Overcoming Peptide-Centric DIA Limitations: Discovering Unsuspected Modifications in Narrow-Window DIA Using Mascot Error Tolerant Search

DIA data hides PTMs — an Error Tolerant search reveals them

WP 622

MATRIX SCIENCE

Richard Jacob[1], Ville Koskinen[2], Patrick Emery[2]
[1]Matrix Science Inc, Boston, MA, USA
[2]Matrix Science Ltd, London, UK
richardj@matrixscience.com matrixscience.bsky.social

COI: Ville Koskinen and Patrick Emery are directors and minority shareholders of Matrix Science Ltd.

How it works: Peptide-centric DIA workflows rely on spectral libraries or in-silico predicted libraries to identify peptides. This imposes fundamental constraints: only pre-specified variable modifications, fixed missed cleavage limits, and known protein sequences can be searched. Unsuspected post-translational modifications, chemical artefacts, non-specific cleavage products, and primary sequence variants are systematically missed. Mascot's spectrum-centric approach searches MS/MS spectra directly against a sequence database. The Error Tolerant search extends this by performing a two-pass search: a standard first-pass search identifies proteins, then a comprehensive second pass iterates through the entire Unimod modification list without prior knowledge of what may be present.

Dataset: We used a publicly available dataset PXD037506 from the PRIDE repository.

Data analysis: We ran an ET search on a single prostate cancer specimen sample, P1. The number of modified sites, variable modification per site and number of arrangements were limited to reduce search space and speed up the Error Tolerant search. 10 files split between cancer and non-cancer patients were analyzed to verify the sequence variants results. All results reported at 1% FDR.

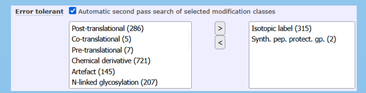
Example results: After scoring improvements (poster WP 449), 17,624 additional PSMs were identified, resolved to 4,227 distinct peptide sequences. An AI agent was used to generate leads and make an initial interpretation. Proteins selected as biomarkers in the publication were analyzed in more depth, comparing identifications to UniProt, GlyConnect, EBI ProtVar, iPTMnet and other resources. Modifications related to sample handling and deamidation were ignored. In total 2480 were identified as known.

Note: limiting the modification search space means site localization is less precise than a targeted search.

Error Tolerant Search

First pass
with standard search parameters and minimum variable modifications

Second pass
- Semi-specific enzyme
- Selected modification categories tested
- All possible substitutions tested
- Only one unsuspected modification per peptide
- Mass delta filter based on fragment ion tolerance

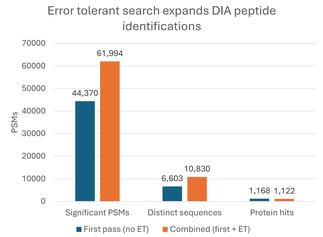


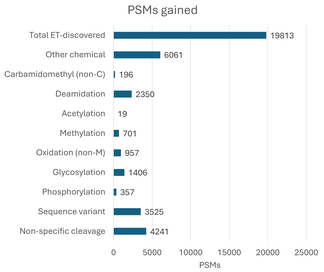
Statistical validation
Each match's identity threshold is calculated from the combined first-pass plus second-pass trial count, and because the two search spaces are disjoint the passes are thresholded independently.

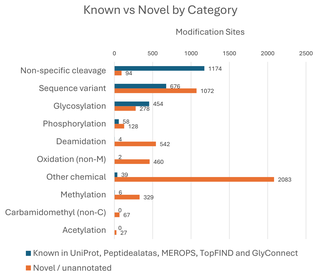
Modification statistics for all protein families

Modification	Delta	Type	Site	Total matches
Carbamidomethyl	57.021464	fixed	C	29221
Oxidation	15.994915	variable	M	3800
Non-specific cleavage		ET	-	3504
Deamidated	0.984016	ET	N	2113
Delta:H(2)C(2)	26.015643	ET	N-term	524
Ammonia-loss	-17.026532	ET	N-term	351
Gln->pyro-Glu	-17.026532	ET	N-term	307
Deamidated	0.984016	ET	Q	289
Acetyl	42.010565	variable	Protein N-term	278
Oxidation	15.994915	ET	P	226
Val->Met	31.972071	ET	V	223
Carboxymethyl	58.005479	ET	C	173
Phospho	79.966331	ET	S	165
Cys->Dha	-33.987721	ET	C	160
Glu->Gln	-0.984016	ET	E	151
Asp->Asn	-0.984016	ET	D	146
Ammonia-loss	-17.026549	ET	N	140
Methyl	14.015643	ET	N-term	126
Ammonium	17.026549	ET	E	114
Pro->Asn	16.990164	ET	P	104
Xle->Asn	0.958863	ET	L	103

Overview of results

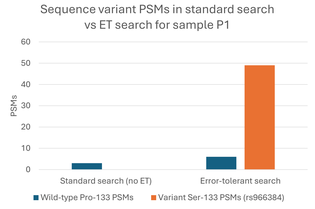




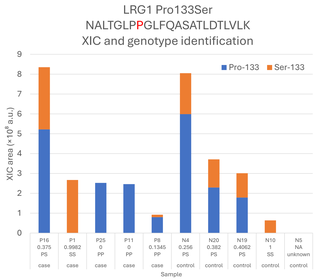


Sequence Variant

Peptides from two biomarker proteins, Leucine-rich alpha-2-glycoprotein 1 (LRG1, P02750, shown) and Serotransferrin (TF, P02787), were identified with sequence variants. These were known East-Asian variants that are expected to be observed in Han-Chinese donors.



From 10 Han-Chinese donors, in LRG1, 2 were Ser/Ser homozygotes (Pro133 peptide essentially absent), 2 were Pro/Pro homozygotes, 5 were heterozygous and one was not measurable.



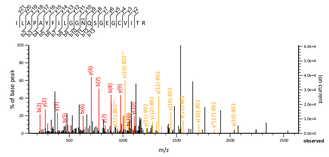
As this only affected one peptide from the ~19 used for protein quantitation it does not change the overall protein quantitation results. However, this peptide should not be used for any targeted MRM/PRM validation.

Glycosylation

Example of known identification of N-linked high-mannose on Monocyte differentiation antigen CD14 (P08571)

Hex:9 HexNAc:2 and Hex:8 HexNAc:2 at Asn 282, max score 162 with 11 PSMs matching an entry in GlyConnect.

There are also plenty of novel sites. Acid ceramidase (A0A1B0GUH5) was identified with Hex(3)HexNAc(2), a Man3-like core at N268 with a canonical N-X-S/T sequon:
ILAPAYFILGGNQSGEGCVITR



There are 7 PSMs with a max score of 117. Additional supporting evidence includes NeuAc-H2O and Hex+HexNAc Oxonium ions in the spectrum, a high-mannose precursor at Man5, part of a trim ladder, and two other complex glycoforms at that site.
The protein also caries two additional N-linked glyco sites at N257 and N340 or N346 although the number of matches and scores are lower.

Phosphorylation

Low-molecular-weight kininogen (P01042-2) has a documented phosphorylation site at S332 as well as documented sites at T326, T327, S329, T337 and T342. The search found 13 PSMs with a high score of 127.

Score	Mr(calc)	Delta	Sequence	Site Analysis
127.4	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho S8 98.89%
107.4	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho E7 0.97%
96.0	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho E10 0.07%
93.3	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho K6 0.04%
85.4	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho S5 0.01%
83.0	2340.8974	-0.0036	ETTCSKESNEELTESCETK	Phospho T3 0.00%

In this case the match to S332 in the peptide ETTCSKESNEELTESCETK was clearly localized.

There are very few novel STY phospho sites in the data set, just 3 out of 21 sites, each with just one PSM. We investigated Maltase-glucoamylase (O43451). The PSM score and expect values are very good, 87 and 3.4e-05, and the presence of neutral-loss -98 Da ions suggests the peptide is phosphorylated. There are five S/T candidates: T823, T828, T829, T830, S833, and only one b ion in the region of interest.
GGYIFPTQQPNTTTLASRKNPLGLIIALDENK
And it is not possible to localize the site from this analysis.

Score	Mr(calc)	Delta	Sequence	Site Analysis
87.0	3549.8232	-0.0258	GGYIFPTQQPNTTTLASRKNPLGLIIALDENK	Phospho S17 20.00%
87.0	3549.8232	-0.0258	GGYIFPTQQPNTTTLASRKNPLGLIIALDENK	Phospho T14 20.00%
87.0	3549.8232	-0.0258	GGYIFPTQQPNTTTLASRKNPLGLIIALDENK	Phospho T13 20.00%
87.0	3549.8232	-0.0258	GGYIFPTQQPNTTTLASRKNPLGLIIALDENK	Phospho T12 20.00%
87.0	3549.8232	-0.0258	GGYIFPTQQPNTTTLASRKNPLGLIIALDENK	Phospho T7 20.00%

Analysis of the site with NetPhos-3.1 indicated and S833 (S17 in the peptide) as the highest predicted site in the sequence phosphorylated by Protein Kinase C (PKC) or an unspecified kinase. Further analysis is required to confirm this site as a phosphorylation.

WP622: Overcoming Peptide-Centric DIA Limitations: Discovering Unsuspected Modifications in Narrow-Window DIA Using Mascot Error Tolerant Search

Richard Jacob1, Ville Koskinen2, Patrick Emery2

1Matrix Science Inc, Boston, MA, USA
2Matrix Science Ltd, London, UK

richardj@matrixscience.com @matrixscience.bsky.social

Introduction

Peptide-centric DIA workflows rely on spectral libraries or in-silico predicted libraries to identify peptides. This imposes fundamental constraints: only pre-specified variable modifications, fixed missed cleavage limits, and known protein sequences can be searched. Unsuspected post-translational modifications, chemical artefacts, non-specific cleavage products, and primary sequence variants are systematically missed. Expanding a spectral library to include all possible variants would be computationally impractical. Mascot's spectrum-centric approach searches MS/MS spectra directly against a sequence database without these constraints. The Error Tolerant search extends this by performing a two-pass search: a standard first-pass search identifies proteins, then a comprehensive second pass iterates through the entire Unimod modification list, semi-specific cleavage, and a residue substitution matrix without prior knowledge of what may be present.

Methods

We investigated prostate cancer specimens from PRIDE PXD037506 that were acquired by DIA using 8 m/z isolation windows. Thermo raw files were processed with Mascot Distiller and Mascot Server using an Error Tolerant search. Search parameters were Trypsin with one missed cleavage, fixed carbamidomethyl(C), variable oxidation(M) and 20ppm tolerance for both precursor and fragment ions. Unimod modifications are split into different classes depending on their nature. All classes were searched except “isotopic labels” and “synthetic peptide protection groups”, a total 1442 modifications. Each modification would require separate, purpose-built spectral libraries under a peptide-centric approach. A target-decoy strategy was applied with a 1% PSM false discovery rate for both passes independently. Assignments were validated against UniProt, GlyConnect, and EBI ProtVar.

Preliminary Data

The error tolerant search of the sample P1.raw identified 6,835 additional significant PSMs and 1,072 additional peptide sequences beyond first-pass results (44,066 vs 37,231 total PSMs at 1% FDR). The majority of these identifications are inaccessible to standard peptide-centric workflows. Non-specific cleavage products accounted for 1,908 matches. There were 1610 +1Da modifications. Phosphorylation at S234 in Osteopontin was identified with strong site localization and confirmed by UniProt. Glycopeptide identifications (383 PSMs) included Hex(1)HexNAc(1)NeuAc(1) on Alpha-2-HS-glycoprotein, verified by GlyConnect. Single residue substitutions (922 PSMs) included a confirmed Pro-to-Ser variant in Immunoglobulin heavy constant alpha 2. Incorporating all these variable modifications into spectral libraries is impractical, but an error tolerant search allows for their identification in a single search.

Novel Aspect

Spectrum-centric error tolerant searching overcomes peptide-centric library constraints, enabling unbiased discovery of modifications and variants from chimeric DIA spectra.

Topic session: Posttranslational Modifications: Qualitative and Quantitative Analysis

COI: Ville Koskinen and Patrick Emery are directors and minority shareholders of Matrix Science Ltd.