PTMScout, a Web Resource for Analysis of High Throughput Post-translational Proteomics Studies*[S]

Kristen M. Naegle‡§, Melissa Gymrek‡§, Brian A. Joughin¶, Joel P. Wagner‡, Roy E. Welsch‖, Michael B. Yaffe‡¶**, Douglas A. Lauffenburger‡¶, and Forest M. White‡¶‡‡

The rate of discovery of post-translational modification (PTM) sites is increasing rapidly and is significantly outpacing our biological understanding of the function and regulation of those modifications. To help meet this challenge, we have created PTMScout, a web-based interface for viewing, manipulating, and analyzing high throughput experimental measurements of PTMs in an effort to facilitate biological understanding of protein modifications in signaling networks. PTMScout is constructed around a custom database of PTM experiments and contains information from external protein and post-translational resources, including gene ontology annotations, Pfam domains, and Scansite predictions of kinase and phosphopeptide binding domain interactions. PTMScout functionality comprises data set comparison tools, data set summary views, and tools for protein assignments of peptides identified by mass spectrometry. Analysis tools in PTMScout focus on informed subset selection via common criteria and on automated hypothesis generation through subset labeling derived from identification of statistically significant enrichment of other annotations in the experiment. Subset selection can be applied through the PTMScout flexible query interface available for quantitative data measurements and data annotations as well as an interface for importing data set groupings by external means, such as unsupervised learning. We exemplify the various functions of PTMScout in application to data sets that contain relative quantitative measurements as well as data sets lacking quantitative measurements, producing a set of interesting biological hypotheses. PTMScout is designed to be a widely accessible tool, enabling generation of multiple types of biological hypotheses from high throughput PTM experiments and advancing functional assignment of novel PTM sites. PTMScout is available at http://ptmscout.mit.edu. ***Molecular & Cellular Proteomics 9:2558–2570, 2010.***

Post-translational modifications (PTMs)[1] regulate cellular signaling networks by modifying activity, localization, turnover, and other characteristics of proteins in the cell. For example, signaling in receptor tyrosine kinase networks, such as those downstream of epidermal growth factor receptor (EGFR) and insulin receptor, is initiated by binding of cytokines or growth factors and is generally propagated by phosphorylation of signaling molecules. Additionally, receptor surface expression can be regulated by ubiquitination, whereas gene expression can be regulated by acetylation of transcription factors and histones. With the increasing utilization of high throughput mass spectrometry (MS) technologies and the ability to enrich for a particular modification from a biological sample, hundreds or even thousands of PTM sites can now be identified in a single experiment and relatively quantified across biological conditions (1). This increase in the number of PTM sites identified in each analysis has led to a rapid and accelerating expansion of known post-translational modifications as evidenced by the number of the entries in a knowledgebase of phosphorylation over the past 5 years: in 2004, when it was first published (2), Phospho.ELM contained 1,703 known phosphorylation sites; in 2009, Phospho.ELM contained 19,649 known sites of phosphorylation, a more than 10-fold increase.

A number of database resources, including Phospho.ELM (3), PhosphoSite (4), PHOSIDA (5), and SysPTM (6), have emerged in response to the large production of phosphorylation data and are expanding to include other PTMs. For instance, PhosphoSite (4), originally established as a compendium of phosphorylation sites, has started to incorporate acetylation, methylation, glycosylation, ubiquitination, and other PTMs. Sites of modification can also be found in large protein compendia, such as UniProtKB (7). In addition to

From the Departments of ‡Biological Engineering and **Biology, ¶Koch Institute for Integrated Cancer Research, and ‖Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received, May 25, 2010

Published, MCP Papers in Press, July 14, 2010, DOI 10.1074/mcp.M110.001206

[1] The abbreviations used are: PTM, post-translational modification; GO, gene ontology; RMA, robust multiarray average; EGFR, epidermal growth factor receptor; FRK, Fyn-related kinase; CBP, cAMP-response element-binding protein (CREB)-binding protein; NCBI, National Center for Biotechnology Information; GNF, Genomic Institute of the Novartis Research Foundation; FAK, focal adhesion kinase; HRG, heregulin; HMEC, human mammary epithelial cell; SOM, self-organizing map; UIM, ubiquitin interaction motif; MF, molecular function.

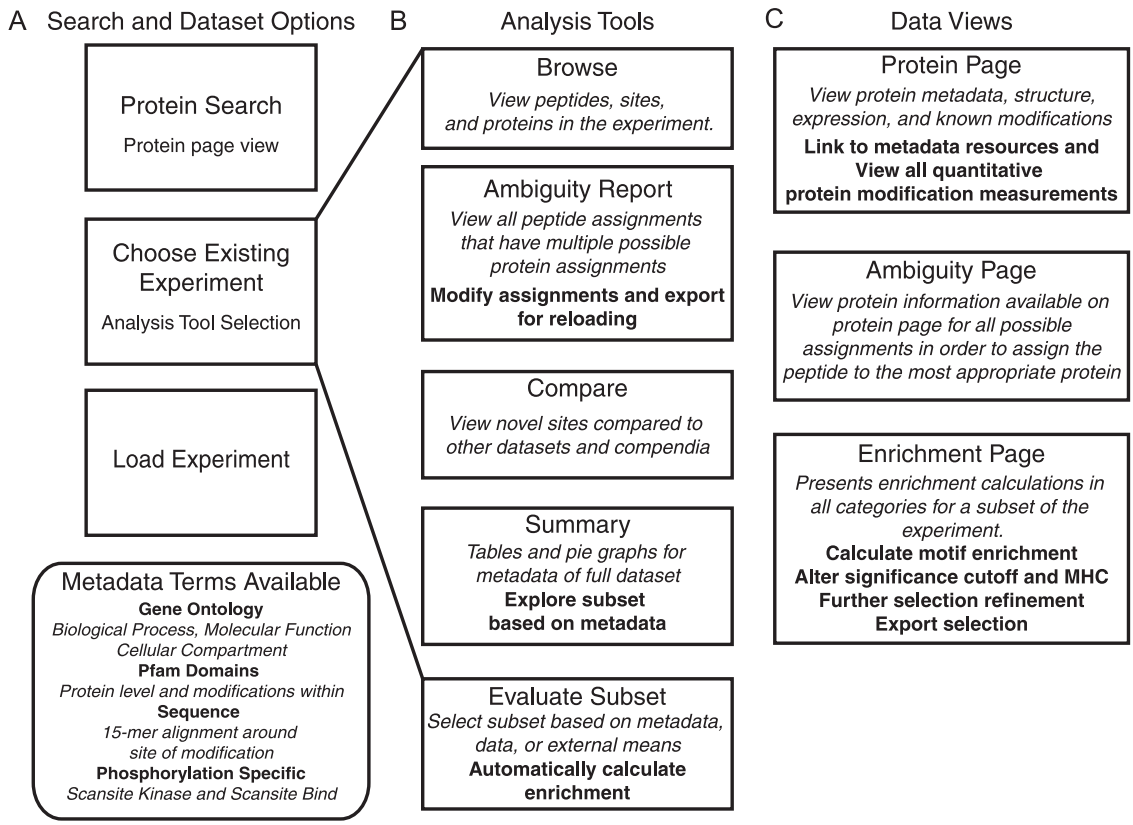


FIG. 1. **Depiction of major features, analysis tools, and page view types available in PTMScout.** *A*, one can choose to search by protein, load a new data set, or analyze an existing data set. The metadata incorporated in PTMScout to annotate the biological molecules of interest are described. *B*, analysis tools for experimental data sets include experiment browsing; full data set ambiguity reports and assignment functions; data set comparisons and novelty analysis; data set summary features, which can be linked directly to subset evaluation; and subset selection and evaluation by metadata or data queries. *C*, most analysis and search tools rely on three fundamental data views: protein pages, ambiguity pages, and enrichment pages of subset selection features.

storing PTM data, SysPTM (6) brings bioinformatics resources to bear on PTMs in the context of the PTM compendium, such as mapping known PTMs onto signaling pathways (6). Other bioinformatics tools specific to phosphorylation exist as well: for example, Phospho.ELM, PHOSIDA (5), Scansite (8), KinasePhos (9), and PPSP (10) contain predictions or annotations of the kinase responsible for the phosphorylation of particular sites. Unfortunately, there is no wide ranging resource that currently allows users to browse the data from diverse high throughput PTM experiments (with the exception of PHOSIDA, which is specific to experiments done by Mann and co-workers (5)).

Despite the daunting volume of PTM measurement by MS, the lack of computational methods for deriving experimental hypotheses from these data sets has become a bottleneck limiting the contribution of high throughput PTM study to biological understanding. To address this limitation, a few studies have implemented unsupervised learning techniques as a method of reducing data dimensionality to elucidate dynamic and functional patterns in phosphoproteomics measurements (1, 11). These methods were successful in highlighting functionality of novel protein modifications, but the vast majority of uncharacterized PTMs remain without putative biological function even after unsupervised learning.

We have developed PTMScout in an effort to bring hypothesis generation tools into proteomics studies of PTMs while leaving each peptide in the larger context of the phosphoproteome, acetylome, etc. PTMScout provides an interface to a novel database of post-translational protein modifications that incorporates functional annotations and mRNA expression profiles. Individual experiments and their quantitative measurements are treated as unique entities uploaded by their producers for availability to the broader scientific community. These data may then be analyzed through PTMScout via subset selection by functional or dynamic annotation along one axis followed by identification of statistically significant enriched annotations along all other axes. Additionally, by utilizing expression profiles and multiple existing data sets, PTMScout provides information regarding the assignment of individual peptides to particular proteins in cases where there are a number of proteins or protein isoforms from which a peptide may have been cleaved.

The main functions of PTMScout are 6-fold (see Fig. 1 for a depiction). First, PTMScout allows browsing of experimental

data alongside publically available annotations, such as gene ontology (GO) terms and protein domain structures. Second, PTMScout allows for direct comparison of a particular experiment with one or more other experiments, including highlighting novel sites of modification. Third, PTMScout highlights potentially contentious assignments of peptides to proteins and gives biologists tools, such as tissue and cellular mRNA expression data, to determine an optimum protein assignment. Fourth, PTMScout allows for data set reduction by subset selection either on the data itself or on its imported annotations. Fifth, PTMScout provides automated statistical significance testing of a number of metrics orthogonal to the selection criteria (including quantitation in dynamic data and external annotations) in selected subsets. Finally, PTMScout provides an interface to unsupervised learning and automatic partition labeling based on statistically significant enriched information.

Example use cases of PTMScout for generating biological insights from a number of previously published data sets will be shown using the data described in Table I, which includes a group of data sets with quantitatively measured conditions in cells stimulated with EGF ligand (data sets EGF4 (1), EGF7 (12), and HER2 (13)) as well as a discovery data set from acetylated intracellular proteins lacking quantitative information (data set AcK (14)). Using the readily available tools within PTMScout (Fig. 1), we were able to construct multiple biological hypotheses regarding the potential functional characterization of multiple PTM sites, including a role for phosphorylation of Tyr-497 on Fyn-related kinase (FRK) in EGFR proliferation by subset selection in the EGF4 (1) data set. We were also able to find potential signal integrators between focal adhesions and the EGFR pathway by using a combination of subset selection and enrichment based on dynamics as well as metadata selection and the ability to view data on a protein across data sets. Moreover, peptide assignment ambiguity tools were used to indicate a preferable protein assignment for the Src family kinase activation loop phosphorylation event. An interface to arbitrary data set clustering was used to recapitulate unsupervised learning results from the EGF4 (1) data set, and subset selection by quantitative data was used to expand the endocytic signaling module in this data set. Using data set comparison tools, the degree to which the AcK (14) data set expands our current knowledge of acetylation is quantified. By using subset selection based on protein sequence and previously described acetyltransferase sequence recognition, we demonstrate that CBP/p300 acetylates both non-histone and histone proteins, and in particular, it targets RNA-binding proteins. Finally, the data set summary view demonstrates that there may be acetylation sites missed by using trypsin as the proteolytic enzyme prior to MS measurement.

EXPERIMENTAL PROCEDURES

Database and Data Sources—The master database underlying PTMScout was built using MySQL. The database schema is outlined in supplemental Fig. 1. External, publically accessible protein information, including sequence, alternate accessions, gene names, and species, is retrieved from NCBI GenPept and RefSeq (15), International Protein Index (16), or Swiss-Prot (7), depending on the accession type given in a new data set. Gene ontology terms are from The Gene Ontology consortium (17). Species-specific annotation files and the current ontology file are downloaded from the Gene Ontology web site, and GO programming packages were used to parse annotation and ontology files. GO terms based on inferred electronic annotations, which have not undergone further curation, are not stored in the PTMScout database. Results in this study were produced using downloaded files from GO version 1.2 and annotation files retrieved December 12, 2009. Protein domain information comes from Pfam (18). When possible, Swiss-Prot identifiers are used to parse domains from the current Pfam release. When lookup in the stand-alone Pfam release is not possible, the Pfam-A hidden Markov model library and the BioPerl Hmmpfam package are used to predict domains in a protein sequence. Predictions with scores less than 10^{-5} are considered, and when there is overlap between domains, the domain with the most stringent score is kept. Results in this study were produced from Pfam Release 23. When a phosphorylation site falls into a predicted structural domain, we include this as a separate annotation of enrichment denoted "Pfam site." Gene expression information comes from the Genomic Institute of the Novartis Research Foundation (GNF) SymAtlas project (19). Expression information, analyzed by gcRMA, for human and mouse tissue types and the NCI60 cell lines was downloaded and placed in PTMScout as expression tables. GNF SymAtlas annotation tables are used to link PTMScout proteins with appropriate tissue/NCI60 mRNA expression.

For phosphorylation sites, PTMScout currently includes predictions of the responsible kinases and binding partners from Scansite (8) when available. Scansite predictions for an input peptide sequence are automatically retrieved, parsed, and then stored in prediction tables of PTMScout. PTMScout Scansite prediction stringencies correspond with suggested scores from Scansite. Ambiguous peptide-protein assignments are identified by exact match of the full-length peptide sequence identified by MS among all of the protein data sources imported to PTMScout, which is also expanded to include proteins within the relevant species by searching the RefSeq (15) database for a peptide match.

Curated data sets of phosphorylation sites were obtained from Phospho.ELM (3) and PhosphoSite (4) by request. Automatic curation of UniProt (7) for phosphorylation and acetylation is performed by searching for both large scale analysis terms (example search, "phosphorylation large scale analysis at") as well as modified residues (search term, "MOD_RES"). UniProt search results were then parsed and placed into a PTMScout loadable format.

Kinase activation loop predictions are based on finding the conserved amino acid sequence "DFG" to the N-terminal side of the modification and a flanking "APE" to the C-terminal side (20). This exact requirement matched ~58% of all kinase domains in the PTMScout database version 1.1 as of January 2009. We used ClustalW2 (21) to align all kinase catalytic domains within PTMScout, which at the time included 306 domains. We found that 180 of them had both the conserved DFG and APE, whereas 70 had only the conserved DFG sequence. Those that do not match the motif exactly usually have partially conserved flanking sequences, such as "DYG" or "SLE." On average, the two surrounding motifs were within 25 amino acids of each other, and 83% of proteins contained the motifs within 22–27 amino acids of each other. Based on the resulting ClustalW2 alignment, we developed a set of rules for identifying

activation loop modifications. First, if the amino acid sequence surrounding the modification site contains a DFG and an APE motif or degenerate sequences "D(F/P/L/Y)G" and "(A/S/P)(P/I/L/W)(E/D)" spanning less than 35 amino acids, it is marked confidently as being within the activation loop. If degenerate matches are made and are more than 35 amino acids apart, then it is marked as *potentially* being within the activation loop.

Calculations—Selection of foregrounds occurs at the level of proteins, the level of experimentally measured peptides, or the level of individual sites of post-translational modification, depending on the category of data or annotation being used for the selection. For example, gene ontology terms and structural domain criteria will select subsets at the protein level, whereas quantitative data will select at the measured peptide level, and enzyme specificity predictions and sequence motif features will select at the PTM level. We define the p value for enrichment of a characteristic in a foreground relative to a background as the probability that a characteristic would be as enriched, or more enriched, if the foreground were randomly selected from the full data. This quantity can be calculated exactly using the hypergeometric distribution. The probability of having k or more labels in the foreground occurring by random chance when we choose any n objects from the background, size N, having a total of K objects with that same label is calculated as follows.

$$p(k') = \sum_{k'=k}^{\min(n,K)} \frac{\binom{K}{k'}\binom{N-K}{n-k'}}{\binom{N}{n}} \qquad \text{(Eq. 1)}$$

To determine k, K, n, and N for a label, translation from the selection criterion specificity to the label specificity of interest is performed. Not all mappings of selection specificity to label specificity are 1:1. For example, quantitative measurement selection may lead to redundant selection at a protein level. A search for significantly enriched amino acid sequence motifs was performed using a previously published greedy search algorithm with a search index of ± 7 amino acids surrounding the site of modification and a branch cutoff term of 0.01 (22).

Categorical multiple hypothesis correction can be user-corrected through the PTMScout interface. Bonferroni correction (23) is the most stringent correction method where the corrected α is the desired p value divided by the number of labels tested. The false discovery rate is implemented according to the method of Benjamini and Hochberg (24).

PTMScout can be found on line at http://ptmscout.mit.edu. A tutorial for using PTMScout to obtain the results presented in this study can be found in the PTMScout documentation. Unless otherwise noted, results are from version 1.2 of PTMScout, which includes Pfam Release 23, gene ontology annotations from version 1.2 downloaded on December 12, 2009, and UniProt compendium results from Release 15.11. All protein records in PTMScout version 1.2 have been retrieved after December 11, 2009. All terms considered enriched in the results have an false discovery rate-adjusted p value of 0.05 or better unless noted otherwise.

RESULTS

PTMScout is a web application that provides access and a computational interface to an underlying MySQL database. The PTMScout database contains data from high throughput studies of protein modifications and existing PTM compendia (Fig. 2*A*). Phosphorylation and acetylation experiments are currently included, but PTMScout has been designed to incorporate additional modification types as they become available experimentally. The database incorporates information at the protein level, such as GO terms (17) and Pfam domain structures (18) (Fig. 2*C*) as well as information at the level of individual sites of modification, such as Scansite (8) predictions of the enzymes responsible for or proteins interacting with modification sites, short peptide sequences aligned around the site of modification, and the predicted Pfam domain in which a site falls if any (Fig. 2*B*). The information contained in the database and the level of its specificity are depicted in Fig. 2 using an example peptide measured on JAK2 (Swiss-Prot accession number O60674) after stimulation by EGF (12). In addition to high throughput data sets of measured modifications, PTMScout incorporates larger, curated data sets of known post-translational modifications (3, 4, 7) for easy comparison of a new experiment with the current state of knowledge for a particular modification or modifications on a particular protein. PTMScout version 1.2, at the time of this writing, included 16 unique data sets, 11 experimental and five compendia, totaling 224,072 modifications across 72 species (133,440 phosphoserine, 38,906 phosphothreonine, 34,149 phosphotyrosine, and 17,577 acetyllysine).

The scale of high throughput proteomics PTM data is approaching that of genomics data, leading to similar problems as it is difficult to derive biological meaning from large data sets without a specific prior hypothesis. To address this challenge, PTMScout tools allow users to partition their data into a more comprehensible format using subset selection in one of four ways. First, users can select a subset of data that is annotated with a particular label from an imported data source (such as a GO term). Second, users can select a subset of data based on quantitative characteristics of the experimentally measured data. Third, users can partition the experiment by an external means, such as unsupervised learning, and import the partitioning scheme to PTMScout. Finally, the user can combine any or all of the first three methods to create a subset of data. Once a data subset is selected by one of these means, the statistical significance of enrichment with respect to the full data set is automatically calculated according to a hypergeometric distribution for all other metadata annotations as well as some qualitative characteristics of data dynamics. The fundamental philosophy of this method is that subset selection, partitioning, or clustering in one feature dimension could produce a biological hypothesis highlighted by a statistically significant enrichment of a term or feature in another dimension.

Features for subset selection and enrichment include GO terms, Pfam domains (at a protein level as well as a site level), kinase and binding domain predictions from Scansite, local sequence features, and measured quantitative features. Many features are categorical, and selection and enrichment significance calculations are straightforward as detailed under "Experimental Procedures." Selection and enrichment of quantitative data and sequence features require special mention because of their increased complexity. We allow for local

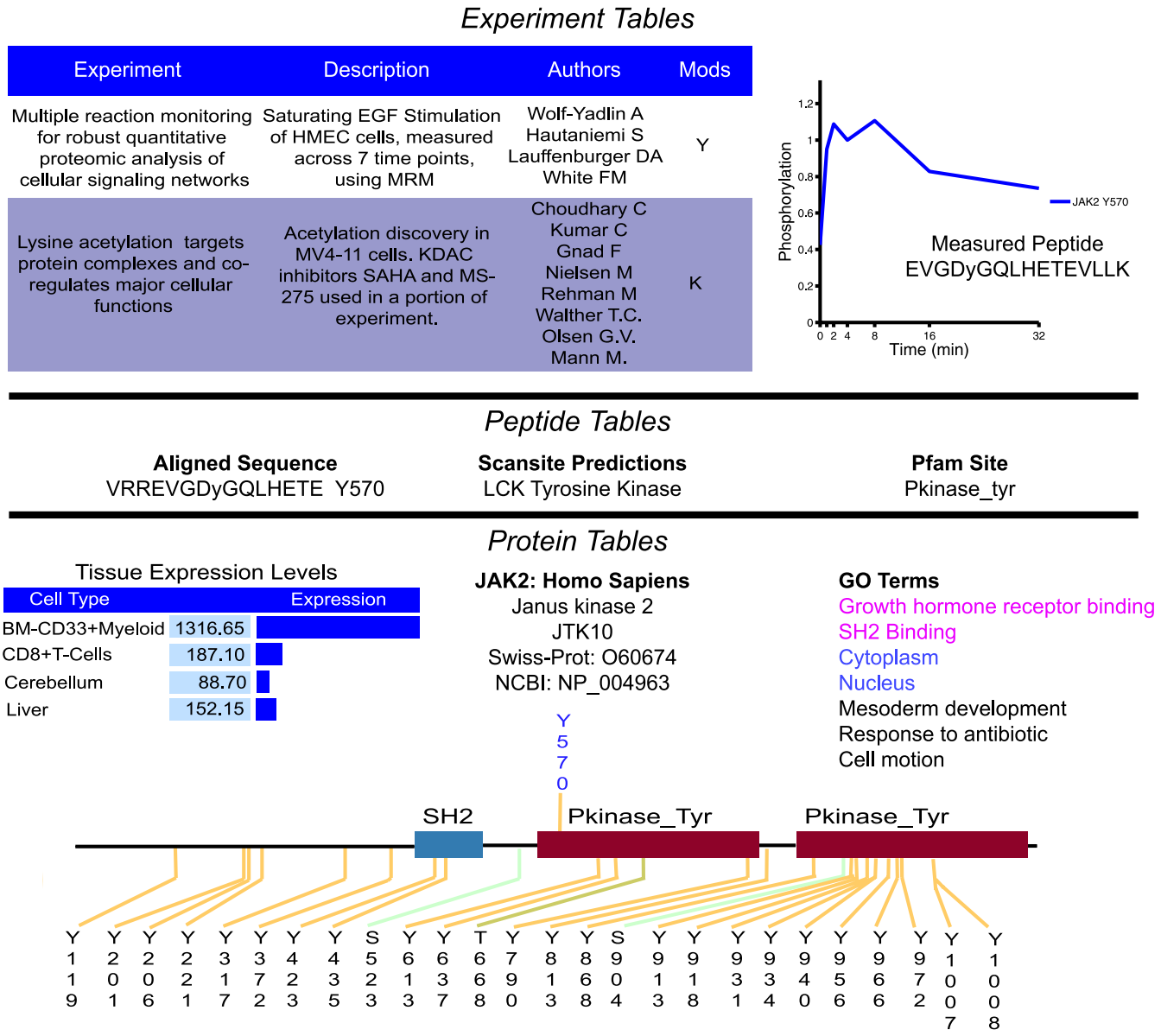


FIG. 2. **Abstracted database schema of PTMScout with example peptide measured from JAK2 in EGF7 experiment (12).** The database consists of three major classes of information. *A*, experimental data are the basic data elements and consist of all measured modified peptides in an experiment and their associated quantification when applicable. *B*, PTMScout site-level information includes the 15-mer sequence of peptides centered on individual sites of modification, the domain within which the site falls in its protein (Pfam_site), and any predictions on the function or regulation of that individual modification. *C*, protein-level information includes GO annotations, Pfam domain structures, GNF SymAtlas mRNA expression information, and a variety of database accessions, gene names, and protein names. Peptide and experiment data in the database are connected through protein identifiers, allowing for direct comparison across data sets and compendia.

sequence feature selection by regular expression queries. For example, to search for a common SH2 binding motif containing a phosphotyrosine with a hydrophobic amino acid in the +3 position, one can choose to search for "y . . (P/L/I/V/M)" (phosphotyrosine (y) followed by any two amino acids followed by a proline, leucine, isoleucine, valine, or methionine). In the reverse problem, searching for meaningful linear amino acid sequences in a subset, *e.g.* to establish an enriched motif, requires an algorithm to reduce the search space to a feasible size. For this purpose, PTMScout uses a previously developed greedy search motif algorithm (22). Additionally, PTMScout implements flexible search queries for quantitative data fields by allowing the user to create simple mathematical expressions as selection criteria. For example, one can search for an "early response" subset of tyrosine phosphorylation events in the EGFR signaling network (*e.g.* the EGF7 experiment (12); see Table I) by requiring a 4-fold change in the 1st min with the query "time(1 min) ÷ time(0 min) ≥ 4." PTMScout searches for quantitative data enrichment in a subset by testing for specific qualitative descriptors of quantitative dy-

TABLE I
Data sets used to demonstrate functionality of PTMScout
Reference name is for quick reference to the data set of interest. NA, not applicable.

Reference name	Data set name	Cell type	Stimulation	Measurements	PTM of interest	Data set size (peptides)
EGF7 (12)	Multiple reaction monitoring for robust quantitative proteomics analysis of cellular signaling networks	HMEC	EGF	0, 1, 2, 4, 8, 16, 32 min	Phosphotyrosine	222
EGF4 (1)	Quantitative proteomics analysis of phosphotyrosine-mediated cellular signaling networks; supplemental Table 1	HMEC	EGF	0, 5, 10, 30 min	Phosphotyrosine	77
HER2 (13)	Effects of HER2 overexpression on cell signaling networks governing proliferation and migration	HMEC, 24H	EGF, HRG	0-, 5-, 10-, 30-min EGF, HRG stimulation; parental, 24H cell lines	Phosphotyrosine	68
AcK (14)	Lysine acetylation targets protein complexes and co-regulates major cellular functions	MV-411	NA	NA	Acetyllysine	3,286

namic features: -fold change, maximum modification, interval of peak up-regulation, and interval of peak down-regulation among each of the quantitative data points. By rigorously defining a "dynamic feature space," we are able to calculate the statistical significance of enrichment of these labels in the same way we might calculate the significance of the representation of a GO term or kinase prediction in a subset. Finally, all query types can be combined, enabling the identification of, for example, a subset of sites that adhere to canonical SH2 binding motifs *and* are up-regulated within the first time point.

Activating Kinase Events in EGFR Pathway—Although *ab initio* prediction of the function of specific phosphorylation sites is difficult, typically phosphorylation within the activation loop of the catalytic domain of protein kinases can be expected to enhance activity of the kinase by driving a structural transition (20). To predict whether a PTM falls within a kinase activation loop, PTMScout searches for the conserved flanking amino acid sequences DFG on the N-terminal side and APE on the C-terminal side of the site of modification (20). This definition of the activation loop conservation was expanded by aligning the kinase catalytic domains of kinases within PTMScout using ClustalW2 (21) (see "Experimental Procedures"). Of the 2,089 tyrosine and serine/threonine kinase domains in PTMScout version 1.2, ~79% of the activation loops are confidently predicted, and with some certainty, another ~3.6% are predicted. Activation loops cannot be predicted for the remaining kinases based on searching for conserved flanking sequences.

We used PTMScout to explore the subset of kinase catalytic domain modifications in the EGF4 experiment (1) by performing a metadata query requesting PTMs that fall into the Pfam domains "Pkinase" (serine/threonine kinase domains) *or* "Pkinase_tyr" (tyrosine kinase domains). Fig. 3 illustrates the query results, which include modifications within the catalytic domains of the kinases CDC2 (Swiss-Prot accession number P06493), EPHA1 (NCBI accession number NP_005223), EPHA2 (Swiss-Prot accession number P29317), GSK3A (Swiss-Prot accession number P49840), INSR (NCBI accession number NP_000199)/IGF1R (Swiss-Prot accession number P08069), MAPK1 (Swiss-Prot accession number P28482), MAPK3 (Swiss-Prot accession number P27361), PRPF4B (NCBI accession number NP_789770), and PTK2 (Swiss-Prot accession number Q05397). PTMScout predicts that all of these except the site on CDC2 fall into the activation loop of their respective kinase and are therefore potentially kinase activating events. Additionally, phosphorylation levels for each of these sites increased, albeit slightly in some cases, upon stimulation by EGF (see Fig. 3*B*). This subset is significantly enriched, based on a false discovery rate-corrected value of 0.01, for proteins that contain Pkinase domains relative to the full EGF4 data set (see Fig. 3*C*). It is interesting to note that although all modifications to serine/threonine kinases in the data set occur within the catalytic domain and are represented in this subset less than half of the measured modifications to tyrosine kinases occur within the catalytic domain.

We examined the remaining phosphorylation events on tyrosine kinases by choosing a subset based on proteins with "Domains = Pkinase_tyr." In addition to the four phosphorylation sites that occur within tyrosine kinase catalytic domains, another 16 sites are found on 10 tyrosine kinases, including Tyr-497 on FRK (Swiss-Prot accession number P42685), a Src family kinase. This site falls on the C-terminal tail of the protein, which, based on relative proximity between the C terminus of the protein and the kinase domain, is similar in location to the negative regulatory site (Tyr-527) of Src (25). Alignment of the 11 Src family kinases indicates all but one member, SRM (Swiss-Prot accession number Q9H3Y6), contain a tyrosine in this region of the protein, and there is evidence for phosphorylation on all of these sites (supplemental Table 1). By extension, it is reasonable to pre-

A

Protein	Site	Activation Loop
CDC2	Y15	No
EPHA1	Y781	Yes
EPHA2	Y772	Yes
GSK3A	Y279	Yes
INSR	Y1189	Yes
MAPK1	Y187	Yes
MAPK3	Y204	Yes
PRPF4B	Y849	Yes
PTK2	Y576	Yes

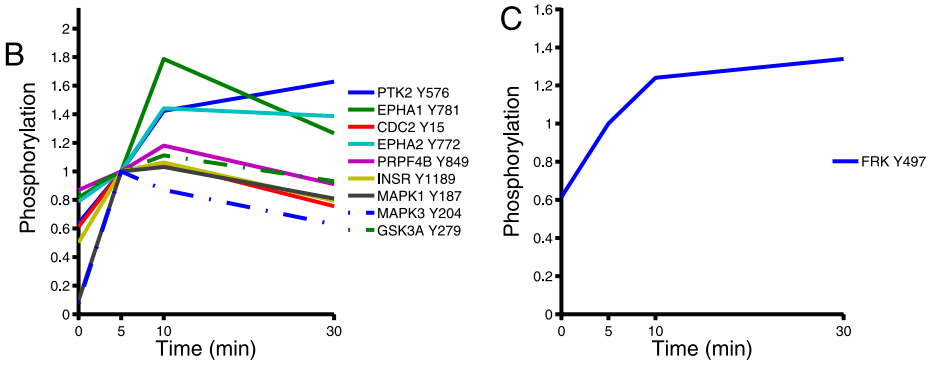
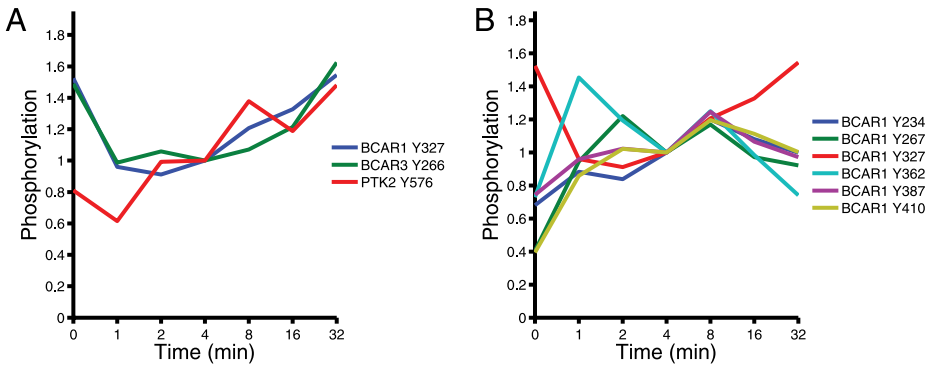


FIG. 3. **Subset selection of kinase catalytic domain phosphorylation sites in EGF4 data set (1).** *A*, nine sites were found to occur within kinase catalytic domains. With the exception of CDC2, all sites are predicted to fall within the activation loop of their respective kinases. *B*, quantitative measurement graph of the nine phosphorylation sites indicates they are all responsive to EGF stimulation to some extent. *C*, quantitative dynamic measurements of another tyrosine kinase phosphorylation site on a Src family kinase, FRK, that falls outside of the kinase catalytic domain on the C-terminal portion of the protein. Because of its similarity to the negative regulation site of Src, this may indicate that negative regulation of FRK increases after stimulation by EGF.

dict that Tyr-497 on FRK may bind the SH2 binding domain of FRK, thereby inhibiting kinase activity of FRK. Among the 10 Src family kinases with known phosphorylation sites on a tyrosine in the C-terminal tail of the protein, the sequence surrounding Tyr-497 on FRK is the most dissimilar, potentially indicating that FRK Tyr-497 may be phosphorylated by a kinase different from the phosphorylating kinase(s) for the analogous sites on other family members. Phosphorylation of FRK Tyr-497 increases 2-fold by 30 min after stimulation by EGF, which indicates that EGF stimulation may cause this particular Src family kinase to decrease in catalytic activity. Gene ontology annotations for FRK indicate nuclear localization and involvement in the negative regulation of cell cycle progression. If suppression of cell cycle progression is dependent on its kinase activity, then this phosphorylation site may be one specific mechanism by which EGF stimulation enhances cell growth and proliferation.

Focal Adhesion Signaling in Response to EGF—The flexible query interface of PTMScout allows users to apply intuitive rules for defining a subset of interest based on the features inherent in any particular data set. For example, using the data in the EGF7 experiment (12), we selected a subset of phosphorylation sites that are immediately down-regulated in response to EGF stimulation, a rare event. To generate this subset, we required that base-line phosphorylation be at least 30% higher than at 1 min after stimulation: "time(0 min) ÷ time(1 min) ≥ 1.3." This query resulted in the selection of only three phosphorylation sites on three proteins: BCAR1 (NCBI accession number NP_055382) Tyr-327, BCAR3 (Swiss-Prot accession number O75815) Tyr-266, and PTK2/FAK (Swiss-Prot accession number Q05397) Tyr-576. PTMScout shows that Tyr-576 falls within the activation loop of the kinase domain of PTK2/FAK. Dynamics for these sites after EGF stimulation (shown in Fig. 4*A*) indicate that these sites immediately decrease upon EGF introduction but recover within 32 min. The other functional annotations enriched in this cluster indicate a function in integrin-mediated signaling and localization at the focal adhesions (GO biological process term "integrin-mediated signaling pathway" and GO cellular component term "focal adhesion"). BCAR1, BCAR3, and FAK have been implicated as important signaling molecules with involvement in both EGFR and focal adhesion signaling pathways (26, 27). Although the mechanism underlying decreased phosphorylation of these sites following EGFR activation is still unclear, our findings may indicate specific phosphorylation events involved in EGFR/focal adhesion cross-talk.

BCAR1 has five additional sites of phosphorylation, which increases in response to EGF treatment in the EGF7 experiment: Tyr-234, Tyr-267, Tyr-362, Tyr-387, and Tyr-410. Phosphorylation events on BCAR1 were chosen as a subset by selecting "protein name = BCAR1" (see Fig. 4*B* for quantitative measurements of all six sites in response to EGF treatment). All of the phosphorylation sites measured on BCAR1 in this data set have high sequence similarity as shown by motif analysis: all have proline in the +3 position relative to the phosphotyrosine, and all but one site additionally contains an aspartic acid in the +1 position. CRK and NCK are known to bind to BCAR1 (28), and enrichment analysis indicates that



C

Motif	Subset*	Full Expt**	P-Value
.....y..PP...	4	4	1.23 x 10-7
.....yD.P....	**5**	**9**	**1.33 x 10-7**
.....yDOP....	4	6	1.82 x 10-6
.....-.yD.P....	4	7	4.22 x 10-6
.....y.OP....	5	16	4.49 x 10-6
.....-.yD.PP...	3	3	9.49 x 10-6
.....y.VPP..	3	3	9.49 x 10-6
.....yD.P...O	3	3	9.49 x 10-6
.....yDVP....	3	3	9.49 x 10-6
.....Vy..PP.A.	3	3	9.49 x 10-6
.....y..P....	**6**	**40**	**1.80 x 10-5**
.....y.VP....	4	10	2.48 x 10-5
.....OyD.P....	3	4	3.76 x 10-5

* out of 6
** out of 234

FIG. 4. **BCAR1 subset selection and enrichment in EGF7 data set (12).** Dynamics are plotted along a $\log_2$-transformed axis for early time point clarity. *A*, dynamics of the only three phosphorylation sites in the EGF7 data set that decrease in the first min by 30% or more. These sites are enriched for integrin-mediated signaling pathway annotation in GO biological process annotations. *B*, dynamics of all six phosphorylation sites on BCAR1, all of which increase in response to EGF except Tyr-327. *C*, motif enrichment of all six phosphorylation sites compared with the full data set. The number of sites matching the motif in the subset and the full data set are given along with the significance of that enrichment. All sites have a +3 proline, and all but one have an aspartic acid in the +1 position.

the majority of CRK and NCK binding events that occur downstream of EGFR in this data set occur on BCAR1 based on enrichment for CRK and NCK binding predictions by Scansite. Given the apparent redundancy of phosphospecific binding functionality of the six phosphorylation sites measured on BCAR1, it is interesting that one site has a completely opposing dynamic response to EGF stimulation. In integrin-mediated signaling complexes, BCAR1 is associated with at least three tyrosine kinases, Src, FAK, and Abl (29). Although specific kinase targets on BCAR1 are not clearly mapped, we see that the rare dynamic of BCAR1 Tyr-327 correlates with a similar decrease in phosphorylation on the activating site of FAK, Tyr-576 (26), possibly indicating a difference in enzymatic control of Tyr-327 compared with the remaining five phosphorylation sites.

Physical aspects of EGF addition (*e.g.* shear stress from addition of the solution containing EGF and swirling of the medium) could be responsible for mechanotransduction-related signaling events at the focal adhesions *versus* a signaling response due to the growth factor itself. PTMScout has the ability to plot all quantitative measurements of modifications on a particular protein across all experiments contained in PTMScout. Across multiple experiments, Tyr-327 is consistently down-regulated in response to stimulation by EGF (see supplemental Fig. 2); however, in the HER2 experiment (13), Tyr-327 decreases in response to EGF but increases in response to HRG. Addition of HRG should produce similar mechanical cues (see above) compared with the introduction of EGF, so these results indicate that the decrease in phosphorylation on these focal adhesion signaling molecules is probably EGF-specific rather than a consequence of mechanical handling.

Assignment of Src Family Kinase Activation Loop Phosphorylation Sites—Proteolytic peptide fragments can present an assignment problem as peptides can often match multiple proteins within a proteome. This ambiguity typically occurs when there are multiple isoforms or multiple gene products with a high degree of similarity surrounding sites of modification. For quantitative data, there is no clear way to deconvolute the degree to which each protein contributed to a particular peptide measurement without further intensive experimentation. To address the issue of ambiguous peptide/protein assignments, PTMScout generates an automated report that allows users to immediately see all peptides within an experimental data set that could have been assigned to multiple proteins. Additionally, while viewing any peptide assignment in the data set throughout PTMScout, potentially ambiguous assignments are highlighted for the user, and information is presented that may help with selecting a particular protein among many choices. Specifically, for every protein that may have contributed to the peptide measurement, PTMScout illustrates 1) the degree to which other data sets and compendia have annotated a protein, 2) the extent of GO annotations, 3) protein domain structure, and 4) tissue

expression available for mouse and human tissues as well as NCI60 cell lines incorporated from GNF SymAtlas (19). New protein assignments can be made using a web form, exported as a new data set, and then reloaded as a child experiment of the original. This process allows scientists to explore a data set using the assignments they prefer while faithfully maintaining the assignments chosen in the initial load of the data set to PTMScout.

To demonstrate the usefulness of the peptide assignment tools of PTMScout, we examined the possible protein assignments of the trypsinized fragment representing activation loop phosphorylation of several Src family kinases from the EGF7 data set (12). Although the initial assignment of trypsinized peptide "LIEDNEyTAR" was to the proto-oncogenic tyrosine kinase LCK (Swiss-Prot accession number P06239), based on its sequence, the measured peptide could belong to any of the proteins LCK, YES1, FYN, or SRC. All of these proteins are Src family kinases, but closer examination of their individual characteristics can help make a more informed protein choice. Although FYN, SRC, and YES1 are expressed ubiquitously across all cell types, LCK is only highly expressed in leukocytes and T-cells according to data imported to PTMScout from SymAtlas (19) (supplemental Fig. 3). Because the EGF7 experiment was performed on human mammary epithelial cells (HMECs), which express an extremely low level of LCK mRNA, it is unlikely that the peptide measured resulted from the cleavage of LCK. Among FYN, SRC, and YES1, based on relatively similar mRNA expression in epithelial type cells, FYN is the protein with the most GO annotations. **There are three possible isoforms of FYN that match the given sequence isoforms A, B and C.** The majority of experiments and compendia have preferentially chosen isoform A, the canonical sequence of FYN. In the absence of any external confirmatory experiments, the combination of all relevant information indicates that the most informative selection for the peptide LIEDNEyTAR in the EGF7 data set is FYN isoform A (Swiss-Prot accession number P06241).

Unsupervised Learning Highlights Roles for Proteins in Endocytosis of EGFR—PTMScout can be used to explore the characteristics enriched in subsets created by unsupervised learning algorithms. The intent of unsupervised learning is to partition the members of a data set into clusters based on similar quantitative measurements. This data set reduction may then highlight interesting points of biology based on shared functionality in the cell. In addition, novel pathway components with unknown function can be hypothesized to share similar function or similar pathway effectors. An example of unsupervised learning applied to a quantitative PTM data set appears in the experimental study of the EGF4 experiment by Zhang *et al.* (1). Using a self-organizing map (SOM) (30), nine potential signaling modules were created based on similar dynamics. Two of these clusters were explored in depth: one cluster included EGFR Tyr-1173 as well as several proteins known to bind directly to the receptor, whereas the second cluster included several proteins known to be involved in endocytosis. The "endocytosis" cluster contained sites whose phosphorylation reached maximum levels relatively late at 10 min and were strongly dephosphorylated at 30 min. Two members of this cluster had no known function in the EGFR network; these proteins were known at the time as Ymer (Swiss-Prot accession number Q8IVM0) and Chr20 ORF18 (NCBI accession number NP_006453), also known as CCDC50 and RBCK1, respectively. Based on their grouping with phosphorylation sites on endocytic proteins STAM1 (Swiss-Prot accession number Q92783), STAM2 (Swiss-Prot accession number O75886), EPS15 (Swiss-Prot accession number P42566), ACK1 (Swiss-Prot accession number Q07912), and ANXA2 (Swiss-Prot accession number P07355), the authors proposed a role for Ymer/CCDC50 and Chr20 ORF18/RBCK1 in endocytosis and trafficking of EGFR. Since that time, RBCK1 was found to be involved in endocytic pathways following cytokine stimulation (31), and Ymer/CCDC50 was found to suppress ligand-mediated down-regulation of EGFR (32), thereby validating the original hypothesis derived from unsupervised learning.

The type of observation made in Zhang *et al.* (1) regarding a common endocytic functionality of several proteins in a cluster involves extensive familiarity with the proteins and intensive manual curation. To determine whether the PTMScout enrichment analysis of clusters could help bypass the onerous task of determining similarity, we evaluated the clustering solution given in Zhang *et al.* (1) through the arbitrary grouping interface of PTMScout. The interface to arbitrary data partitioning currently involves exporting a data set, appending it with cluster assignments, and then importing the appended file for enrichment analysis. The endocytic cluster highlighted in Zhang *et al.* (1) was enriched for proteins containing the domains "UIM" (ubiquitin interaction motif) and "VHS," which are both indications of endocytosis according to Pfam (18). Additionally, this cluster was enriched for dynamic terms "peak phosphorylation" at 10 min and "peak down-regulation" between 10 and 30 min. Therefore, PTMScout is capable of circumventing some of the time-consuming tasks of determining the shared functionality in a data set partition of unsupervised learning.

To further investigate the functional assignments of phosphorylation sites with temporal dynamics featuring peak activation at 10 min followed by a quick dephosphorylation between 10 and 30 min, we created a subset of data with these features using the data-driven subset selection interface. Specifically, we required phosphorylation at 10 min to be 10% greater than phosphorylation at 5 min and 50% greater than phosphorylation at 30 min. This produced the subset shown in Fig. 5, which includes the seven phosphorylation sites from the SOM endocytic cluster as well as phosphorylation sites on PTPN18 (Swiss-Prot accession number Q99952), TFRC (Swiss-Prot accession number P02786), and CAV1 (Swiss-Prot accession number Q03135). TFRC and

A

Protein	Site
ANXA2	Y30
CAV1	Y14; Y6; Y25
CCDC50	Y145
EPS15	Y849
PTPN18	Y389
RBCK1	Y288
STAM	Y381
STAM2	Y374
TFRC	Y20
TNK2	Y859; Y860

B

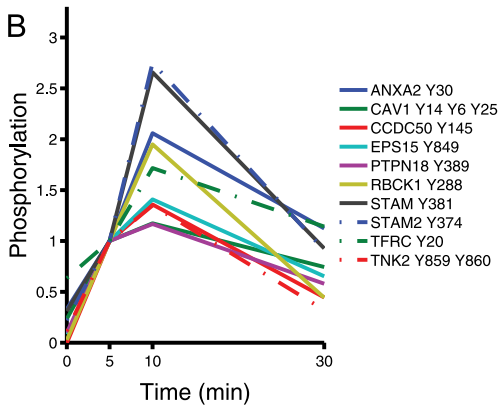


FIG. 5. **Extended endocytic subset from EGF4 data set.** *A*, the subset of the EGF4 data set where phosphorylation at 10 min is at least 10% higher than at 5 min and phosphorylation at 10 min is at least 50% higher than at 30 min. This subset includes all members of the SOM "trafficking" cluster from Ref. 1 as well as sites on PTPN18, CAV1, and TFRC. *B*, dynamics of the extended endocytic subset.

CAV1 are both known to be involved in endocytosis, but functional characterization for PTPN18 in the EGFR pathway has not yet been elucidated. Based on the inclusion in this cluster, PTPN18 may also participate in the endocytic pathway. Another possibility is that phosphorylation of Tyr-389 on PTPN18 may play a role in the negative regulation of EGFR because the temporal profile for this site follows so closely with the phosphorylation dynamics of the negative regulation machinery of endocytosis.

Trypsin Is Potentially Limiting in Measurement of Acetylation, and "(G/S)k" Is an Acetylation Motif Specific to RNA-binding Proteins—The AcK data set (14) published in August 2009 is one of the most recently loaded experiments in PTMScout. This single data set was larger than all other large scale measurements of acetylation recorded in UniProt (version 15.8, released in September 2009). Upon comparing it with curated acetylation data sets using the comparison tool implemented in PTMScout, we found that only ~2.5% of the ~3,200 acetylated peptides in the AcK data set have been previously detected. If we include UniProt annotation records that extend acetylation knowledge by predicted similarity among species, protein families, and other non-strict annotations, this fraction increases only to 5%. Although the AcK data set as originally published (14) contains 3,885 acetylated peptides, PTMScout contains only 3,286 acetylated peptides because the remaining peptides were given nucleotide record identifiers. PTMScout handles only proteomics accession types.

PTMScout allows users to view a breakdown of their data set by annotation terms incorporated in the database (such as GO annotations, domain structures, kinase predictions, etc.) through the Experiment Summary functionality of PTMScout. Table II represents the top terms for GO molecular function (MF) and predicted Pfam domains of the AcK data set (14). One of the top MF annotation terms is "none" (*i.e.* no MF annotation), indicating that many of the proteins acetylated in this data set are not yet annotated with regard to function. The domain information in Table II represents the number of proteins containing the indicated domains in the data set. As can be seen from this table, acetylation frequently occurs on proteins containing "RRM_1" domains, which are thought to be an indication of an RNA-binding protein (18). This information is consistent with the prevalence of the "RNA binding" term in the GO MF breakdown. Acetylation of histone proteins is present as expected, but there is also a significant degree of acetylation on signaling proteins as indicated by domains such as Pkinase, "SH3_1," "PH," and "SH2." Fig. 6 illustrates a motif logo (33) for the entire AcK data set, indicating the amino acid frequencies surrounding the site of acetylation.

TABLE II

Terms for GO molecular function and Pfam domains for the AcK experiment (14) with highest incidence

The number of terms present in the data set represents the total number of proteins with that GO term or at least one of the indicated domains. There are 1,662 unique proteins in the data set.

GO: MF	No. proteins	Pfam domain	No. proteins
Protein binding	664	None	56
None	401	RRM_1	55
RNA binding	62	WD40	38
DNA binding	52	Pkinase	37
ATP binding	50	Helicase_C	37
Protein homodimerization activity	44	PHD	28
Molecular function	44	DEAD	26
Transcription factor activity	41	AAA	20
Identical protein binding	40	SH3_1	20
Structural constituent of ribosome	39	PH	19
Transcription coactivator activity	31	Bromodomain	17
ATPase activity	29	zf-C2H2	15
Transcription factor binding	27	TPR_1	15
Unfolded protein binding	26	CH	15
Protein N terminus binding	23	efhand	15
Protein C terminus binding	23	SH2	14
Zinc ion binding	22	SAP	13
Transcription activator activity	20	Histone	13
Calcium ion binding	19	Myb_DNA-binding	11
Enzyme binding	18	UQ_con	11
Transcription corepressor activity	18	HEAT	11
GTPase activity	17	Ank	11
Protein heterodimerization activity	16	zf-C3HC4	10
Ubiquitin-protein ligase activity	16	PCI	10
Actin binding	16	Filament	10
Translation initiation factor activity	16	Cpn60_TCP1	10
Single-stranded DNA binding	16	PWWP	9
Protein-serine/threonine kinase activity	15	Mito_carr	9

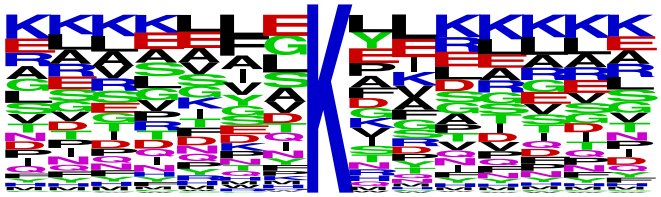


FIG. 6. **Summary.** The at-a-glance feature for the AcK data set (14) includes a frequency motif logo (33) representation of all singly acetylated sites aligned on the central modified residue. There is a high frequency of lysines in all positions except those immediately proximal to the central residue.

Surprisingly, there is an abundance of lysines surrounding the central modified lysine with the exception of those positions most proximal (−3 to +2) to the central residue. This systematic enrichment of lysine in the vicinity of acetyllysine indicates that trypsin, which cleaves peptides to the C-terminal side of lysine and arginine residues, may not be the most efficient protease for high throughput analysis of the acetylome because it may be producing peptides too small to be analyzed and sequenced by reverse-phase liquid chromatography-mass spectrometry (LC-MS). Clearly this technique was successful in identifying thousands of sites; however, the motif analysis would suggest that an improved approach might be to combine several samples processed using different proteolytic enzymes to achieve a more comprehensive coverage of the acetylome.

Acetyltransferases, like kinases, are thought to recognize linear amino acid sequences surrounding the site of modification (34). The motif (G/S)k (an acetyllysine (k) preceded by a glycine or serine) was found to be a consensus sequence for two related acetyltransferases, CBP and p300, using direct substrate identification with recombinant acetyltransferase (35). In addition to identification of a motif for these acetyltransferases, the authors looked for an expanded role for acetyltransferases beyond the canonical roles of histone and transcription factor modification and found proteins, such as Rch1, a nuclear importin, to be acetylated by CBP. We chose to look at the subset of acetylated sites in the large AcK experiment (14) that matched the consensus motif of CBP/p300 by searching for (G/S)k sequences. This subset selection returned 656 acetyllysine sites, ~20% of the entire data set, of which 292 were "Sk" sites and 364 were "Gk." Histone proteins, as identified by the presence of a "histone" domain, are not significantly enriched in this subset in agreement with the hypothesis that CBP/p300 can acetylate both histones and non-histone proteins. However, proteins responsible for acetylation of histone proteins are enriched in the subset of acetylation sites possessing the (G/S)k motif as indicated by enrichment of GO biological process terms histone H3/H4-K5/H4-K8/H4-K12 acetylation and GO cellular compartment term "histone acetyltransferase complex." Additionally, acetylation of RRM_1 domain-containing proteins is enriched within this subset specifically for acetylation within the domain itself. If (G/S)k is indeed specific to CBP/p300 recognition, our PTMScout results indicate that CBP/p300 is responsible for acetylation of RNA-binding proteins and proteins indicated to be involved in histone acetylation.

DISCUSSION

PTMScout provides uniform, web-based access to MS-measured PTM data and automates much of the feature selection and information extraction that are currently performed manually following MS analysis of biological samples. For example, residue position assignment and comparison with PTM data compendia for discovery of novel PTM measurements are intensive manual operations that are performed automatically in PTMScout. Additionally, programmatic access of protein databases by PTMScout during data set loading allows for protein assignment error checking, thereby correcting typical errors in protein assignment, including redirected records, updated records with significantly changed sequence information, and species assignment errors. Although PTMScout can automatically handle most protein record redirections, protein errors causing terminal failures due to sequence mismatch between the peptide and the assigned protein are reported in an error log; erroneous species assignments can be seen easily in the data set summary function of PTMScout. Furthermore, PTMScout allows for user-defined uploading of their own mass spectrometry data sets. Currently, PTMScout does not support private data in the database, but it is a goal for future versions of PTMScout to allow for a central repository and private repositories of experimental data sets.

Data analysis by subset selection and subsequent enrichment have proven to be useful tools for deriving biological hypotheses. However, hypothesis generation is currently limited by metadata annotations. For example, the endocytic cluster found in the EGF4 experiment (1) had only a few GO annotations indicating a role in endocytosis despite several reports demonstrating that the majority of proteins in the cluster participate in the endocytic pathway. Despite these limitations, the cluster featured enrichment of UIM and VHS domain-containing proteins, thereby enabling hypotheses regarding phosphorylation of specific sites and regulation of endocytosis. An expanded endocytic cluster was generated in PTMScout through use of relative quantitative dynamics, leading to identification of another protein that may be involved in EGFR endocytosis. Although the richness of hypotheses and observations is expanded by inclusion of relative quantitation across multiple conditions, PTMScout is also successful at deriving insight from data sets without quantitation as demonstrated with the AcK data set (14).

By considering the composition of the entire data set, enrichment testing provides a way to uniquely label a data set partition. An interesting cluster highlighted in the EGF4 study was an "early response cluster," which was composed of several known EGFR-binding proteins. Interestingly, enrich-

ment for quantitative dynamic features failed to corroborate this feature as being specific to that cluster alone. Despite the fact that all members of the cluster experienced a large increase in phosphorylation within the first 5 min of stimulation, this early response label is applicable to more than two-thirds of the entire data set, and therefore, although this label is correct, it is not a unique feature of that cluster compared with the remaining data set. By performing enrichment of subsets compared with the background of the data set itself *versus* the entire phosphoproteome or acetylome, experimental biases, such as antibody specificity or MS fragmentation patterns, are eliminated.

PTMScout is a widely and readily accessible, user-friendly, web-based PTM database with multiple bioinformatics tools to enable automated feature selection and subset generation. Here we have demonstrated the application of PTMScout to multiple published phosphorylation and acetylation data sets, leading to multiple hypotheses regarding the potential functionality of various proteins and PTM sites. As more experimental data sets are loaded into PTMScout, additional biological insight not currently available from individual data sets will emerge as we are able to compare the regulation and response of individual phosphorylation sites under a variety of conditions. Application of PTMScout to quantitative PTM data sets will facilitate the main data analysis challenge facing high throughput PTM proteomics and will provide putative functional assignments to a greater percentage of previously uncharacterized sites.

Acknowledgments—We thank Stephen Goldman of the BioMicro Center for server administration support, John Naegle for web site design consultation, and members of the White and Lauffenburger laboratories for helpful discussion.

* This work was supported, in whole or in part, by National Institutes of Health Grants U54-CA112967 and R01-CA096504.

S This article contains supplemental Figs. 1–4 and Tables 1 and 2.

§ Both authors contributed equally to this work.

‡‡ To whom correspondence should be addressed: Massachusetts Inst. of Technology, 77 Massachusetts Ave. 56-787A, Cambridge, MA 02139. Tel.: 617-258-8949; Fax: 617-258-0225; E-mail: fwhite@mit.edu.

REFERENCES

1. Zhang, Y., Wolf-Yadlin, A., Ross, P. L., Pappin, D. J., Rush, J., Lauffenburger, D. A., and White, F. M. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol. Cell. Proteomics* **4,** 1240–1250
2. Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5,** 79
3. Diella, F., Gould, C. M., Chica, C., Via, A., and Gibson, T. J. (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucleic Acids Res.* **36,** D240–D244
4. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4,** 1551–1561
5. Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8,** R250
6. Li, H., Xing, X., Ding, G., Li, Q., Wang, C., Xie, L., Zeng, R., and Li, Y. (2009) SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol. Cell. Proteomics* **8,** 1839–1849
7. UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37,** D169–D174
8. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31,** 3635–3641
9. 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
10. Xue, Y., Li, A., Wang, L., Feng, H., and Yao, X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* **7,** 163
11. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
12. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D. A., and White, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 5860–5865
13. Wolf-Yadlin, A., Kumar, N., Zhang, Y., Hautaniemi, S., Zaman, M., Kim, H. D., Grantcharova, V., Lauffenburger, D. A., and White, F. M. (2006) Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol. Syst. Biol.* **2,** 54
14. Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehman, M., Walther, T. C., Olsen, J. V., and Mann, M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* **325,** 834–840
15. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65
16. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
17. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29
18. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2008) The Pfam protein families database. *Nucleic Acids Res.* **36,** D281–D288
19. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 6062–6067
20. Nolen, B., Taylor, S., and Ghosh, G. (2004) Regulation of protein kinases: controlling activity through activation segment conformation. *Mol. Cell* **15**, 661–675
21. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948
22. Joughin, B. A., Naegle, K. M., Huang, P. H., Yaffe, M. B., Lauffenburger, D. A., and White, F. M. (2009) An integrated comparative phosphoproteomic and bioinformatic approach reveals a novel class of MPM-2 motifs upregulated in EGFRvIII-expressing glioblastoma cells. *Mol. Biosyst.* **5,** 59–67
23. Dudoit, S., and van der Laan, M. J. (2008) *Multiple Testing Procedures with Applications to Genomics*, Springer, New York
24. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57,** 289–300

25. Ayrapetov, M. K., Wang, Y. H., Lin, X., Gu, X., Parang, K., and Sun, G. (2006) Conformational basis for SH2-Tyr(P)527 binding in Src inactivation. *J. Biol. Chem.* **281,** 23776–23784
26. Calalb, M. B., Polte, T. R., and Hanks, S. K. (1995) Tyrosine phosphorylation of focal adhesion kinase at sites in the catalytic domain regulates kinase activity: a role for Src family kinases. *Mol. Cell. Biol.* **15,** 954–963
27. Schrecengost, R. S., Riggins, R. B., Thomas, K. S., Guerrero, M. S., and Bouton, A. H. (2007) Breast cancer antiestrogen resistance-3 expression regulates breast cancer cell migration through promotion of p130Cas membrane localization and membrane ruffling. *Cancer Res.* **67,** 6174–6182
28. Guan, J. L. (1997) Focal adhesion kinase in integrin signaling. *Matrix Biol.* **16,** 195–200
29. Mitra, S. K., and Schlaepfer, D. D. (2006) Integrin-regulated FAK-Src signaling in normal and cancer cells. *Curr. Opin. Cell Biol.* **18,** 516–523
30. Kohonen, T. (1990) The self-organizing map. *Proc. IEEE* **78,** 1464–1480
31. Tian, Y., Zhang, Y., Zhong, B., Wang, Y. Y., Diao, F. C., Wang, R. P., Zhang, M., Chen, D. Y., Zhai, Z. H., and Shu, H. B. (2007) RBCK1 negatively regulates tumor necrosis factor- and interleukin-1-triggered NF-kappaB activation by targeting TAB2/3 for degradation. *J. Biol. Chem.* **282,** 16776–16782
32. Tashiro, K., Konishi, H., Sano, E., Nabeshi, H., Yamauchi, E., and Taniguchi, H. (2006) Suppression of the ligand-mediated down-regulation of epidermal growth factor receptor by Ymer, a novel tyrosine-phosphorylated and ubiquitinated protein. *J. Biol. Chem.* **281,** 24612–24622
33. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190
34. Kouzarides, T. (2000) Acetylation: a regulatory modification to rival phosphorylation? *EMBO J.* **19,** 1176–1179
35. Bannister, A. J., Miska, E. A., Görlich, D., and Kouzarides, T. (2000) Acetylation of importin-α nuclear import factors by CBP/p300. *Curr. Biol.* **10,** 467–470