Tutorial Overview

Outline

In this series of 5 tutorials we will learn how to use various tools for the analysis of DIA data (see figure 1). These tutorials can be grouped into 3 conceptual stages in the analysis workflow (red boxes in figure 1):

Library Generation
The first component required for analysis of DIA data is library generation. We will do this using two different approaches. In tutorial 1 we will perform a standard database search[1–5] of DDA data generated from the samples of interest to build a spectral library from which we will generate peptide query parameters to later analyze the DIA data. In tutorial 3, we will similarly build a spectral library, however, this will be done directly from the DIA data using DIA-Umpire[6].

DIA data analysis
We will use Skyline[7] (tutorial 2) or OpenSWATH[8] / PyProphet[9] / TRIC[10] (tutorial 4) to perform peptide-centric analysis of the DIA data using the libraries we have built in the first part.

Statistical analysis
In tutorial 5, the output from the DIA data analysis will be subjected to comparative statistical analysis using the R package MSstats[11]. To prepare the data for MSstats analysis we will use another R package, SWATH2Stats[12], to perform some filtering and reformatting.

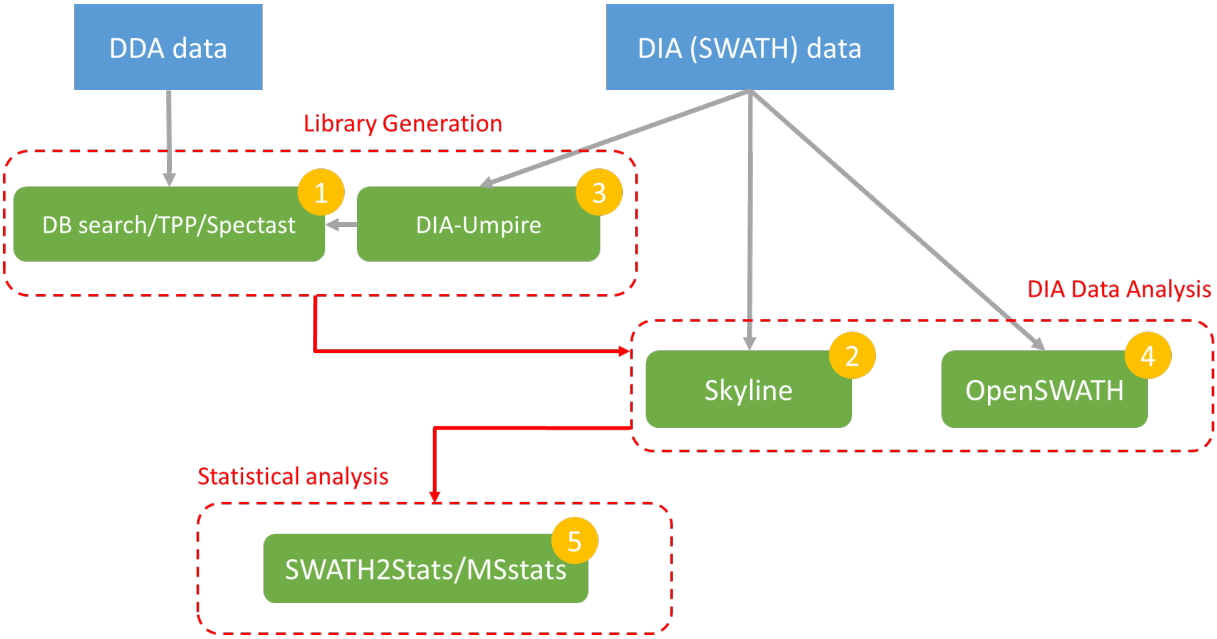


Figure 1 – DIA/SWATH Course tutorial overview

In addition to these tutorials where you will follow a written step-by-step procedure, we will have a walk-through session where Brendan MacLean, principle Skyline developer, will demonstrate the analysis of another data set in Skyline online with participants following his analysis.

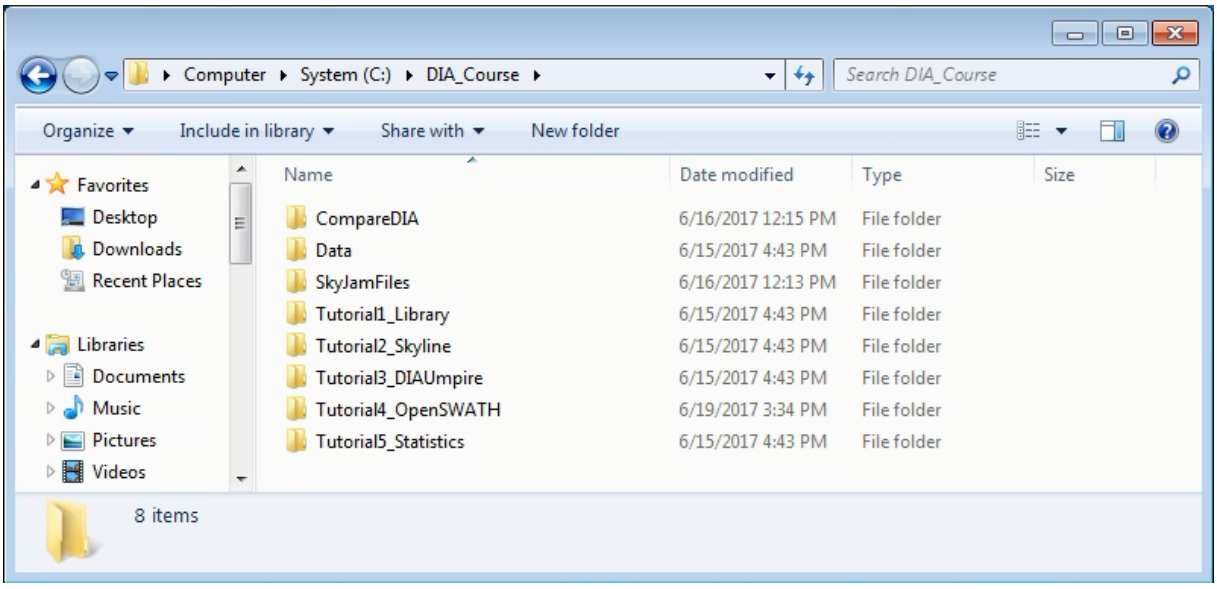
Dataset

The data we will analyze come from the LFQBench study[13] where quantitative benchmarking samples were created by mixing proteomes of 3 organisms in defined ratios (figure 2). The DIA (SWATH-MS) data were acquired in technical triplicate on a QqTOF instrument (6600 TripleTOF, AB Sciex) using a 64 variable width window precursor isolation scheme. Similarly, DDA data were acquired by analyzing samples from each of the 3 species in separate injections for the purpose of library generation. By using this data set we have a ground truth regarding the quantitative differences between samples A and B. This will allow us to determine whether out analysis is performing well or not.

Figure 2 – DIA/SWATH Course data overview (adapted from[13])

Directory Structure

All of the data file you need to perform the analysis are already loaded onto the virtual machines we will use for the analysis in the directory “C:\DIA_Course” (a copy of all data with the same structure is provided on your USB stick). In this directory you will find individual folders for each of the 5 tutorials which contain some files specific to each tutorial. The DDA and DIA data you need are in the “Data” directory along with some other general files that will be required. The “CompareDIA” and “SkyJamFiles” directories will be needed for the Skyline walk-through session(s). There is also a directory called “backup”. This directory contains all of the intermediate and final files from the completed tutorials. In case you experience some problems in a given section, you should be able to use the files in here to continue the tutorials successfully.



Appendix 1 – Software used in the tutorials

In the information email sent before the course we included a complete list of software and dependencies that had been installed on the virtual machines fort the purpose of the tutorials. Here we summarize again the tools required and provide links where they can be downloaded with detailed instructions for installation.

- Trans Proteomic Pipeline

Link:
http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP
http://tools.proteomecenter.org/wiki/index.php?title=Windows_Installation_Guide

A broad collection of software tools from the Institute for Systems Biology, Seattle focused primarily on DDA analysis of various types. This software is relatively easy to install on Windows (requires pre-installation of Perl and some MSVC/.Net Frameworks components) and is supported on Mac and Linux. The components we are using: Comet, PeptideProphet, InterProphet, SpectraST

- Skyline

Link:
https://skyline.ms/wiki/home/software/Skyline/page.view?name=default

Skyline is a freely-available and open source Windows client application for building Selected Reaction Monitoring (SRM) / Multiple Reaction Monitoring (MRM), Parallel Reaction Monitoring (PRM - Targeted MS/MS), Data Independent Acquisition (DIA/SWATH) and targeted DDA with MS1 quantitative methods and analyzing the resulting mass spectrometer data. Skyline is very well tested and supported and is easy to use due to a rich graphical user interface.

- DIA-Umpire

Link:
http://diaumpire.sourceforge.net/

DIA-Umpire is an open source Java program for computational analysis of data independent acquisition (DIA) mass spectrometry-based proteomics data. It enables untargeted peptide and protein identification and quantitation using DIA data.

- OpenSWATH / PyProphet / TRIC

Link:
http://openswath.org/en/latest/

The OpenSWATH Workflow enables targeted data analysis of data-independent acquisition (DIA) or SWATH-MS proteomic data. OpenSWATH is part of, and distributed with, the OpenMS (http://www.openms.de/) project. PyProphet and TRIC are Python tools. These tools require Python and several other dependencies to installed. A tutorial focused specifically on these tools has recently been published[14].

- SWATH2stats

Link:
https://bioconductor.org/packages/release/bioc/html/SWATH2stats.html

SWATH2stats is an R package intended to transform SWATH data from the OpenSWATH software into a format readable by other statistics packages while performing filtering, annotation and FDR estimation. SWATH2stats requires R and is distributed by the Bioconductor project (https://bioconductor.org/)

- MSstats

Link: http://msstats.org/msstats-2/

MSstats is an open-source R package for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. It requires R and is also available from the Bioconductor project (https://bioconductor.org/).

Appendix 2 – Raw data file conversion and centroiding

The raw data (DIA and DDA) has been converted to the mzXML[15] format. The data has further been centroided, and had an absolutely intensity threshold applied. The commands to perform these steps, using the ProteoWizard[16] tool MSConvert, are given below (note: there is also a GUI version of this tool, MSConvertGUI, available if you prefer this to the command line).

Note: While we do routinely analyze data either in profile or centroid mode, we would not typically apply the absolute intensity threshold. This thresholding has been applied for the purpose of the course to reduce the file size and reduce the analysis time. While the quality of the analysis should not be significantly affected by this step. Examples of how centroiding and thresholding affect the data are shown in Figure 3 below.

- Commands for converting to mzXML and centroiding

```
"c:\Program Files\ProteoWizard\ProteoWizard 3.0.10765\msconvert.exe"
--mzXML --filter "peakPicking true 1-" *.wiff
```

- Commands for converting to mzXML, centroiding, and thresholding (filter order is important!)

```
"c:\Program Files\ProteoWizard\ProteoWizard 3.0.10765\msconvert.exe"
--mzXML --filter "peakPicking true 1-" --filter "threshold absolute
2 most-intense" *.wiff
```

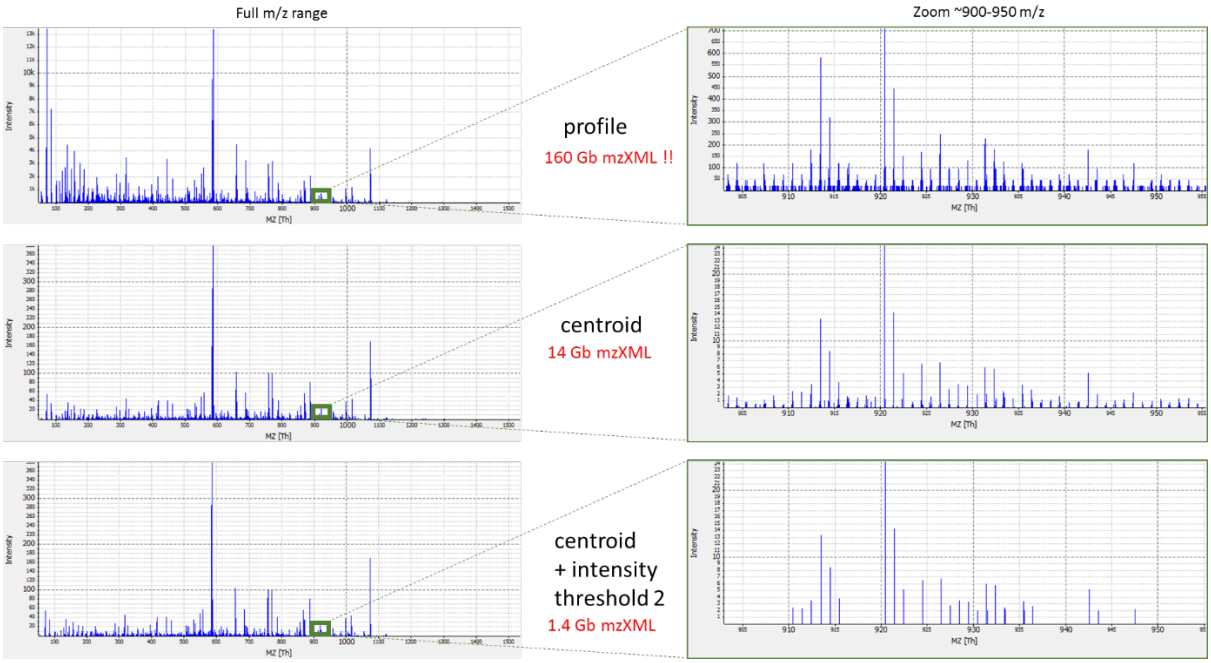


Figure 3 – effects of centroiding and intensity thresholding

References

1. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* **13,** 22–24 (2013).
2. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–92 (2002).
3. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10,** M111 007690 (2011).
4. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–67 (2007).
5. Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10,** 426–441 (2015).
6. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12,** 258–264 (2015).
7. Maclean, B. *et al.* Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **26,** (2010).
8. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32,** 219–223 (2014).
9. Teleman, J. *et al.* DIANA - algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* btu686 (2014). doi:10.1093/bioinformatics/btu686

10. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **advance online publication,** (2016).
11. Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30,** 2524–2526 (2014).
12. Blattmann, P., Heusel, M. & Aebersold, R. SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools. *PLOS ONE* **11,** e0153160 (2016).
13. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34,** 1130–1136 (2016).
14. Röst, H., Aebersold, R. & Schubert, O. Automated SWATH Data Analysis Using Targeted Extraction of Ion Chromatograms. in *Proteomics* (eds. Comai, L., Katz, J. E. & Mallick, P.) 289–307 (Springer New York, 2017). doi:10.1007/978-1-4939-6747-6_20
15. Pedrioli, P. G. *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–66 (2004).
16. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–6 (2008).

We would like to thank SystemsX for supporting the Zurich DIA / SWATH Course 2017.

SystemsX.ch
The Swiss Initiative in Systems Biology