A novel computational pipeline for immunopeptidomics with DIA-LC-MS/MS

Rui Qiao, Haibo Bian, Benjamin Keunen, Qing Zhang, Lei Xin and Baozhen Shan

Bioinformatics Solutions Inc., Waterloo, Ontario, Canada

Introduction

The immunopeptidome workflow has become increasingly crucial in analyzing data-independent acquisition (DIA) mass spectrometry data, particularly in immunology and cancer research. Immunopeptides, short peptide fragments presented on the cell surface by major histocompatibility complex (MHC) molecules, play a vital role in the immune system's ability to recognize and respond to foreign or altered-self antigens. These peptides, often present in low abundance and exhibiting high diversity, pose significant challenges for identification and quantification using traditional data-dependent acquisition (DDA) methods. DIA mass spectrometry offers a distinct advantage in this context. It consistently captures MS2 spectra of low-abundance peptides across multiple technical replicates, enhancing the possibility of their identification and quantification. However, the intricate nature of DIA spectra, coupled with the vast search space of HLA peptides, presents new difficulties in accurately identifying and quantifying these important molecules.

In this work, we propose an efficient workflow for DIA immunopeptidome data analysis. Our approach not only identifies peptides from canonical databases but also discovers novel peptides through de novo sequencing. This comprehensive method aims to unlock the full potential of DIA data in immunopeptidome research, potentially leading to significant advancements in our understanding of immune responses and the development of targeted immunotherapies.

Methods

We propose a streamlined workflow for analyzing DIA immunopeptidomic datasets:

1. Generate pseudo scans from MS1 features in the given DIA immunopeptide mass spectrometry dataset (similar to the DIA-Umpire method).

2. Perform parallel analyses:
   a. Search the raw DIA scans against a list of HLA peptides commonly reported in the Immune Epitope Database (IEDB), to calibrate and optimize search parameters.
   b. Conduct an unspecific database search on the pseudo scans created in step 1.
   c. Execute de novo peptide sequencing on the pseudo scans created in step 1.

3. Create a comprehensive spectral library using the identified precursors from steps 2a, 2b, and 2c. Perform an additional round of spectral library searching for false discovery rate (FDR) control and quantification.

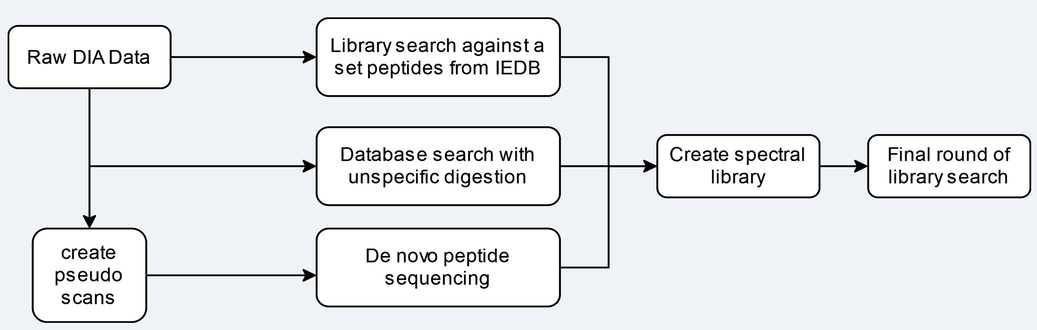

Figure 1. Flow chart of the DIA Peptidome workflow in PEAKS 12.

Results

The sensitivity and accuracy of a DIA peptidome workflow heavily relies on the DIA library search engine it employs. We benchmarked the identification and quantification performance of PEAKS 12 using a recently published dataset [1] (PXD046444) comprising three species (human, yeast, and E. coli) obtained from a Thermo Astral mass spectrometer. For our analysis, we selected three technical replicates under identical experimental conditions and performed DIA direct database search with match-between-run and high-precision quantification strategies. Table 1 presents the number of quantifiable precursors and protein groups (defined as those with a coefficient of variation < 20%). Our results demonstrate that PEAKS 12 quantified a comparable number of precursors and protein groups to state-of-the-art methods. To further validate our findings, we selected an additional set of three technical replicates from a different experimental condition to assess whether the protein quantity ratios aligned with expectations between groups. Figure 2 illustrates the protein ratio histogram derived from this comparison, with the expected ratio for different species shown as dashed vertical lines. As we can see, PEAKS 12 achieves low CV for precursor quantity estimation while the protein quantity ratio estimate remains to be accurate.

	# Quantifiable precursors	# Quantifiable protein groups
DIANN 1.9	162805	12203
PEAKS 12	171736	13382

Table 1. Number of quantifiable precursors and protein groups on three species dataset

We evaluated our DIA peptidome workflow using HLA datasets obtained from RA957 cell lines [2]. In a previous analysis of this dataset, AlphaPeptDeep [3] reported the identification of 36,947 unique sequences. Our workflow, however, identified 45,777 unique sequences at a 1% sample-level false discovery rate (FDR), representing a significant improvement in peptide identification. Figure 3 presents a Venn diagram illustrating the overlap of sequences identified from two technical replicates. Notably, we observed a 90% overlap between the replicates, demonstrating the high consistency and reproducibility of identifications achieved by PEAKS 12.

This substantial increase in unique sequence identifications, coupled with the high degree of overlap between replicates, underscores the enhanced sensitivity and reliability of our DIA peptidome workflow powered by PEAKS 12. These results highlight the potential of our approach to provide more comprehensive and accurate immunopeptidome profiling, which could have significant implications for various immunological studies and

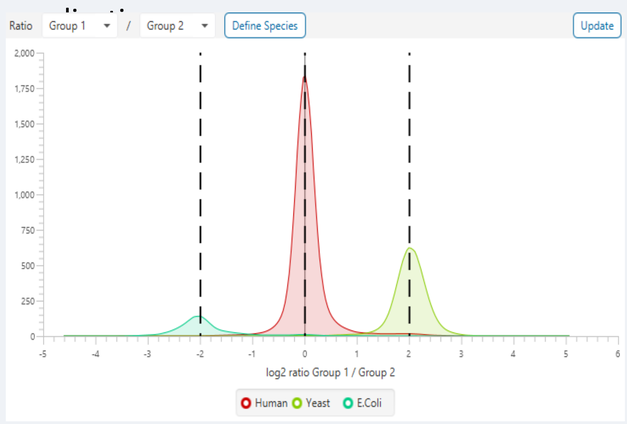

Figure 2. Protein quantity ratio distribution on three species dataset.

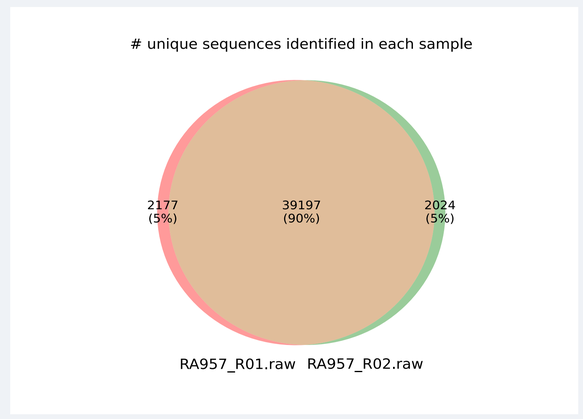

Figure 3. Number of unique sequences identified on two technical replicates.

In addition to identifying peptides from canonical databases, our workflow demonstrates the capability to discover sequences absent from these databases through de novo sequencing of pseudo scans. For instance, in this dataset, our approach uncovered 103 sequences not present in the reference database.
Historically, validating the existence of de novo peptides lacking database representation has been challenging due to the absence of gene support. Our software addresses this issue by presenting results not only with a q-value but also through an interactive interface that visualizes precursors' MS1 profiles and MS2 fragment ion signals across all samples. This feature allows users to validate identification results by verifying consistent peptide occurrence at predicted retention times across samples and assessing the correlation between MS1 profiles and fragment ion signals.
Figure 4 illustrates a de novo sequenced peptide without a protein accession. Notably, this peptide exhibits highly correlated MS1 and MS2 signals across four samples, providing additional confidence in our peptide identification even in the absence of gene support. This advanced capability of our workflow enhances the potential for novel peptide discovery and offers a more comprehensive view of the immunopeptidome. By combining database searching with de novo sequencing and providing robust visualization tools, our approach enables researchers to explore and validate peptide identifications with greater confidence, potentially uncovering biologically significant sequences that might otherwise be overlooked.

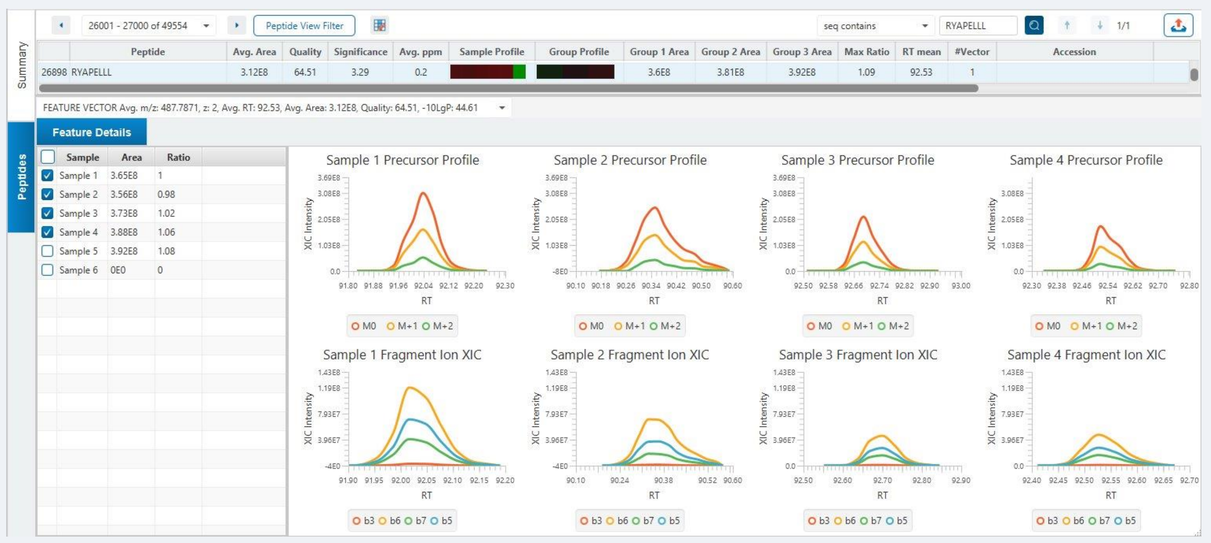

Figure 4. Peaks 12 peptide view for result demonstration

References

1. Guzman, Ulises H., et al. "Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition." *Nature biotechnology* (2024): 1-12.
2. Pak, HuiSong, et al. "Sensitive immunopeptidomics by leveraging available large-scale multi-HLA spectral libraries, data-independent acquisition, and MS/MS prediction." *Molecular & Cellular Proteomics* 20 (2021).
3. Zeng, Wen-Feng, et al. "AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics." *Nature Communications* 13.1 (2022): 7238.

Scan QR to…

@PEAKSmassspec @PEAKSProteomics Bioinformatics Solutions Inc.

Summary

- PEAKS 12 demonstrate comparable identification and quantification performance to the state-of-the-art methods in DIA data analysis.
- PEAKS 12 DIA Peptidome workflow incorporates multiple steps for immunopeptidome identification with several quality-related attributes filter, which provides accurate and sensitive automated identification of immunopeptidome.