

Mass Dynamics 2.0: A modular web-based platform for accelerated insight generation for proteomics.

Anna Quagliari¹, Aaron Triantafyllidis¹, Bradley Green¹, Mark R. Condina^{1,2}, Paula Burton Ngov¹, Giuseppe Infusini¹ and Andrew I. Webb^{1,3,4}

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



Introduction

Despite the rapid expansion of both volume and complexity of proteomics data, the biological knowledge surfaced from these datasets is still largely underutilized due to the complexity of the analysis pipelines requiring teams with multiple specialised expertise.

The **Mass Dynamics (MD)** cloud- and web-based platform is engineered to streamline the analysis and interpretation of quantitative proteomics data by bringing together the right people in one platform.

MD 2.0² reframes the analysis and decision-making process for proteomics with:

- Easy to use 'drag and drop' analysis and visualisation modules;
- Ability to interrogate multiple uploaded experiments into a unified analysis;
- Ability to re-analyse new datasets using various statistical methods;
- Sharing and collaboration capabilities with multiple users on the same experiment.

Case studies: How MD 2.0 facilitates analysis

The visualisations in this case study were created using the LFQ DDA datasets with PRIDE identifiers PXD016433, PXD016447, and PXD019678² 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 ProteoScopeTM, 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

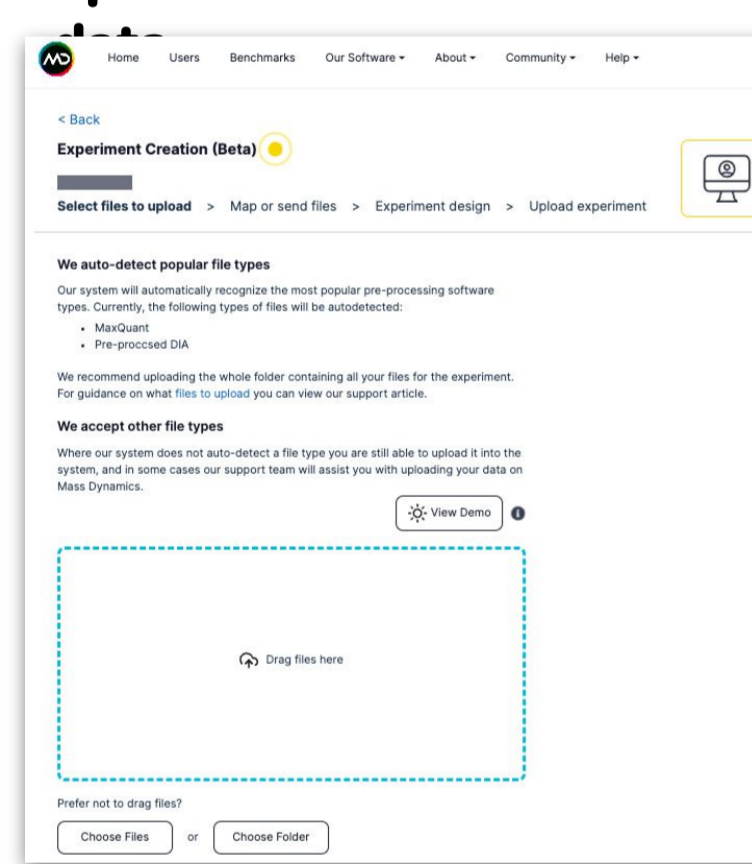
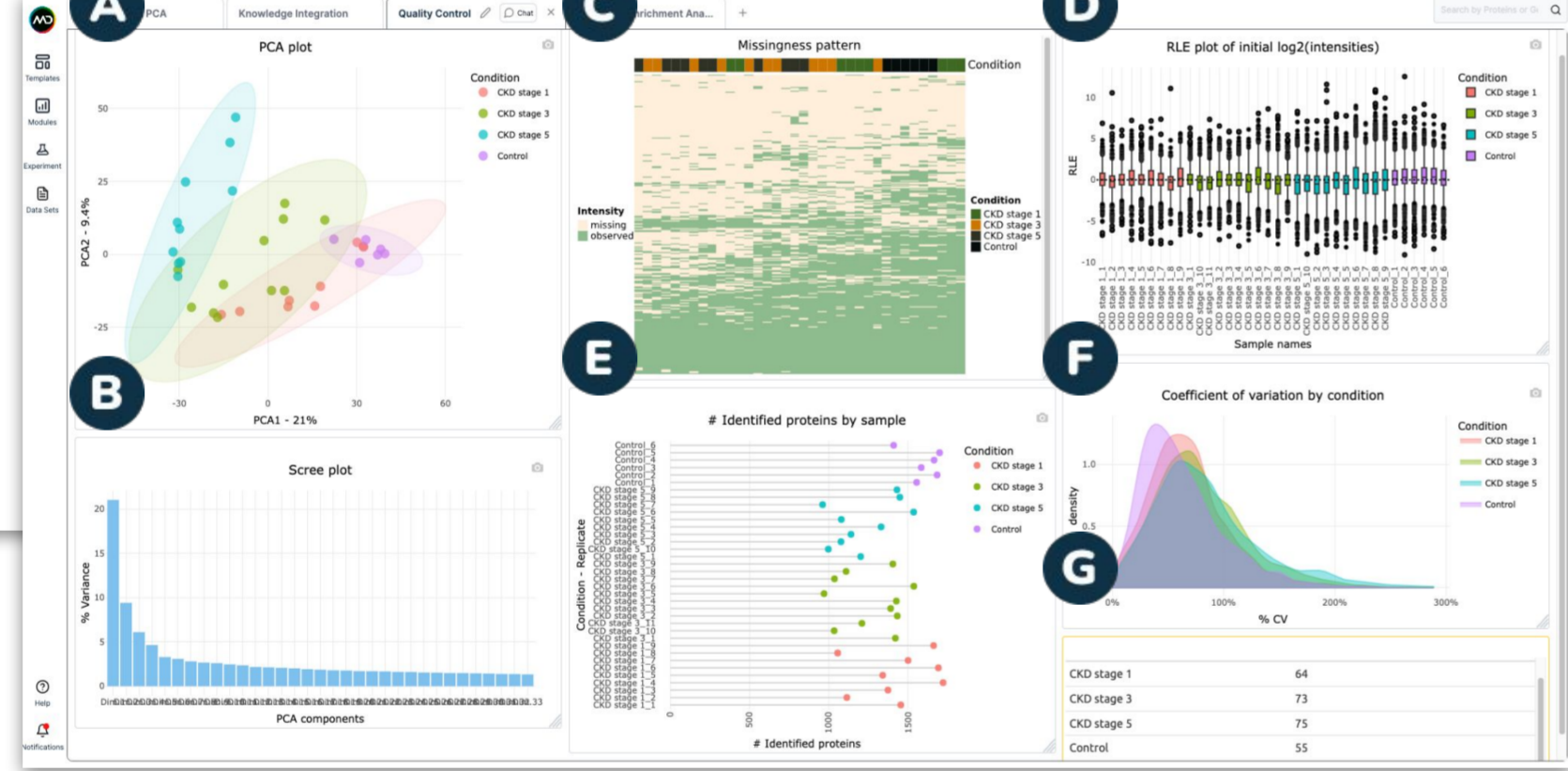


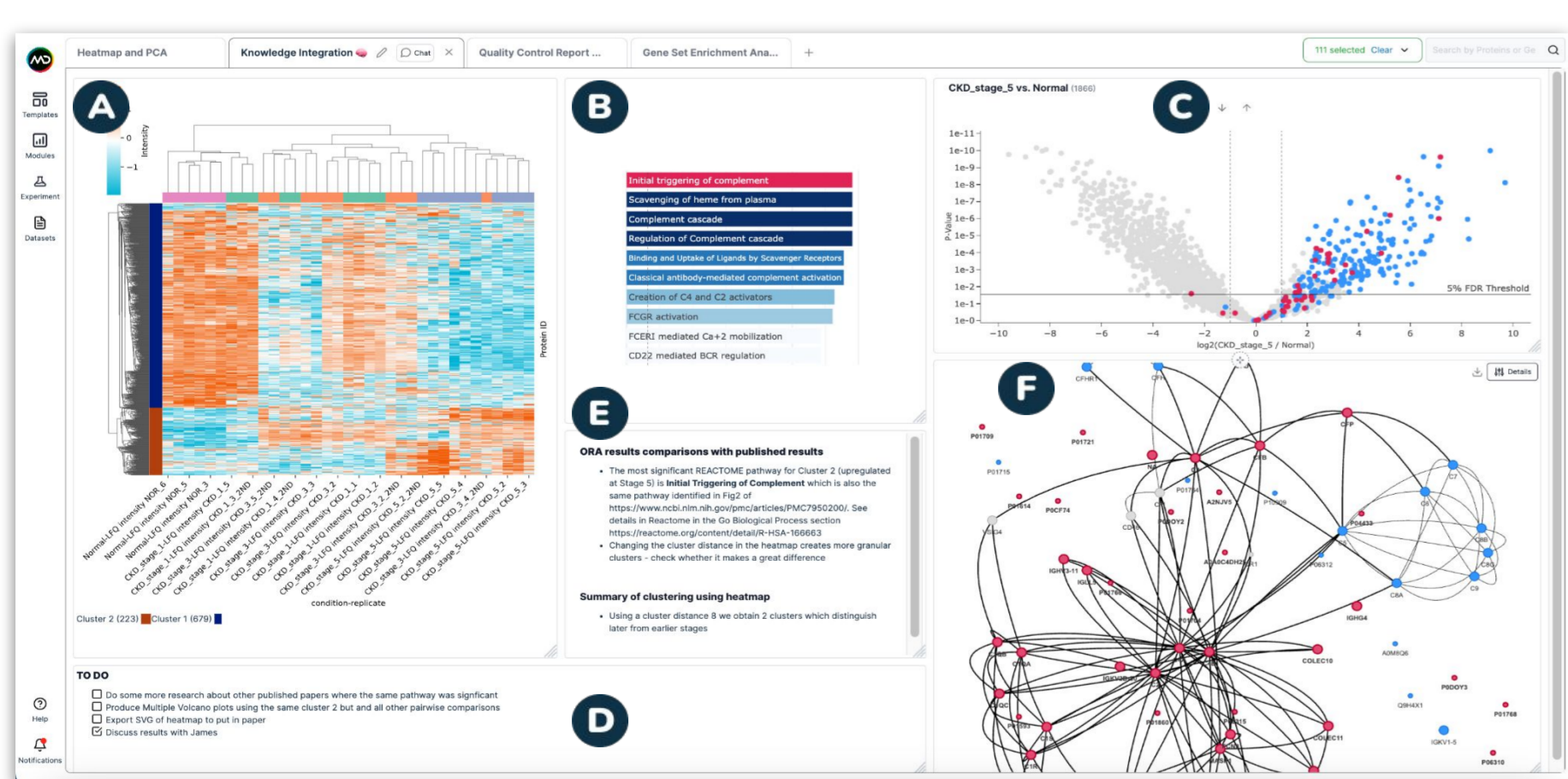
Figure 2: Example output from MD2.0 after uploading the PXD016433 (CKD) dataset via Generic Format upload



Interactively explore quantitative proteomics data

- Directly run **state-of-the-art statistical methods** for differential expression and knowledge interpretation such as limma³ and CAMERA⁴;
- 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 Plotly⁵, Seaborn⁶, Matplotlib⁷ and UpSetPlot⁸;
- 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 API⁹ database and **gene set enrichment analysis with CAMERA⁴** 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 **UniProt¹⁰, Gene Ontology (GO)¹¹, Reactome, MsigDB¹²**;
- Generated protein lists can be interrogated against **STRING** protein-protein Interaction network database¹³.

References

1. Quagliari 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, G.K. 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 Montreal (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., Strobl, H., Vuilleumot, 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-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47(D1):D607-D613 (2019).

Interpret the effect of various processing steps by combining multiple datasets

- Re-process your experiment, leveraging different processing & analysis options (e.g. normalization, imputation approaches, dose-response curve);
- 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);

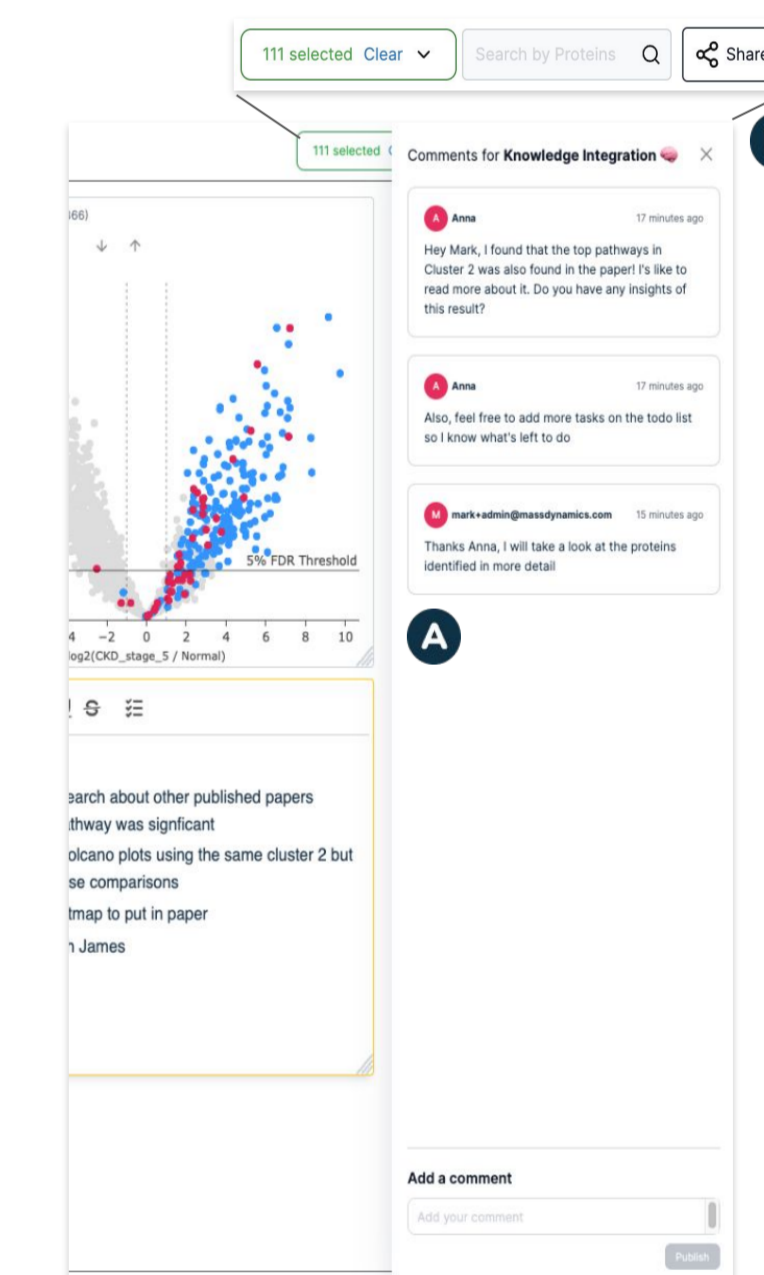
Figure 5: Comparing different experiments using the multi-experiment option - PXD016433 and PXD016447.



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 described¹. (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).