msdata2: proteomics benchmarking datasets

Chong Tang, Laurent Gatto

Computational biology and bioinformatics lab, de Duve Institute, UCLouvain

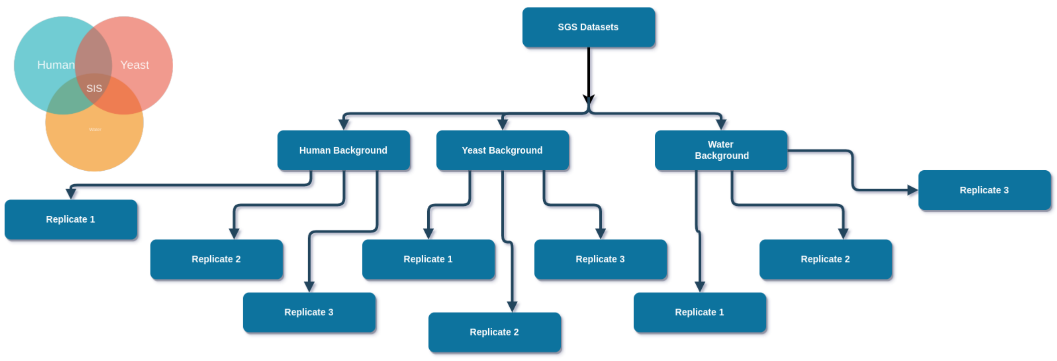
chong.tang@uclouvain.be – laurent.gatto@uclouvain.be

Summary

The `msdata2` package contains a set of public quantification/identification proteomics datasets. The purpose of the package is to provide standard and curated datasets to facilitate the benchmarking of identification/quantitative proteomics workflows.

The SGS Datasets

The first dataset in `msdata2` is OpenSWATH output from SGS dataset. The SWATH-MS Gold Standard (SGS) dataset consists of 90 SWATH-MS runs of 422 synthetic stable isotope-labeled standard (SIS) peptides in ten different dilution steps, spiked into three protein backgrounds of varying complexity (water, yeast and human), acquired in three technical replicates [1].



The SGS dataset was manually annotated, resulting in 342 identified and quantified peptides with three or four transitions each. In total, 30,780 chromatograms were inspected and 18,785 were annotated with one true peak group, whereas in 11,995 cases no peak was detected. The data were processed and converted into `MSnSet` objects [2] (see below and on the right):

Slot	Information
assayData	quantitative matrix with XIC values
phenoData	sample metadata
featureData	feature metadata (identification data, peptides sequences, ...)
experimentData	experimental methods and general annotations
processingData	processing information and log

Protein level overview

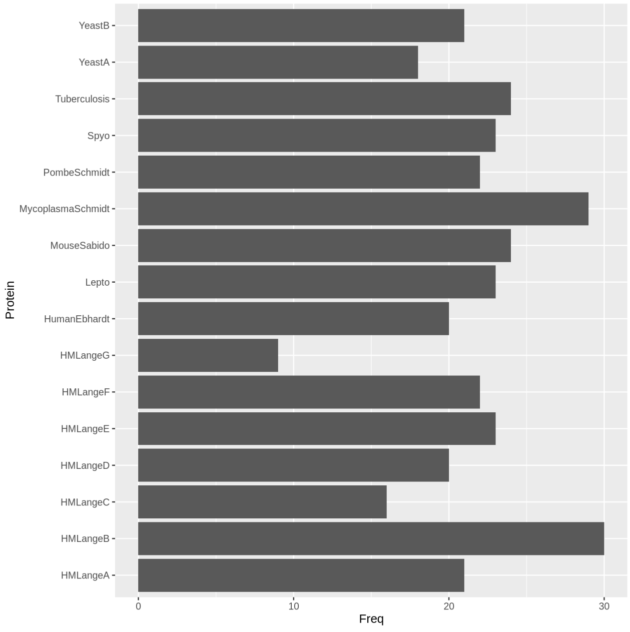


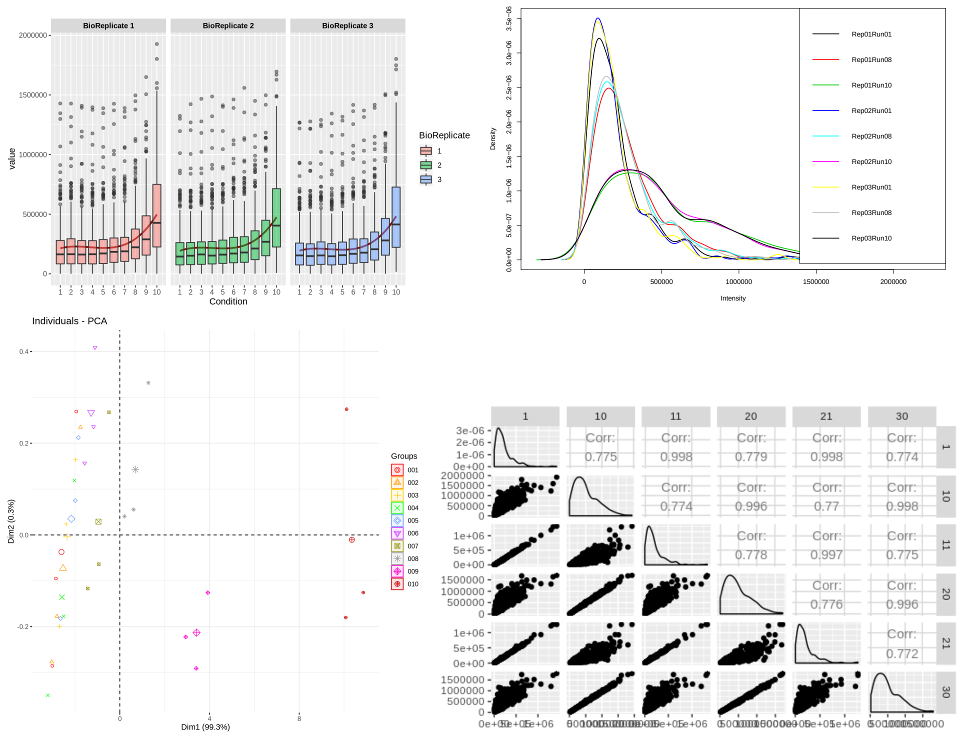
Figure 1: **Unique peptide count** for each protein in the human background.

Conclusion and Future Work

Our package `msdata2` will provide formatted labelled and label-free quantification and raw data for proteomics. In future, we will also include large raw MS data with `msdata2`, with the help of ExperimentHub. This package will be a fundamental module for a benchmarking work on different computational workflows.

Data exploration

Data visualization on the `MSnSet` with Human background: We can notice the similarities between each replicates on these plots. PCA on quantification values indicates *Run*08 - dilution 4X or dilution steps higher concentration would provide higher accuracy on identification/quantification.



Data preparation

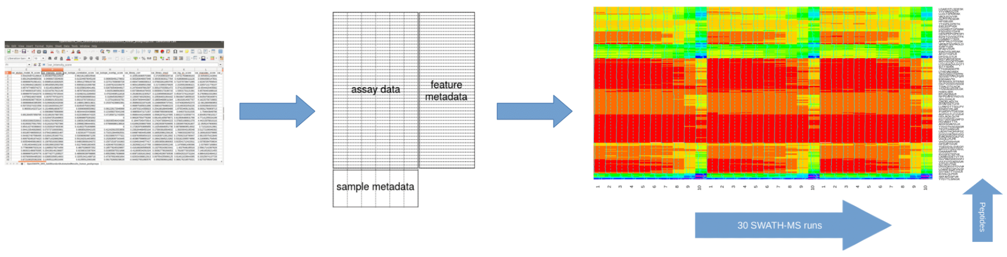


Figure 2: The OpenSWATH outputs for SGS dataset (human background) are cleaned, formatted and compressed as `MSnSet` object in `msdata2` [2].

Acknowledgements This research is supported by Wallonie-Bruxelles International (WBI), F.R.S.-FNRS and China Scholarship Council (CSC) co-funding fellowship.

References

[1] Röst,H.L. *et al.* (2014): *OpenSWATH enables automated, targeted analysis of data-independent acquisition ms data. Nat. Biotechnol.*, 32, 219–223.

[2] Gatto L, Lilley K (2012): *MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics*, 28, 288-289.