Tissue-based absolute quantification using large-scale TMT and LFQ experiments

Hong Wang[1], Chengxin Dai[1], Julianus Pfeuffer[2], Timo Sachsenberg[2], Aniel Sanchez[3], Mingze Bai[1], and Yasset Perez Riverol[4]

[1]Affiliation not available
[2]University of Tübingen
[3]Lund University
[4]European Bioinformatics Institute

April 17, 2023

Abstract

Relative and absolute intensity-based protein quantification across cell lines, tissue atlases, and tumour datasets is increasingly available in public datasets. These atlases enable researchers to explore fundamental biological questions, such as protein existence, expression location, quantity, and correlation with RNA expression. Most studies provide MS1 feature-based label-free quantitative (LFQ) datasets; however, growing numbers of isobaric tandem mass tags (TMT) datasets remain unexplored. Here, we compare traditional intensity-based absolute quantification (iBAQ) proteome abundance ranking to an analogous method using reporter ion proteome abundance ranking with data from an experiment where LFQ and TMT were measured on the same samples. This new TMT method substitutes reporter ion intensities for MS1 feature intensities in the iBAQ framework. Additionally, we compared LFQ-iBAQ values to TMT-iBAQ values from two independent large-scale tissue atlas datasets (one LFQ and one TMT) using robust bottom-up proteomic identification, normalisation, and quantitation workflows.

Tissue-based absolute quantification using large-scale TMT and LFQ experiments.

Hong Wang [1], Chengxin Dai [1,2], Julianus Pfeuffer [3], Timo Sachsenberg[4,5], Aniel Sanchez [6], Mingze Bai[1,2], Yasset Perez-Riverol [7, *]

[1] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China.

[2] State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing, 102206, China.

[3] Algorithmic Bioinformatics, Freie Universität Berlin, Berlin, Germany.

[4] Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen, 72076, Germany

[5] Institute for Biological and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany

[6] Section for Clinical Chemistry, Department of Translational Medicine, Lund University, Skåne University Hospital Malmö, Malmö, Sweden

[7] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

Abstract

Relative and absolute intensity-based protein quantification across cell lines, tissue atlases, and tumour datasets is increasingly available in public datasets. These atlases enable researchers to explore fundamental biological questions, such as protein existence, expression location, quantity, and correlation with RNA expression. Most studies provide MS1 feature-based label-free quantitative (LFQ) datasets; however, growing numbers of isobaric tandem mass tags (TMT) datasets remain unexplored. Here, we compare traditional intensity-based absolute quantification (iBAQ) proteome abundance ranking to an analogous method using reporter ion proteome abundance ranking with data from an experiment where LFQ and TMT were measured on the same samples. This new TMT method substitutes reporter ion intensities for MS1 feature intensities in the iBAQ framework. Additionally, we compared LFQ-iBAQ values to TMT-iBAQ values from two independent large-scale tissue atlas datasets (one LFQ and one TMT) using robust bottom-up proteomic identification, normalisation, and quantitation workflows.

Proteomics is a powerful tool for understanding the underlying biology of cells and tissues. Large-scale cell lines, tumour datasets, or tissue atlases enable researchers to ask fundamental questions about the proteome, such as protein existence, expression location and correlation with RNA expression [1-3]. The number of publicly available datasets continues to expand every year [4], facilitating their reuse [5, 6] and integration into protein expression resources [7, 8]. Label-free intensity-based absolute quantification (iBAQ) is a robust and common method to estimate the expression of proteins without the need for a standard reference sample [9, 10]. This method measures relative protein abundances within a sample and can be converted to approximate absolute scales, like copy number when certain assumptions are met. iBAQ protein expression has been only explored for the label-free data-dependent (DDA) [9] and independent acquisition (DIA) methods using MS1 [10].

MS2 methods [11, 12], such as spectral counting, can serve as a proxy for absolute quantification in bottom-up proteomics experiments. Spectral-counting algorithms offer some advantages because they can be applied directly to the data commonly collected for identification purposes including TMT (multiplex) experiments. In 2011, Colaert et. al. [12] explored three MS2-based quantitative methods: Exponentially modified Protein Abundance Index (EmPAI) [13], Normalized Spectral Abundance Factor (NSAF) [14], and normalized Spectral Index (SIn) [15]. Their findings indicated that the NSAF method outperformed both EmPAI and SIn in terms of accuracy and precision [12]. However, spectral counting-based quantification has limitations because it does not use chromatography peak attributes such as height or area potentially limiting its accuracy and dynamic range [16, 17]. Ahrné et al. [18] undertook a distinct intensity-based strategy to calculate iBAQ values in TMT datasets, treating them as label-free datasets. This involved distributing MS1 intensities of all TMT-labelled features among the individual samples based on the relative reporter ion intensities. However, this approach is more complex, as the datasets need to be analyzed as label-free experiments and precursor ion intensities must be extracted. Furthermore, this approach has not been applied to a large-scale dataset or benchmarked across different datasets.

Here, we explored an alternative approach to perform absolute protein expression analysis on TMT datasets using the direct reporter ion intensities. To assess the accuracy of this method, we employed a gold-standard mix-proteome dataset (PXD007683) [19] analyzed with both LFQ and TMT methods. We then calculated iBAQ values based on either MS1 feature or reporter ion intensities (respectively) and compared the correlation for all quantified proteins. Additionally, we applied robust normalization and quantitation workflows to analyze two large-scale tissue datasets from Jian et al. (TMT – PXD016999) [1] and Wang et al. (LFQ – PXD010154) [2].

Intensity-based absolute quantification (iBAQ) values were estimated using the MS1 intensities for label-free experiments, and the reporter ion intensities in the case of TMT datasets. Feature intensity tables for all analyzed datasets were generated using the quantms (https://quantms.readthedocs.io/) workflow which enables the analysis of DDA, DIA label-free, and TMT datasets. Each generated feature was the combination of a peptide sequence, modifications, charge state, sample, fraction, and technical or biological replicate. Feature intensities were normalized using quantile normalization, the highest intensity for each feature was selected across replicates. Finally, feature intensities were averaged (mean) at the peptide sequence level.

iBAQ is computed by dividing the sum of peptide intensities by the number of theoretically observable peptides of the protein. Each iBAQ value was normalized to the sum of all iBAQ values for the same sample (riBAQ) [20, 21]. All analysis steps are included in a Python package (https://github.com/bigbio/ibaqpy).

We tested the TMT-iBAQ approach using a mix-proteome dataset comprising both Human and Yeast samples in multiple concentrations [19]. The primary objective of the dataset and the original study was to evaluate the capability of TMT and LFQ approaches in accurately quantifying fold changes of 3-, 2-, and 1.5-fold across the entire dataset. All parameters for the reanalysis were annotated using the SDRF file format [22] (**Supplementary Note 1**). In the present study, we did not explore the differential expression across samples (as originally designed by O'Connell et. al. [19]) but compared the expression of the Human proteins when using TMT-iBAQ or LFQ-iBAQ.

In the PXD007683 dataset, we quantified a total of 94,804 peptides and 8,401 proteins. There were 33,321 peptides and 6,273 proteins commonly identified using TMT and LFQ approaches; while 18,524 peptides from 392 proteins and 42,959 peptides from 1,736 proteins were quantified using only LFQ or TMT approaches, respectively. The peptide intensity between both approaches is statistically significantly correlated for all samples ($R > 0.44$, p-value $< 2.2e\text{-}16$ –**Supplementary Note 2**). The log-scale iBAQ values for both TMT and LFQ approaches of the PXD007683 dataset were compared, as shown in Figure 1A-B. First, we evaluated the reproducibility of the two methods across all 11 sample replicates for both approaches (**Figure 1A**). Samples analysed with the label-free method showed a higher coefficient of variation (average CV = 15%), while TMT samples had an average CV=11%. The iBAQ values displayed a similar distribution across the 11 samples, with a higher median intensity observed for TMT experiments than LFQ in all samples (Figure 1A). The iBAQ Pearson correlation between the TMT and LFQ approaches is remarkably high ($R > 0.83$, p-value < 2.2 e-16). These results demonstrate that the iBAQ values obtained from both LFQ and TMT approaches in this benchmark dataset are highly consistent and reliable. In fact, this result is supported by the long use of MS2 (based on fragment ion intensities) data for quantification in proteomics experiments by using MRM, DIA or having found good correlations between precursors and their reporters in DDA experiments [23].

While previous authors [16, 19, 24] have found that LFQ and TMT methods offer similar performance in terms of accuracy when analysing the same sample, comparisons of these methods for proteome characterization between different studies with similar tissue remains unexplored. We tested this in reanalysis of two large-scale human tissue datasets from Jian et al. (TMT – PXD016999) [1] and Wang et al. (LFQ – PXD010154) [2] (**Supplementary Note 1**). Both datasets were analysed using the same database (UniProt human Swiss-Prot 092022), the quantms workflow, and the corresponding datasets parameters (**Supplementary Note 1**). For PXD010154, a total number of 340,306 peptides and 14,602 proteins were quantified, while the number of quantified peptides and proteins for PXD016999 were 173,678 and 10,351, respectively. Figure 2A shows the distribution of iBAQ values for all shared tissues between both datasets (adrenal gland, liver, lung, ovary, pancreas, prostate, spleen, stomach, and testis), while median intensity is higher for TMT experiments compared with LFQ for all tissues except prostate. Figure 2B shows the iBAQ correlation between both experiments for the shared tissues, and all tissues show a correlation coefficient higher than 0.80. The iBAQ values obtained by LFQ and TMT of these 9 tissues had a strong correlation and high consistency. Previously, Betancourt *et. al.* [25] integrated TMT results with LFQ using the three most abundant peptides for each protein quantified (TOP3), but the reproducibility and the correlation between both technologies were never explored. Using the transformed normalized intensities as suggested by Jiang et. al. [1], instead of the iBAQ values from reporter ion intensities (as suggested in this research), could negatively affect the correlation between relative proteome abundances obtained with LFQ or TMT.

In summary, intensity-based absolute quantification (iBAQ), as previously reported, is a robust and common method for estimating the relative/absolute expression of proteins. This study explored and extended the capabilities of the LFQ-iBAQ approach to perform proteome-wide quantification in TMT datasets using direct reporter ion intensities. The results showed that the iBAQ correlation between the TMT and LFQ approaches in different datasets is high, indicating the potential of the direct reporter ion intensity method for

relative protein abundance analyses in TMT datasets. This new approach can enable the future integration public TMT and LFQ proteomics datasets using intensity-based methods instead of less accurate spectral counting which could improve the accuracy and reproducibility of proteomics meta-analyses.

Data Availability Statement

The Raw data of the three reanalysed datasets can be found in ProteomeXchange with the original accessions: PXD016999, PXD010154, and PXD007683. The iBAQ values for all samples can be found on GitHub: https://github.com/ypriverol/quantms-research/tmt-lfq-ibaq.

Acknowledgements

YPR would like to acknowledge funding from EMBL core funding, Wellcome grants (208391/Z/17/Z, 223745/Z/21/Z), and the EU H2020 project EPIC-XS [823839]. JP would like to acknowledge Forschungscampus MODAL (project grant 3FO18501). CD and MB are supported by the National Key Research and Development Program of China (2017YFC0908404, 2017YFC0908405) and the Natural Science Foundation of Chongqing, China (cstc2018jcyjAX0225). The authors acknowledge Dr. Phillip Wilmarth for helpful discussions and assistance with manuscript preparation.

References

[1] Jiang, L., Wang, M., Lin, S., Jian, R.*, et al.* , A Quantitative Proteome Map of the Human Body. *Cell* 2020,*183* , 269-283 e219.

[2] Wang, D., Eraslan, B., Wieland, T., Hallstrom, B.*, et al.* , A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* 2019, *15* , e8503.

[3] Di Meo, A., Sohaei, D., Batruch, I., Alexandrou, P.*, et al.* , Proteomic Profiling of the Human Tissue and Biological Fluid Proteome. *J Proteome Res* 2021, *20* , 444-452.

[4] Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D.*, et al.* , The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022,*50* , D543-D552.

[5] Claeys, T., Menu, M., Bouwmeester, R., Gevaert, K., Martens, L., bioRxiv 2022.

[6] Prakash, A., Garcia-Seisdedos, D., Wang, S., Kundu, D. J.*, et al.* , Integrated View of Baseline Protein Expression in Human Tissues. *J Proteome Res* 2023, *22* , 729-742.

[7] Lautenbacher, L., Samaras, P., Muller, J., Grafberger, A.*, et al.* , ProteomicsDB: toward a FAIR open-source resource for life-science research. *Nucleic Acids Res* 2022, *50* , D1541-D1552.

[8] Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., von Mering, C., Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 2015,*15* , 3163-3168.

[9] Schwanhausser, B., Busse, D., Li, N., Dittmar, G.*, et al.* , Global quantification of mammalian gene expression control.*Nature* 2011, *473* , 337-342.

[10] He, B., Shi, J., Wang, X., Jiang, H., Zhu, H. J., Label-free absolute protein quantification with data-independent acquisition.*J Proteomics* 2019, *200* , 51-59.

[11] Arike, L., Peil, L., Spectral counting label-free proteomics.*Methods Mol Biol* 2014, *1156* , 213-222.

[12] Colaert, N., Vandekerckhove, J., Gevaert, K., Martens, L., A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision. *Proteomics* 2011,*11* , 1110-1113.

[13] Ishihama, Y., Oda, Y., Tabata, T., Sato, T.*, et al.* , Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 2005, *4* , 1265-1272.

[14] Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S.*, et al.* , Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors.*Proc Natl Acad Sci U S A* 2006, *103* , 18928-18933.

[15] Griffin, N. M., Yu, J., Long, F., Oh, P.*, et al.* , Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol* 2010, *28* , 83-89.

[16] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G.*, et al.* , Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 2005,*4* , 1487-1502.

[17] Chen, Y. Y., Chambers, M. C., Li, M., Ham, A. J.*, et al.* , IDPQuantify: combining precursor intensity with spectral counts for protein and peptide quantification. *J Proteome Res* 2013,*12* , 4111-4121.

[18] Ahrne, E., Martinez-Segura, A., Syed, A. P., Vina-Vilaseca, A.*, et al.* , Exploiting the multiplexing capabilities of tandem mass tags for high-throughput estimation of cellular protein abundances by mass spectrometry. *Methods* 2015, *85* , 100-107.

[19] O'Connell, J. D., Paulo, J. A., O'Brien, J. J., Gygi, S. P., Proteome-Wide Evaluation of Two Common Protein Quantification Methods.*J Proteome Res* 2018, *17* , 1934-1942.

[20] Shin, J. B., Krey, J. F., Hassan, A., Metlagel, Z.*, et al.* , Molecular architecture of the chick vestibular hair bundle.*Nat Neurosci* 2013, *16* , 365-374.

[21] Jarnuczak, A. F., Najgebauer, H., Barzine, M., Kundu, D. J.*, et al.* , An integrated landscape of protein expression in human cancer. *Sci Data* 2021, *8* , 115.

[22] Dai, C., Fullgrabe, A., Pfeuffer, J., Solovyeva, E. M.*, et al.* , A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat Commun* 2021, *12* , 5854.

[23] Krey, J. F., Wilmarth, P. A., Shin, J. B., Klimek, J.*, et al.* , Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. *J Proteome Res* 2014,*13* , 1034-1044.

[24] Wang, Z., Karkossa, I., Grosskopf, H., Rolle-Kampczyk, U.*, et al.* , Comparison of quantitation methods in proteomics to define relevant toxicological information on AhR activation of HepG2 cells by BaP. *Toxicology* 2021, *448* , 152652.

[25] Betancourt, L. H., Gil, J., Sanchez, A., Doma, V.*, et al.* , The Human Melanoma Proteome Atlas-Complementing the melanoma transcriptome. *Clin Transl Med* 2021, *11* , e451.

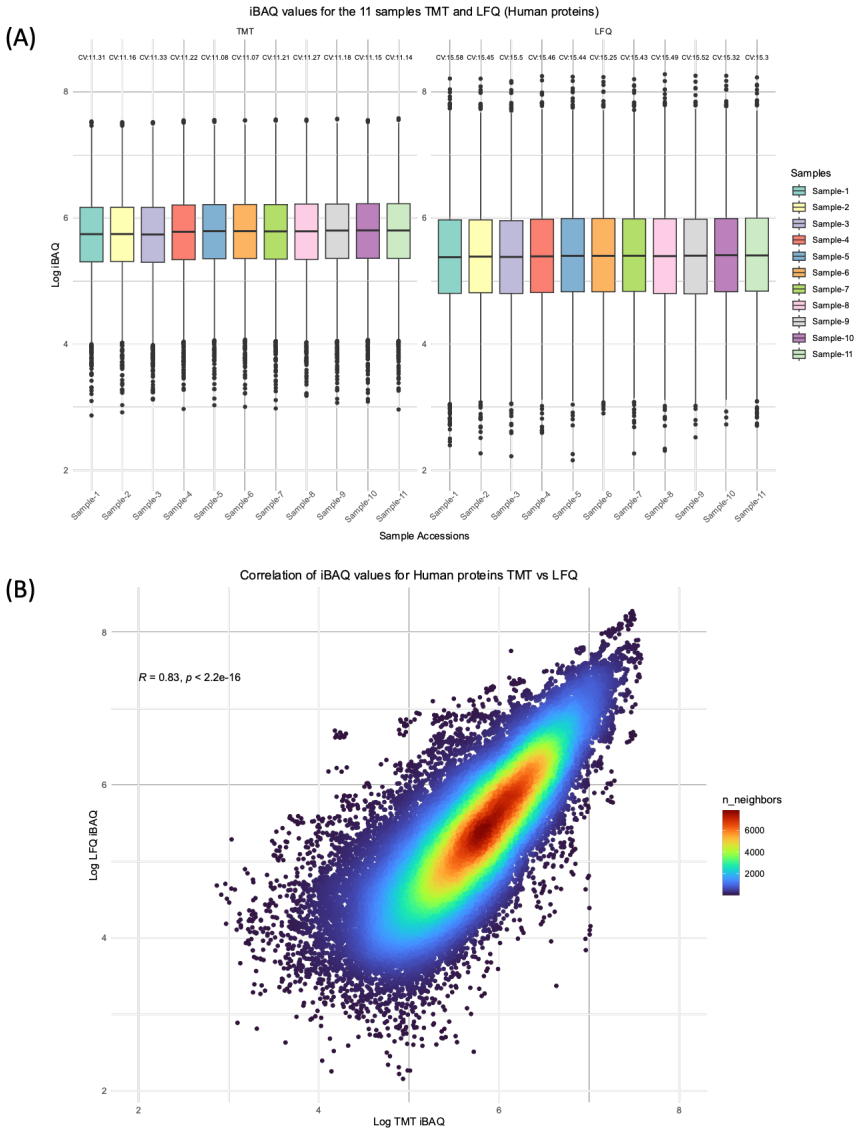

Figure 1 : (A) Boxplot of iBAQ Log-transformed for the 11 samples dataset PXD007683, for both TMT and LFQ approaches. (B) Correlation between iBAQ values for all quantified proteins between the TMT and LFQ approaches, for dataset PXD007683.

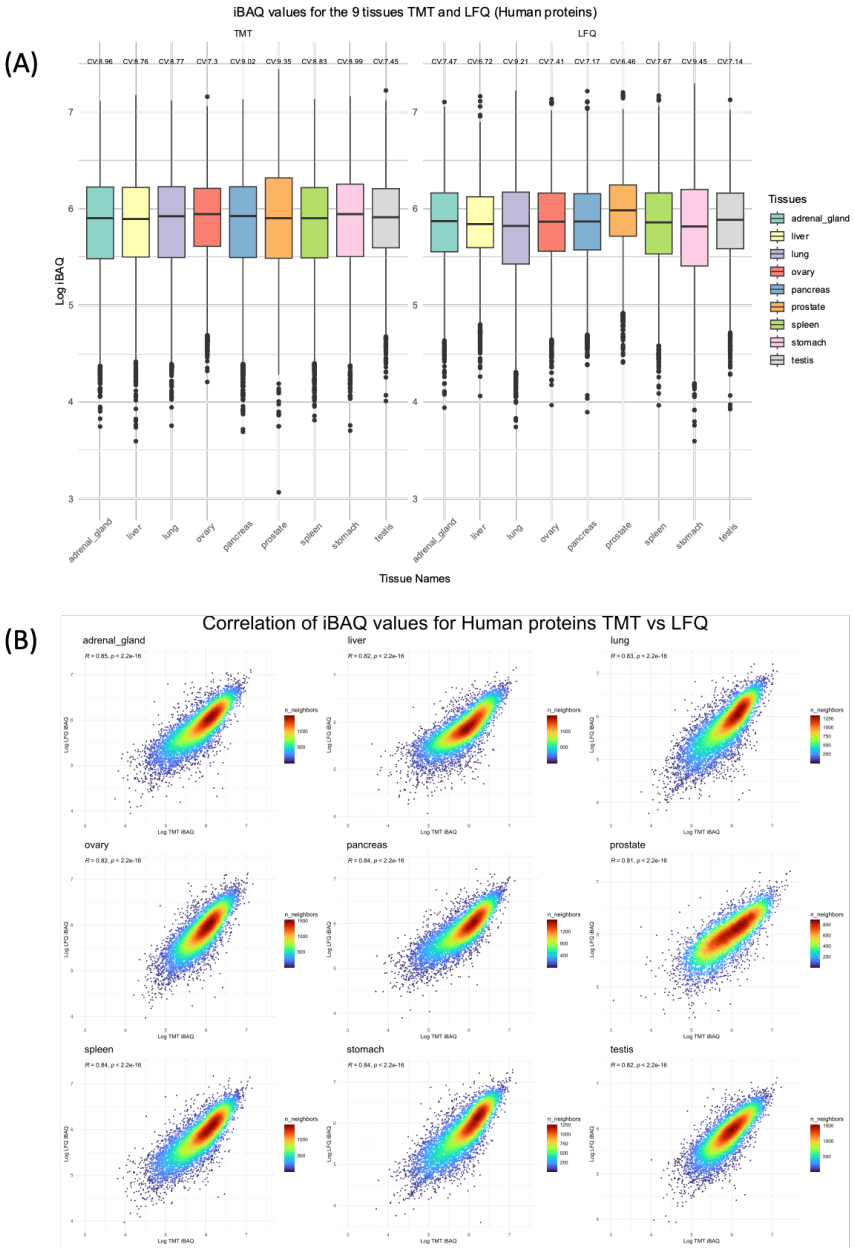

Figure 2 : (A) Boxplot of iBAQ log-transformed for all tissues shared between datasets PXD016999 and PXD010154. (B) Correlation between iBAQ values for all quantified proteins between PXD016999 and PXD010154 datasets.