Production of a large scale phosphorylation site map in *D. melanogaster* Kc167 cells

As described in the main text we extensively separated the peptide samples derived from *D. melanogaster* Kc167 cells and used three different phosphopeptide isolation methods. We observed, in accordance with previous results from more limited analyses that the three methods isolated distinct, but overlapping parts of the phosphoproteome(Bodenmiller *et al*, 2007a). This underscores the need for using different phosphopeptide isolation methods combined with extensive peptide fractionation if a given phosphoproteome is to be extensively mapped.

Nevertheless the aim of mapping a whole phosphoproteome has clearly not been achieved. In the LC-MS/MS analyses we detected at least twice the number of phosphopeptide precursor ions than were identified. This is based on the number of precursor ions undergoing a neutral loss of -98 Da which is indicative for a phosphoserine or phosphothreonine containing peptide. Many of these peptides could not be conclusively identified due to the reduced phosphopeptides backbone fragmentation upon CID in ion traps(Aebersold and Goodlett, 2001; Bodenmiller *et al*, 2007b). A further obstacle to fully cover the phosphoproteome is that part of it is not detectable by mass spectrometry if trypsin is used as protease. Many kinase motives are lysine and arginine rich (e.g. the GSK-3, ERK1, ERK2, CDK5 substrate motif KpSPXXK or MLCK kinase substrate motif KKRXXpSX[R/K][R/K])(Peri *et al*, 2003). Therefore the resulting phosphopeptides are too short to be detected in the mass spectrometer. Also, many phosphopeptides are too long to be detected with standard mass spectrometric methods. For example the tryptic peptide containing S505 of dAKT1 is 57 amino acids long. Furthermore phosphorylations on tyrosine residues which are known to be of very low abundance (as in the case of CHICO) and are therefore very difficult to detect without prior enrichment using phosphotyrosine specific antibodies.

Thus the application of inclusion lists(Picotti *et al*, 2007; Stahl-Zeng *et al*, 2007) to sequence the non-identified phosphopeptides by MS2/MS3 experiments(Beausoleil *et al*, 2004),

multistage activation(Schroeder *et al*, 2004) and electron transfer dissociation(Syka *et al*, 2004) combined with the usage of further proteases for protein digestion and further phosphopeptide enrichment strategies would increase the number of identified phosphorylation sites, but will also require a considerable amount of mass spectrometry resources.

Considering the fact that we observed over 50 % of the *D. melanogaster* Kc167 cells proteome to be phosphorylated, it can be assumed that, even though we mixed *D. melanogaster* Kc167 cell samples of different conditions, the assumption that 30 % of all proteins are phosphorylated at a given time in a cell(Mann *et al*, 2002) is probably correct. As we detected at least another 10,000 phosphopeptides as precursor ions to which no definitive sequence could be assigned, it can even be hypothesized that at a given time even more than 30 % of all proteins in a cell are phosphorylated.

Assignment of the phosphorylated amino acid residue

In order to determine the certainty of the assignment of a phosphate group to a hydroxyamino acid the dCn was used as it has been shown recently that it directly correlates with the certainty of phosphorylation site assignment(Beausoleil *et al*, 2006) (see Material and Methods). The data in Supplementary Figure S2 indicates that the percentage of ambiguously assigned sites increased from approximately 1 % at a dCn value of 0.4 to approximately 18 % at a dCn value of 0. At a dCn of 0.1 around 9 % of the sites were considered ambiguous.

We therefore decided to consider a phosphopeptide with a dCn higher than 0.1 to have correctly assigned phosphorylation site (> 90 % certainty).

Consensus spectra

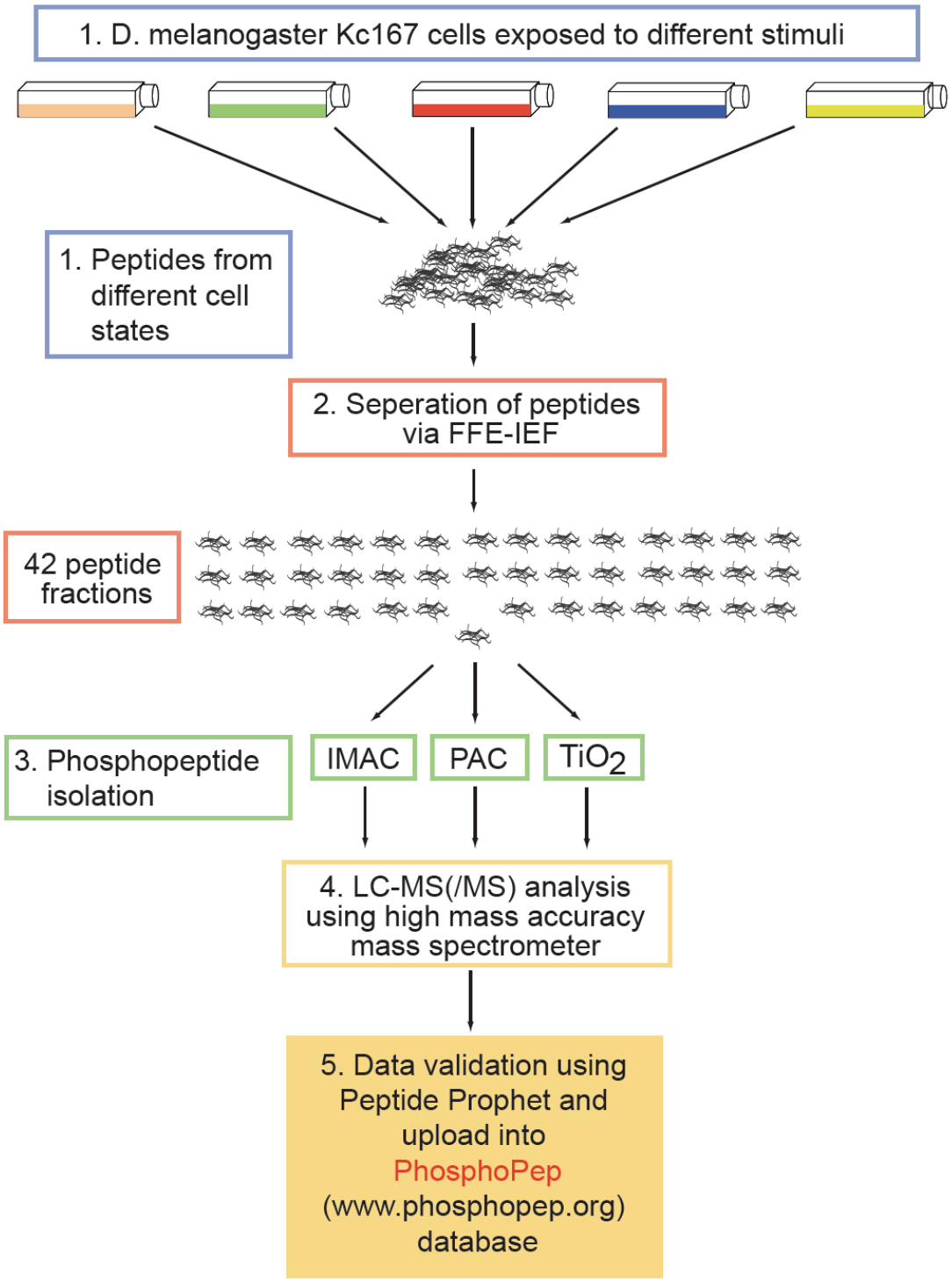
SpectraST identifies fragment ion spectra by spectral matching. The search tool output is further processed and validated by PeptideProphet(Keller *et al*, 2002) and the TPP(Keller *et al*, 2005)**.**

Briefly, multiple MS/MS observations (if available) of each distinct peptide ion were combined to create a "consensus" spectrum containing only the consistently observed spectral features. The consensus spectra thus created are noise-reduced, high-quality spectral representations of their corresponding peptide ions, which allow rapid and accurate MS/MS identifications by spectral matching, and facilitate the selection of appropriate transitions for MRM studies (Lam *et al*, 2007; Stein and Heller, 2006; Stein and Scott, 1994).

To demonstrate the increased performance in terms of search speed and accuracy of search results of spectral matching compared to traditional database searching, we searched two phosphopeptide datasets (one generated with a high mass accuracy mass spectrometer, LTQ-FT, and one generated with a lower mass accuracy mass spectrometer, LTQ) with Sequest(Eng *et al*, 1994) against FlyBase(Grumbling and Strelets, 2006) database and with SpectraST against PhosphoPep. Results of both searches are processed and validated with PeptideProphet [39], and then compared in terms of number of identified spectra and sensitivity. Search speed was increased by 86 times for data acquired on the high mass accuracy (LTQ-FT) mass spectrometer and 745 times for data acquired on the lower accuracy (LTQ) mass spectrometer (Supplementary Figure S3**A**). The superior performance of spectral matching to the consensus spectrum library is apparent from improvements in the sensitivity/error plot computed by PeptideProphet (Supplementary Figure S3**B**). These improvements resulted in significantly higher numbers of conclusively identified spectra (~ 3-96 % increase in identified spectra with a PeptideProphet p value > 0.9, see Supplementary Table I).

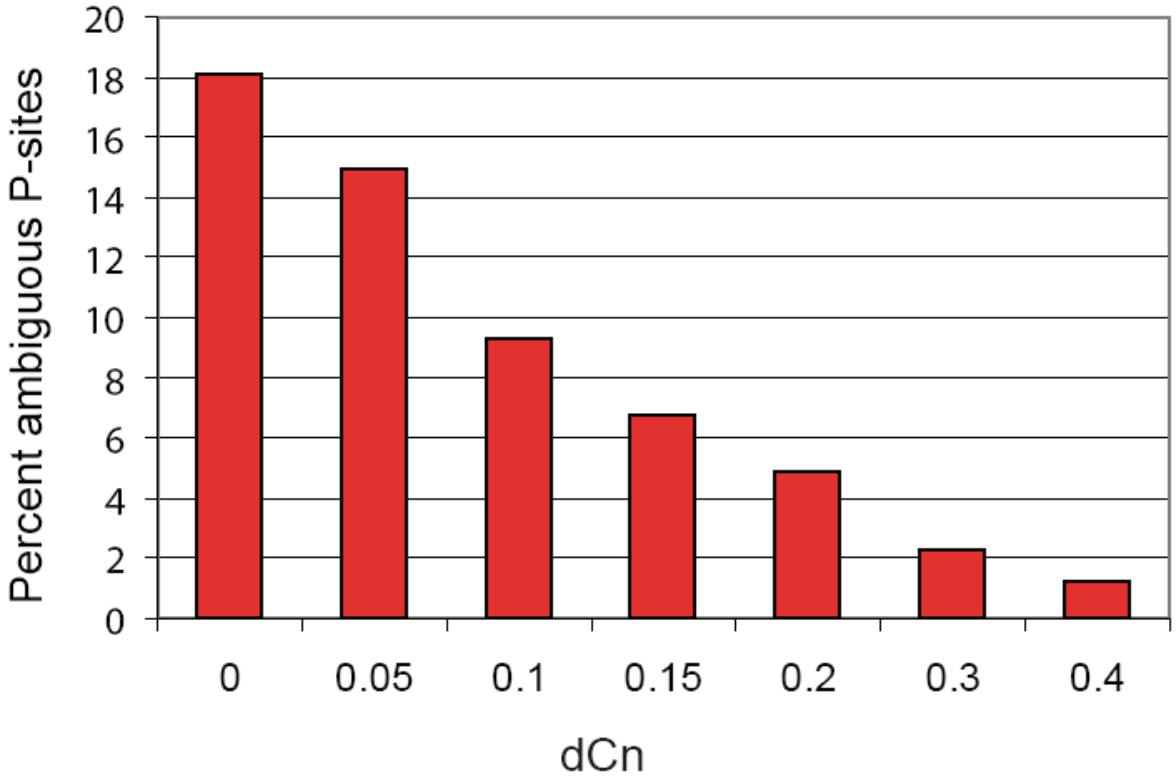
It must be noted that as it is difficulty to identify multiply phosphorylated peptides from MS2 spectra with sequence search engines like Sequest it will also be difficult to construct the corresponding consensus spectra. Furthermore, the performance of spectral library searches for multiply phosphorylated peptides is due to the lack of available datasets not yet fully assessed.

Supplementary Figure S1



Experimental pipeline for the generation of large scale phosphoproteomic datasets Schematic illustration of the experimental pipeline which was used to generate the high confidence large scale phosphopeptide dataset of *D. melanogaster*. Full details are given in the text.

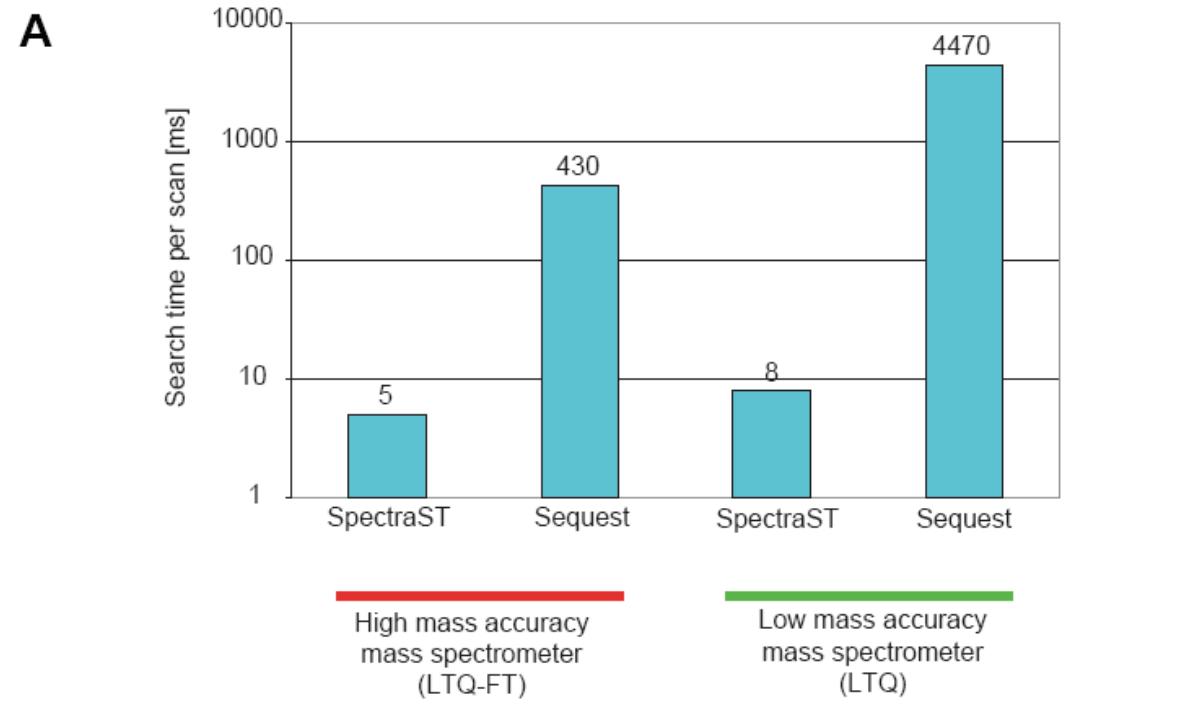
Supplementary Figure S2

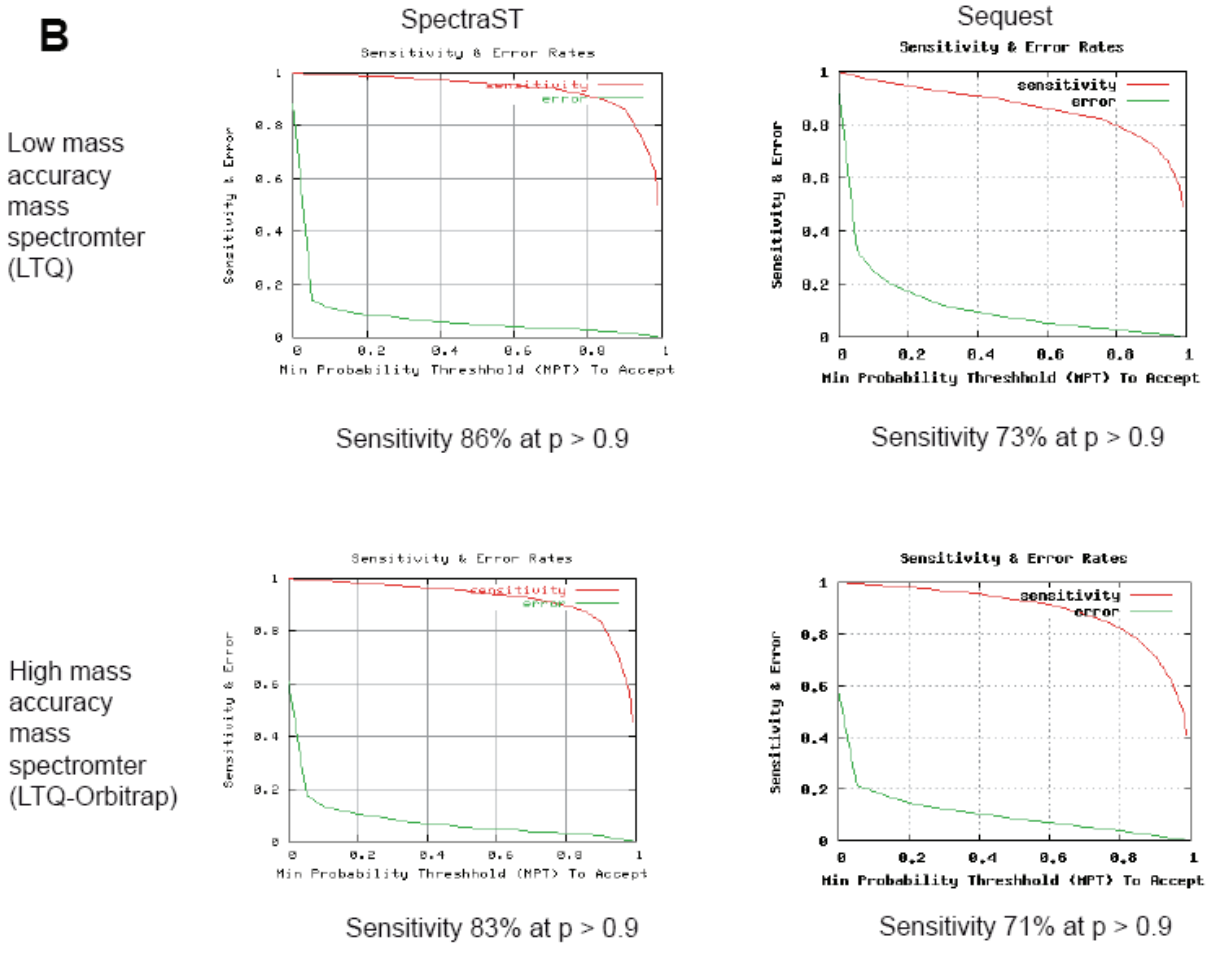


Assessment of the certainty of phosphorylation site assignment

For the certainty of the assigned phosphorylation amino acid the dCn value was used(Beausoleil *et al*, 2006). For each dCn the percentage of ambiguously assigned phosphorylation sites is shown. As can be seen a dCn value of > 0.1 yields an accumulated phosphorylation site certainty of >90%.

Supplementary Figure S3





Performance comparison between SpectraST and Sequest

In **A** the average time (in milliseconds) is shown which each algorithm needed to search one MS2 spectrum derived from a high mass accuracy mass spectrometer (LTQ-Orbitrap, 10166 spectra used) or from a low mass accuracy mass spectrometer (LTQ, 27556 spectra used). Note a logarithmic scale is used. In **B** the sensitivity and error curves for the two datasets is shown as computed by PeptideProphet(Keller *et al*, 2002).

Supplementary Table I

Performance comparison between SpectraST and Sequest

Test dataset	Sequest*	SpectraST*	% Increase in identifications*
IMAC	51842	53244	2.7
TiO_2	38512	50751	31.8
PAC	21262	41719	96.2

*Spectra identified with a PeptideProphet p value > 0.9

Performance comparison between SpectraST and Sequest.

The IMAC, TiO_2 and PAC dataset was searched with both Sequest and SpectraST. For both search algorithms the number of identified spectra with a PeptideProphet p value > 0.9 is shown.

Supplementary Table IIA

Results of Chi Square Test for Figure 2A

	Internal reference[y] vs. FlyBase	Phosphoproteins[y] vs. FlyBase	Phosphoproteins vs. internal reference[z]
Isomerase	35.21	0.45	11.5
Miscellaneous function	3.24	2.07	0.1
Transferase	4.90	9.75	22.9
Transfer/carrier protein	2.01	6.06	1.6
Ion channel	23.37	10.07	7.6
Transporter	32.57	26.24	0.1
Synthase and synthetase	60.76	0.06	22.4
Extracellular matrix	9.72	12.22	0.8
Phosphatase	0.00	0.78	0.7
Defense/immunity protein	3.02	3.94	0.2
Transcription factor	36.03	10.37	129.9
Hydrolase	0.66	15.19	10.3
Ligase	15.34	2.09	3.3
Receptor	87.04	39.14	31.7
Select regulatory molecule	14.75	44.20	7.1
Membrane traffic protein	14.82	21.37	0.6
Cytoskeletal protein	16.37	35.94	3.4
Select calcium binding protein	0.03	0.31	0.5
Lyase	6.69	9.24	20.6
Oxidoreductase	0.69	52.67	44.7
Protease	22.34	33.20	2.8
Signaling molecule	8.55	2.71	2.2
Chaperone	22.19	8.29	1.3
Nucleic acid binding	90.66	139.61	6.2
Cell junction protein	0.53	0.58	2.9
Cell adhesion molecule	16.15	10.25	1.2
Kinase	0.80	39.99	27.6

[y] = Values greater than 6.635 (P = 0.01) are considered to be significant

[z] = Values greater than 5.024 (P = 0.025) are considered to be significant

Supplementary Table IIB

Results of Chi Square Test for Figure 2B

	Internal standard[y] vs. FlyBase	Phosphoproteins[y] vs. FlyBase	Phosphoproteins vs. internal reference[z]
Neuronal activities	8.12	2.33	1.3
Signal transduction	5.90	0.03	4.1
Sulfur metabolism	1.08	7.38	4.4
Developmental processes	3.98	0.23	5.4
Other metabolism	1.40	10.81	5.8
Non-vertebrate process	27.95	24.05	0.9
Cell proliferation and differentiation	2.22	2.05	0.0
Coenzyme and prosthetic group metabolism	0.27	4.17	5.7
Cell structure and motility	14.79	23.34	1.8
Immunity and defense	0.14	2.60	3.6
Apoptosis	1.40	5.30	1.4
Oncogenesis	5.33	7.31	0.4
Muscle contraction	0.01	0.01	0.0
Transport	57.58	36.42	0.4
Phosphate metabolism	0.53	0.03	0.2
Carbohydrate metabolism	2.34	22.44	13.4
Nucleoside, nucleotide and nucleic acid metabolism	44.21	55.37	2.5
Homeostasis	0.10	0.03	0.0
Protein metabolism and modification	11.88	2.21	1.9
Cell cycle	62.71	64.82	1.0
Nitrogen metabolism	0.07	0.41	0.6
Intracellular protein traffic	26.73	21.85	0.0
Cell adhesion	0.05	0.01	0.0
Lipid, fatty acid and steroid metabolism	37.88	45.30	4.1
Sensory perception	29.52	15.41	1.7
Electron transport	23.48	41.24	11.5
Amino acid metabolism	3.36	7.14	1.6
Protein targeting and localization	10.38	13.34	0.5
Miscellaneous	7.86	2.10	1.7

[y] = Values greater than 6.635 (P = 0.01) are considered to be significant

[z] = Values greater than 5.024 (P = 0.025) are considered to be significant

References

Aebersold R, Goodlett DR (2001) Mass spectrometry in proteomics. *Chem Rev* **101:** 269-295.

Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li JX, Cohn MA, Cantley LC, Gygi SP (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *P Natl Acad Sci USA* **101:** 12130-12135.

Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **24:** 1285-1292.

Bodenmiller B, Mueller LN, Mueller M, Domon B, Aebersold R (2007a) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Meth* **4:** 231-237.

Bodenmiller B, Mueller LN, Pedrioli PG, Pflieger D, Junger MA, Eng JK, Aebersold R, Tao WA (2007b) An integrated chemical, mass spectrometric and computational strategy for (quantitative) phosphoproteomics: application to Drosophila melanogaster Kc167 cells. *Mol BioSyst* **3:** 275-286.

Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectr* **5:** 976-989.

Grumbling G, Strelets V (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res* **34:** D484-488.

Keller A, Eng J, Zhang N, Li XJ, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1:** 2005 0017.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74:** 5383-5392.

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7:** 655-667.

Mann M, Ong SE, Gronborg M, Steen H, Jensen ON, Pandey A (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol* **20:** 261-268.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference

database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363-2371.

Picotti P, Aebersold R, Domon B (2007) The Implications of Proteolytic Background for Shotgun Proteomics. *Mol Cell Proteomics* (In press).

Schroeder MJ, Shabanowitz J, Schwartz JC, Hunt DF, Coon JJ (2004) A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* **76:** 3590-3598.

Stahl-Zeng J, Lange V, Ossola R, Aebersold R, Domon B (2007) High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol Cell Proteomics* (In press).

Stein SE, Heller DN (2006) On the risk of false positive identification using multiple ion monitoring in qualitative mass spectrometry: large-scale intercomparisons with a comprehensive mass spectral library. *J Am Soc Mass Spectrom* **17:** 823-835.

Stein SE, Scott DR (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom.* **5**: 859-866.

Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *P Natl Acad Sci USA* **101:** 9528-9533.