BRIDGING COMPUTATIONAL COMPLEXITY AND BIOLOGICAL DISCOVERY: INTERACTIVE DIFFERENTIAL ABUNDANCE ANALYSIS IN MD 3.0

Mansi Aggarwal[1], Sara Ceccacci[2,3], Kevin Roger2, Ida Chiara Guerrera2, Anna Quaglieri1

[1]MassDynamics, Melbourne,Victoria3000,Australia, [2]Necker Proteomics,Université Paris Cité - Structure Fédérative de Recherche Necker, INSERM US24/CNRS UAR3633, Paris 75015, France, 3Université Paris Cité, INSERM, UMR 1163, Institut Imagine, Laboratory of Genetic Skin Diseases, F-75015 Paris, France

Mass Dynamics

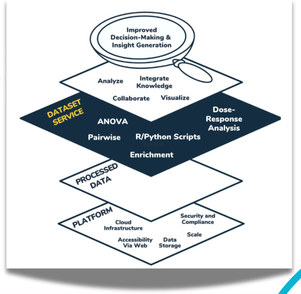
Introduction

Differential abundance (DA) analysis is a fundamental tool in clinical research, especially useful for understanding disease mechanisms and identifying biomarkers. However, the complexity of these analyses often requires specialized statistical and computational expertise, making it difficult for teams without these technical skills to fully leverage the power of DA. MD 3.0 addresses this challenge by providing a **flexible, intuitive platform** that enables users to **perform complex analyses** in an interactive web-based, cloud-powered interface, without the need for coding expertise. With MD 3.0, researchers can now focus on **collaborative exploration** and **interpretation of biological data,** rather than navigating software and programming languages. This study demonstrates how MD 3.0 empowers researchers to interactively define experimental designs, integrate control variables, and tweak model parameters, making the **full potential of DA analysis** accessible to users, regardless of their computational expertise. We applied this workflow to a **clinically relevant** dataset from the **Necker Proteomics Platform**.

Flexible DA workflow with MD 3.0 Dataset Service

The MD 3.0 Dataset Service streamlines the orchestration and scaling of scientific workflows, providing users with an intuitive interface for easy access. The Pairwise Dataset is now a part of the Dataset Service, enabling users to:

- Scale DA proteomic workflows using the limma1 framework;
- Easily and interactively define contrasts;
- Adjust for covariates;
- Customise various limma parameters;



In the following sections, we demonstrate how the MD 3.0 Dataset Service enhances **accessibility and reproducibility**, showcasing how users can seamlessly **execute end-to-end DA workflows** and leverage interactive visualization modules within the app.

1

Data exploration using interactive modules

For this study, we used a **clinically relevant** dataset from the **Necker Proteomics Platform**, focusing on a skin disease. The dataset includes samples from 24 patients, each providing a lesional and a non-lesional skin sample. The study demonstrates how MD 3.0 enables the easy fitting and comparison of multiple models, **with and without adjustment for confounding variables**, to identify key proteins and pathways distinguishing lesional from non-lesional skin samples.

After uploading the dataset into the app, we used the interactive dimensionality reduction module to explore the variations in the data associated with patient-specific metadata, such as patient ID, age and sex. The platform's interface allowed us to easily visualize these variations and gain insights into the patterns present in the data.

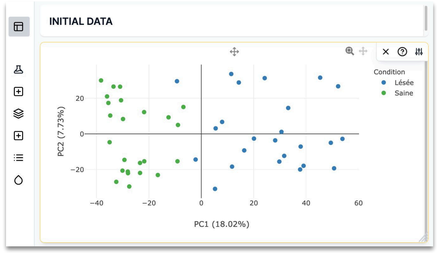


Figure 1. Principal Component Analysis (PCA) using the MD 3.0 Dimensionality Reduction module.
PCA with initial dataset, prior to adjusting for patient ID using the limma1 removeBatchEffect() function. The separation between "Lésée" (diseased) and "Saine" (healthy) groups is less distinct than when we remove patient-related variability in the next PCA (Fig 2A).

2

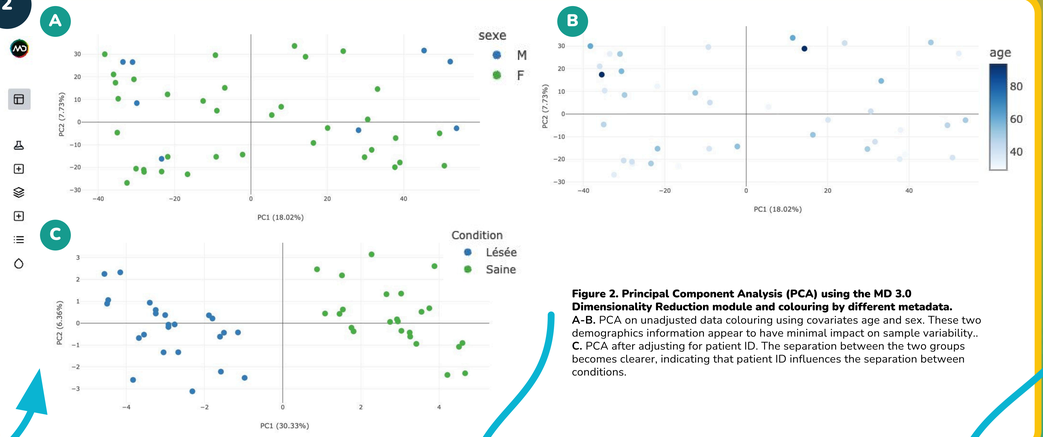


Figure 2. Principal Component Analysis (PCA) using the MD 3.0 Dimensionality Reduction module and colouring by different metadata.
A-B. PCA on unadjusted data colouring using covariates age and sex. These two demographics information appear to have minimal impact on sample variability..
C. PCA after adjusting for patient ID. The separation between the two groups becomes clearer, indicating that patient ID influences the separation between conditions.

3

Custom design setups for differential abundance analyses

The Pairwise Comparison dataset simplifies the process of performing DA analysis by allowing easy incorporation of control variables into the experimental design matrix. This enables the comparison of different models interactively, providing insights into proteins with varying expression patterns between the lesional and nonlesional skin regions.

In this study, **we compared three distinct models**: 1) without any adjustments, 2) using a paired design to account for patient variability, 3) incorporating patient ID with the additional covariates,
age and sex. All other parameters are set to default for all three models. The results of the three datasets can then be accessed by interactive modules in the app.

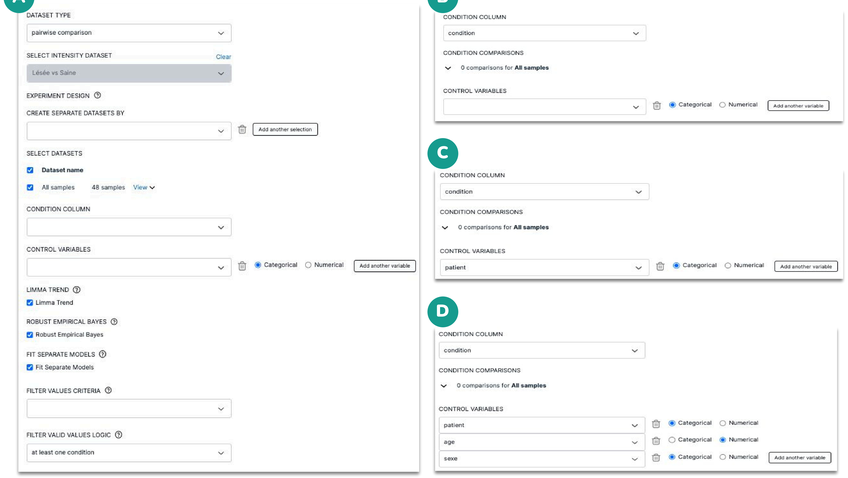


Figure 3. Step-by-step setup of the three differential abundance datasets via the interactive user interface in MD 3.0. A. The Pairwise Comparison form allows users to setup and trigger the creation of differential abundance analysis. Users can choose the *condition* columns, specify comparisons of choice, incorporate several *control variables*, adjust limma parameters, and apply different types of missing value filters. **B.** Setup for the first model — no adjustments. **C.** Setup for the second model — paired design with patient ID. **D.** Setup for the third model — patient ID along with additional covariates i.e., age and sex.

4

Results

The results of the three datasets were easily compared within MD 3.0's interactive visualization ecosystem. Pairwise volcano plots (Fig 4A-C) detected broad changes in protein significance, while upset plots (Fig 4D) revealed the overlap and unique differentially expressed proteins across models. Dimensionality reduction (Fig 1 vs Fig 2C) confirmed that **specifying the paired design** with Patient ID increased separation between healthy and diseased samples, leading to more differentially regulated proteins. **Adjusting for age and sex further improved analysis power**, identifying even more significant proteins (Fig 4C). Reactome over-representation analysis of common upregulated proteins across the three models revealed enriched biological pathways. This entire analysis was efficiently performed interactively in MD 3.0 in **under an hour.**

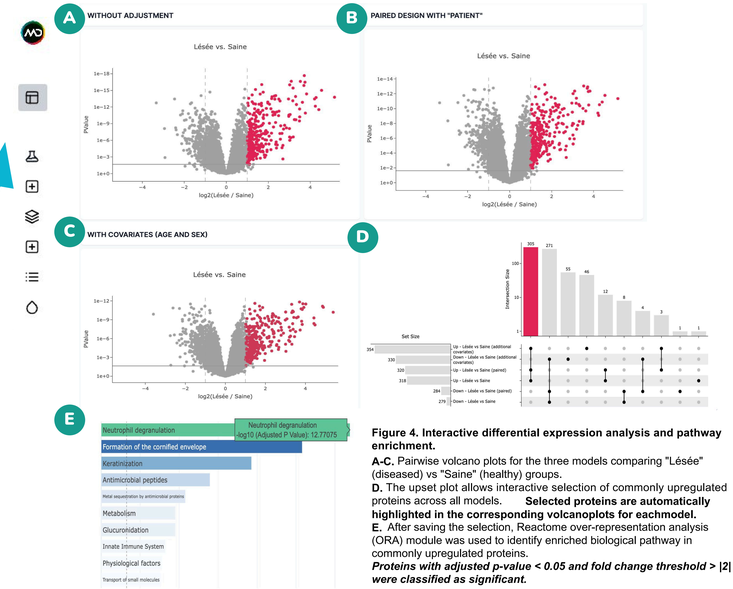


Figure 4. Interactive differential expression analysis and pathway enrichment.
A-C. Pairwise volcano plots for the three models comparing "Lésée" (diseased) vs "Saine" (healthy) groups.
D. The upset plot allows interactive selection of commonly upregulated proteins across all models. **Selected proteins are automatically highlighted in the corresponding volcanoplots for eachmodel.**
E. After saving the selection, Reactome over-representation analysis (ORA) module was used to identify enriched biological pathway in commonly upregulated proteins.
Proteins with adjusted p-value < 0.05 and fold change threshold > |2| were classified as significant.

Future Directions

MD 3.0simplifies differential abundance analysis, making it accessible to all researchers through an intuitive interface and interactive visualizations. It enables easy model comparison and supports biological discovery by automating dataset generation and ensuring reproducibility.

Next, we plan to incorporate advanced differential expression techniques like **limpa2** and **DEqMS3**, and other cutting-edge methods for improved model comparison. With the upcoming support for peptide data in our app, these techniques will leverage peptide-level uncertainty improving protein quantification and biological inferences without inflating false discoveries.

References

1. **Ritchie, M. E. et al. (2015).** *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 43(7), e47. https://doi.org/10.1093/nar/gkv007

2. **SmythLab. (2025).** *SmythLab/limpa*. GitHub. https: //github.com/SmythLab/limpa

3. **Zhu, Y. et al. (2020).** *DEQMS: a method for accurate variance estimation in differential protein expression analysis.* Molecular & Cellular Proteomics, 19(6), 1047–1057. https: //doi.org/10.1074/mcp.tir119.001646

Conflicts of interest

The authors Mansi Aggarwal and Anna Quaglieri are employees of Mass Dynamics, a for-profit enterprise delivering software as a service for the processing, analysis, and sharing of proteomics data. The authors Sara Ceccacci, Kevin Roger, Ida Chiara Guerrera are affiliated with INSERM, which is a client of Mass Dynamics.