nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-020-0955-0

Check for updates

MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets

Meena Choi[1], Jeremy Carver[2], Cristina Chiva[3,4], Manuel Tzouros[5], Ting Huang[1], Tsung-Heng Tsai[1], Benjamin Pullman[2], Oliver M. Bernhardt[6], Ruth Hüttenhain[7], Guo Ci Teo[8], Yasset Perez-Riverol[9], Jan Muntel[6], Maik Müller[10], Sandra Goetze[10,11], Maria Pavlou[10], Erik Verschueren[7], Bernd Wollscheid[10,11], Alexey I. Nesvizhskii[8], Lukas Reiter[6], Tom Dunkley[5], Eduard Sabidó[3,4], Nuno Bandeira[2]✉ and Olga Vitek[1]✉

MassIVE.quant is a repository infrastructure and data resource for reproducible quantitative mass spectrometry–based proteomics, which is compatible with all mass spectrometry data acquisition types and computational analysis tools. A branch structure enables MassIVE.quant to systematically store raw experimental data, metadata of the experimental design, scripts of the quantitative analysis workflow, intermediate input and output files, as well as alternative reanalyses of the same dataset.

Quantitative mass spectrometry data analysis currently has multiple unmet reproducibility goals[1]. At the minimum, the mass spectrometry–based workflows must provide enough information to enable its full independent replication[2]. Beyond that, conclusions of data analysis should not be dependent on particular tuning parameters or software tools. Data analysis should demonstrate that alternative and equally appropriate parameter settings or software lead to qualitatively similar conclusions.

In mass spectrometry–based proteomics, data analysis is broadly categorized into peptide ion identification and quantification. Much progress in identification has been made in terms of open availability of tools and transparency of their algorithms. Archival resources MassIVE, PRIDE[3], Panorama[4,5], the PASSEL[6] component of Peptide Atlas, and jPOST in ProteomeXchange[7,8] store raw data, peak lists, search engine output, identification results and corresponding mass spectra.

Unfortunately, reproducibility and transparency of data analysis for relative protein quantification is less satisfactory. First, given the great diversity of biological objectives and experiments, quantitative analyses require richer metadata describing experimental design and biological samples. Second, quantitative experiments require many data processing steps, which are distinct from similar steps in quantitative transcriptomic investigations. These include detection and identification of chromatographic peaks and reporter ions, and propagating those identities across multiple runs. Finally, existing analysis tools (such as Skyline[9], MaxQuant[10], OpenMS[11], OpenSWATH[12], DIA-Umpire[13], Proteome Discoverer or Spectronaut[14]) integrate, in their own unique ways, diverse functionalities for identification and quantification. These tools offer various parameters and options, and output different details in various storage formats. Many tools also offer graphical user interfaces, for which analyses are difficult to document.

Benchmarking of individual analysis strategies and tools for quantitative proteomics workflows has become increasingly prevalent[15], but lack infrastructure to store, document, annotate and reanalyze the full diversity of analyses.

To meet these reproducibility needs, we implemented MassIVE.quant, an infrastructure that supports quantitative mass spectrometry–based proteomics experiments. MassIVE.quant is integrated with an existing repository, the mass spectrometry interactive virtual environment (MassIVE).

MassIVE.quant systematically stores the intermediate output files of every tool and workflow in a way that allows the user to easily inspect, reproduce or modify any component of the workflow, beginning with well-defined intermediate files. To accomplish this, we first developed a series of steps that represent a quantitative proteomics experiment with any experimental design, data acquisition and data analysis tools (Fig. 1). These steps consist of (1) annotations of experimental design; in particular, descriptors of biological samples and conditions; (2) strategies of sample preparation and data acquisition; (3) peptide ions identification; (4) quantification and (5) statistical analysis. At each step, MassIVE.quant provides the infrastructure to store all intermediate descriptions, annotations, analysis scripts and results.

MassIVE.quant does not prescribe a standard format, but meets scientists where they are by directly accommodating the diverse nature of existing workflows. Each dataset contains links to the original publications or to metadata, which can be used to gain deeper insight into the biological context of the experiment. While

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [3]Proteomics Unit, Center for Genomics Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain. [4]Proteomics Unit, Universitat Pompeu Fabra, Barcelona, Spain. [5]Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, Hoffmann-La Roche Ltd, Basel, Switzerland. [6]Biognosys, Zurich, Switzerland. [7]Department of Molecular and Cellular Pharmacology, University of California, San Francisco, San Francisco, CA, USA. [8]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [9]Proteomics Services, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [10]Department of Health Sciences and Technology, Institute of Translational Medicine, ETH, Zurich, Switzerland. [11]Swiss Institute of Bioinformatics, Lausanne, Switzerland. ✉e-mail: bandeira@ucsd.edu; o.vitek@northeastern.edu

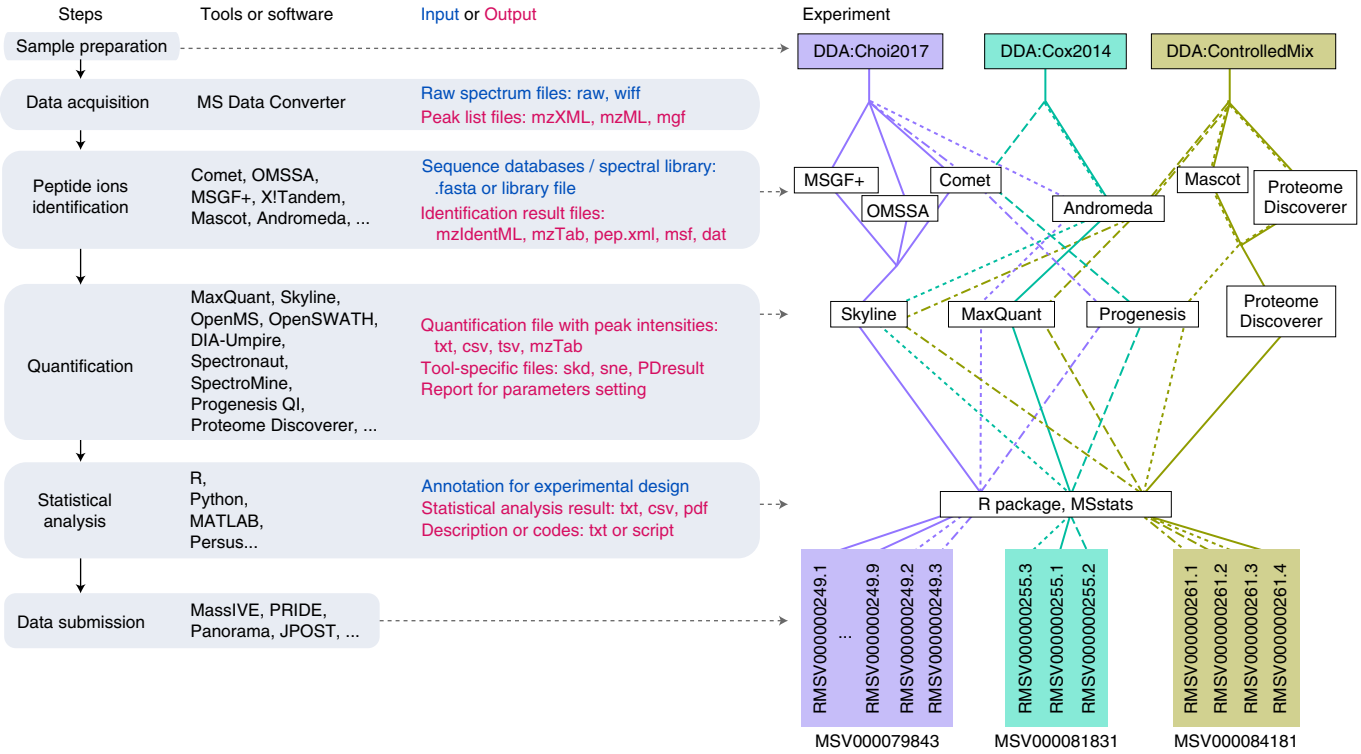


Fig. 1 | Outline of MassIVE.quant repository structure and reanalysis of three DDA-based experiments. Each step can be performed with multiple algorithms and software tools, generating tool-specific files in diverse formats. For the experiments in the figure, MassIVE.quant stores the intermediate outputs from combinations of algorithms and tools for peptide ion identification and quantification. For example, DDA:Choi2017 was processed with eight combinations of parameter settings in Skyline. Each reanalysis is saved with a unique reanalysis ID, prefixed by RMSV, under the experiment repository prefixed by MSV in MassIVE.quant.

analyte identification is represented by existing standard formats (such as mzTab, thousands of examples available for MassIVE datasets) and some of these can represent some quantitative information (such as the quant section in mzTab files, for example, the mzTab file for RMSV000000249.18), the output files produced by quantitative analyses tools can be of any nature. None of the formats mentioned before support reporting of the results of statistical analyses of quantitative data. We chose a tabular format (that is, .csv) as a common representation of the output of quantitative and statistical analyses for all tools. This format emphasizes biologically relevant aspects of the output, such as the identity of differentially abundant peptides or proteins, the magnitudes of fold changes and the associated variation.

Next, at each step, a branch structure enables the user to view reanalyses of each experiment. The reanalyses can be performed by the user offline with any combination of software tools and settings. MassIVE.quant stores the intermediate files and allows the user to check for the presence of script files, accuracy of parameters and completeness of documentation.

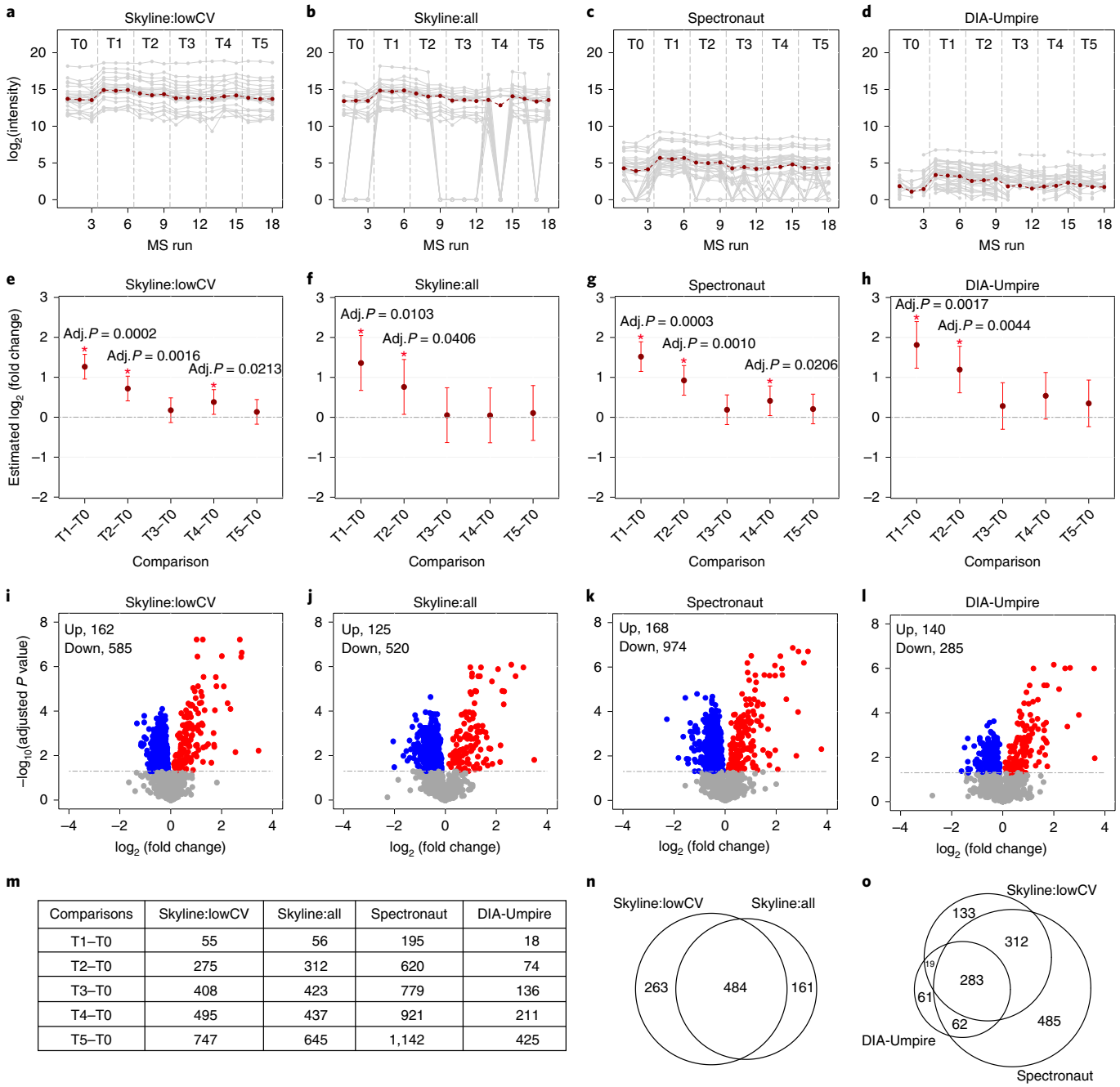
To scale the submission procedure and to ensure the reproducibility of a quantitative workflow, MassIVE.quant maintains datasets with four levels of curation (bronze, silver, gold and platinum), reflecting the documentation and the reproducibility of the quantitative workflow. During the submission, the infrastructure checks whether the submission of the dataset or reanalysis meets the minimal requirements for the entry level of curation. The submitter can then request the advanced review to level up.

MassIVE.quant further automates the statistical analysis of quantified proteins with an online MSstats workflow accessible with a user-friendly interface. This workflow can be used to reproduce the statistical analysis steps in MassIVE.quant reanalyses, as well as to analyze new private or public datasets. The MSstats Comparison workflow in MassIVE.quant automatically compares MSstats outputs across alternative reanalyses, and produces figures such as Fig. 2o. This enables the user to evaluate the implications of alternative reanalyses on conclusions regarding differential protein abundance.

The online user documentation clearly describes the structure and the vocabulary used by MassIVE.quant, and provides detailed instructions for contributing data, reanalyses and comparisons (Supplementary Note 1). Supplementary Figs. 1 and 2 and Supplementary Table 1 describe the infrastructure of MassIVE.quant and give an example of the reanalysis submission workflow.

The impact of the choice between various analysis options is best understood in the presence of some notion of ground truth. Therefore, we populated MassIVE.quant with a collection of ten datasets with controlled mixtures with known changes in protein abundance. These include three datasets collected using label-free data-dependent acquisition (DDA), four datasets collected using data-independent acquisition (DIA), one dataset collected using selective reaction monitoring (SRM) with heavy labeled isotope peptides and two DDA dataset collected using chemical labeling (TMT) (Supplementary Table 2). These datasets vary in background proteomes, number of conditions and replicates, and type and number of differentially abundant proteins.

We also populated MassIVE.quant with 95 reanalyses of these ten controlled datasets using multiple software tools, performed by the developers of the tools or by expert users. All the DDA experiments in Fig. 1 were processed with up to six tools for identification and four tools for quantification. For example, data from



Comparisons	Skyline:lowCV	Skyline:all	Spectronaut	DIA-Umpire
T1–T0	55	56	195	18
T2–T0	275	312	620	74
T3–T0	408	423	779	136
T4–T0	495	437	921	211
T5–T0	747	645	1,142	425

Fig. 2 | Reanalyses of DIA:Selevsek2015, profiling changes in proteome abundance of *Saccharomyces cerevisiae* over six time points: T0 (0 min), T1 (15 min), T2 (30 min), T3 (60 min), T4 (90 min) and T5 (120 min), $n = 3$ biologically independent samples per each time point, in response to osmotic stress (RMSV000000251). a–d, Discrepancies of quantification of protein YKL096W across data processing tools. Gray lines, fragments reported by each tool. Red lines, protein quantification summarized by MSstats. **a**, Skyline:lowCV used Skyline to quantify a subset of the fragments with low coefficient of variation. **b**, Skyline:all used Skyline to quantify all detectable peptides, with a maximum of six fragments each. **c**, Data processed by Spectronaut. **d**, Data processed by DIA-Umpire. **e–h**, Discrepancies in detecting differential abundance for protein YKL096W across data processing tools, with statistical analysis by MSstats: Skyline:lowCV (**e**), Skyline:all (**f**), Spectronaut (**g**) and DIA-Umpire (**h**). Dark red dot, center for error bars, model-based estimates of $\log_2$(fold change) of protein abundance, as determined by MSstats. Error bars, 95% confidence intervals for the $\log_2$(fold change), as determined by MSstats. *Adjusted $P < 0.05$. **i–l**, Volcano plots, summarizing differential abundance between T5 and T0: Skyline:lowCV (**i**), Skyline:all (**j**), Spectronaut (**k**) and DIA-Umpire (**l**). Dashed line, FDR = 0.05; blue dots, significantly down-regulated proteins; red dots, significantly up-regulated proteins (counts are shown at the top left corner; other time points are shown in Supplementary Fig. 3). **m**, Number of differentially abundant proteins across all time points and all tools, FDR = 0.05. **n**, Venn diagram of differentially abundant proteins between two processing approaches by Skyline, comparing T5 versus T0. **o**, Venn diagram of differentially abundant proteins across all tools, comparing T5 versus T0 (other time points are shown in Supplementary Fig. 4).

Choi et al.[16] (DDA:Choi2017) were processed with eight different combinations of parameter settings in Skyline. The combinations of algorithms, tools and settings generated ten distinct quantification reports. Finally, up to five different types of downstream statistical analysis per dataset using MSstats[17] generated 22 distinct tests for differential protein abundance for DDA:Choi2017.

To demonstrate the use of documentation, reanalysis and curation in basic biology investigations, we further populated MassIVE.quant with a collection of biological datasets, at the time of publication, including eight DIA/SWATH, seven SRM, 12 DDA and six experiments collected DDA-TMT acquisition, analyzed with multiple tools; 25 datasets with platinum level of curation and 18 datasets with gold (Supplementary Table 3). For example, the DIA experiment by Selevsek et al.[18] (DIA:Selevsek2015) was reanalyzed four times using different analysis strategies and different processing tools and parameter settings. Figure 2, Supplementary Table 4 and Supplementary Figs. 3 and 4 illustrate how changes in data processing propagated themselves into discrepancies in the number of quantified proteins, frequency of missing values and lists of differentially abundant proteins. Figure 2a–d illustrates these discrepancies in the special case of one protein. The analysis strategies and processing tools affected protein-level summaries in terms of scale, variation and patterns of missing values. This in turn affected the estimates of fold changes (Fig. 2e–h) and tests for differential abundance (Fig. 2i–l). Analysis with filtering in Skyline, applied to limit the DIA features to those known to be informative a priori (Skyline:lowCV, RMSV000000251.1) detected a smoother, and therefore more biologically plausible, pattern of differential abundance in time (Fig. 2m). While the true differential abundance is unknown, changes identified by most tools are more likely to be real (Fig. 2n–o). Such comparisons help curate the results of biological investigations.

To summarize, MassIVE.quant provides an opportunity for large-scale deposition of heterogeneous experimental datasets and facilitates a community-wide conversation about the benefits of its use. We hope that the community will find the resource useful and welcome user-driven submissions of both new datasets and documented reanalyses of the existing datasets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0955-0.

Received: 10 September 2019; Accepted: 13 August 2020;
Published online: 14 September 2020

References

1. Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226–1227 (2011).
2. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
3. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
4. Sharma, V. et al. Panorama: a targeted proteomics knowledge base. *J. Proteome Res.* **13**, 4205–4210 (2014).
5. Sharma, V. et al. Panorama public: a public repository for quantitative data sets processed in skyline. *Mol. Cell Proteom.* **17**, 1239–1244 (2018).
6. Farrah, T. et al. PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* **12**, 1170–1175 (2012).
7. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
8. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).
9. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
10. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
11. Rost, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
12. Rost, H. L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219–223 (2014).
13. Tsou, C. C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
14. Bruderer, R. et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell Proteom.* **14**, 1400–1410 (2015).
15. Navarro, P. et al. A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* **34**, 1130–1136 (2016).
16. 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.* **16**, 945–957 (2017).
17. Choi, M. et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
18. Selevsek, N. et al. Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-mass spectrometry. *Mol. Cell Proteom.* **14**, 739–749 (2015).

Methods

Manual and tutorials. MassIVE.quant user manuals and video tutorials explain how to (1) create an account in MassIVE.quant, (2) upload files to MassIVE account via FTP, (3) submit quantification reanalysis to MassIVE.quant, (4) access reanalyses, (5) run MSstats workflow in MassIVE.quant and (6) compare the results of statistical analysis by MSstats. User manuals are available at MassIVE.quant section in https://ccms-ucsd.github.io/MassIVEDocumentation/. The video tutorial is available in https://www.youtube.com/channel/UCPeNhYFItiabsoOITPZBc5Q.

Data analysis for DIA:Selevsek2015. For Skyline:All, all files for ProteomeXchange ID PXD001010 were downloaded, along with the forward and reversed FASTA and mzXML files ($n=46$) used in the library peptide search (requested from the authors). Using Skyline (daily v.3.7.1.11571), first a spectral library was built using the iProphet score cut-off 0.0242 suggested in the paper as achieving 1% false discovery rate (FDR) at the peptide-spectrum matches (PSM) level, with 1,031 ambiguously matched peptides excluded, resulting in 82,439 unique peptides (104,993 entries). Because the Biognosys iRT standard peptides used were not included in the search, an iRT library was created from these files by adding all detected peptides with the iRT standards added as targets in Skyline and importing the mzXML files for MS1 extraction. The iRT values were then calculated using both the extracted peaks for the iRT standards and target peptide MS1 peaks where the peak contained a matching tandem mass spectrometry ID, because the runs used fractionated samples and all peptides were not expected in all runs. For DIA, allowing unique peptides of length 7–45 amino acids resulting from semitryptic cleavage with up to two missed cleavages, with Carbamidomethyl (C) and optionally Oxidation (M), precursors of charge states 2, 3 or 4, from 400–1,200 *m/z* (the range covered by the DIA method), with six product transitions found in the library, of y or b type and 1 or 2 charge state (excluding y1, y2, b1 and b2). Chromatogram extraction was set to use time of flight extraction at 18,000 resolving power with high-selectivity extraction applying to all tandem mass spectra within 10 min of predicted retention times using the iRT library. Importing the FASTA file and then removing duplicate peptides and empty proteins resulted in targets for 4,603 proteins, 68,910 peptides, 87,042 precursors and 522,252 fragment ions, at 32% protein, 2.7% unique peptide FDR by reversed sequence decoy counting (decoys/targets). The protein FDR is likely overstated because the FASTA file contains only 6,717 protein sequences, which means as many as two-thirds of false peptides can be expected to occur in a true protein, while the same is not true for detections of reversed peptides. Even at 10% protein FDR, however, this target set seemed to contain a higher error rate than we felt desirable. For these experiments, we decided to rebuild the library using the iProphet score cut-off 0.9, with 361 ambiguously matched peptides excluded, resulting in 64,501 unique peptides (84,245 entries). For our most inclusive method, we chose to include only fully tryptic peptides and no variable modifications (dropping oxidation (M)), which resulted in targets for 4,152 proteins, 36,889 peptides, 48,082 precursors and 288,492 fragment ions, at 2.6% protein, 0.29% unique peptide FDR by reversed sequence decoy counting. An equal number of shuffled sequence decoys were generated for mProphet model generation. The 18 runs were then imported into the template an mProphet model trained and applied. The MSstats report was exported for further analysis.

For Skyline:lowCV, using Skyline (daily v.3.7.1.11571) and starting from the settings for the broadly inclusive test, we restored semitryptic cleavage and the variable modification oxidation (M), expecting unstable peptides to be filtered by our first experiment. Importing the forward and reversed FASTA file resulted in targets for 4,246 proteins, 58,168 peptides, 74,314 precursors and 445,884 fragment ions, at 4.5% protein, 0.34% unique peptide FDR by reversed sequence decoy counting. Reversed sequences, accounting for 193 proteins and 200 peptides, were left in the targets list to be carried through the subsequent 'reproducibility' experiment. An equal number of shuffled sequence decoys were generated for mProphet model generation. The runs ($n=8$ reported as $n=4$ in the paper) acquired for the reproducibility experiment were then imported into the template, an mProphet model trained and applied. Next, the targets were filtered for peptides detected at q values less than 0.01 in at least four (of eight) runs, and where the CV of the detected peak areas, median normalized, was less than 20%, resulting in targets for 2,212 proteins, 13,744 peptides, 18,910 precursors and 113,460 fragment ions, at 1% protein, 0.18% unique peptide FDR by reversed sequence decoy counting. The remaining peptides comprised 249 of 2,367–10.5% (oxidation (M), 679 of 18,479–3.7%) unmodified semitryptic and 12,816 of 37,322–34.3% unmodified tryptic peptides. As expected, oxidation (M) and semitryptic peptides made it through this filter at much lower rates than unmodified fully tryptic peptides. An equal number of shuffled sequence decoys of matching types were generated for mProphet model generation. The 18 runs were then imported into the template an mProphet model trained and applied. The MSstats report was exported for further analysis.

For Spectronaut, a spectral library was generated using all available raw files from the original publication using the Pulsar search engine integrated in Spectronaut 11(11.0.15038) with default settings. The uniprot yeast reference proteome was used for the spectrum centric data analysis for library generation. DIA data were then analyzed using default settings and exported using the built in MSstats Report (v.3.7.3) export schema.

For DIA-Umpire, the raw files were converted to mzXML files with centroiding. The resulting mzXML files were processed by the signal extraction module of DIA-Umpire to generate pseudo-tandem mass spectra. The generated pseudo-MS/MS spectra were searched using X! Tandem, Comet and MSGF+ search engines. The output files from the search engines were further analyzed by PeptideProphet and combined by iProphet. FDR filtering was done with PeptideProphet and ProteinProphet. DIA-Umpire's Quant module was for the quantification analysis. The outputs for all-level quantification (FragSummary, PeptideSummary, ProtSummary) were used in further analysis.

Statistical analysis for DIA:Selevsek2015. R package MSstats v.3.10.6 was used to preprocess the output from Skyline, Spectronaut and DIA-Umpire before statistical analysis, to have protein quantification and to perform differential abundance analysis. MSstats estimated $\log_2$(fold change) and the standard error by linear mixed effect model for each protein. To test two-sided null hypothesis of no changes in abundance, the model-based test statistics were compared to the Student t-test distribution with the degrees of freedom appropriate for each protein and each dataset. The resulting P values were adjusted to control the FDR with the method by Benjamini–Hochberg. Parameter settings as well as the R code used to analyze DIA:Selevsek2015 are available in reanalysis container, RMSV000000251 in MassIVE.quant.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All the datasets that support this study are publicly available in MassIVE.quant (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp) with MassIVE and ProteomeXchange identifiers. Additionally, identifiers for all the datasets are listed in Supplementary Tables 2, 3 and 5.

Acknowledgements

This work was supported in part by NSF CAREER award no. DBI-1054826, grant no. NSF DBI-1759736 and the Chan-Zuckerberg foundation to O.V., grant no. NIH-NLM 1R01LM013115 to N.B. and O.V., NSF award no. ABI 1759980, NIH award nos. P41GM103484 and R24GM127667 to N.B. and the Personalized Health and Related Technologies (grant no. PHRT 0-21411-18) strategic focus area of ETH to B.W. The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech) and it is a member of the ProteoRed PRB3 consortium that is supported by grant no. PT17/0019 of the PE I+D+i 2013–2016 from the Instituto de Salud Carlos III (ISCIII) and ERDF. We acknowledge support from the Spanish Ministry of Science, Innovation and Universities, 'Centro de Excelencia Severo Ochoa 2013–2017', SEV-2012–0208 and Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (grant no. 2017SGR595). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 823839 (EPIC-XS). Y.P.-R. acknowledges the Wellcome Trust (grant no. 208391/Z/17/Z). We thank the MacCoss laboratory (Department of Genome Sciences, University of Washington) for the Skyline analyses and contributing the processed data, the Slavov laboratory (College of Engineering, Northeastern University) for providing the data and the Guo laboratory (School of Life Sciences, Westlake University, China) for providing the data.

Author contributions

M.C., J.C., N.B. and O.V. designed the research. M.C. and T.H. collected datasets and performed statistical analysis. J.C. and B.P. implemented MassIVE.quant. T.-H.T. performed statistical analysis. C.C., E.S. and M.T. experimented with new controlled mixtures. C.C., M.T., R.H., G.C.T., Y.P-R., J.M., M.M., S.G., M.P., E.V., B.W., O.M.B., A.I.N., L.R., T.D. and E.S. analyzed data up to quantification. M.C., N.B. and O.V. wrote the manuscript, with input from all authors.

Competing interests

O.M.B., J.M. and L.R. are employees of Biognosys AG. Spectronaut is a trademark of Biognosys AG. M.T. and T.D. are employees of Hoffmann-La Roche Ltd. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0955-0.

Correspondence and requests for materials should be addressed to N.B. or O.V.

Peer review information Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

nature research

Corresponding author(s): Olga Vitek

Last updated by author(s): Aug 12, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	This manuscript does not introduce new experiments, but analyses publicly available data. Descriptions of all the data acquisition steps corresponding to each dataset are available in reanalysis repository for each experiment. The list of links are available in MassIVE.quant (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp).
Data analysis	Currently MassIVE.quant has 180 reanalyses, each with its own identifier. It is not practical to list them here. Instead we provided this information in the supplementary tables, and pointed to it from the 'Data Availability' section. All this information is available in MassIVE.quant (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the datasets that support this study are available in MassIVE. Identifiers for all the datasets are listed in Supplementary Table 2, 3, and 5, and MassIVE.quant (https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	NA (This manuscript does not introduce new experiments, but analyses publicly available data.)
Data exclusions	NA
Replication	NA (This manuscript does not introduce new experiments, but analyses publicly available data.)
Randomization	NA (This manuscript does not introduce new experiments, but analyses publicly available data.)
Blinding	NA (This manuscript does not introduce new experiments, but analyses publicly available data.)

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging