive HF was set to positive mode in a top 12 configuration. Full MS scans of mass range 400–1600 were collected at 60,000 resolution to hit an AGC target of 3e6. The maximum injection time was set to 100 ms. MS/MS scans were collected at 30,000 resolution, AGC target of 1e6, and maximum injection time of 55 ms. The isolation width was set to 1.5 m/z with a normalized collision energy of 27. Only precursors charged between +2 and +4 that achieved a minimum AGC of 1e4 were acquired. Dynamic exclusion was set to “auto” and to exclude all isotopes in a cluster.

Thermo .RAW files were converted to .mzXML format using ReAdW and searched against a Uniprot Human FASTA database (downloaded July 1 2014 to maintain consistency with Lawrence *et al* (17), 87,613 entries) with Comet (version 2015.02v2), allowing for variable methionine oxidation, protein N-terminal acetylation, and phosphorylation at serines, threonines, and tyrosines. Cysteines were assumed to be fully carbamidomethylated. Searches were performed using a 50 ppm precursor tolerance and a 0.02 Da fragment ion tolerance using fully tryptic specificity (KR|P) permitting up to two missed cleavages. Search results were filtered to a 0.6% PSM-level (Peptide to Spectrum Match-level) FDR using Percolator (version 3.1), which we determined in this experiment to closely track to a 1% peptide-level FDR. Site localization was performed using an in-house implementation of Ascore (8) that was modified to not compete positional isomers against each other in order to have a higher chance of detecting overlapping isomers. We set Ascore to use a 0.02 Da fragment ion tolerance and we filtered for phosphopeptides with at least one corresponding PSM that produced an Ascore value >= 20 (p-value<0.01).

DIA / PRM Acquisition and Processing:

The Thermo Q-Exactive HF mass spectrometer was configured to acquire 20 MS/MS scans at 30,000 resolution, AGC target 1e6, maximum injection time 55 ms, using overlapping precursor isolation windows of 20 m/z units and centered at: [500.4774, 520.4865, 540.4956, 560.5047, 580.5138, 600.5229, 620.5319, 640.541, 660.5501, 680.5592, 700.5683, 720.5774, 740.5865, 760.5956, 780.6047, 800.6138, 820.6229, 840.632, 860.6411, 880.6502, 900.6593, 490.4728, 510.4819, 530.491, 550.5001, 570.5092, 590.5183, 610.5274, 630.5365, 650.5456, 670.5547, 690.5638, 710.5729, 730.582, 750.5911, 770.6002, 790.6093, 810.6183, 830.6274, 850.6365, 870.6456, and 890.6547]. Full MS scans (mass range 485–925, resolution 30,000, AGC target 3e6, maximum injection time 100 ms) were interspersed every 18 scans. MS/MS scans were programed with normalized collision energy of 27 and an assumed charge state of +2.

For PRMs, the Thermo Fusion tribrid mass spectrometer was configured to collect MS/MS scans corresponding to 62 precursor targets in the PI3K/AKT signaling pathway scheduled with 10-minute retention time windows using Phosphopedia. The large 10-minute window enables both scheduling from Phosphopedia without additional calibration runs and the detection of alternate positional isomers that may elute far away from the target. Full MS scans (mass range 400–1600) were collected at 60,000 resolution to hit an AGC target of 3e6. The maximum injection time was set to 100 ms. MS/MS scans were collected at mass range of 100–1600, resolution of 30,000, AGC target of 1e6, and maximum inject time of 55 ms. The isolation width was set to 0.7 m/z with a normalized collision energy of 27.

A Bibliospec (20) HCD spectrum library of tryptic phosphopeptides was created from the Thermo Q-Exactive data previously published in Lawrence *et al.* (17) using Skyline (version 3.1.0.7382) (21). This .BLIB library and accompanying .iRTDB normalized retention time database were used to search the .mzMLs for peptides. Thermo .RAW files were converted to .mzML and .mzXML formats using the ProteoWizard package (version 3.0.10922) where they were peak picked using vendor libraries and deconvoluted using Prism in "overlap_only" mode. We used EncyclopeDIA (16), a spectrum library search engine to detect peptides in a peptide-centric approach. Our engine searches DIA data using b- and y-ion fragments of charges +1 and +2, and includes phosphate neutral losses that can be found in library spectra. We applied the following settings: 30,000 resolution (effectively +/− 16.7 ppm tolerance) for precursor, fragment, and library. Detected features were assigned and corrected to <0.01 FDR using Percolator version 3.1.

In addition to Thesaurus, DIA experiments were analyzed with PIQED and IPF. PIQED 0.1.2 using DIA-Umpire v2.0 was configured for 10 ppm mass tolerances and 30,000 resolution, and to extract +2 to +4 charged peptides. DIA-Umpire produces three .MGFs for each .mzXML; all three were searched with the Comet v2017012 pipeline using the Trans Proteome Pipeline v5.1.0. Comet was configured using the same parameters as for DDA. After Comet analysis, PIQED ran PeptideProphet, iProphet, and PTMProphet through the xinteract interface using the following command line options: "-p0.01 -l6 -PPM -OH -i -M-STY:79.966,M:15.995,n:42.010565-MZTOL=0.1-PPMTOL=10" to combine and localize the three searches per sample. IPF was run using OpenMS v2.4.0-nightly-2018-08-22. The .BLIB library was converted to TraML and filtered to use the top 6 transitions for detection. For iRT anchors we used the most highly abundant phosphopeptide in each of 11× 10-minute retention time bins that were also found localized in every experiment by Thesaurus (both HeLa and MCF-7). OpenSwath was configured to use 30,000 resolution (+/− 16.67 ppm tolerance) using the following command line options: "-tr transitionlist_6transitions_decoys.PQP -swath_windows_file dia_ranges_analysis.tsv -tr_irt common_high_abundance_phosphopeptides.TraML -sort_swath_maps -batchSize 1000 -readOptions cacheWorkingInMemory -tempDirectory temp -use_ms1_traces -mz_extraction_window 16.67 -ppm -mz_correction_function quadratic_regression_delta_ppm -RTNormalization:alignmentMethod lowess -Scoring:stop_report_after_feature 5 -enable_uis_scoring -Scoring:Scores:use_mi_score -Scoring:Scores:use_ms1_mi -Scoring:Scores:use_total_mi_score". PyProphet was configured to use small MS1-level fitting parameters (lambda: 0.001, 0, 0) to ensure convergence for MS1-level scoring, and no additional parameters for MS2 and transition-level scoring. Finally, IPF was configured to use "--no-ipf_ms1_scoring --no-ipf_ms2_scoring", and configured to export results using the default --max_transition_pep=0.7 and --ipf_max_peptidoform_pep=0.4 filters.

Primary Scoring Using All Fragment Ions:

Site-specific fragment ions can distinguish positional isomers, but applying current site localization tools originally designed for DDA (8, 22–25) to DIA can be problematic because they assume a constant background noise level inconsistent with the high level of background interference ions found in DIA data (Supplementary Figure 2). We account for this by calculating a background frequency distribution to estimate the likelihood of detecting an interfering signal as the frequency of each m/z in the raw file for a given precursor isolation window. Our approach allows us to quickly query the distribution of ions for every fragment ion individually, enabling the use tight mass tolerances (measured in ppm) to assess the likelihood of interference.

For each queried phosphopeptide, we determine the set of combinatorial permutations corresponding to the number of phosphorylations and the number of potential phosphoacceptor sites. If a positional isomer permutation is not present in the library, then a synthetic library spectrum is generated from the anchor by shifting fragment ion peak intensities for each b-type, y-type, and neutral loss ions to the appropriate m/z's, using an approach similar to that used in SpectraST (26).

First, we extract chromatograms for fragment ions in a window +/− 10% of the total acquired chromatographic time from the retention time anchor. At every retention time point in that window we calculate a score (Primary Score) based on the X!Tandem HyperScore where the function is the dot product of the intensities in the acquired spectrum (I) and the library spectrum (P) multiplied by the factorial of the number of matching ions:

$$Primary\,Score = -Log_{10}\left(\left(\sum_{i=0}^{n} I_i \cdot P_i\right) \cdot n!\right)$$

Iterative Localization Scoring:

We begin the localization process by comparing the highest scoring isomer and retention time point to every other alternate isomer (j) that either a) has not been detected, b) has been detected nearby this RT, or c) scores higher at this RT. We calculate a p-value as the probability of finding all of the detected site-specific ions (n) by chance from the background frequency distribution (m) and the total number of site-specific ions considered (N). The final localization p-value is the max (least significant) of these values across alternate isomers (j):

$$Localization\,p - value = max_j\left(\prod_{i=0}^{n} p\left(null \middle| m_{ij}\right)\right)$$

The localization score is the -log$_{10}$(p-value) to produce a positive score for higher confident localizations. This score is smoothed across time by Gaussian weighting, where the Gaussian standard deviation is estimated from the expected peak width (here we used 25 sec = 6 standard deviations). Thesaurus then extracts the chromatographic shape of the localizing fragment ions (for the target isomer compared to the alternate isomer with the least significant p-value) to calculate the IonCount score, a measure of the number of co-eluting fragments. The peak shape of every co-eluting b- and y-ion (i) associated with the target isomer is compared to the shape defined by the localizing fragment ions using Pearson's correlation (c_i). The IonCount score is calculated as:

$$IonCount\,Score = \sum_{i=0}^{n} \left(c_i\right)^2$$

The sum of squares of correlation values is used to heavily downweight the impact of ions that poorly correlate with the target isomer. Only ions with positive correlation scores are used. The target isomer is considered detected if the apex p-value p <=0.01 and the IonCount score is >=3 (these thresholds are user adjustable). If the target isomer does not pass these thresholds, the retention time window is blacklisted for this isomer and up to one more attempt can be made to localize the peptide. The localization process is iterated until all isomers have been detected or ruled out.

Localization Post Processing:

A parallel process is performed using decoy-generated spectra and the scoring features for both are fed into Percolator 3.1 to generate Q-values. Of peptides that pass the detection Q-value threshold, an additional localization Q-value is calculated using the Benjamini-Hochberg method and the top reported localization p-values for each isomer. Detected isomers are filtered to a user-settable Q-value (typically 0.01 or 0.05) using both thresholds to ensure high confidence detections.

Positional isomer searching can be performed in an "uncalibrated" manner (where retention times in the library are assumed to be precise) if the DIA data was searched directly with a DIA library search tool or if DDA experiments were run concurrently (SWATH). Alternatively, Thesaurus supports searching in a "calibrated" manner, which assigns relative retention time ordering by searching each peptide anchor across the entire experiment window using the Primary Score if retention times are unknown (e.g. importing NIST libraries) or need to be calibrated (e.g. with spectrum libraries acquired on different platforms, gradients, or HPLC columns). Alternatively, searches can be performed across the entire acquisition window when analyzing targeted PRM data. We have also enabled options for only calculating localization scores and estimating FDR, skipping the detection of positional isomers that are not found in the library.

Quantitation and Statistical Analysis:

We used strict criteria to consider a localized peptide quantifiable. In addition to the localization scoring requirements, we also required at least three quantitative fragment ions and that the localized isomer was observed in every replicate of at least one condition. Thesaurus uses the site-specific fragment ions to determine the shape of the peak and assigns quantitative fragment ions as those that match that shape for quantification with Pearson's correlation coefficients greater than 0.9. Quantification was performed by summing the background-subtracted peak areas of site-specific fragment ions or all fragment ions, depending on the level of peptide separation. Background subtraction removes the trapezoidal area below the peak integration window. Integrated intensities were normalized within each replicate group, and across groups to the control intensity median. Statistical analysis was performed on the MCF-7 dataset globally with Benjamini-Hochberg FDR corrected one-way ANOVAs from six conditions with six replicates.

PRM results were further validated with follow-up analysis in Skyline-Daily (version 3.6.1.10615). Skyline was configured to extract all +1 and +2 b- and y-ions, including neutral losses of phosphate, as well as precursor traces for the monoisotopic, first and second isotopes. After initially importing the runs, peptides were hand-curated to match the retention time boundaries determined by site-localizing analysis. Fragment ions that appeared to be interfered with were removed from the analysis.

Reporting Summary:

Detailed justification of the experimental design, as well as information on our methods, including cell culture validation and sources is available in the online Life Sciences Reporting Summary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We would like to thank members of the Villén and MacCoss labs for useful discussions. We would also like to thank G. Rosenberger, H. Röst, J. Meyer and B. Schilling for insightful conversations about detecting positional isomers. B.C.S. is supported by F31 GM119273. R.T.L. was supported by a Samuel and Althea Stroum Endowed Graduate Fellowship. This work is supported by NIH grants P41 GM103533, R21 CA192983, and U54 HG008097 to M.J.M.; and R35 GM119536, R01 AG056359, and a research grant from the W.M. Keck Foundation to J.V.

REFERENCES

1. Boersema PJ, Foong LY, Ding VM et al. In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. Mol Cell Proteomics. 2010;9:84–99. [PubMed: 19770167]
2. Villén J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A. 2007;104:1488–1493. [PubMed: 17242355]
3. Schweiger R, Linial M. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biol Direct. 2010;5:6. [PubMed: 20100358]
4. Huang CY, Ferrell JE. Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci U S A. 1996;93:10078–10083. [PubMed: 8816754]
5. Nash P, Tang X, Orlicky S et al. Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. Nature. 2001;414:514–521. [PubMed: 11734846]
6. Chiu JC, Ko HW, Edery I. NEMO/NLK phosphorylates PERIOD to initiate a time-delay phosphorylation circuit that sets circadian clock speed. Cell. 2011;145:357–370. [PubMed: 21514639]
7. Liu YF, Herschkovitz A, Boura-Halfon S et al. Serine phosphorylation proximal to its phosphotyrosine binding domain inhibits insulin receptor substrate 1 function and promotes insulin resistance. Mol Cell Biol. 2004;24:9668–9681. [PubMed: 15485932]
8. Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006;24:1285–1292. [PubMed: 16964243]

9. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol Cell Proteomics. 2012;11:1475–1488. [PubMed: 22865924]
10. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004;1:39–45. [PubMed: 15782151]
11. Rosenberger G, Liu Y, Röst HL et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat Biotechnol. 2017;35:781–788. [PubMed: 28604659]
12. Röst HL, Rosenberger G, Navarro P et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol 2014;32(3):219–223. [PubMed: 24727770]
13. Meyer JG, Mukkamalla S, Steen H, Nesvizhskii AI, Gibson BW, Schilling B. PIQED: automated identification and quantification of protein modifications from DIA-MS data. Nat Methods. 2017;14:646–647. [PubMed: 28661500]
14. Tsou CC, Avtonomov D, Larsen B et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods. 2015;12:258–64, 7 p following 264. [PubMed: 25599550]
15. Peckner R, Myers SA, Jacome ASV et al. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat Methods. 2018;15:371. [PubMed: 29608554]
16. Searle BC, Pino LK, Egertson JD et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun. 2018;9:5128. [PubMed: 30510204]
17. Lawrence RT, Searle BC, Llovet A, Villén J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nat Methods. 2016;13:431–434. [PubMed: 27018578]
18. Yi Z, Luo M, Carroll CA, Weintraub ST, Mandarino LJ. Identification of phosphorylation sites in insulin receptor substrate-1 by hypothesis-driven high-performance liquid chromatography-electrospray ionization tandem mass spectrometry. Anal Chem. 2005;77:5693–5699. [PubMed: 16131083]
19. Luo M, Langlais P, Yi Z et al. Phosphorylation of human insulin receptor substrate-1 at Serine 629 plays a positive role in insulin signaling. Endocrinology. 2007;148:4895–4905. [PubMed: 17640984]
20. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem. 2006;78:5678–5684. [PubMed: 16906711]
21. MacLean B, Tomazela DM, Shulman N et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26:966–968. [PubMed: 20147306]
22. Olsen JV, Blagoev B, Gnad F et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127:635–648. [PubMed: 17081983]
23. Savitski MM, Lemeer S, Boesche M et al. Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics. 2011;10:M110.003830.
24. Taus T, Köcher T, Pichler P et al. Universal and confident phosphorylation site localization using phosphoRS. Journal Proteome Res. 2011;10:5354–5362. [PubMed: 22073976]
25. Fermin D, Walmsley SJ, Gingras AC, Choi H, Nesvizhskii AI. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. Mol Cell Proteomics. 2013;12:3409–3419. [PubMed: 23918812]
26. Ma CW, Lam H. Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring. J Proteome Res. 2014;13:2262–2271. [PubMed: 24661115]

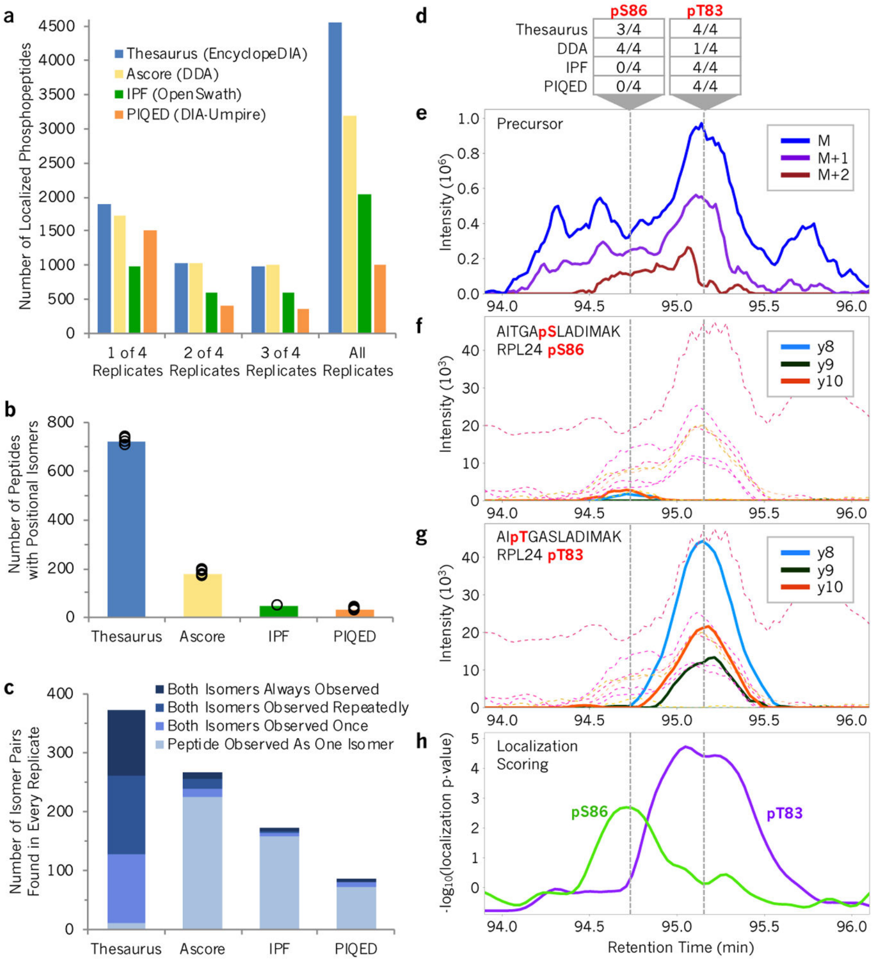
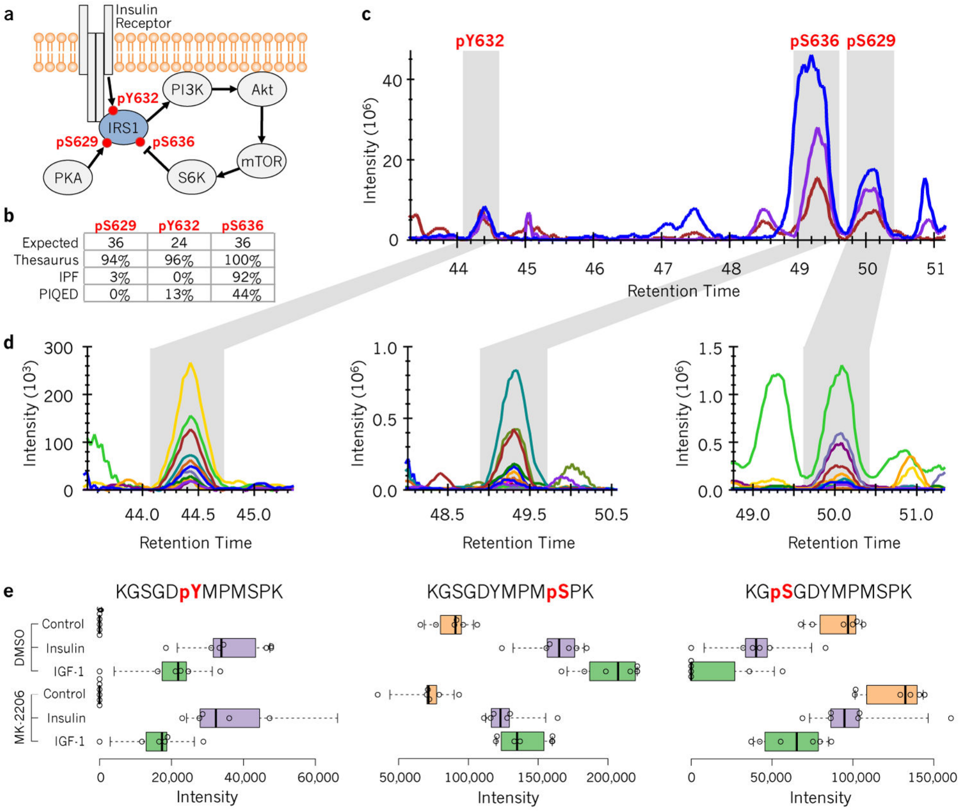


Figure 1. An approach for detecting phosphopeptides with Thesaurus.

(**a**) The number of localized HeLa phosphopeptides detected in four technical replicates from DIA data with Thesaurus, IPF, and PIQED, or from DDA data with Ascore. (**b**) The average (bars) and number (circles) of phosphopeptides detected with multiple positional isomers from the same samples (N=4). (**c**) The number of singly phosphorylated peptides with two acceptor residues that were detected in all four technical replicates. To be included in this chart both isomers of the phosphopeptide must have been observed in the same replicate by at least one analysis approach. (**d**) The number of times the singly phosphorylated RPL24 peptide AITGASLADIMAK was independently observed as pT83 and pS86 using Thesaurus, IPF, PIQED, and Ascore (with DDA) (N=4), where a representative case is shown in (**e-h**) for illustrative purposes. (**e**) Precursor extracted ion chromatogram for the singly phosphorylated peptide AITGASLADIMAK. Dashed grey lines indicate the peak apex for the individual isomers. (**f,g**) Site-specific y8, y9, and y10 ions (solid) and other y-ions (dashed) for pS86 (**f**) and pT83 (**g**). (**h**) Localization p-values using Thesaurus for pS86 (green) and pT83 (purple).



	pS629	pY632	pS636
Expected	36	24	36
Thesaurus	94%	96%	100%
IPF	3%	0%	92%
PIQED	0%	13%	44%

Figure 2. Detection and quantification of IRS1 phosphorylation.

(**a**) Diagram of IRS1 phosphorylation at sites S629, Y632, and S636. (**b**) The expected number of total observations (4 or 6 conditions, N=6) and the actual detection rates of each positional isomer from each sample with independent analyses using either Thesaurus, IPF, or PIQED, where a representative case is shown in (**c-d**) for illustrative purposes. (**c**) Precursor traces for three singly phosphorylated positional isomers of the IRS1 peptide KGSGDYMPMSPK in insulin-stimulated MCF-7 cells. (**d**) Corresponding fragment ions indicating phosphorylation at Y632 by INSR, S636 by S6K, and S629 by PKA. Thesaurus detected the pY632 positional isomer (absent from our library) using the pS629 and pS636 isomers as anchors. (**e**) Box plots and values indicating summed fragment ion intensities for the three phosphosites on IRS1 across six cell culture replicates after stimulation with insulin, IGF-1, or unstimulated (control); with and without the AKT inhibitor MK-2206. Boxes indicate quartiles and medians, while whiskers indicate the estimated 5% and 95% ranges.