A Multi-Instrument, Skyline-Based Comparison of DIA Peptide Detection and Statistical Confidence Tools

Dario Amodei[1], Don Marsh[2], Hannes Rost[3], Lucia Espona Pernas[3], George Rosenberger[3], Reudi Aebersold[3], Parag Mallick[1], Michael J Maccoss[2], Brendan MacLean[2]

1 Canary Center for Early Cancer Detection, Stanford University 2 Department of Genome Sciences, University of Washington
3 ETH Zurich, Zurich, Switzerland

Introduction and Background 1

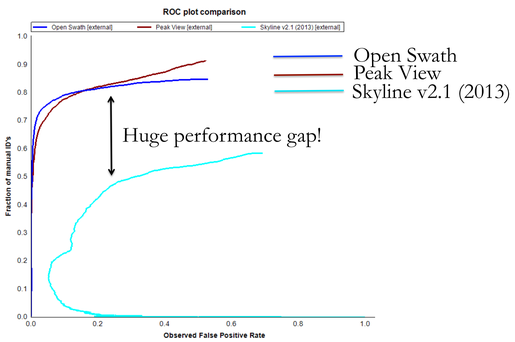
Users of mass spectrometry-based proteomics have traditionally been faced with a choice between two basic approaches: data dependent analysis (DDA, also known as shotgun proteomics), or Selected Reaction Monitoring (SRM, also known as targeted proteomics). DDA is a powerful tool for initial broad proteome-wide discovery but faces serious issues with reproducibility and sensitivity. SRM gives highly sensitive and reproducible assays but is limited to 10's or 100's of peptides. In the last three years, a new acquisition methodology, Data Independent Acquisition (DIA), potentially offers to combine precision of SRM with the wide scope of DDA.

Unfortunately, wide adoption of DIA has been limited by a number of practical challenges. In particular, peptide detection (peak picking) is both more challenging and larger scale in DIA relative to SRM data, necessitating an automated solution in order to successfully identify peptides in large DIA runs. Algorithms for DIA peak picking have been developed and even implemented in some software tools, including Open Swath, Spectronaut, and Peak View, but these tools are narrow, difficult to use, require special kits, are not compatible with all vendors, and/or lack the rich processing, visualization, and customization environment of Skyline.

Here, we implement advanced, mProphet-based DIA peak picking algorithms in Skyline, and compare its performance against other software tools for processing DIA data. We find that Skyline performs equivalent to or slightly better than other tools including Open Swath, Spectronaut, and ABSCIEX Peak View on a diverse range of datasets covering over ½ million peptide-run pairs with peptides spiked into human cell, yeast, and depleted plasma backgrounds, on both unlabeled and isotope pair data, using data acquired from both ABSCIEX 5600 and Thermo Q-Exactive instruments. Finally, we discuss future plans for incorporating retention time alignment and transition correlation across runs in order to increase peak picking accuracy and consistency.

Skyline Circa 2013 2

Dataset from Rost et al, *Nat. Biotech*, 2014 -- 344 heavy peptides spiked-in human background
10 dilution points x 3 replicates -- ABSCIEX 5600 SWATH 25 m/z windows
10350 total peptide-runs
80% of peptides had manually curated peak

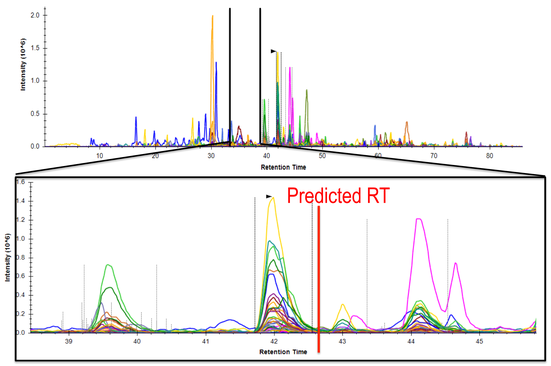


"Sensitivity": What fraction of all curated peaks were correctly identified (automated pick within curated bounds)?

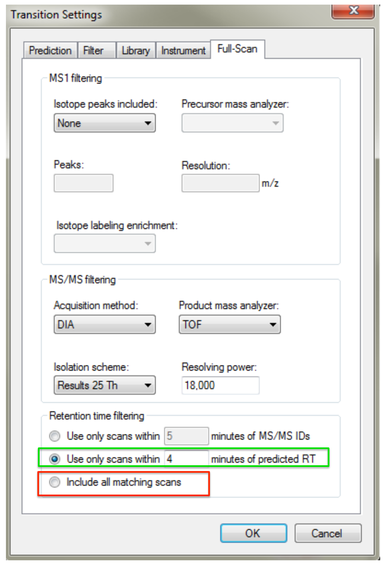
"1- Specificity": Fraction of incorrect peaks (outside bounds) or spurious peaks (manual curation says no peak)

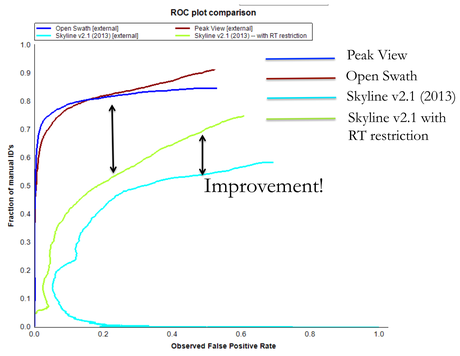
Retention Time Range Restriction 3

- Initially, Skyline did not restrict the range to the predicted retention window by default
- Most users ended up importing the whole chromatogram making peak picking difficult



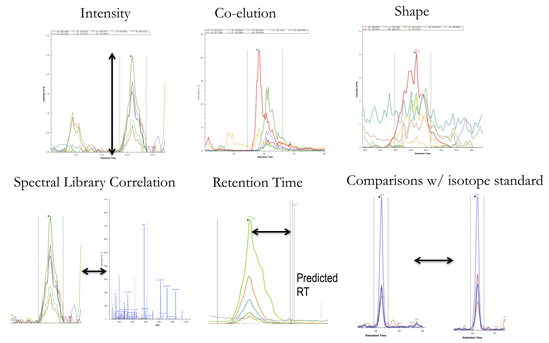
In response, we implemented retention range restriction around a predicted RT, and made this the default option in Skyline

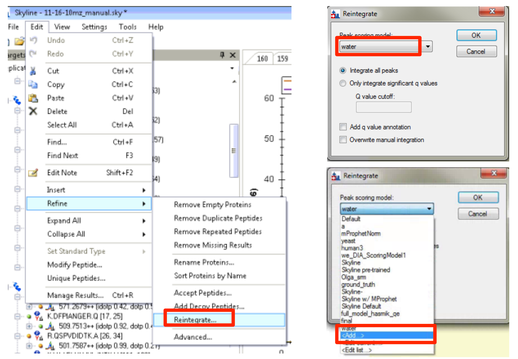


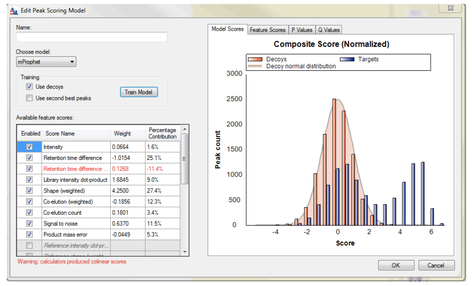


Implementing mProphet In Skyline 4

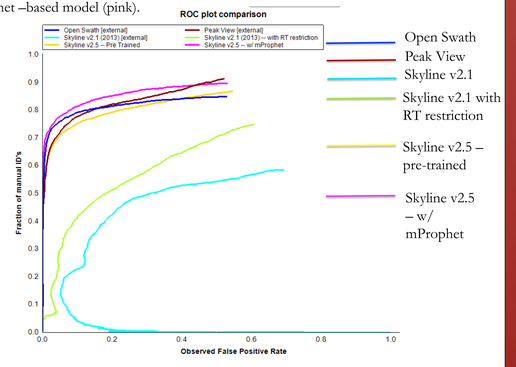
The mProphet algorithm: Use machine learning (linear discriminant analysis) to form a composite score based on these (and other) indicators of peak quality, using decoy peptides to train [1].





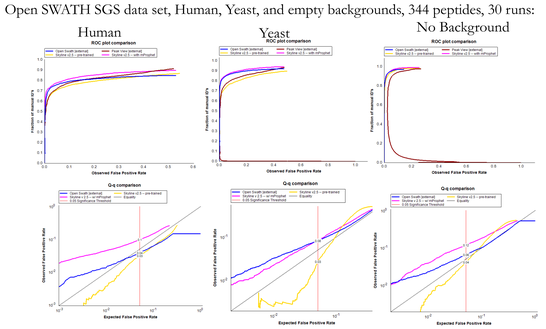


After implementation of mProphet, Skyline matches or exceeds other tools. We tested both the default Skyline model (yellow, the model which is used to score peaks automatically on import, which has fixed weights), and a custom-trained mProphet –based model (pink).

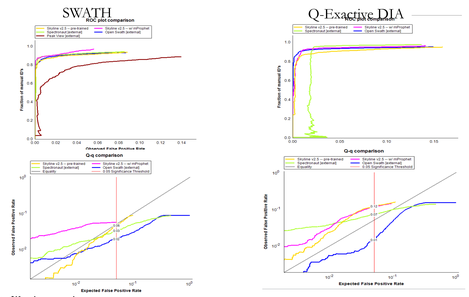


Results On Other Datasets 5

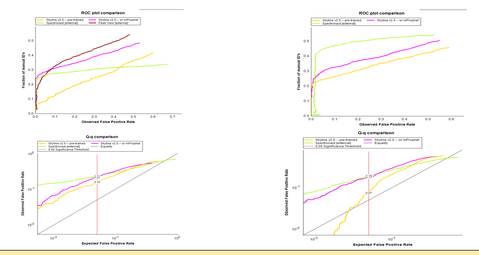
Open SWATH SGS data set, Human, Yeast, and empty backgrounds, 344 peptides, 30 runs:



Light/Heavy pairs spiked into depleted plasma, 136 peptides, 30 runs, 0.01-100 fm/ul. Heavy constant intensity standards:



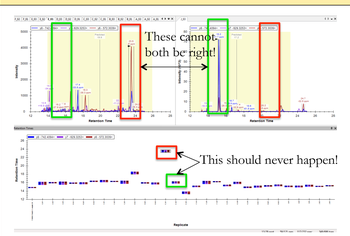
Light dilution series:



Conclusions 6

- Skyline's peak picking is now roughly equivalent to other tools for processing DIA data
- Proper treatment of isotope standards requires modification of published algorithms
- Most tools overstate statistical confidence

Future Directions 7



- Multi-replicate peak picking (ensuring the same molecule is measured across all runs)
- Comparison of Skyline with other tools on large number of DIA data sets
- Getting the statistics right
- Better feeding of DDA assay development pre-runs into DIA

References

[1] Reiter et al, mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods*, 2011
[2] Rost et al, *Nature Biotechnology*, 2014