omics.data.edu.au

Proteomics Requirements Summary

This document is a high level summary of the Proteomics requirements and information captured during the discoverability phase of the RDS Omics project

Summary

- The Proteomics analyses undertaken by QIMR Berghofer Protein Discovery Centre (QPDC) and by APAF aim at identifying and quantifying all the proteins present in the samples under investigation.
- Initially QPDC and APAF will use two different label free methods to quantify the relative protein abundance. QPDC will use the Precursor signal intensity based method whilst APAF will use a more recent technique named “Sequential window acquisition of all theoretical mass spectra” (SWATH-MS).

Groups Consulted

QIMR Berghofer Protein Discovery Centre (QPDC)

- Jeffrey Gorman, Group Leader Protein Discovery Centre
- Marcus Hastie, Proteomics Researcher
- Emma Norris, Bioinformatician

APAF (Australian Proteome Analysis Facility) Macquarie University, Sydney

- Prof Mark Molloy, APAF Director
- Dana Pascovici, Biostatistician
- Natasha Care, Scientific Officer (Mass Spectrometry)
- Dr Christoph Krisp, Research Fellow

Work Pattern

The general workflow consist of the following steps

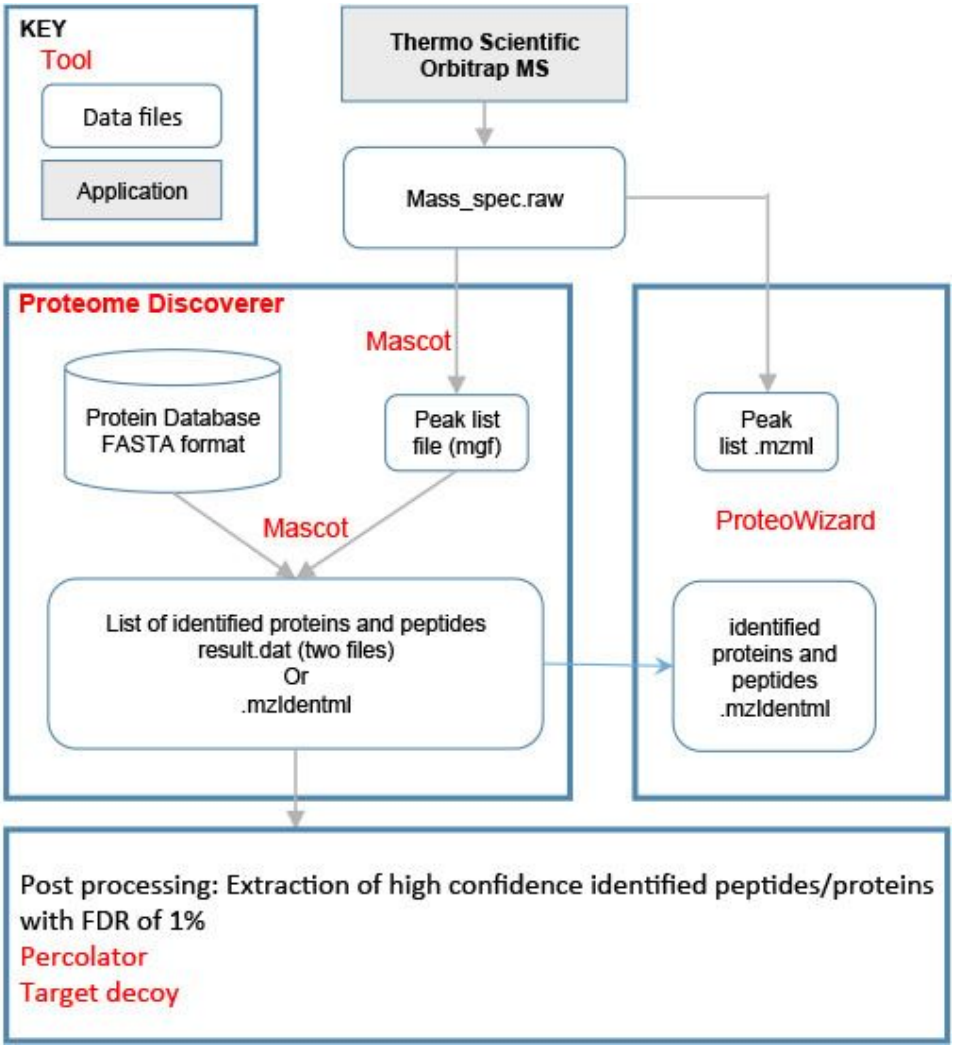
- Protein extraction
- Proteolytic digestion of proteins to generate peptides
- Desalting of the peptide mixture
- Injection of the peptides to an online LC-MS/MS system
- Bioinformatic analysis: protein identification and quantification
- Statistical analysis and data visualization

omics.data.edu.au

Work pattern at QIMR

- Depending on the experimental design and objectives of the analysis, Facility staff may choose to use the 5600 from ABSciex or one of the Orbitraps from Thermo Fisher Scientific.
- Raw data coming off the machines are saved on the Lab workstations and then on QIMR server.
- The Facility staff select the most suitable set of software depending on the instrument that was used and the nature of analysis that is required.
- Raw and Processed data are sent to the Clients
- Raw and Processed data might be submitted to international repositories such as PRIDE or Massive

Typical data/work flow at QIMR

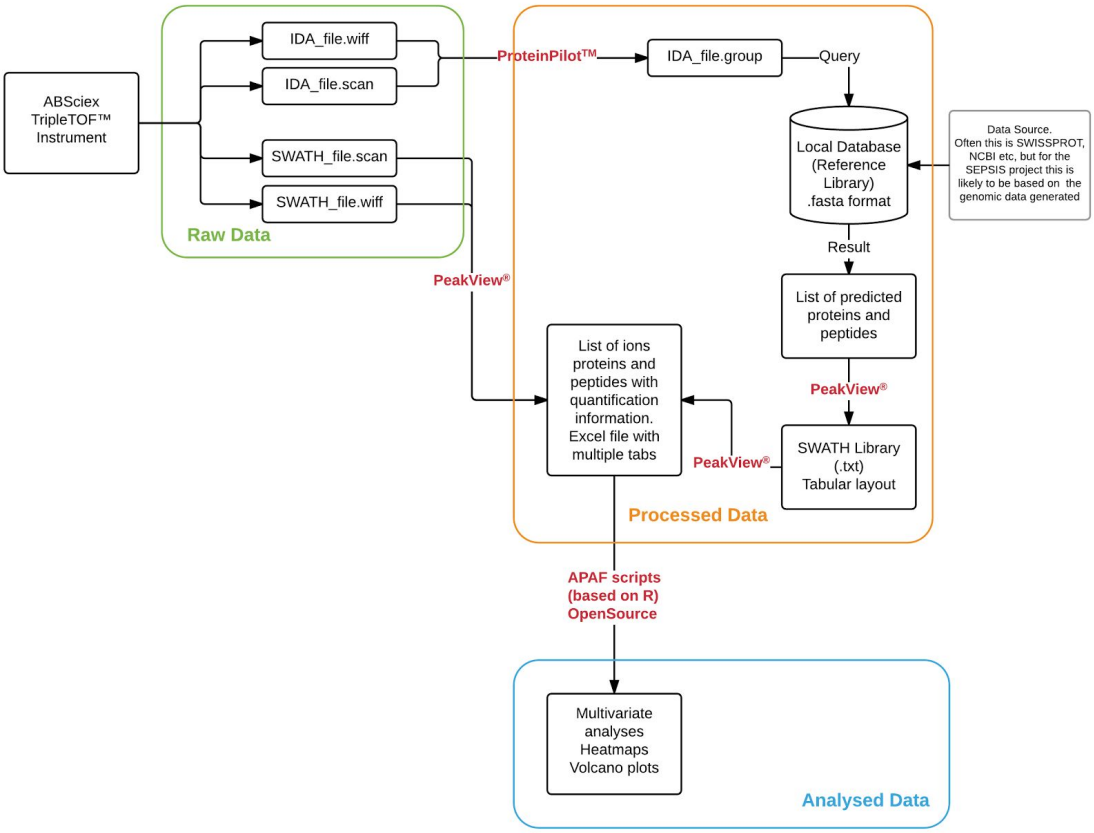


omics.data.edu.au

Work pattern at APAF

- ABSciex TripleTOF instrument used to generate raw SWATH-MS data.
- Raw data coming off the machines are saved on the Lab workstations and then on APAF server.
- APAF staff use AbSciex ProteinPilot and PeakView software to identify peaks in MS data, and generate a list of predicted proteins and peptides (with quantification data) - i.e. generate ‘processed’ data.
- APAF staff use an R-based script ("SWATH-pairs and overall") for statistical analyses on the processed data to generate ‘analysed’ data
- Raw, Processed and Analysed data are sent to the Clients
- See also diagram below:

APAF SWATH-MS Data/work flow



Data Overview

A detailed review of file formats commonly used in Mass Spectrometry proteomics is available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518119/

Raw data
The formats of raw data are binary and vendor specific.
The file size is around 2.5GB per sample.

- Thermo Scientific instruments: . raw extension

These files may contain profile mode spectra or centroided spectra as selected by the user (or both)

- SCIEX instruments: .wiff extension

These files may contain all information in a run or alternatively might contain only metadata and be paired with a file having a .wiff.scan extension that contains the spectra. (2-3GB per sample)

Peak list
The raw data need to be transformed into peak list before it can be processed. These files encode multiple MS/MS spectra in a single file via m/z, intensity pairs separated by headers.

- mzXML (open format) ~5GB per sample
- Mascot Generic Format (text file) ~1GB per sample

Processed files: Protein identification and quantification

- mzIdentML (open format)
- Mascot: .dat ~0.5GB to 1GB

Search space
Sequence search engines such as Mascot or X!Tandem require a list of potential protein sequences. Fasta format is mainly used.

Software/Applications used

- **Xcalibur software from Thermo Scientific**

Xcalibur is a Windows only software that provides for method setup, data acquisition, data processing, and reporting. It can export the MS data as .RAW extension file for subsequent analysis. The .RAW extension file output from Xcalibur software can use as an input for a wide range of software to convert data into peak lists. This includes Mascot Distiller and Mascot Daemon.
Xcalibur™ Software - Thermo Scientific

- **Proteome Discoverer from Thermo Scientific**

Proteome Discoverer that can read the raw data from Xcalibur, perform data conversion and protein identification. Integrate other search engines such as Mascot.
http://www.thermoscientific.com/en/product/proteome-discoverer-software.html

- **Proteowizard**

The ProteoWizard Library and Tools are a set of modular and extensible open-source, cross-platform tools and software libraries that facilitate proteomics data analysis. However, the conversion from raw data to mzXML can only be done on windows.
http://proteowizard.sourceforge.net

- **Mascot from Matrix Science**

Mascot allows searches against a range of sequence databases such as SwissProt. Matches are evaluated statistically by comparing observed and calculated peptide fragments.
http://www.matrixscience.com/server.html

- **ProteinPilot from Sciex**

Allows for an automatic peptide identification – searches hundreds of biological and other modifications, genetic variants, and unexpected cleavages simultaneously with the Paragon™ Algorithm, without an explosion of false positives, cost in expanded search time, or the complexity of multi-stage analyses that plague alternate approaches. Similar functionality to Mascot but works on Windows only. See
http://sciex.com/products/software/proteinpilot-software

- **PeakView® from Sciex**

Stand-alone software application for Windows mainly used for SWATH quantification. See
http://sciex.com/products/software/peakview-software

- **“SWATH-pairs and overall” R-based script**

APAF biostatisticians have developed an R-based script for statistical analyses on the processed data to generate ‘analysed’ data. It includes option for : multivariate analysis and a number of visualisation outputs: volcano plots and heatmaps. These scripts are compiled in GenePattern (http://www.broadinstitute.org/cancer/software/genepattern/), and are currently only available for internal APAF users (from the instance of GenePattern on the APAF server) for use.

Deficiencies in processes and tools

All software mentioned during the interviews are commercial software requiring to pay a license fee.

X! Tandem is an open source software that can perform the protein identification (search).

Note that ABSciex is increasingly moving towards inclusion of their peak identification and quantification software within a cloud-based analysis set-up - hosted on Illumina BaseSpace infrastructure: https://basespace.illumina.com/home/index http://sciex.com/applications/life-science-research/multi-omics-bioinformatics.

Note also that it is APAF's opinion that currently, interpretation of SWATH-MS data is beyond the reach of 90% of researchers, and requires specialist interpretation by experts in SWATH analysis. They would not recommend deployment of PeakView and ProteinPilot for SWATH analysis by the average researcher.

Relationship with other Streams

An iterative process often happen when genomics data is available for the sample under investigation. A customised search space can be created from the genome which allows for more accurate matches during protein identification. In turn, unmatched proteins might be the result of incorrect annotations (gene models) in the genome, providing feedback for improvement of such models.

APAF's area of expertise is squarely focussed on proteomics analysis, with exposure to other omics areas being limited.