The msdata2 data package for proteomics benchmarking

Chong Tang, Laurent Gatto

Computational biology and bioinformatics lab, de Duve Institute, UCLouvain

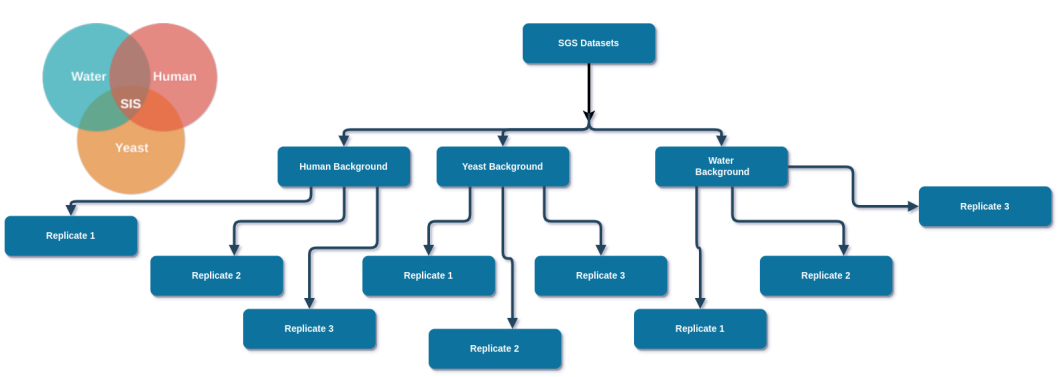
chong.tang@uclouvain.be – laurent.gatto@uclouvain.be

Summary

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

The SGS Datasets

The first dataset in `msdata2` is a SWATH experiment processed with 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 (1, 2, 4, 8, ..., 512 times), 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

Handling raw MS data

The raw MS data in `msdata2` will be handled using the `Spectra` package. The new `Spectra` package provides a common, flexible and a high-performance infrastructure to represent and handle MS data in R [3]. The spectrum data (m/z - intensity pairs) is represented by a matrix handled by a dedicated *Backend* class while keeping a consistent user experience. Efficient data processing is obtained through lazy and parallel evaluation.

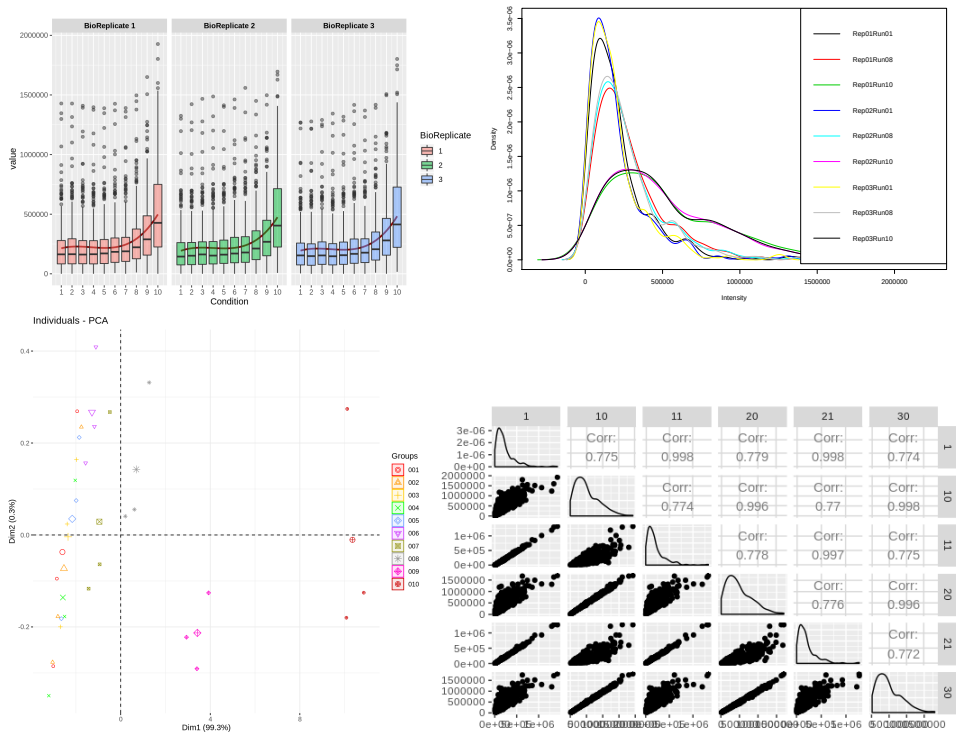
class	source files	storage	writeable
MsBackendDataFrame	manual	in-memory	yes
MsBackendHdf5Peaks	hdf5	on-disk	yes
MsBackendMzR	mzML, mzXML, CDF	on-disk	no
MsBackendHmdbXml	XML	on-disk	yes
MsBackendRawFileReader	Thermo .raw	on-disk	no

Conclusion and Future Work

Our package `msdata2` will provide formatted labelled and label-free quantification and raw data for proteomics. In the future, we will also include large raw MS data, making use of the ExperimentHub cloud infrastructure. Our goal is for `msdata2` to become a benchmarking tools on different computational proteomics workflows.

Data exploration

Data visualization on the `MSnSet` with Human background. We can notice the good consistency between replicates. The principal component analysis confirms good separation for runs 10 to 7 (up to dilution 8x). Further dilutions become much more difficult to tell apart.



Data preparation

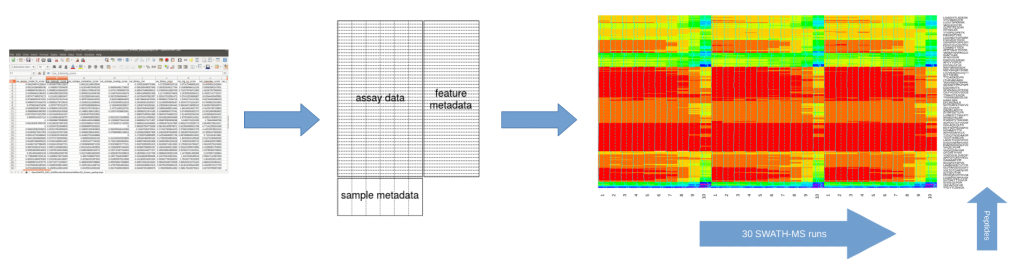


Figure 1: The OpenSWATH outputs for SGS dataset (human background) are cleaned, formatted and compressed as `MSnSet` object in `msdata2` [2].

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. http://lgatto.github.io/MSnbase.

[3] Laurent Gatto, Johannes Rainer and Sebastian Gibb (2019). **Spectra**: Spectra Infrastructure for Mass Spectrometry Data. R package version 0.3.0. https://rformassspectrometry.github.io/Spectra/.