High-Throughput PEAKS Workflow and SPIDER Algorithm for Large Scale Clinical Proteomics and Variant Identification using PEAKS Online X

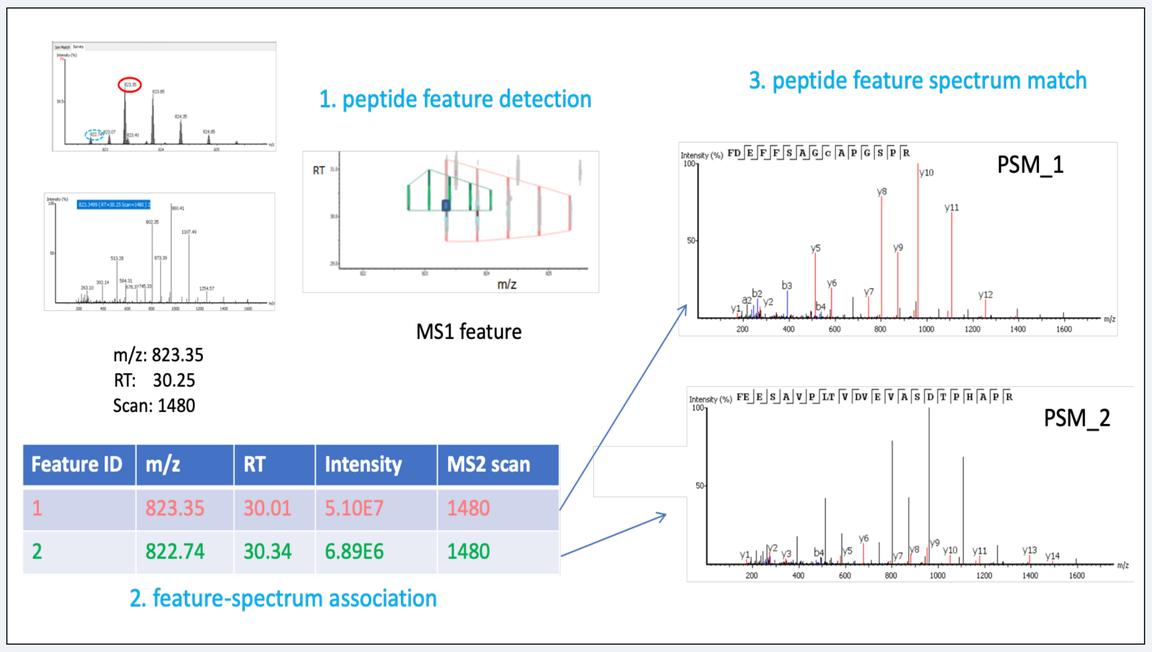
Jonathan R Krieger, Lei Xin, & Baozhen Shan
Bioinformatics Solutions Inc., Waterloo, Ontario, Canada

Introduction

The field of clinical proteomics is rapidly advancing and new mass spectrometry technology as well as streamlined sample preparation workflows are making the use of mass spectrometry in the clinical lab ever more attractive. Large scale clinical projects, whether research or clinical, often require quantitation of thousands of proteins, either by label free (LFQ) or by using isobaric tags such as TMT. PEAKS Online X is a new high-throughput protein sequencing software solution that runs on a shared resource, is flexible to scale, and is fully parallelized with the ability to run on any cluster or multi-cluster CPU machine. Herein we describe the use of PEAKS online in analyzing two published data sets; one employing LFQ, and the other employing TMT.

PEAXS Online X

Feature-based peptide identification



Feature ID	m/z	RT	Intensity	MS2 scan
1	823.35	30.01	5.10E7	1480
2	822.74	30.34	6.89E6	1480

Figure 1 - PEAKS Online X uses a feature-based identification method into its unique *de novo* assisted identification workflow.,

MS1 scans are extracted and pre-processed, from which the peptide features and their elution profiles are detected. Then according to the isolation window and the retention time, the relevant features are associated with the MS2 spectra. Therefore, each MS2 can be associated with multiple features and they are all analyzed in the database searching.

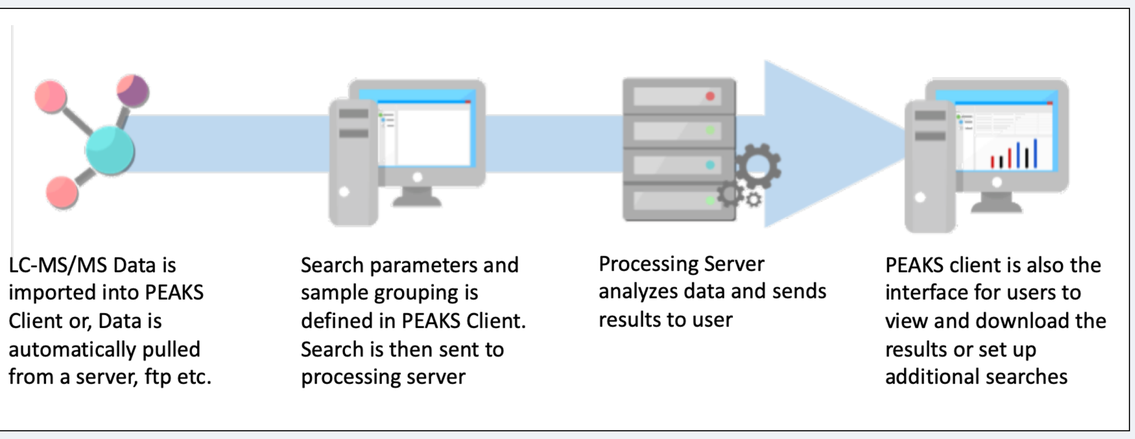


Figure 2 – Overview of the PEAKS Online X Architecture

PEAKS Online X is Scalable and offers Unparalleled Speed

- PEAKS Online X is a new high throughout search tool that is based on the same proven algorithms used in PEAKS Studio Software but on a larger scale.
- PEAKS Online X is specifically designed for large projects, that:
  - allows concurrent access from multiple users and;
  - can process multiple projects in parallel based on priority
- Samples can be added and removed, or parameters changed “on the fly” allowing for changes to be made without re-searching entire sample set
- High Throughput, high capacity
  - Reduce processing time 50x
  - Can handle 1000 samples or more

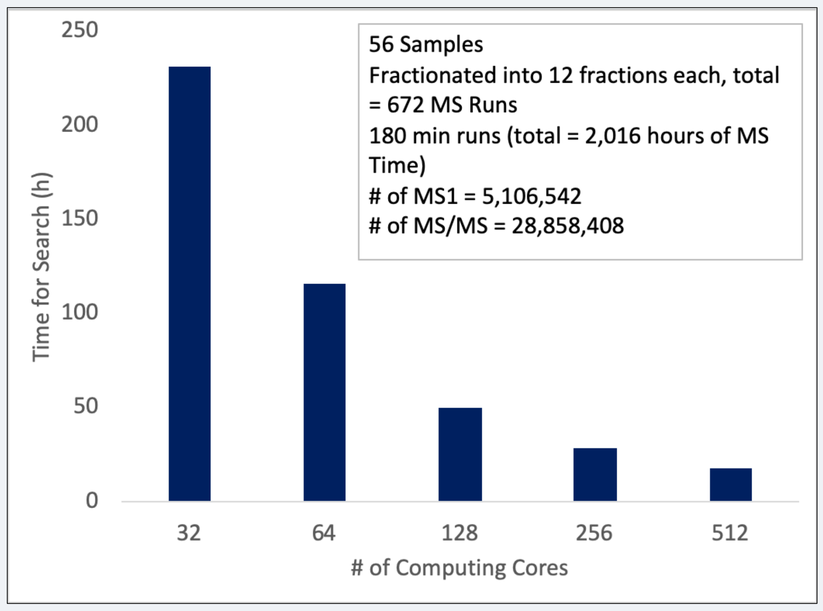


Figure 3 – PEAKS Online X performance increases linearly (and search speed inversely) with the number of computing cores available, making Peaks Online X scalable.

Contact Us

Bioinformatics Solutions Inc.
http://www.bioinfor.com
peaks@bioinfor.com
1-855-885-8288

References

[1]O’Connell J, *et al* 2018. Journal of Proteome Research)
Data downloaded via PRIDE Exchange: PXD007683

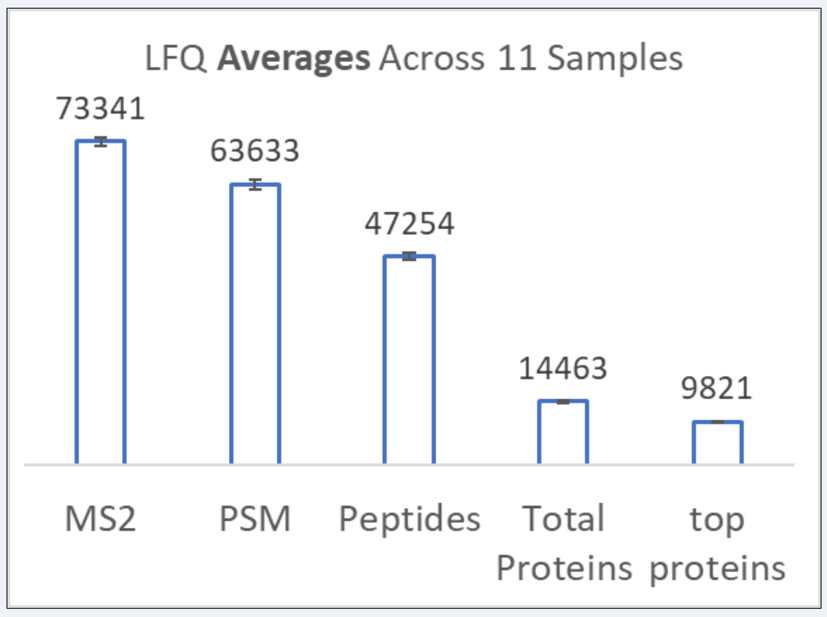
[2]Clinical Proteomic Tumor Analysis Consortium (NCI/NIH) (CPTAC) Clear Cell Renal Cell Carcinoma (CCRCC) Discovery Study
Data downloaded via NCI Proteomic Data Commons

Study Design & Results

Study Design:

- A previously generated data set was used for this study[1]
- Yeast lysate was spiked into human lysate at 3 different percentages relative to total protein content: Group 1 – 10%; Group 2 – 5%; and Group 3 – 3.3%. This was done to achieve a 2x-fold change between the first and second group, a 3x-fold change between group 1 and group 3, and a 1.5x-fold change between the final two groups.
- Samples were run in 2 ways:
  - 11 samples were run on 180min gradients for Label Free Quantification
  - 11 samples were multiplexed into a TMT 11plex, fractionated and run for TMT Quantification (MS3)

Results:



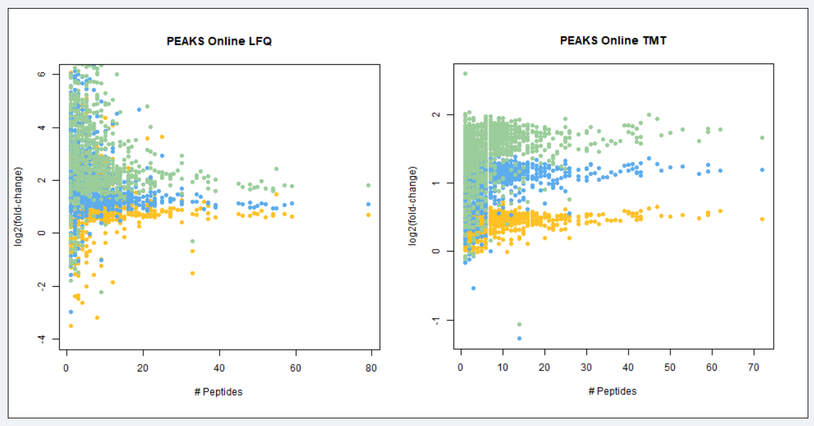
◀ Figure 4 – Results from Peaks Online X Search of LFQ Samples

▼ Table 1 – Summary of total unique peptides and proteins observed using PEAKS Online X compared to the original publication

Sample	Unique Peptides	Proteins	% Increase Observed with Online
LFQ Online	95251	12671	163%
LFQ Published	58383	7731	
TMT Online	89575	11384	122%
TMT Published	74791	9312	

- Searching this dataset using the new Peaks Online X algorithm resulted in a substantial increase in the number of identified unique peptides as well as proteins compared to the original searches.
- Similar to the conclusions obtained by the original authors of this dataset, increased precisions is obtained using the TMT method when compared to a LFQ method

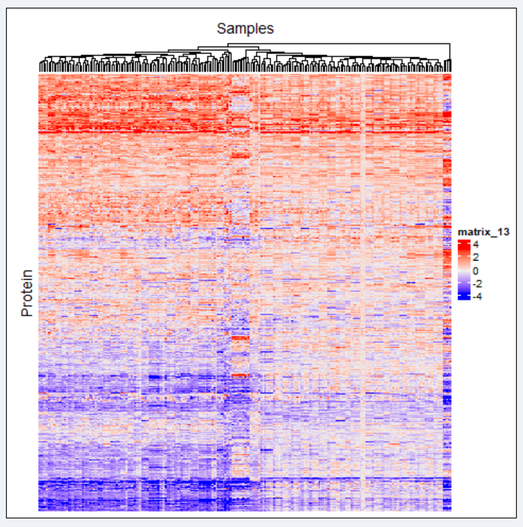
Figure 5 – Overview of yeast proteins spiked in at different ratios. In order to assess whether Peaks Online X was in fact quantifying proteins correctly, we examined the yeast proteins by looking at the fold change. Scatter plots of the average fold change are shown for each yeast protein for the 3-fold (green), 2-fold (blue) and 1.5 fold (yellow) against the number of peptides for each protein.



Applications and Sequence Variance Analysis

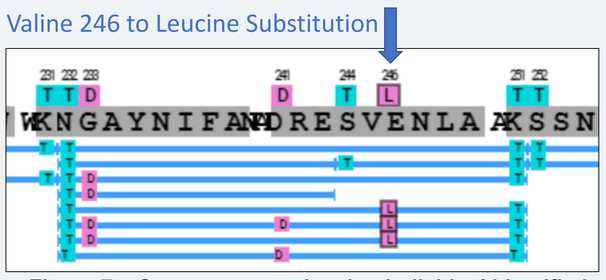
Study Design:

- A CPTAC[2] dataset of 110 tumor samples from 110 patients with Clear Cell Renal Cell Carcinoma (CCRCC) were analyzed by TMT following CPTAC guidelines
- Each TMT group was fractionated using high pH fractionation into 25 samples
- This data was re-run with PEAKS Online X

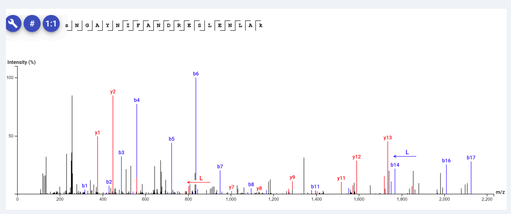


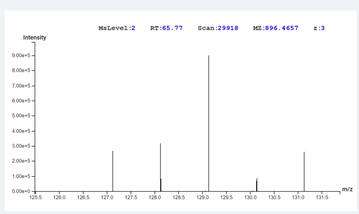
▲ Figure 6 – Heatmap of identified proteins. The pooled control samples all cluster to the far right as expected.

- 21,613,809 MS2, resulting in 7,401,366 PSMs
- 457,239 unique peptides, resulting **13,452** proteins identified, in 24 hours of search time.
- Proteins cluster into 6 groups
- **We used SPIDER algorithm to identify individual point mutations:**



▲ Figure 7 – Coverage map showing individual identified peptides of this region of the protein. In 3 identified peptides, V246L is identified. T represents TMT Tag modification while D represents deamidation





▲ Figure 8 – MS2 level evidence of V246L mutation.

◀ Figure 9 – MS3 level quantification of peptides containing the V246L mutation. Quantification suggests this mutation is more prevalent in some patients (129C) compared to others (126)