Journal of proteome research

pubs.acs.org/jpr

Article

Mass Dynamics 1.0: A Streamlined, Web-Based Environment for Analyzing, Sharing, and Integrating Label-Free Data

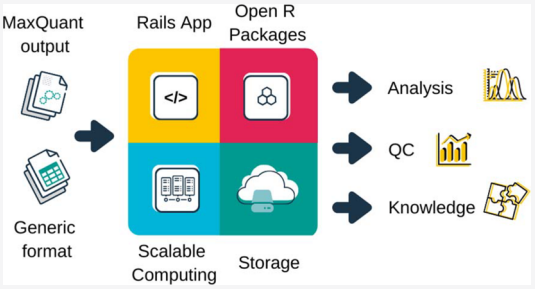
Joseph Bloom, Aaron Triantafyllidis, Anna Quaglieri, Paula Burton Ngov, Giuseppe Infusini,* and Andrew Webb*

Cite This: *J. Proteome Res.* 2021, 20, 5180−5188

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Label-free quantification (LFQ) of shotgun proteomics data is a popular and robust method for the characterization of relative protein abundance between samples. Many analytical pipelines exist for the automation of this analysis, and some tools exist for the subsequent representation and inspection of the results of these pipelines. Mass Dynamics 1.0 (MD 1.0) is a web-based analysis environment that can analyze and visualize LFQ data produced by software such as MaxQuant. Unlike other tools, MD 1.0 utilizes a cloud-based architecture to enable researchers to store their data, enabling researchers to not only automatically process and visualize their LFQ data but also annotate and share their findings with collaborators and, if chosen, to easily publish results to the community. With a view toward increased reproducibility and standardization in proteomics data analysis and streamlining collaboration between researchers, MD 1.0 requires minimal parameter choices and automatically generates quality control reports to verify experiment integrity. Here, we demonstrate that MD 1.0 provides reliable results for protein expression quantification, emulating Perseus on benchmark datasets over a wide dynamic range. The MD 1.0 platform is available globally via: https://app.massdynamics.com/.



KEYWORDS: *MaxQuant, automated data analysis, label-free quantification, benchmarking, web-based software tool*

INTRODUCTION

Proteomics can be defined as the application of technologies for identification and quantification of the protein content of complex biological samples. Over the past few decades, proteomics research and the complexity of experimental research questions addressed by it have developed rapidly and are having a growing impact in biological and medical research. In particular, areas such as understanding mechanisms of action in disease progression and therapeutic intervention as well as detection of diagnostic markers, identifying candidates for vaccine production, and understanding pathogenic mechanisms and gene expression patterns have been of growing importance in advancing many areas of medically related research.[1,2]

Mass spectrometry via liquid chromatography (LC)/MS−MS is the leading technology in proteomics research, with label-free quantification (LFQ) and isotope-based labeling methods being two approaches that facilitate interrogation and measurement of many analytes in even very complex samples. Isotope-based labeling methods such as SILAC and TMT have provided the gold standard for protein quantification but are limited by their applicability to types of samples, which cannot be easily labeled, and by the associated cost of reagents required. By contrast, LFQ is a simpler, more economical, and scalable method that requires a considered experimental design to achieve robust biological insights.[3]

An exemplar of the growing use and adoption of LFQ-based approaches is the sheer variety of analytical software packages that have been developed to support LFQ experiments, which have been comprehensively covered recently by Al Shweiki *et al.*,[4] the most widespread of which is MaxQuant.[5] MaxQuant's success can be attributed at least in part to the holistic nature of its analytical tools, covering a breadth of steps including feature extraction, database search, protein identification, and quantification, which must otherwise be achieved using a combination of other tools.

While data processing is essential for the success of proteomic analysis, objective quality control reporting and reproducible downstream analysis are equally important. The Perseus computational platform is the analytical counterpart to MaxQuant and provides users with a highly flexible framework for post-processing and visualizing results.[6,7] Despite the diverse suite of tools offered by Perseus (accompanied by supporting documentation and online tutorials), the sheer amount and breadth of functionality can be overwhelming for

Received: August 23, 2021
Published: October 14, 2021

users requiring straightforward analysis of LFQ proteomics data. For some users, who commonly repeat identical analytical procedures frequently for collaborators, Perseus can be more labor- and time-intensive than other tools designed to automate the same process such as LFQAnalyst.[8]

Since the publication of LFQAnalyst in 2019, there have been numerous publications of automated LFQ pipelines including Eatomics[9] and ProVision.[10] While each of these resources differs in their scope and degree of automation, they are among a growing list of attempts to shift toward standardization through automation of straightforward statistical analyses and visualization provided as an output of the LFQ pipeline. While these attempts have undoubtedly made headway toward more accessible and reliable analytical processes, there remain large challenges such as scale, storage, sharing, and platform sustainability that will inhibit broad adoption among the community.

MD 1.0 addresses these needs by meeting multidimensional criteria in reliability, ease of use, and transparency while offering functionality designed to facilitate collaboration and sharing. A qualitative survey of the tools mentioned above, MD 1.0 is provided in Table S1.

There are several measures that can be taken to ensure that tools facilitating LFQ proteomics maintain reliability of analysis and interpretability of results. First, data analysis and statistical approaches should be demonstrated to produce accurate and reliable results in accordance with the accepted best practices of statistical testing. Second, by demonstrating that the specific implementation of these methods has been successful and continues to be over the life of the tool, this ensures that the specific code works and is not accidentally compromised during further development. Third, by incorporating automated quality control metrics and figures, these provide confidence that results in non-benchmark datasets that can be reasonably interpreted. By meeting these three criteria, any scientific tool can be safely shared with users of varying degrees of expertise.

However, reliability of results is only one side of the coin when it comes to developing useful, empowering tools for the proteomics field. Interviews with 100 scientists in the field of mass spectrometry[11] found that "free" tools often imposed significant hidden costs in the time spent learning to use the software and in manual tuning of parameters. Though opportunity costs of users are hard to quantify, they represent a genuine cost of use.

It is likely that an awareness of these hidden costs is in part a large driver for the number of tools created for automating analysis of LFQ data, although calls for robust, user-friendly automation in proteomics have been present since 1999.[12] Automated tools not only are less time-consuming but also reduce the number of points of failure that must be investigated in development, benchmarking, and comparison. For many practical reasons, automation simplifies tools, making them more accessible to users and scientific developers. Moreover, codifying analysis, open-sourcing analytical workflows, and benchmarking allow for standardization that addresses reproducibility difficulties that plague the broader scientific field.

Tools that automate highly complicated tasks inevitably end up comprising large amounts of code that must be available for the tool to be reproducible (as a necessary, but not sufficient condition). Tools such as OpenMS[13] achieve reproducibility by open-sourcing their code, enabling a thorough and ongoing peer review as tools grow over time.

MD 1.0 attempts to meet these standards of reliability, automation (thereby ease of use), and transparency in the specific domain of LFQ analysis. However, MD 1.0 provides further functionality designed to assist with resource integration, annotation, sharing, and collaboration.

Traditional and ubiquitous methods for sharing LFQ data likely include email or messenger service (e.g., Slack)-type transfers of text and use of file storage tools like Dropbox or Google Drive. These tools present obstacles to effective collaboration such as separating data from quality control metrics, which can obscure interpretability. By allowing an entire experiment to be shared with ease, MD 1.0 attempts to make supervision and collaboration easier for scientists to ask for and receive assistance while performing their experiments.

METHODS

Benchmarking Datasets

Different LFQ benchmarking datasets are chosen to verify that MD 1.0 recovers comparable results to Perseus. These include two datasets with PRIDE[14] identifiers PXD000279[3] and PXD010981,[15] which have ground truth and one "real world" scenario, with PRIDE identifier PXD002057.[16]

PXD000279 ("dynamic range dataset") contains raw data for two groups (four replicates each) enriched with one of two "Universal Protein Standards" (1 and 2), which test LFQ accuracy over a large dynamic range.

PXD002057 ("HER2 dataset") contains raw data for an experiment with two groups, each with three samples. These two groups come from two cancer cell lines, a parental SKBR3 cell line and another cell line derived from the first, which is resistant to human epidermal growth factor receptor 2 (HER2)-targeted therapy.

Last, PXD010981 ("iPRG2015 dataset") contains raw data for the iPRG2015 benchmarking dataset. This dataset is composed of four samples with 200 ng of tryptic digests of *Saccharomyces cerevisiae* (ATCC strain 204508/S288c). Each was spiked with different quantities of six individual digested proteins (ovalbumin, myoglobin, phosphorylase b, beta-galactosidase, bovine serum albumin, and carbonic anhydrase) according to a schema in that publication (present in the benchmarking github repository "lfq_benchmarking").

To ensure that certain spiked proteins (P02768 and P06396 for the dynamic range dataset and P44683 for the iPRG2015 dataset) were not removed as contaminants, the proteinGroups.txt output was manually edited prior to benchmarking analysis. The resulting input files are provided in the Supporting Information.

Raw Data Processing with MaxQuant

MaxQuant v1.6.17.0 was used with default parameters except for an LFQ min ratio of 1 and enabling match between runs. Output files used for quality control included msms.txt, peptides.txt, modificationSpecificPeptides.txt, proteinGroups.txt, and evidence.txt, while only proteinGroups.txt was used for quantification. While MaxQuant labeled spiked proteins as potential contaminants, proteinGroups.txt was manually edited to remove this label so these rows would be present in subsequent processing.

Results and parameters files can be downloaded from the Supporting Information and from the Mass Dynamics platform at the following addresses:

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188

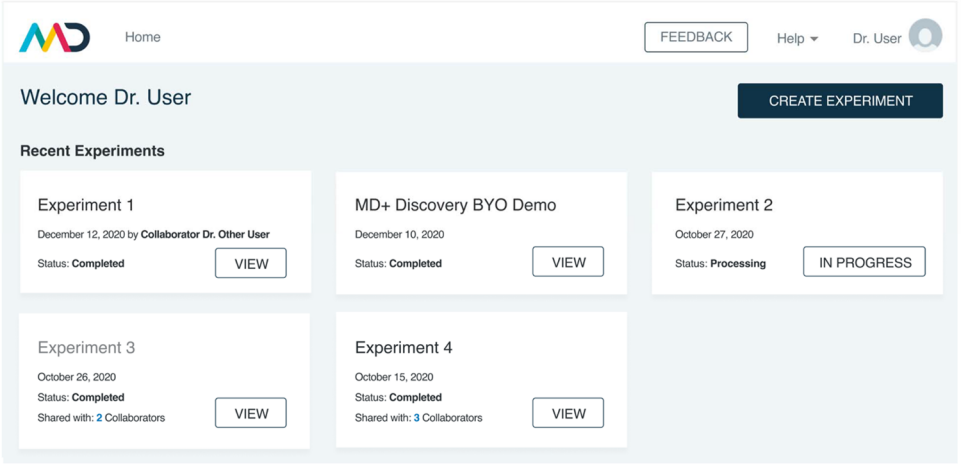


Figure 1. MD 1.0 landing page. The user interface begins at the experiments page after users sign up and log in. Users can see their experiments, with associated status ("in progress" or "view" for finished experiments), dates, and owners. Experiments can be the users' own (such as experiment 2) or shared via a collaborator (experiment 1). Users can see which of their experiments have been shared with other users and who those users are via the "shared with" link in bold. Experiments are either in progress or completed in which case the view button is accessible. The "create experiment" button allows users to upload data for a new LFQ analysis.

iPRG2015: https://app.massdynamics.com/p/56a6d7f7-c129-4d7a-a221-d2ef3b8ff4c3

Dynamic Benchmark Range Dataset: https://app.massdynamics.com/p/152bd07f-ddd6-4943-bf62-298648b43bd3

HER2: https://app.massdynamics.com/p/b832f17a-a9f3-4890-8533-2d69835f814e

Please note that public experiment links provide a limited feature set due to "read only access". Downloading input files from the Supporting Information and uploading as new experiments will allow access to all features.

Additionally, all MaxQuant and MD 1.0 outputs are contained within PRIDE repository PXD028038.

MD 1.0 LFQ Processing

Statistical analysis is performed using R version 4.1.0.

An experiment design file is generated from user input during the experimental setup in the application prior to processing, which is thereafter automatic. Samples can be grouped into two or more experimental groups, and all pairwise statistical comparisons will be generated between those groups by the following workflow. The samples or experimental groups uploaded over several submissions will not be automatically compared.

The following steps are then taken to perform the analysis:

1. Proteins corresponding to reverse sequences, potential contaminants, and proteins only identified by site are filtered out.
2. Intensities provided inside proteinGroups.txt are converted to the log2 scale.
3. Missing values are imputed using the MNAR ("missing not at random") method with a mean shift of −1.8 and a standard deviation of 0.3 as recommended in the Perseus protocol.
4. Protein groups where more than 50% of intensities are imputed for both conditions are excluded from the quantitative analysis.
5. Differential expression (DE) analysis is performed using linear models with the Bioconductor package limma,[17] in particular using the limma-trend method.[18,19] *P*-values are calculated using the robust empirical Bayes procedure to compute moderated *t*-statistics.
6. The Benjamini−Hochberg correction is used to account for multiple testing.

MD 1.0 does not currently take into account covariates or paired experiment design.

All the codes used to reproduce the abovementioned workflow using the MaxQuant output are provided in the LFQProcessing R package, available on GitHub.[20] To extend the use of the workflow to output from software other than MaxQuant, the analysis steps were also implemented for a generic format of summarized protein intensities and are available in the MassExpression R package on GitHub.[21]

Volcano plots are used to summarize DE results in the Rails application. Guides are provided in these plots to indicate a false discovery rate (FDR) cutoff of 5% and an absolute fold change of at least 2.

Implementation

MD 1.0 is composed of two separate components, a web component using a modern software stack (JavaScript and rails) and a processing component built using R and Elastic Compute Cloud (ec2) on Amazon Web Services (AWS). The combination of these two components ensures that processing is repeatable and runs in a computing environment using identical software, parameters, and code while maintaining the privacy and security of user data. Users have the option to share their data with specific users or share publicly.

The R code bases and packages responsible for MaxQuant and generic format processing are public to enable complete reproducibility of the Mass Dynamics 1.0 analysis. The AWS infrastructure, the JavaScript, and Rails components are not open sources but are only responsible for storing, moving files, making API calls (to Reactome), and rendering the user interface.

Perseus Processing. All proteinGroups.txt were analyzed with Perseus version 1.6.14.0. Proteins corresponding to reverse sequences, contaminants, and proteins only identified by site were removed. Intensities were transformed to the log2 scale, and missing values were imputed as in the MD 1.0 protocol. Student *t*-tests were performed for each pairwise

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188

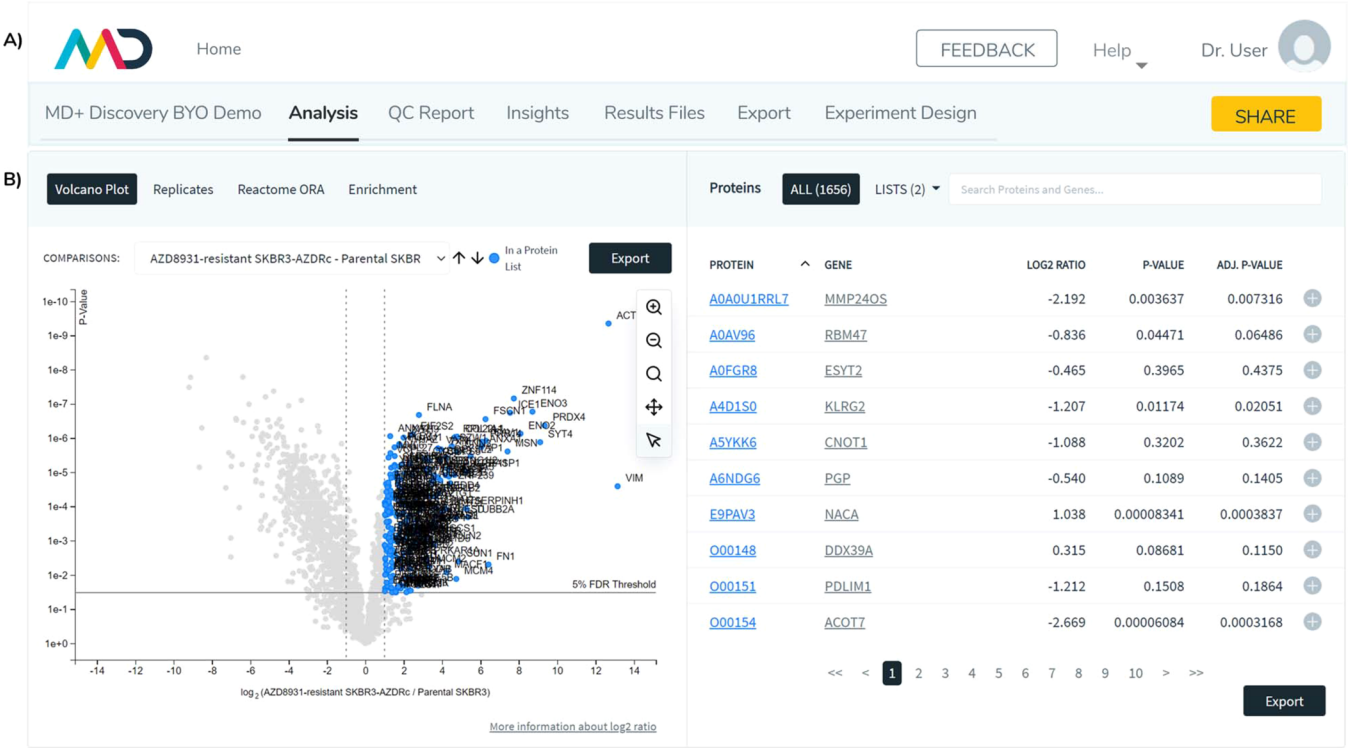


Figure 2. MD+ Discovery experiment view header (A) and experiment view volcano plot (B). In an experiment view (A), users can choose between the following tabs: analysis, QC report, insights, results files, export, and experiment design. Users also can share their experiment. Inside the analysis tab (B), users can choose between tabs for viewing the volcano plots, filtering by protein observations by replicates and the Reactome ORA tab. The volcano plot and table allow users to dynamically search for, select, and manipulate protein lists by adding, removing, and annotating proteins as they complete their analysis.

comparison with an S value of 0 and a Benjamini−Hochberg FDR correction were used for multiple testing adjustment. Tests were only performed when more than 50% of values are not imputed in at least one group. Session files are available in the Supporting Information.

Figure Generation and Results Comparison

All results figures and tables were calculated using bespoke python scripts utilizing packages including pandas for data manipulation, plotly and matplotlib for graphics, and scipy for Pearson correlation calculations. All the codes used to perform the comparisons are available on GitHub at https://github.com/MassDynamics/lfq_benchmark or Zenodo (https://doi.org/10.5281/zenodo.5516668).

RESULTS AND DISCUSSION

Overview of MD+ Discovery, User Interface, Experiment Creation, and Sharing

Mass Dynamics 1.0 is a web-based, integrated, and automated analysis and collaboration environment that facilitates LFQ experiments. MD 1.0 enables users to upload MaxQuant output files and visualize the quantitative analysis. The user interface begins at the experiments page after users sign up and log in (Figure 1).

To create a new experiment, users can click the create experiment button on the landing page, which takes them to the experiment creation page. Here, they are given the choice of uploading a generic format for protein intensity data, MaxQuant.txt folder outputs, or use of the "demo" HER2 dataset. TMT data can be processed using the same statistical pipeline with the addition of protein level median normalization for each channel in each sample.[22]

After uploading their LFQ data, experiment files are presented to the user, who is prompted to allocate them into replicate groups. They can then add an experiment description and complete the experiment creation step. Computation time depends on the size of the experiment. Users are sent an email notification when their experiment has been completed.

In an experiment view (Figure 2), users can choose between the following tabs: analysis, quality control (QC) report, insights, results files, export, and experiment design. Analysis contains protein expression volcano plots and tables. The QC report contains several experimental quality control plots. The insights tab contains a list of comments and notes introduced by the user. The results files tab contains a list of output files that can each be downloaded, which includes the tab separated files produced by the quantitative analysis scripts. The export tab enables users to export the volcano plot graph with all "candidate proteins" (proteins that have been selected and added to at least one selection list) highlighted and annotated. Last, the experiment design tab indicates which files have been assigned to each experimental condition.

As the experiment data is stored securely in the cloud, sharing is as simple as giving collaborators access to the same website. After an experiment is completed, users can select the share button and enter the email address of someone they would like to share their experiment with. This email address is then sent an invitation to join MD 1.0, which contains a link to the shared experiment. Insights associated with proteins, which are accessible at all protein level analysis interfaces, perpetuate

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188

Figure 3. The Reactome ORA tab enables users to perform over-representation analysis (ORA) using the Reactome API. A volcano plot is shown with all proteins in that pathway (in gray) and that experiment and the "candidate proteins" blue and labeled, with a corresponding estimated false discovery rate (FDR) provided.

between users accessing the same experiment and are thus shared with the experiment data.

Experiments uploaded to the MD 1.0 platform will be stored for at least 5 years from upload and after that pending further correspondence. Users can delete experiments on request.

MD 1.0 Facilitates Analysis and Annotation of Label-Free Quantitation Results

After an experiment has finished processing, the analysis tab can be used to visualize results in the form of a volcano plot (Figure 2B). Next to the volcano plot is a table containing a list of each protein with the associated gene name, estimated fold change, *p*-value, and adjusted *p*-value. From this table, proteins can be added to the list that can be used to filter results to proteins of interest or into groups of up- or downregulated proteins. A protein in a list can be given an annotation, where the user is able to comment on a protein with text and/or hyperlinks, which are then presented in the insights tab.

MD 1.0 Allows Users to Perform Over-Representation Analysis with Reactome

If protein accession codes are contained within input files, then MD 1.0 will automatically link table views to Uniprot.org,[23] and the Reactome ORA (Figure 3) provides further integration with an external knowledge base. The Reactome ORA tab uses the Reactome[24] API content service to provide over-representation analysis (ORA) results to users. For each candidate list, one API call is made to perform ORA, while a second one is used to retrieve the complete list of proteins in each resulting pathway.

As no background can be specified using the Reactome API, the hypergeometric test is performed using the ratio calculated as the number of entities in the pathway and, in the candidate list, divided by the total number of associated entities known to Reactome. The provided FDR is calculated using the Benjamini–Hochberg method.

MD 1.0 Quality Control Report Produces Diagnostic Figures to Assess Experiment Health

A feature of MD 1.0 is the automated generation of a quality control report accessible in the quality control (QC) report (see the Supporting Information) tab in the experiment view.

The report contains three sections, experiment health, feature completeness, and identifications. Experiment health contains principal component analysis scatter plots of the first two principal components and a scree plot for all proteins, differentially expressed proteins, modification-specific peptides, and all peptides. Quantitative CV (coefficient of variation) distributions and sample intensity correlation plots are then produced for proteins, peptides, and modification-specific peptide tables. The feature completeness section provides the percentage of missing measurements in a histogram at the protein, peptide, and modification-specific peptide levels and a histogram of the percentage of all measurements missing at the LC–MS run (file) level. Last, the identifications section reports counts per file for detected PSMs, modification-specific peptides, peptides, and proteins. Complete QC reports are provided both in application and in the Supporting Information.

Due to the automatic nature of the QC report and data processing, users can quickly review and assess experimental results and confidently interpret results or share with collaborators via other features.

MD 1.0 Reproduces Perseus Results Reliably on Sample Datasets

To determine the reliability of the MD 1.0 automated workflow for LFQ quantification, we analyzed the same experiments on Perseus and MD 1.0. Two of these experiments, the iPRG2015, and dynamic range datasets contained ground truth data where we have expected DE proteins and fold changes. The last dataset from a study on

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180–5188

breast cancer resistant cells constitutes a more realistic real-world scenario with no ground truth.

We compared Perseus and MD 1.0 results on these datasets using continuous and discrete measures of accuracy (for the ground truth datasets) and by similarity (for all datasets).

Tables 1 and 2 show the confusion matrices for Perseus and MD 1.0 DE performance across the dynamic range and iPRG2015 datasets, respectively. We defined a "true positive" if the estimated and true protein abundance ratios were both larger than 2 in the same direction, and an adjusted *p*-value of less than 0.05 was produced. "False negatives", where the true protein abundance ratio was greater 2 but the estimated value was not or the adjusted *p*-value was greater than 0.05, were rare but occurred once in the dynamic range dataset (in common between the two experiments) and three times for Perseus in the iPRG2015 dataset. Tables 1 and 2 provide more details on these missed detections.

Table 1. Binary Evaluation of Differential Expression Predictions between Perseus and MD 1.0 on the Dynamic Range Benchmark Dataset (Cox et al.)[a]

	MD 1.0		Perseus	
	expected true	expected false	expected true	expected false
observed true	39	4	39	4
observed false	1	2198	1	2198
sum	40	2202	40	2202

[a]Confusion matrices for the dynamic range dataset were identical between MD 1.0 and Perseus. Both methods produced results within the 1% standard for an acceptable false discovery rate. One false negative was produced by both Perseus and MD 1.0 pertaining to gamma-synuclein (UniProt ID: O76070), resulting from higher adjusted *p*-values of 0.578975 and 0.351076, respectively.

Table 2. Binary Evaluation of Differential Expression Predictions between Perseus and MD 1.0 on iPRG2015 Benchmarking Dataset (Choi et al.)[a]

	MD 1.0		Perseus	
	expected true	expected false	expected true	expected false
observed true	30	14	27	3
observed false	0	19,528	3	19,539
sum	30	19,542	30	19,542

[a]Confusion matrices for the iPRG2015 results. Both methods produced results within the 1% standard for an acceptable false discovery rate. The Perseus protocol used failed to detect three protein abundance ratios greater than 2 in all cases due to lower confidence than required by the definitions used. The comparisons were phosphorylase b in samples 1 vs 2 (adjusted *p*-value of 0.058) and ovalbumin in samples 2 vs 3 and 3 vs 4 (adjusted *p*-values of 0.0594 and 0.0880, respectively).

It is noted that, for the iPRG2015 dataset that MD 1.0 produced, 14 false positives as opposed to three were produced by Perseus. This discrepancy and others described above may be due to the differences in the randomly imputed values via the MNAR imputation used in both workflows and possibly because the MD 1.0 workflow uses the empirical Bayes *t*-test from the limma package to perform a modified *t*-test as opposed to the standard *t*-test used by the Perseus protocol. By sharing information across tests, the empirical Bayes method can better estimate the underlying variance in observations and therefore gain greater confidence about differentially expressed proteins. This may explain why MD 1.0 detected all 30 expected positives in Table 2, while the Perseus protocol only detected 27.

The Pearson correlation was used to measure the similarity between estimated log fold changes and the true log fold changes (according to the benchmark dataset descriptions). Scatter plots of these values are provided in Figure 4A–D. In all cases, the Pearson correlation was greater than 0.9, suggesting very high accuracy.

The Pearson correlations between MD 1.0 and Perseus estimated log fold changes were 0.998, 1.0, and 0.988 for the dynamic range dataset, iPRG2015, and HER2 studies, respectively (Figures 4–6), whereas the associated −log10 adjusted *p*-values varied slightly more, with Pearson correlations of 0.925, 0.980, and 0.901 for the dynamic range dataset, iPRG2015, and HER2 studies, respectively, showing that MD 1.0 produces results consistent with what can be achieved by those using the Perseus platform, except with a completely automated workflow.

CONCLUSIONS AND OUTLOOK

Mass Dynamics 1.0 is a web-based, integrated, and automated analysis and collaboration environment that facilitates label-free quantitative experiments. It currently accepts the MaxQuant output and is designed to be adaptable to accept data from other pre-processing tools. The output of MD 1.0 is served to the user via web pages, which contain interactive figures, tables, and downloads. The analysis performed by MD 1.0 is inherently reproducible both in the platform and elsewhere and can be easily shared or published to the community.

With the intention of broadening access to a wider range of users from computational and non-computational backgrounds, this environment provides many features that facilitate quality control, reproducibility, automation, and transparency.

We demonstrated that MD 1.0 reproduces results of the Perseus protocol reliably and accurately with respect to known benchmarking datasets while producing comprehensive quality control reports that allow the user to be confident about the quality of the experiment. MD 1.0 therefore constitutes a reliable, straightforward, and streamlined alternative to the Perseus platform when performing LFQ.

Mass Dynamics is well placed to begin expanding MD 1.0 in terms of analysis available beyond LFQ. Creating further interfaces to enable analysis of peptides, modified peptide statistics may enable users to perform peptide level or PTM analysis such as for phosphoproteomic analysis. Further development of the processing may facilitate more complex experimental designs that, for example, could handle paired samples.

Enrichment analysis such as over-representation analysis (ORA) and gene set enrichment analysis (GSEA) might be achieved via integrations of third party databases such as GO, DAVID, KEGG, or Drugbank. User interface improvements may involve more opportunities for annotation and sharing utilities or allow comparisons between multiple different experiments contained within the platform.

Last, leveraging the cloud-storage element of MD 1.0 implementation, insights may be gained by cross-referencing experiment data and insights.

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188

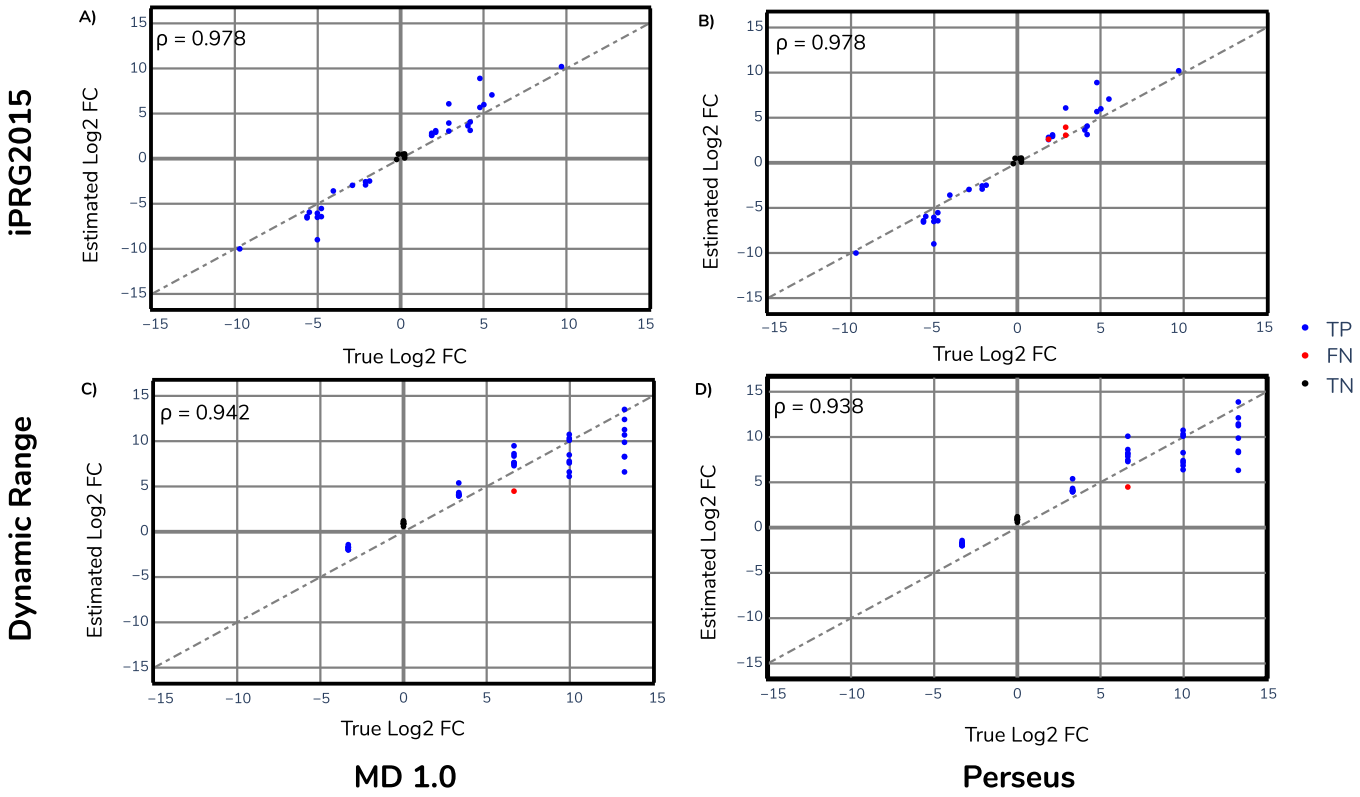


Figure 4. Scatter plots of the true and estimated log fold changes produced by Perseus MD 1.0 for iPRG2015 and the dynamic range datasets. (A) iPRG2015 Perseus, (B) iPRG2015 MD 1.0, (C) dynamic range dataset MD 1.0, (D) and dynamic range dataset Perseus. Only spiked proteins are plotted. The Pearson correlation is shown.

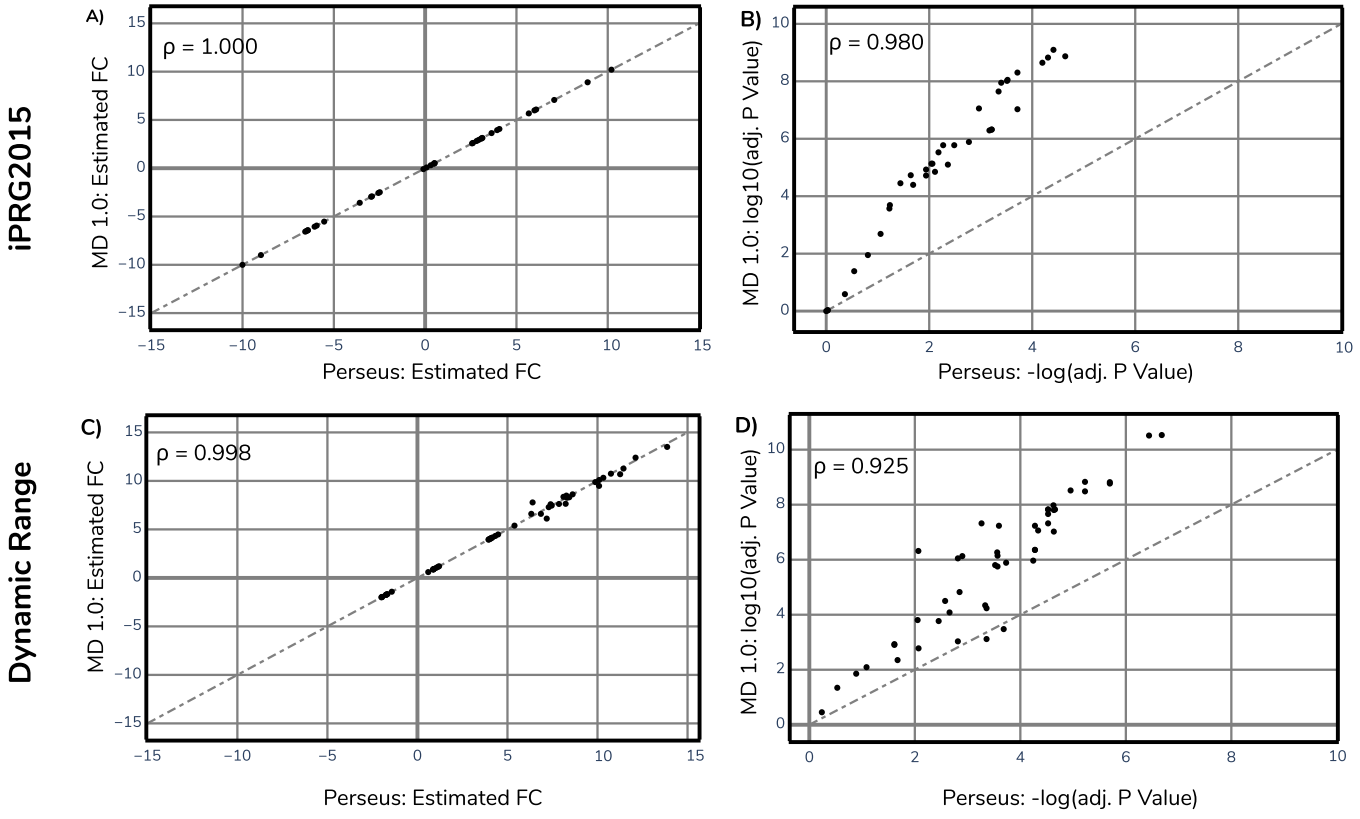


Figure 5. Comparison analysis between LFQ results produced by Perseus and MD+ Discovery MD 1.0. using the iPRG2015 and dynamic range datasets. (A) iPRG2015 fold change estimate comparison. (B) iPRG2015 −log10 adjusted *p*-value estimate comparison. (C) Dynamic range dataset fold change estimate comparison. (D) Dynamic range −log10 adjusted *p*-value estimate comparison.

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188

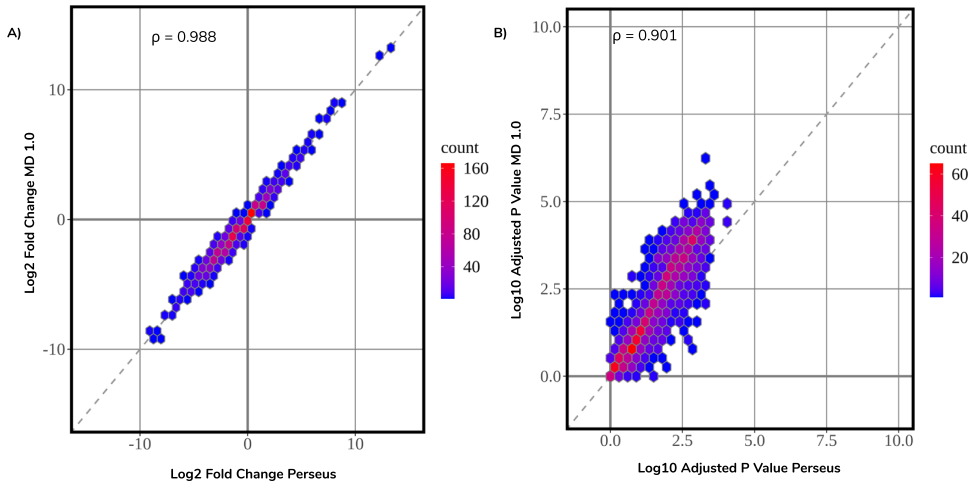


Figure 6. Comparison Analysis between LFQ results produced by Perseus and MD 1.0 using the HER dataset. (A) 2D density scatter plot showing log2 fold change estimates produced by Perseus on the *x* axis and MD 1.0 on the *y* axis. (B) 2D density scatter plot showing log10 adjusted *p*-value estimates produced by Perseus on the *x* axis and MD 1.0 on the *y* axis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00683.

Table S1 qualitative comparison of LFQ statistics and visualization resources (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Giuseppe Infusini − *Mass Dynamics, Melbourne, Victoria 3000, Australia; The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia; Department of Medical Biology, University of Melbourne, Melbourne, Victoria 3010, Australia;* orcid.org/0000-0001-5425-1698; Email: peppe@massdynamics.com

Andrew Webb − *Mass Dynamics, Melbourne, Victoria 3000, Australia; The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia; Department of Medical Biology, University of Melbourne, Melbourne, Victoria 3010, Australia;* orcid.org/0000-0001-5061-6995; Phone: +03 9345 2832; Email: webb@wehi.edu.au

Authors

Joseph Bloom − *Mass Dynamics, Melbourne, Victoria 3000, Australia;* orcid.org/0000-0002-3275-1103

Aaron Triantafyllidis − *Mass Dynamics, Melbourne, Victoria 3000, Australia;* orcid.org/0000-0001-5882-3665

Anna Quaglieri − *Mass Dynamics, Melbourne, Victoria 3000, Australia;* orcid.org/0000-0002-3660-6990

Paula Burton Ngov − *Mass Dynamics, Melbourne, Victoria 3000, Australia;* orcid.org/0000-0001-5783-6528

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00683

Author Contributions

G.I. wrote the initial MD 1.0 processing pipeline in which J.B. further developed into the R package ("LFQProcessing)." A.Q. built the "MassExpression" R package with assistance from J.B. A.T. built the web interface and set up the Amazon Web Services (AWS) compute and cloud storage infrastructure. J.B. performed the benchmarking analysis, prepared the figures, and wrote the manuscript. A.Q., P.B., G.I., and A.W. reviewed the manuscript. P.B. and A.T. interviewed and researched the existing work in this space.

Notes

The authors declare the following competing financial interest(s): The authors Giuseppe Infusini, Aaron Triantafyllidis, Paula Burton, and Andrew Webb declare that they are founders of Mass Dynamics, a for-profit enterprise, delivering software as a service in the processing, analysis and sharing of proteomics data. Joseph Bloom and Anna Quaglieri are employees of Mass Dynamics.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD028038. The MD 1.0 platform is available globally via https://app.massdynamics.com/.

■ ACKNOWLEDGMENTS

The authors would like to thank members of the Future Industries Institute at the University of South Australia for their feedback and support during the development of MD+ Discovery. The authors would also like to thank the proteomics community for their ongoing development of comprehensive tools and repositories for the analysis and sharing of proteomics data because without them, the work in this manuscript would be impossible. Finally, the authors would like to thank the Mass Dynamics team for their assistance with graphics, back-end development, and feedback.

■ REFERENCES

(1) Albulescu, R. et al. Mass Spectrometry for Cancer Biomarkers. In *Proteomics Technologies and Applications*; ed. Abdurakhmonov, I. Y.; IntechOpen, 2019.

(2) Wilcken, B.; Wiley, V.; Hammond, J.; Carpenter, K. Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N. Engl. J. Med.* **2003**, *348*, 2304−2312.

(3) Cox, J.; et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13*, 2513−2526.

(4) Al Shweiki, M. R.; et al. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and

Natural Variability of Protein Abundance. *J. Proteome Res.* **2017**, *16*, 1410−1424.
(5) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301−2319.
(6) Tyanova, S.; et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **2016**, *13*, 731−740.
(7) Rudolph, J. D.; Cox, J. A Network Module for the Perseus Software for Computational Proteomics Facilitates Proteome Interaction Graph Analysis. *J. Proteome Res.* **2019**, *18*, 2052−2064.
(8) Shah, A. D.; Goode, R. J. A.; Huang, C.; Powell, D. R.; Schittenhelm, R. B. LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant. *J. Proteome Res.* **2020**, *19*, 204−211.
(9) Kraus, M.; Mathew Stephen, M.; Schapranow, M.-P. Eatomics: Shiny Exploration of Quantitative Proteomics Data. *J. Proteome Res.* **2021**, *20*, 1070−1078.
(10) Gallant, J. L.; Heunis, T.; Sampson, S. L.; Bitter, W. ProVision: a web-based platform for rapid analysis of proteomics data processed by MaxQuant. *Bioinformatics* **2020**, *36*, 4965−4967.
(11) Smith, R. Conversations with 100 Scientists in the Field Reveal a Bifurcated Perception of the State of Mass Spectrometry Software. *J. Proteome Res.* **2018**, *17*, 1335−1339.
(12) Quadroni, M.; James, P. Proteomics and automation. *Electrophoresis* **1999**, *20*, 664−677.
(13) Pfeuffer, J.; et al. OpenMS - A platform for reproducible analysis of mass spectrometry data. *J. Biotechnol.* **2017**, *261*, 142−148.
(14) Perez-Riverol, Y.; et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47*, D442−D450.
(15) Choi, M.; et al. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J. Proteome Res.* **2017**, *16*, 945−957.
(16) Creedon, H.; et al. Identification of novel pathways linking epithelial-to-mesenchymal transition with resistance to HER2-targeted therapy. *Oncotarget* **2016**, *7*, 11539−11552.
(17) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, No. e47.
(18) Law, C. W.; Chen, Y.; Shi, W.; Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **2014**, *15*, R29.
(19) Phipson, B.; Lee, S.; Majewski, I. J.; Alexander, W. S.; Smyth, G. K. ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *Ann. Appl. Stat.* **2016**, *10*, 946.
(20) Bloom, J.; Infusini, G.. *MassDynamics/lfq_processing: 0.0.39A*; Zenodo, 2021, DOI: 10.5281/zenodo.5516670.
(21) Bloom, J.; Brady, S.; Quaglieri, A. *MassDynamics/MassExpression*; Zenodo, 2021, DOI: 10.5281/zenodo.5516664.
(22) Herbrich, S. M.; et al. Statistical inference from multiple iTRAQ experiments without using common reference standards. *J. Proteome Res.* **2013**, *12*, 594−604.
(23) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D155−D119.
(24) Reactome - a curated knowledgebase of biological pathways, 10.3180/50.

https://doi.org/10.1021/acs.jproteome.1c00683
J. Proteome Res. 2021, 20, 5180−5188