Global Post-translational Modification Discovery

*Qiyao Li[1]; Michael R. Shortreed[1]; Craig D. Wenger[2]; Brian L. Frey[1]; Leah V. Schaffer[1], Mark Scalf[1]; Lloyd M. Smith[1,]**

[1]Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, WI, 53706

[2]no affiliation

[*]corresponding Author

Table of Contents

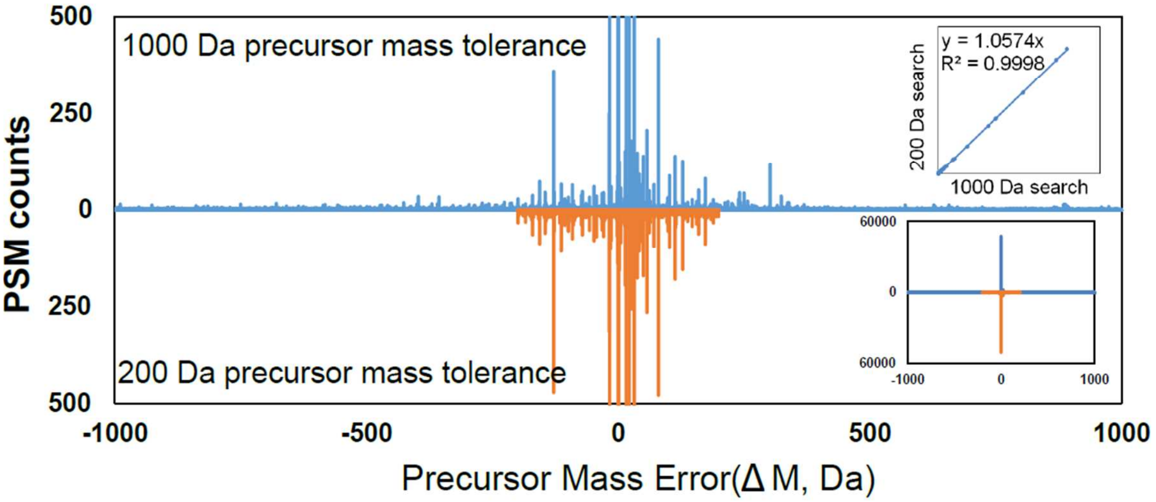
Supplementary Notes

G-PTM-D can provide amino acid specificity information for unknown modifications.

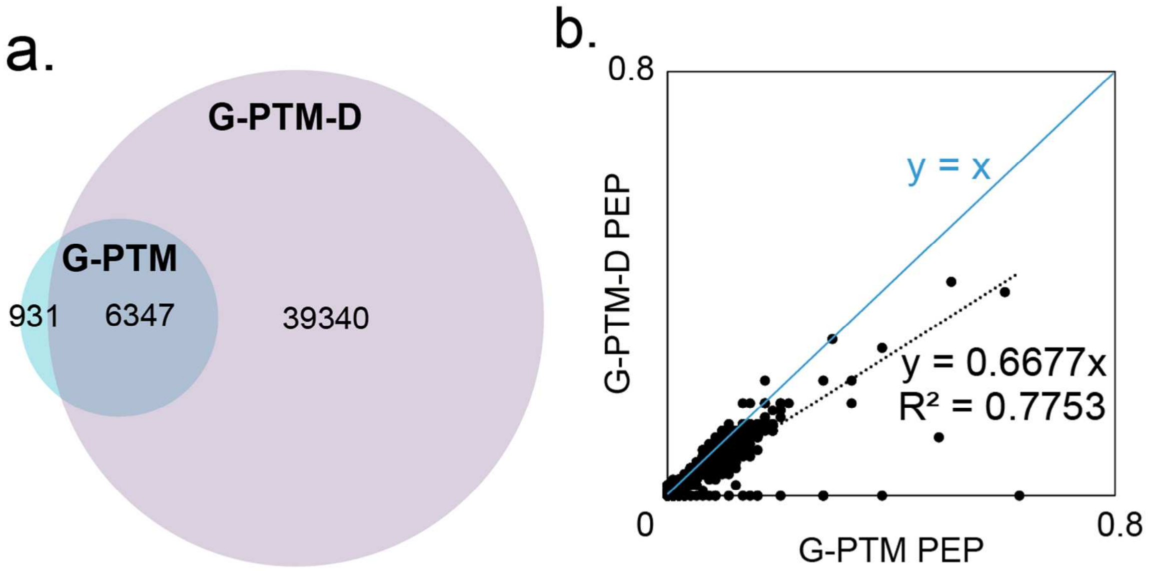
In the Jurkat ΔM histogram, several ΔMs (e.g. -91.007, -73.001, +249.978, +301.978, +323.960) do not correspond to any known modifications to our knowledge. To find out whether they are specific to certain amino acid (AA) residues, we assigned them to any AA in the peptides that had the respective ΔM, then incorporated them into the new XML database, and performed the second-round search with regular mass tolerances. About half of the peptides with the -91.007 Da or -73.001 Da mass differences had the modification on cysteine, which had carbamidomethylation (+57.021 Da) as a fixed modification due to the addition of iodoacetamide during sample preparation. Therefore, the modification -91.007 Da could be -91.007 + 57.021 = -33.986 Da, corresponding to the conversion of cysteine to dehydroalanine (DHA). The -73.001 Da modification likely results from the addition of H_2O to DHA. The mass differences of +249.978 Da, +301.978 Da, and +323.960 Da, occur more frequently on hydrophobic (Ile, Val, Phe) and acidic (Asp, Glu) AAs. The molecular formula for 301.978 could be $C_8H_6N_4O_5S_2$, but its chemical structure is unknown. The mass shift of +323.960 Da is likely +301.978 Da plus a sodium adduct. These examples illustrate the use of G-PTM-D to provide amino acid site localization information for previously unknown modification types.

G-PTM-D results in minimum bias for identification of modified peptides.

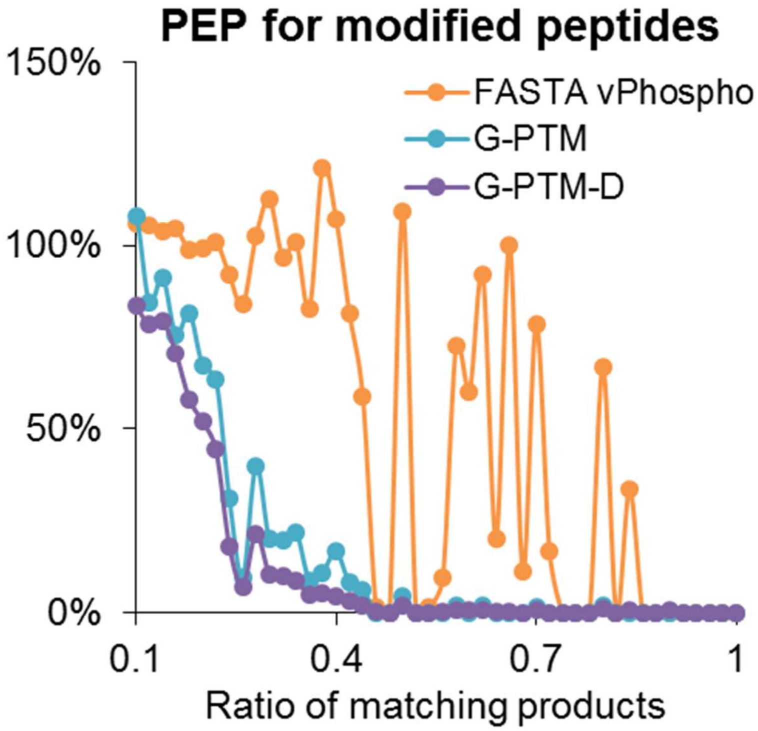
In Supplementary Figure S5, we showed near complete overlap in the distributions of mass error, Q-value, and Morpheus score, for our unmodified and modified PSMs. To further test for the existence of bias towards identification of modified peptides, we performed an additional experiment. In this experiment, high-scoring decoy hits with PTM-characteristic mass shifts and Q-values below 1% were added to the forward target database and used in the second-pass narrow tolerance G-PTM-D search. There were 536 decoys that met this criterion in the HeLa dataset. A second-round search was performed with the same parameters described in the main text. The search resulted in 313,763 target PSMs and 26,890 modified target PSMs based on 1% global FDR (nearly the same as the original search without the high-scoring decoys– 314,414 target PSMs and 26,949 modified target PSMs). Furthermore, the local false discovery rate for modified peptides remains near 1% (1.05%) even when modified decoys from the first-round search are included in the XML database for the second-round search of the G-PTM-D strategy.



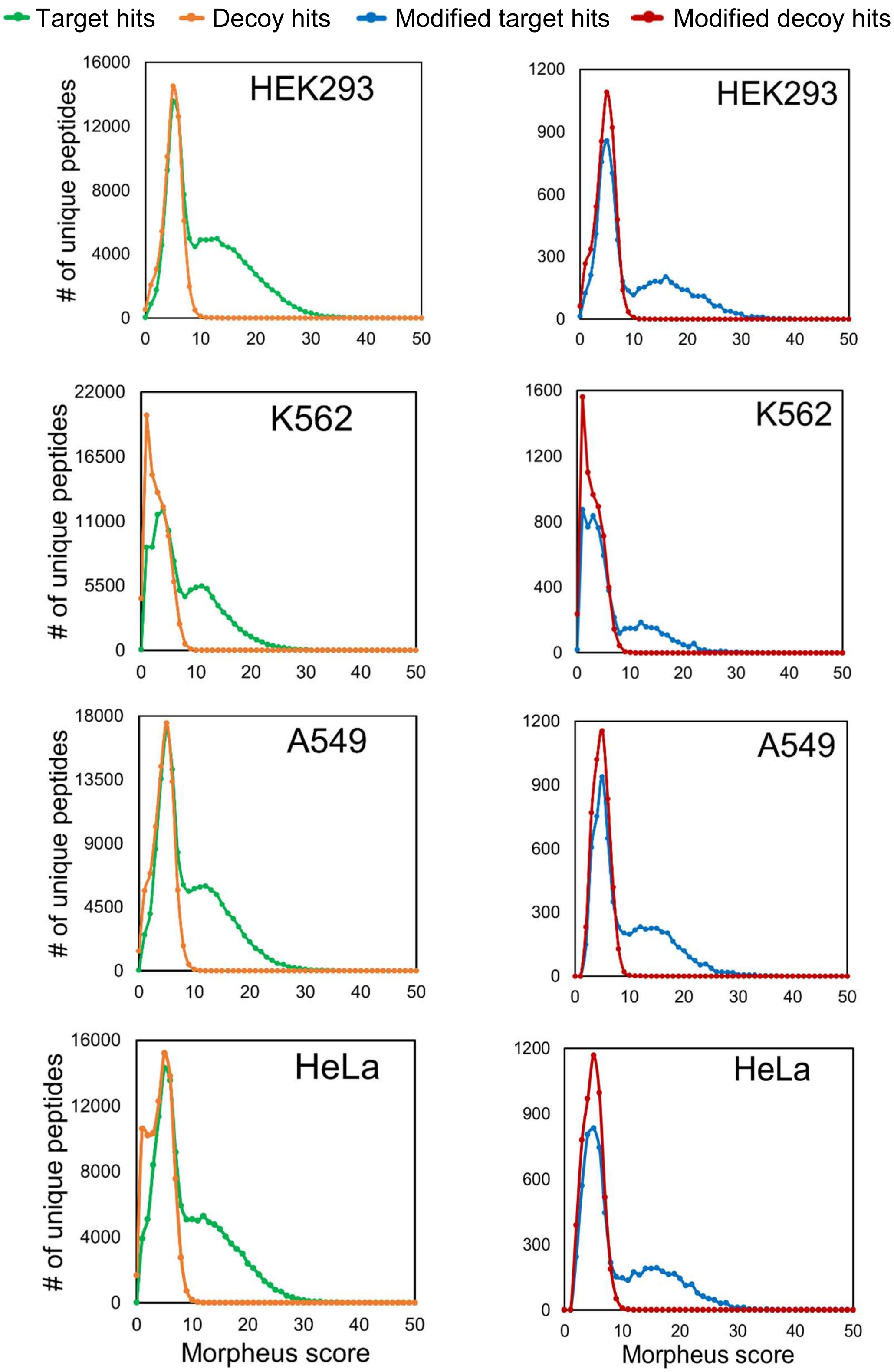
Supplementary Figure S1. Histograms of precursor mass error (ΔM) searched with ±1000 Da or ±200 Da precursor mass tolerances. The lower-right inset shows the unexpanded plot. The upper right inset shows excellent correlation between the peak heights obtained from the two searches, indicating that the sensitivity in identifying potential modifications is not compromised by the size of the mass tolerance, although modifications exceeding 200 Da cannot be observed in the ±200 Da search.



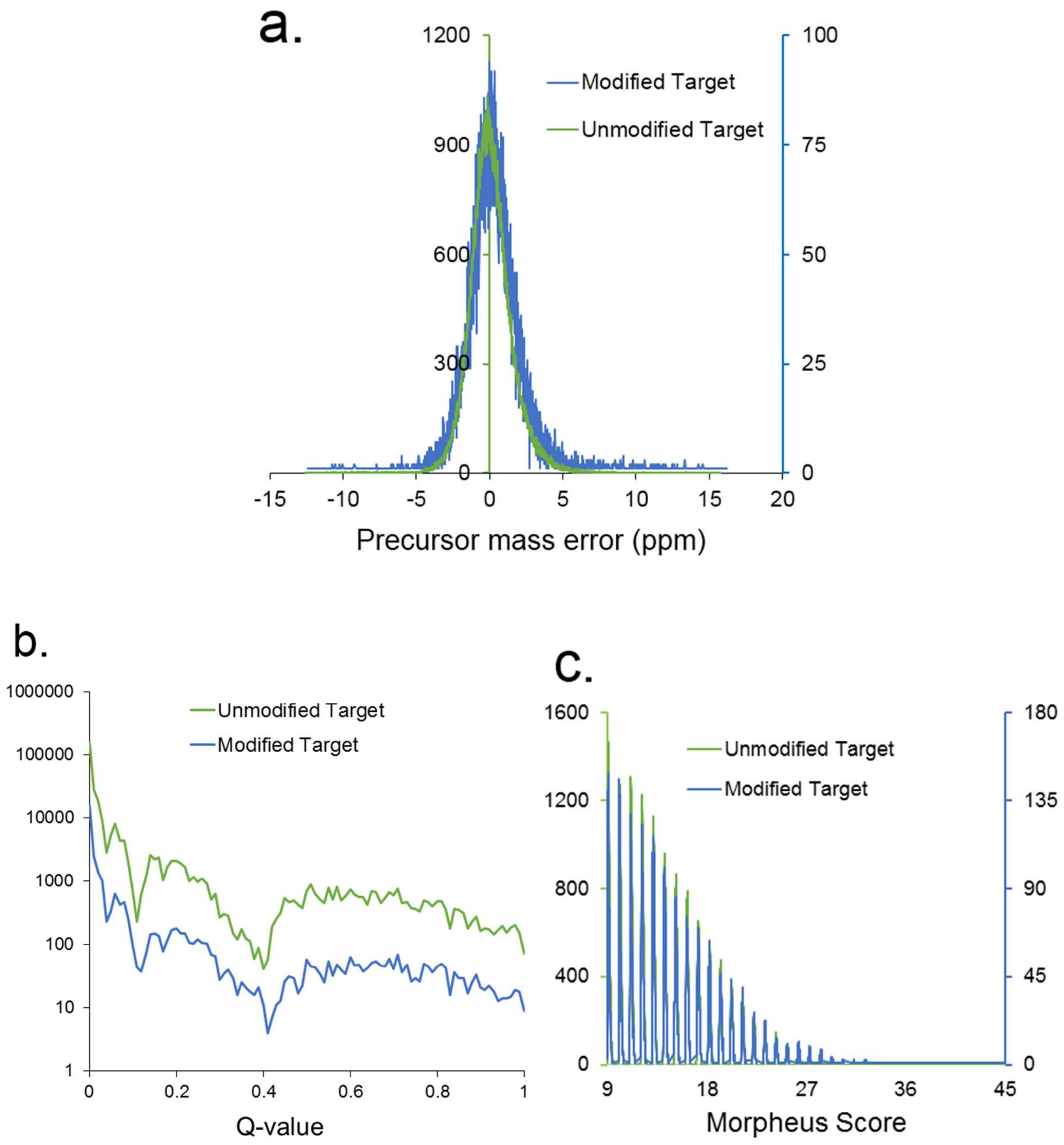
Supplementary Figure S2. Comparison of modified spectra identification by G-PTM or G-PTM-D, from the Jurkat dataset, with 1% global FDR. a) Overlap between the 7278 modified spectra identified by G-PTM and the 45,687 modified spectra identified by G-PTM-D shown in Fig. 2a. b) Correlation of PEP values from G-PTM-D vs. from G-PTM, for the 6347 commonly identified spectra.



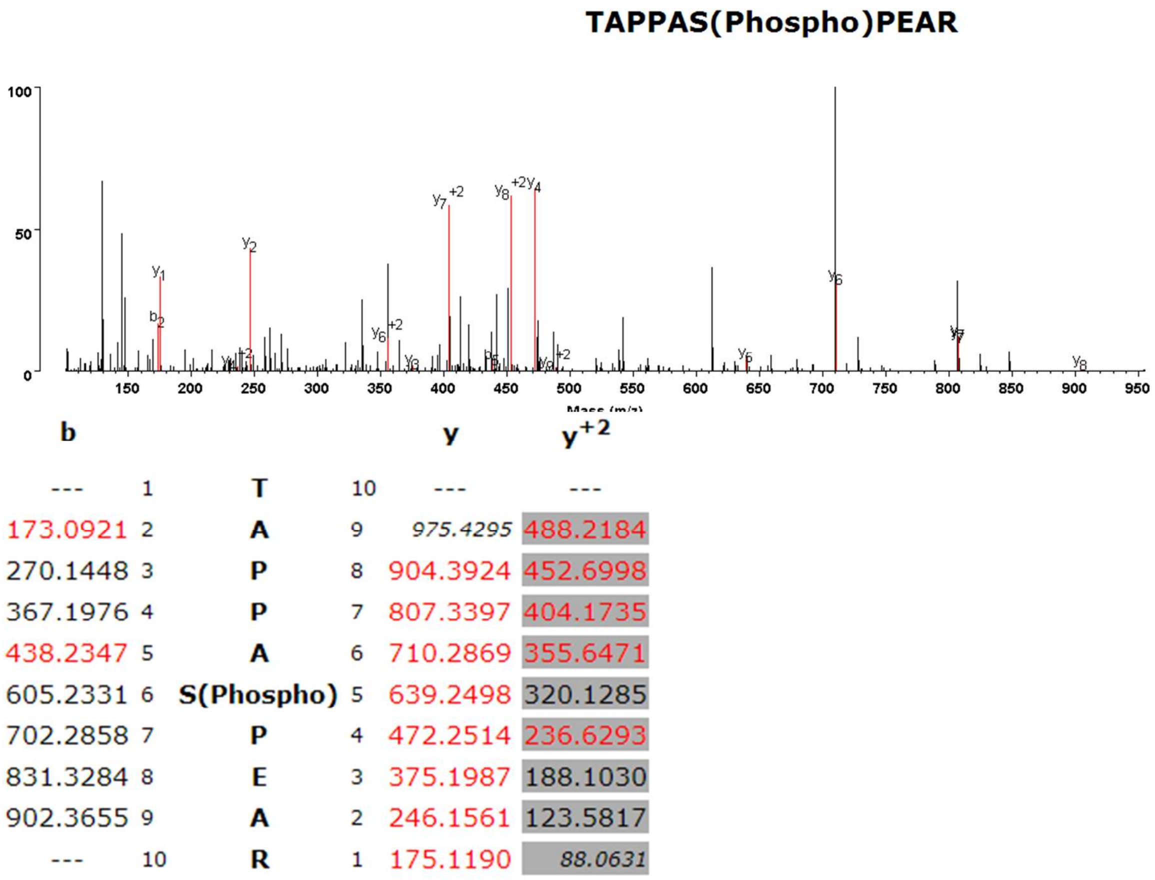
Supplementary Figure S3. Posterior error probability (PEP) for modified peptides as a function of the ratio of matching products, from three types of searches of the Jurkat cell dataset: a vPhospho search (using the UniProt FASTA database with phosphorylation as a variable modification), a G-PTM search (using the PTM-curated UniProt database), and a G-PTM-D search. The PEP at a certain ratio is the number of modified decoy PSMs divided by the number of modified target PSMs among the PSMs that have a ratio within half of the ratio bin size (0.01) from that particular ratio. Results are based on all identifications, without a global FDR cutoff.


Target hits
Decoy hits
Modified target hits
Modified decoy hits
HEK293
K562
A549
HeLa
# of unique peptides
Morpheus score
16000
12000
8000
4000
0
1200
900
600
300
22000
16500
11000
5500
1600
800
400
18000
13500
9000
4500
0
10
20
30
40
50

Supplementary Figure S4. Histograms of the number of unique peptides (all or modified, target or decoy) against Morpheus score, for four human cell lines (score bin size = 1). Note that in situations where the score of a spectrum match is identical for both decoy and target peptide, Morpheus automatically assigns the match to decoy, which is why the number of decoy hits exceeds the number of target hits at low scores.



Supplementary Figure S5. Distributions of (a) precursor mass error, (b) Q-value, and (c) Morpheus score for unmodified (green) and modified (blue) PSMs. Peaks in (a) and (c) were normalized to maximum peak height for easy comparisons of the two distributions, which results in the appearance of higher noise in the Modified Target distributions.



b				y	y^{+2}
---	1	T	10	---	---
173.0921	2	A	9	*975.4295*	488.2184
270.1448	3	P	8	904.3924	452.6998
367.1976	4	P	7	807.3397	404.1735
438.2347	5	A	6	710.2869	355.6471
605.2331	6	S(Phospho)	5	639.2498	320.1285
702.2858	7	P	4	472.2514	236.6293
831.3284	8	E	3	375.1987	188.1030
902.3655	9	A	2	246.1561	123.5817
---	10	R	1	175.1190	*88.0631*

Supplementary Figure S6. An annotated spectrum that is assigned as TAPPAS(Phospho)PEAR in the second-round search, but was assigned (incorrectly) as a decoy peptide in the wide tolerance first-round search of G-PTM-D of the Jurkat dataset. Red font is used to represent ion matches.