Towards exploratory data science for multi-omics applications.

Edoardo Coronado[1,4], Javier López[1,4], Christian Heyd[2], Michael Nguyen[2], Nikolas Kessler[1], Heiko Neuweger[1], Dennis Trede[1], Jonathan Krieger[3] and Tharan Srikumar[3]

[1]Bruker Daltonik GmbH, Fahrenheitstraße 4, 28359 Bremen, Germany, [2]Bruker Scientific LLC, Billerica MA USA , [3]Bruker Ltd., 2800 Highpoint Dr. Suite#206, Milton, ON L9T 6P4, Canada,
[4]Centro de Investigación Mestrelab (CIM), Santiago de Compostela, Spain

Introduction:

Bruker ProteoScape (BPS), formerly PaSER, has been transforming into a comprehensive proteomics data analysis platform that can integrate third-party tools while utilizing the concept of data streaming to realize fully customizable real-time processing workflows including on-the fly decision making based on the data generated. With workflows for the analysis of dda-PASEF, dia-PASEF and diagonal-PASEF supported in BPS, we here focused on the next steps of transforming data to knowledge, by focusing on visualization and post-quantitative analysis tools integrated into BPS. In addition to several QC plots for the dataset, tools to perform dimension reduction analysis (using PCA, t-SNE or LDA) as well as differential expression analysis and visualize the data in an interactive volcano plot have been added. These additional visualizations made directly available within BPS reduces the time to knowledge.

To illustrate this functionality, we utilized 2 publicly available datasets. MSV000092140[1], characterizes Bacillus subtilis cells and spores using a multi-omics (metabolomics and proteomics) approach. While the second data set (PXD023175[2]) a COVID-19 multi-omics study was chosen to be able to illustrate a more typical use case, where the initial data science exploration could be carried out within BPS. In both cases, after the completion of quantitative analysis, we were able to generate publication ready figure panels within seconds that recapitulated the data shown in the original publications. In addition to the proteomics data, the interactive viewer accepts uploads of other annotated datasets including metabolomics and lipidomics searched with MetaboScape and seamlessly appends the data for combined multi-omic analysis. The combined analysis highlighted both complementary and agnostic metabolites\lipids and protein groups into a single visualization.

Taken together, this advancement to the Bruker ProteoScape suite allows for even quicker understanding of complex datasets without the need to engage with 3rd party tools or complex post-processing analysis.

Acknowledgements:

Bruker thanks the creation of the Mestrelab Research Center (CIM) subsidized by the Galician Innovation Agency. Operation funded by the Xunta de Galicia, through the business aid program for the creation and integration of new business research centers 001_IN853D_2022.

[1]https://doi.org/10.1021/acs.jproteome.3c00386
[2]https://doi.org/10.1093/nar/gkac1019

Fig. 1: Visualizations are available for any quantitative analysis conducted within BPS. (A) ID summaries by sample or (B) sample groups along with CV data (C) are provided at three levels: protein groups, precursors or stripped peptides. (D) Upset plots by sample group (or by sample) are readily generated. Interactive volcano plots can be generated when 2 samples groups are selected. A variation of the volcano plots (E) can accept the import of metabolomics\lipidomics data from MetaboScape. Allowing for a first glimpse into the combined multi-omics data.

Results:

Raw data from Huang et. al.[1] was reprocessed in ProteoScape and Metaboscape respectively for proteomics and metabolomics data. A mapping file was generated per the publication. An export was generated from Metaboscape and uploaded directly to ProteoScape, along with the mapping file. This allowed the concatenation of metabolite data to the proteomics data. Differential expression analysis, visualized in an interactive Volcano plot showed correlated protein and metabolite targets, as shown in the publication. Similarly, raw data from Wang et. al[2] was also reprocessed in ProteoScape

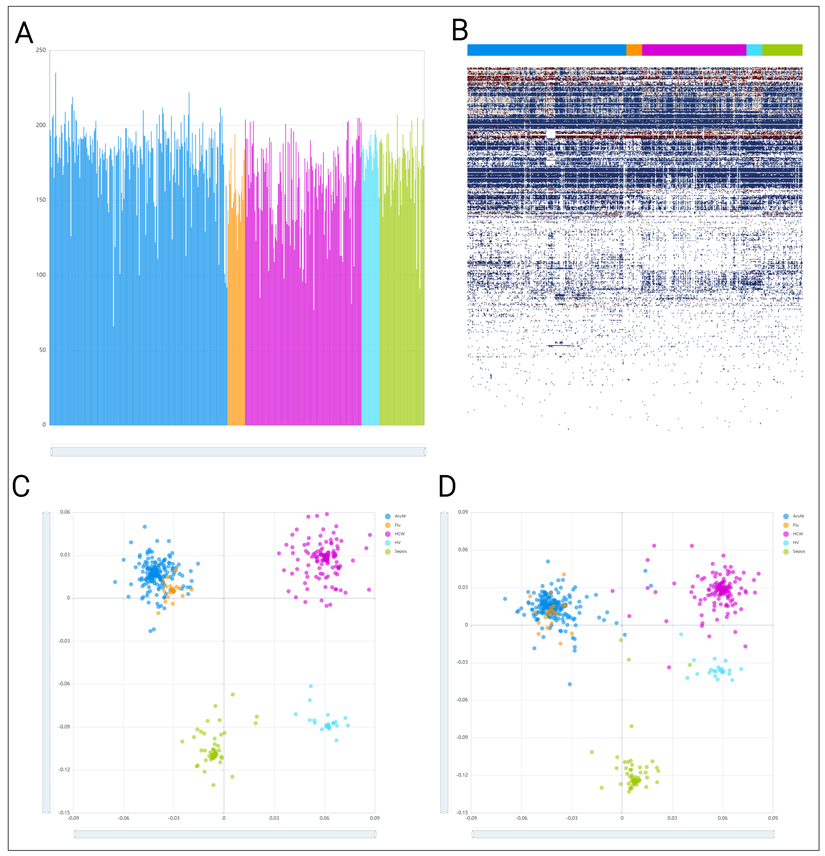

Fig. 2: Data from PXD023175, covering 467 un-depleted patient plasma samples were analyzed in ProteoScape and Metaboscape. The distribution of identified proteins (A) across all samples as well as K-means clustered heatmap (B) are readily generated with ProteoScape. Dimension reduction analysis (using LDA) on the proteomics data (C) and combined with lipidomics data (D) are both able to cluster patients by known symptomatic phenotypes.

and Metaboscape (lipidomics data). Dimension reduction of the dataset, using LDA (linear discriminant analysis), showed good separation based on observed patient segregation at the clinic. Addition of lipidomics data led to tighter clustering by symptomatic phenotypes showing the complementary nature of the lipidomics and proteomics data. Principal component analysis has been conducted, and primary influencers are under further characterization. Taken together, BPS now provides many commonly utilized visualizations in a facile manner.

Conclusion:

- Post-quantitative visualizations allow users to explore their data further
- Dimension reduction analysis and interactive differential analysis can be conducted seamlessly
- Metabolomics\Lipidomics data can be concatenated easily, allowing a first glimpse into the combined multi-omics data.

Technology