Mass Dynamics 2.0: The New Science of Collaboration

Anna Quaglieri[1], Joseph Bloom[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

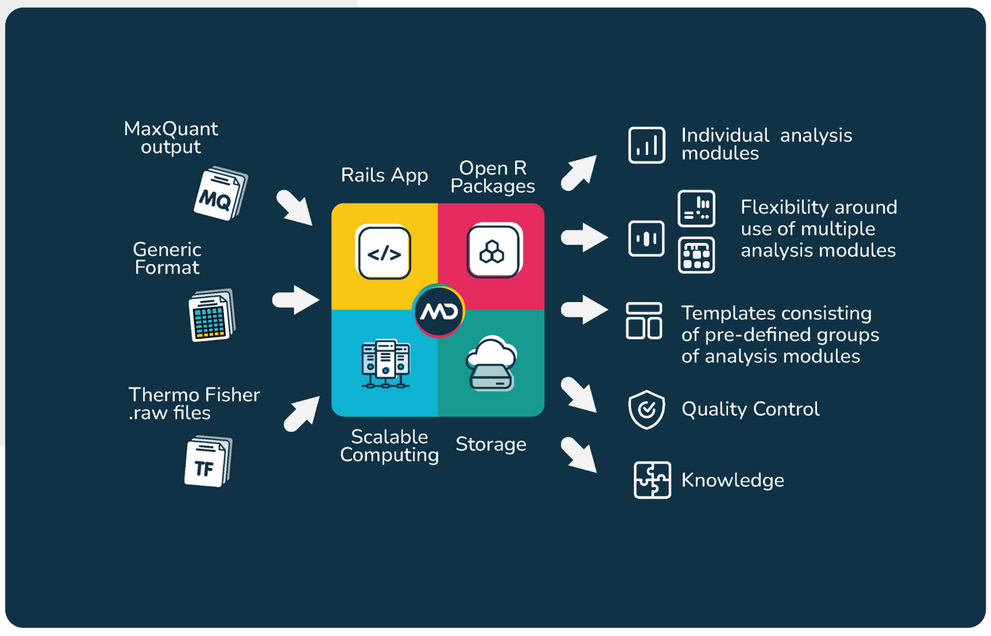
Introduction

The rapid expansion of both volume and complexity of proteomics data has created challenges in surfacing understanding and insights with currently available software tools.

Mass Dynamics 2.0 (MD 2.0) creates a seamless bridge between MS experts and life science researchers, augmenting researcher capabilities to provide access to state-of-the-art analysis techniques without the need to code or have access to deep statistical knowledge.

MD 2.0 reframes the data analysis and insight generation process into a new modular framework

- Greater functionality through 'drag and drop' analysis and visualisation modules
- Templates to create personalised workspaces
- Further standardization of data processing, and
- Gained ease of integration of new data and analysis modules as they become available in the community.



Case study: 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 identifier PXD016433[1] **containing LFQ proteomics data of human urine samples (chronic kidney disease CKD stages 1, 3, and 5 vs healthy controls). The dataset was** uploaded into the new interface using the Generic Format import and processed using the LFQ workflow previously described for MD1.0[2]

1

Accurately assess the quality of your experiment

- **Upload** data from raw LC-MS output or from pre-processed analysis (e.g. MaxQuant, DIA-NN, MSFragger);
- **Combine modules** or use **templated analyses** to determine data quality prior to further result exploration;
- Instantly leverage metadata with **advanced statistical tools** such as RLE plots, Missingness and CV distributions to assess the quality of your experiment health;
- Streamline experiment quality assessment prior to sharing to all collaborators.

Figure 1: Example output from MD2.0 after uploading the PXD016433 dataset via Generic Format upload.

2

Instantly explore quantitative proteomics experiments

- Directly access peer-reviewed, **state-of-the-art statistical methods** for differential expression and knowledge interpretation such as limma[3] and CAMERA[4];
- Work seamlessly with **alternate data visualisations** like heatmaps, upset plots, violin plots etc. orchestrated with **human centered design** principles in a flexible and **customisable workspace** to improve insight generation;
- Plots are generated using Plotly[5], Seaborn[6], Matplotlib[7] and UpSetPlot[8].

Figure 2: ORA analysis of PXD016433.

3

Rapidly derive insights and knowledge

- **Overrepresentation 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 directly connect your analysis results with external knowledge databases and quickly **gain biological insights** from your data;
- The gene set libraries are assembled from publicly available knowledge bases including UniProt[12], Gene Ontology (GO)[13] and Reactome[9].

4

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** for improved collaboration;
- It enables all stakeholders to independently assess data results;
- It allows **export** entire reports or specific modules to *.SVG as required;
- **Analysis** of results in app **can be made public** to allow interactive assessment of results by reviewers and community.

Figure 3. Example taking notes, setting tasks, collaborating live with chat box and sharing options.

5

Future Directions

- **Broaden upload options** for various pre-processed outputs;
- **Broaden statistical analyses** options, e.g. time series and dose response analyses;
- Allow **cross experiments** comparisons;
- More seamless integration with other **knowledge bases** (STRING-DB, PRIDE, etc.);
- More flexibility with **customised templates** and ability to integrate new templates and analysis with community-based input.

Figures

Figure 1: 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 2: ORA analysis of 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 and (C) strip plot showing the results of the Reactome ORA analysis. The analysis reveals significant representation of pathways such as complement activation, as previously described[1]. (D) MD 2.0 allows users to link selected pathways and their proteins in a pairwise comparison results with controls.

Figure 3: Example taking notes, setting tasks, collaborating live with chat box and sharing experiment on MD 2.0. (A) Text box to take notes; (B) Checklist box to create to-do lists; (C) Chat to for live collaboration with collaborators that have access to the experiment; (D) "Share" button to share experiment with collaborators (covered by the live chat in main panel).

References

1. Kim, J. E. et al. Multisample Mass Spectrometry-Based Approach for Discovering Injury Markers in Chronic Kidney Disease. Mol. Cell. Proteomics 20, 100037 (2021). **2**. Bloom, J. et al. Mass Dynamics 1.0: A Streamlined, Web-Based Environment for Analyzing, Sharing, and Integrating Label-Free Data. J. Proteome Res. 20, 5180–5188 (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**. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015). **11**. Geistlinger, L., Csaba, G. & Zimmer, R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinformatics 17, 1–11 (2016). **12**. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489 (2021). **13**. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 47, D330–D338 (2019).