Enhanced insight generation through automated transformation of historical experiments into Quantitative Knowledge Base

Anna Quaglieri[1], Aaron Triantafyllidis[1], Bradley Green[1], Mark R. Condina[1,2], Paula Burton Ngov[1], Giuseppe Infusini[1] and Andrew I. Webb[1,3,4]

[1]Mass Dynamics, Melbourne, Victoria 3000, Australia, [2]Clinical & Health Sciences, University of South Australia, Adelaide, SA 5095, Australia, [3]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia, [4]Department of Medical Biology, University of Melbourne, Melbourne, Victoria 3010, Australia

Introduction

Despite the rapid expansion of both volume and complexity of proteomics data, the ability to easily leverage existing studies to enhance the interpretation of one's experimental results is still largely overlooked due to the computational complexity of the problem. This step is instead often left to the researcher's own capabilities, knowledge and bias, potentially losing key results.

The **Mass Dynamics (MD)** platform is engineered to build a quantitative knowledge-base from a user's or a lab's historical data as they analyze their experiments. It also provides the capacity to interrogate and explore the analysis of proteins across the user's past experiments and those publicly available on the platform.

Here, we present a novel set of features within the MD 2.0[1] application, enabling users to interrogate multiple experiments into a unified analysis, to enhance:

- The results interpretation of single experiments by combining multiple datasets within the application;
- Re-analyse new datasets using various statistical methods;
- Conveniently validate the various dataset's quality through interactive visual tools and define parameters for subsequent multi-experiment evaluations.

Case studies: How MD 2.0 facilitates analysis

Here we present how MD2.0 facilitates proteomics analyses from data imports, processing and knowledge generation. The visualisations in this case study were created using the LFQ DDA datasets with PRIDE identifiers PXD016433, PXD016447, and PXD019678[2] containing:

- 36 LFQ human urine samples (chronic kidney disease [CKD] stages 1, 3, and 5 vs healthy controls);
- LFQ analysis kidney tissue samples from a rat CKD model following filter-aided sample preparation (FASP); and
- Tandem mass tag (TMT)-labeled MS analysis of human primary glomerular endothelial cells (GECs) and proximal tubular epithelial cells (PTECs) before and after inducing 24-h hypoxia injury.

Accurately assess the quality of your experiment

- **Upload** data from pre-processed analysis (e.g. MaxQuant, Bruker ProteoScape™, Spectronaut, DIA-NN, MSFragger);
- **Combine modules** or use **templated analyses** to determine data quality prior to further result exploration;
- Produce **interactive statistical visualizations** such as RLE plots, Missingness and CV distributions to assess the quality of your data;
- Streamline experiment quality assessment prior to sharing to all collaborators.

Figure 1: Improved upload options for DDA and DIA MS data.

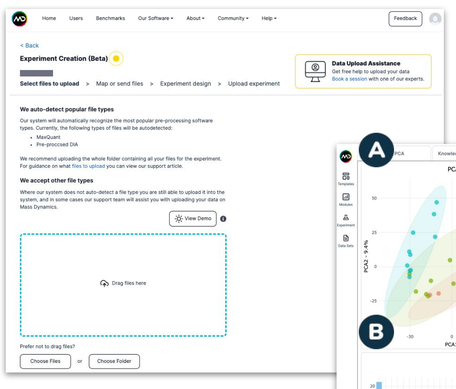
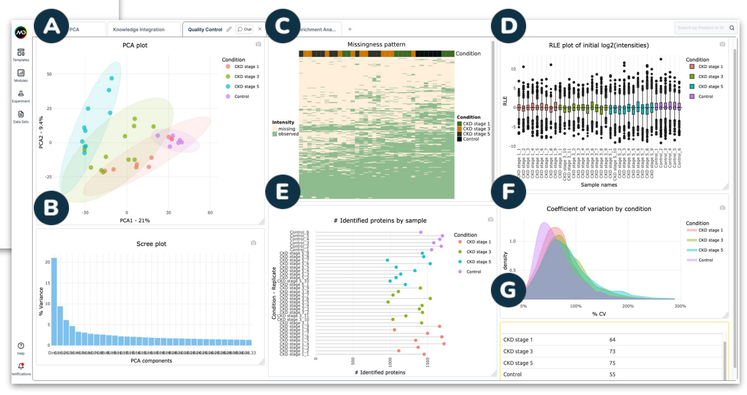


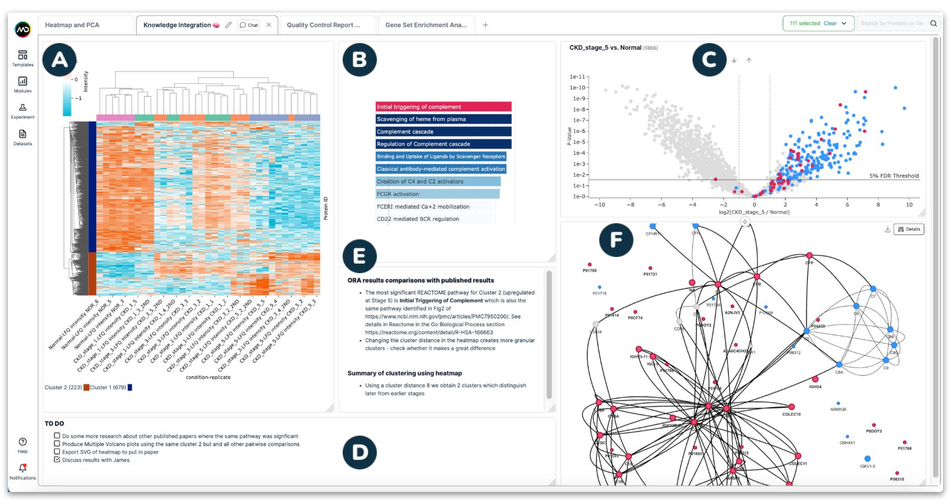
Figure 2: Example output from MD2.0 after uploading the PXD016433 (CKD) dataset via Generic Format upload.



Interactively explore quantitative proteomics data

- Directly run **state-of-the-art statistical methods** for differential expression and knowledge interpretation such as limma[3] and CAMERA[4];
- Work dynamically with **alternate data visualizations** like heatmaps, upset plots, violin plots etc. orchestrated with **human centered design** principles in a **customisable workspace**;
- Plots are generated using Plotly[5], Seaborn[6], Matplotlib[7] and UpSetPlot[8];
- Modules are **interactive**, allowing you to easily track proteins of interest and observe how they behave across different visualizations.

Figure 3: Knowledge Integration for PXD016433.

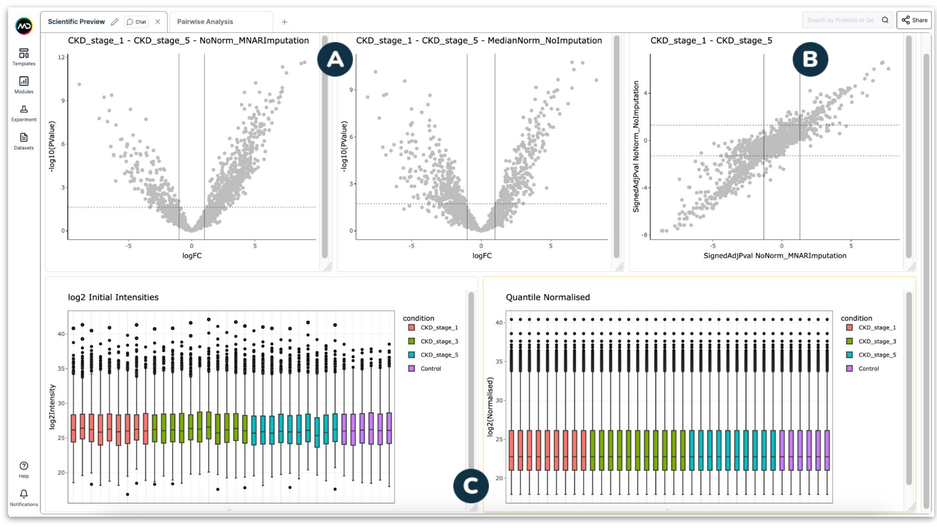


- **Over representation analysis (ORA)** with the Reactome API[9] database and **gene set enrichment analysis with CAMERA**[4] can be performed with the click of a button to connect your analysis results with external knowledge databases;
- The gene set libraries are assembled from publicly available knowledge bases including **UniProt**[10], **Gene Ontology** (GO)[11], **Reactome**, **MsigDB**[12];
- Generated protein lists can be interrogated against **STRING** protein–protein Interaction network database[13].

Interpret the effect of various processing steps by combining multiple datasets

- The **new Dataset service** allows you to re-process your experiment, leveraging different processing steps (e.g. normalization, imputation approaches, experimental design changes);
- Easily generate side-by-side visuals to interrogate the response of differential protein abundance due to processing;
- Compare across different analyses modalities within one experiment (Figure 4);

Figure 4: Comparing different datasets within one experiment for PXD016433.



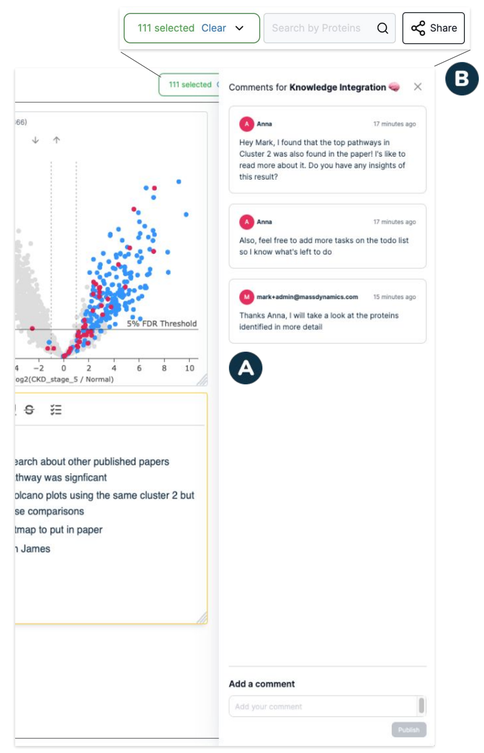
- Compare across different datasets and experiments (Figure 5);

Figure 5: Comparing different experiments using the multi-experiment option - PXD016433 and PXD016447.

Share, collaborate and publish, allow independent analysis

- MD 2.0 has a **cloud-based infrastructure**, not requiring any downloads or licences;
- It has **sharing and commenting features,** with direct notifications in app and by email and ability to define user access rights;
- It allows **notes taking and checklists** for improved collaboration;
- It allows **export** entire reports or specific modules to *.SVG, *.PNG as required;
- **Analysis** of results in app **can be made public** to allow interactive assessment of results by reviewers and community.

Figure 6. Example taking notes, setting tasks, collaborating live with chat box and sharing options.



Future Directions

- **Broaden upload options** for various pre-processed outputs;
- **Broaden statistical analyses** options, e.g. time series and dose response analyses;
- Increase support to more **knowledge bases** (EnrichR, Alphafold, etc.);
- More flexibility with **customized templates** and ability to integrate new templates and analysis with community-based input;
- Workflows to support post-translational modifications (PTMs), including phosphorylation.

Figures

Figure 1: Extended upload options for DDA and DIA MS data.

Figure 2: Example output from MD2.0 after uploading the PXD016433 dataset via Generic Format upload. The user has defined a summary of visuals to assess data quality which include: (A) Principal Components Analysis (PCA); (B) Scree plots of PCA; (C) Missingness heatmap; (D) Relative Log Expression (RLE) plots; (E) Number of identified proteins; (F) CV distribution plot coloured by condition; (G) CV distribution table by condition.

Figure 3: Knowledge Integration for PXD016433. (A) Heatmap to identify 2x main clusters. Cluster 2 (n=223) consisted of proteins that sequentially increased with increasing CKD severity and selected for ORA. (B) Barplot showing the results of the Reactome ORA analysis. The analysis reveals significant representation of pathways such as complement activation, as previously described[1]. (C) MD 2.0 allows users to link selected pathways and their proteins in a pairwise comparison results with controls. (D, E) Mass Dynamics has dedicated check-list (D) and text (E) modules that allow you to setup checklists for you or your team and take notes/information around insights you have made from the analyses in the tab. (F) STRING-DB results from a generated protein list obtained from a pathway identified from ORA..

Figure 4: Comparing different datasets within one experiment for PXD016433; The new 'dataset' service allows one to interrogate how results change when different processing is used. (A) Volcano plot modules compare pairwise analysis between Stage 1 and Stage 5 CKD, where no normalization + minimum not at random (MNAR) is compared against median normalization with no imputation. (B) Log Log plot - A 2D plot that shows the signed adjusted p-values on the log scale for two sets of selected pairwise comparisons, one comparison on the x-axis and one on the y-axis. The sign is derived from the log Ratio. (C) Boxplots showing the change of log expression before and after quantile normalization.

Figure 5: Comparing different experiments using the multi-experiment option - PXD016433 and PXD016447. (A) Volcano plot module showing proteins of interest selected for further interrogation across other experiments. (B) Multi Experiment Trend Analysis allows interrogation of selected proteins across any other experiments loaded into Mass Dynamics. (C) The new multi-experiment functionality allows you to generate volcano plots showing pairwise comparisons from other experiments.

Figure 6: Example taking notes, setting tasks, collaborating live with chat box and sharing experiment on MD 2.0. (A) Add comments for live collaboration with collaborators that have access to the experiment; (B) "Share" button to share experiment with collaborators (covered by the live chat in main panel).

References

1. Quaglieri A, et al. Mass Dynamics 2.0: An improved modular web-based platform for accelerated proteomics insight generation and decision making. BioRxiv (2022) 2. Kim JE, et al. Multisample Mass Spectrometry-Based Approach for Discovering Injury Markers in Chronic Kidney Disease. MCP (2021). 3. Phipson, B. et al. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. Annals of Applied Statistics 10, 946-963 (2016). 4. Wu, D, and Smyth, GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research 40, e133 (2012). 5. Inc, P. T. Collaborative data science. Montreal: Plotly Technologies Inc Montral (2015). 6. Waskom, M. seaborn: statistical data visualization. J. Open Source Softw. 6, 3021 (2021). 7. Hunter, J. D. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95 (2007). 8. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. 20, 1983–1992 (2014). 9. Fabregat A, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics;18(1):142 (2017). 10. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489 (2021). 11. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 47, D330–D338 (2019). 12. Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS (2005). 13. Szklarczyk D. et. al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome- wide experimental datasets. Nucleic Acids Res 47(D1):D607–D613 (2019).