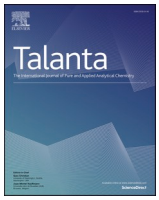
Talanta 266 (2024) 124953

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Normalization methods in mass spectrometry-based analytical proteomics: A case study based on renal cell carcinoma datasets

Luis B. Carvalho [a,b,1], Pedro A.D. Teigas-Campos [a,b,1], Susana Jorge [a,b], Michele Protti [c], Laura Mercolini [c], Rajiv Dhir [d], Jacek R. Wiśniewski [e], Carlos Lodeiro [a,b], Hugo M. Santos [a,b,d,**], José L. Capelo [a,b,*]

[a] BIOSCOPE Group, LAQV-REQUIMTE, Chemistry Department, NOVA School of Science and Technology, FCT NOVA, Universidade NOVA de Lisboa, 2829-516, Caparica, Portugal
[b] PROTEOMASS Scientific Society, Madan Park, 2829-516, Caparica, Portugal
[c] Research Group of Pharmaco-Toxicological Analysis (PTA Lab), Department of Pharmacy and Biotechnology (FaBiT), Alma Mater Studiorum - University of Bologna, Via Belmeloro 6, 40126, Bologna, Italy
[d] Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
[e] Biochemical Proteomics Group, Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Martinsried, Germany

ARTICLE INFO

Keywords:
Normalization methods
Mass spectrometry
Proteomics
Renal carcinoma

ABSTRACT

Normalization is a crucial step in proteomics data analysis as it enables data adjustment and enhances comparability between datasets by minimizing multiple sources of variability, such as sampling, sample handling, storage, treatment, and mass spectrometry measurements. In this study, we investigated different normalization methods, including Z-score normalization, median divide normalization, and quantile normalization, to evaluate their performance using a case study based on renal cell carcinoma datasets. Our results demonstrate that when comparing datasets by pairs, both the Z-score and quantile normalization methods consistently provide better results in terms of the number of proteins identified and quantified as well as in identifying statistically significant up or down-regulated proteins. However, when three or more datasets are compared at the same time the differences are found to be negligible.

1. Introduction

Some analytical chemists encounter significant challenges in analyzing large-scale proteomics data using bioinformatics tools, often because these topics are not sufficiently developed in undergraduate or post-graduate programs. As a result, analytical chemists may need to acquire additional skills and knowledge to effectively handle these types of data. From an analytical chemistry point of view, the so-called shot-gun-based proteomics workflows are characterized by multiple sources of variability, including sampling, sample handling, sample storage, sample treatment and mass spectrometry measurements [1,2]. For instance, analytical chemists know that large-scale and longitudinal studies often encounter a source of variation associated with the difference in storage time and ageing of the samples [2,3]. Samples are typically collected and prepared at different times during the study, introducing an environmental factor that may affect the data. This factor may be compounded by other variables, such as the collection of samples in different centres or countries, which may be subjected to varying numbers of freeze-thaw cycles. In addition, the sample collection protocol may be applied with different levels of attention by different operators, which can lead to unwanted differences in the samples. Furthermore, transporting the samples to the laboratory where they will be stored or analysed can also introduce noise into the data, as the transportation conditions may vary, potentially altering the samples differently. Finally, in a typical shot-gun proteomics workflow, samples are collected and treated in the analytical laboratory before being stored

* Corresponding author. BIOSCOPE Group, LAQV-REQUIMTE, Chemistry Department, NOVA School of Science and Technology, FCT NOVA, Universidade NOVA de Lisboa, 2829-516, Caparica, Portugal.
** Corresponding author. BIOSCOPE Group, LAQV-REQUIMTE, Chemistry Department, NOVA School of Science and Technology, FCT NOVA, Universidade NOVA de Lisboa, 2829-516, Caparica, Portugal.
E-mail addresses: hmsantos@fct.unl.pt (H.M. Santos), jlcm@fct.unl.pt (J.L. Capelo).
[1] These two authors have equally worked for this manuscript.

https://doi.org/10.1016/j.talanta.2023.124953
Received 17 May 2023; Received in revised form 7 July 2023; Accepted 14 July 2023
Available online 17 July 2023

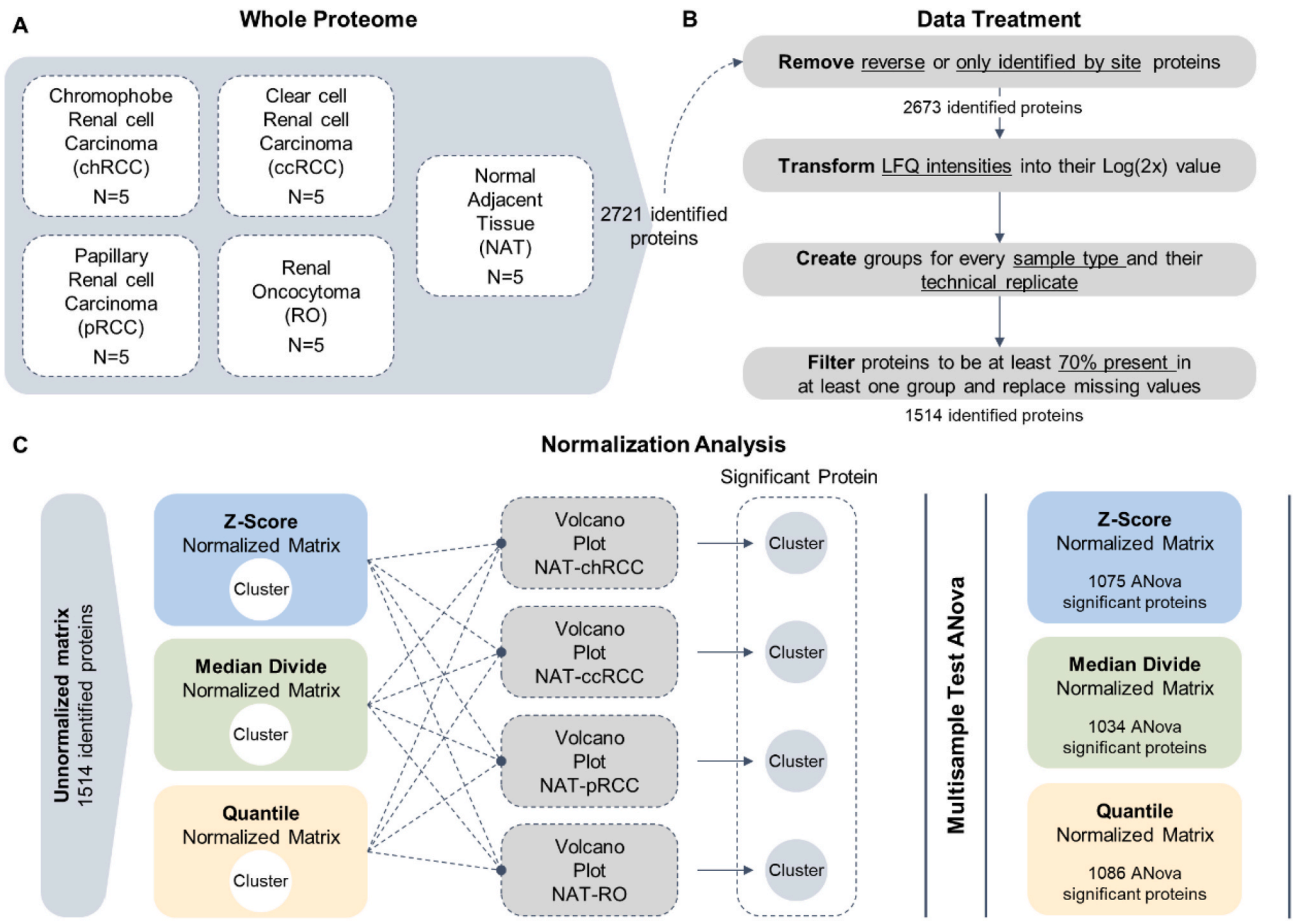


Fig. 1. Schematic representation of the data process applied to the proteomic data when comparing Neat Adjacent Tissue, NAT (control group) against multiple renal tumour types. **A:** Type and number of tumours (description of patients is given in Table SM1). **B:** Data treatment for the proteins quantified using the five datasets. **C:** Types of normalization made: Z-Score, median divide, Quantile and types of comparisons done: Volcano plots, Cluster analysis and Multisample Test Anova.

or investigated using mass spectrometry. The samples are then cleaved with various enzymes depending on the experimental strategy, and the resulting cleavage proteome composed of thousands of peptides is subjected to chromatography-mass spectrometry analysis. The noise generated during the pipeline of work, from sampling to mass spectrometry, makes shot-gun proteomics challenging to compare samples, extract meaningful information and derive accurate conclusions. The ultimate tools to overcome this challenge of variability in shotgun proteomics workflows are the so-called normalization methods [4–6].

Normalization is a method used to adjust the data obtained from different samples, platforms, or batches to remove unwanted variability and enhance comparability between the datasets. It involves applying a mathematical correction to the data to reduce the effects of systematic technical variation and other sources of bias. The choice of normalization method depends on the nature of the data and the research question at hand [7–9]. However, this choice is not as simple as it seems at first glance. The importance of selecting an appropriate normalization method is critical to obtain well-grounded results. Therefore, selecting the optimal way to normalize data has important implications, as it significantly impacts downstream analyses, delivering different results depending on the normalization method selected [1,10,11]. Although several studies have evaluated different proteome normalization approaches, primarily focusing on their ability to enhance data repeatability, there is no consensus about which normalization method must be used, or whether the use is conditioned by the number of groups or the number of samples in each group [10,12–15].

There are several normalization methods, but the most popular ones include Z-score normalization, median divide normalization, and quantile normalization.

The z-score normalization method creates an induced normal matrix by subtracting to each protein in each sample their respective median across all samples in the dataset and then dividing each protein value by the standard deviation of the respective protein across all samples. This process forces the dataset to follow a normal distribution, which can improve the accuracy of downstream statistical analyses. By normalizing the data in this way, the z-score normalization method can address variations in the data, making it a valuable tool for data normalization in proteomic studies [10].

The median divide normalization approach involves dividing all results by the median, which reduces the impact of bias caused by equipment variations and makes the data more comparable to other assays. This method is particularly useful when dealing with large datasets that are collected over time and across different batches, as it reduces the variation in the data introduced by the technical aspects of the assay. The median divide normalization method is a reliable approach for comparing protein expression levels between different samples and provides an accurate representation of the overall distribution of the data. However, this method may not be appropriate for datasets with extreme values or when most values are close to zero [7].

The quantile normalization involves substituting the value of each data point in a sample with the mean of the corresponding quantile, thereby modifying the data distribution of each sample to be identical. This normalization method is particularly useful for comparing samples with different distributions. By standardizing the distribution of each sample, quantile normalization can improve the accuracy and reproducibility of downstream analyses, such as differential expression

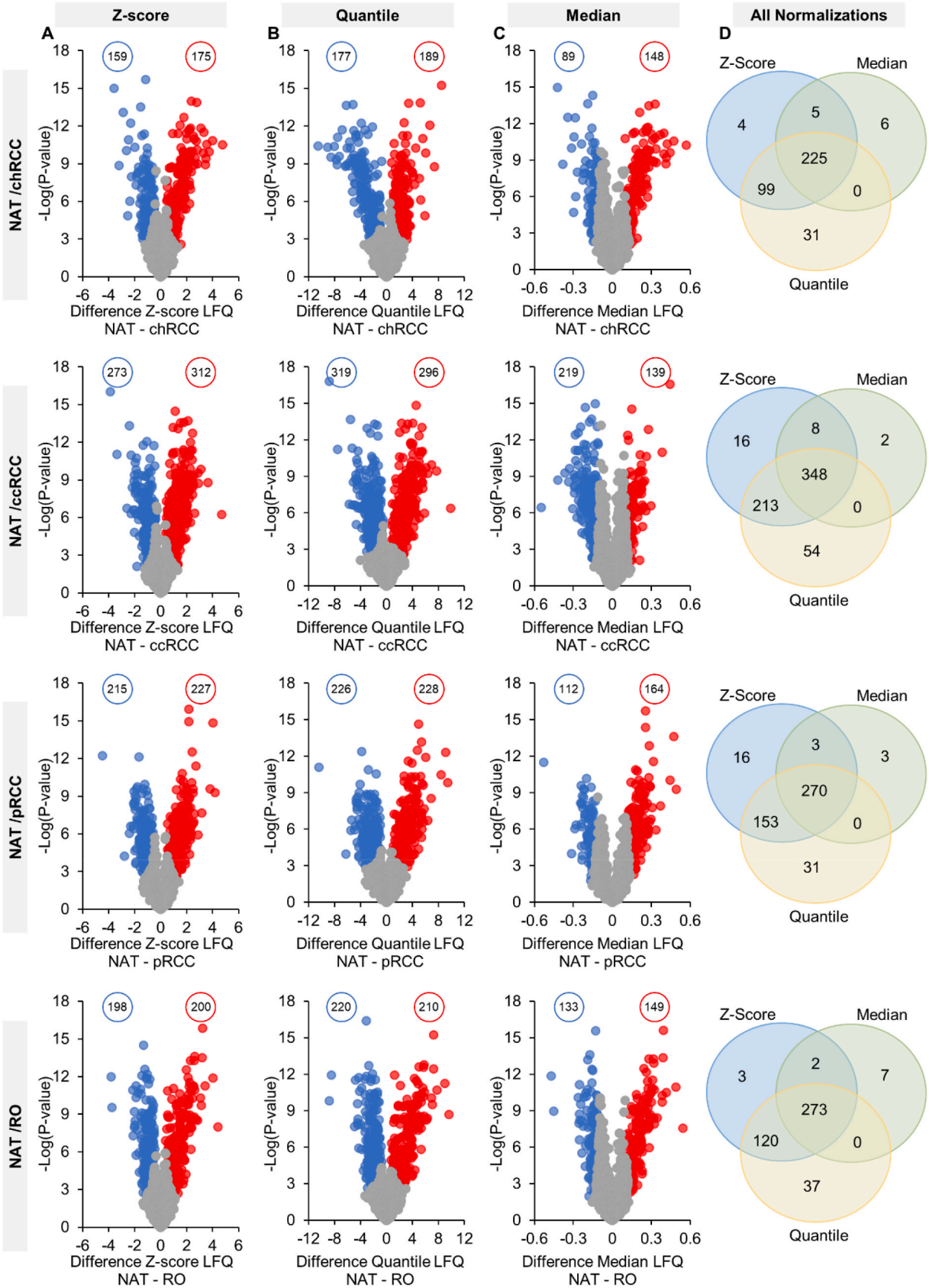


Fig. 2. Comparison of the numbers of proteins found statistically differentially expressed using each normalization method when comparing the near adjacent tissue (NAT) against each renal tumour type. Each column represents the volcano plots obtained using one normalization method: **A:** Z-score; **B:** Quantile; **C:** Median. Each row shows the comparison by pairs (Volcano plot, according to Student's *t*-test, FDR 0.05 and S_0 of 0.1, red dots: overexpressed proteins in NAT; blue dots over-expressed proteins in tumours). chRCC: chromophobe renal cell carcinoma; ccRCC: clear cell renal cell carcinoma; pRCC papillary renal cell carcinoma; RO: Renal oncocytoma. **D:** column summarizes the common and uncommon proteins found among pair comparisons for each normalization method.

analysis and pathway analysis [6,16,17].

In a previous work [18], we applied Z-Score normalization and the Total Protein Approach (TPA) to proteomic data derived from renal carcinoma solid biopsies. Thus, near adjacent tissue (NAT), Papillary Renal Cell Carcinoma (pRCC), Clear Cell Renal Carcinoma (ccRCC), Chromophobe Renal Cell Carcinoma (chRCC) and Renal Oncocytoma (RO), samples were interrogated with the main aim of obtaining new immunohistochemical markers for diagnosis and prognosis. The best biomarkers found via mass spectrometry, namely perilipin-2 (PLIN2), beta-tubulin III (TUBB3), lysosomal-associated membrane protein-1 (LAMP1) and hexokinase-1 (HK1) were later successfully validated using tissue microarrays.

In the present work, the complete mass spectrometry datasets obtained for the different types of renal carcinomas and NAT are interrogated using bioinformatic pipelines differing only in the selected normalization methods, namely Z-score, median divide and quantal normalization. Selecting these specific tumour types allows for a comprehensive analytical assessment of the normalization methods in a clinically relevant context. Through this evaluation, we aim to provide insights into the selection process of appropriate normalization techniques that can enhance the accuracy and reproducibility of proteomic analyses in tumour research. The results obtained were compared with those previously validated using tissue microarrays.

2. Methods

2.1. Patients and sample collection

The experimental workflow developed to interrogate the renal carcinoma and the NAT tissue samples, along with the extended study design and patient sampling collection is described elsewhere [19], furthermore the validation of the biomarkers found using tissue microarrays is also described elsewhere [18] A brief patient summary is provided in Supplementary material Table SM1.

2.2. Mass spectrometry raw data

This study used frozen tissue biopsies of human renal tissues from four different tumour types (ccRCC, chRCC, pRCC, and RO) and a control group (NAT). The samples were properly treated and analysed by mass spectrometry to identify and quantify the protein expression levels [19]. Mass spectrometry raw data was accessed in ProteomeXchange Consortium [20] using the Proteomics IDentications Database (PRIDE) [21] with the unique identifier PXD023296.

2.3. Identification and quantification of proteins

Protein identification and relative label-free quantification (LFQ) was newly performed used the previous described raw data using the precursor signal intensity method and delayed normalization, MaxLFQ, on MaxQuant software V2.0.3.0 [12]. The newly generated proteins group table containing protein identification and quantification can be found in Supplementary material 2. All raw files were processed in a single run with default parameters [22,23]. Database searches were performed using peptide search engine Andromeda against the human UniProt UP000005640_9606 and UniProt UP000005640_9606_additional database [24]. Searches were configured with cysteine carbamidomethylation as a fixed modification and *N*-terminal acetylation and methionine oxidation as variable modifications. We set the false discovery rate (FDR) to 0.01 for protein and peptide levels with a minimum length of seven amino acids. The FDR was determined by searching a reverse database. Trypsin enzyme specificity was set as *C*-terminal to arginine and lysine with a maximum allowance of two missed cleavages.

2.4. Bioinformatics data processing

Bioinformatic data processing was conducted using Perseus V1.6.0 [25,26] as explained in Fig. 1B. Briefly, the data processing can be roughly divided into stages: (i) Elimination of protein groups that were reversely identified or only identified by site from the list of proteins; (ii) Annotation of sample condition and technical replicates; (iii) Transformation of LFQ intensities into Log (2x) value to lessen the impact of outliers and to understand the protein changes proportional across the conditions; and finally (iv) Proteins were filtered so that they were at least present in 70% of the samples in at least one of the created groups and the remaining missing values were imputed according with the respective sample normal distribution with a width of 0.3 and down shift of 1.8.

2.5. Bioinformatic data analysis and normalization comparisons

Bioinformatic data analysis was carried out in Perseus V1.6.0 [25,26] as explained in Fig. 1C in which the analytical comparison of three different types of commonly used normalizations was conducted (Z-Score normalization, Median divide normalization and Quantile normalization). Z-Score and Median divide normalization were obtained using the standard tools in Perseus while Quantile normalization was acquired by using the R package limma add-on for Perseus [27,28]. To access the impact of normalization, the NAT (control group) data set of protein expression was compared with the renal tumours using the three different types of normalization. The number of proteins whose levels were found to be statistically different between the group comparison were accessed by using a volcano plot analysis (Student's *t*-test, FDR 0.05 and S_0 of 0.1). Finally, all the datasets were compared together for each type of normalization using a Multisample Test ANOVA with a FDR of 0.05. All comparison resulting in statistically different proteins were used respectively in a hierarchical clustering analysis with an average linkage, no constraint, pre-processing with k-means, and Euclidean distance. The proteomic data was also processed (i) by comparing the NAT group of samples against two, three or four tumour types at the same time, as shown in Fig. 1 and (ii) by comparing each tumour condition and the control group at a time, as shown in Supplementary material Fig. SM1. This approach was adopted to ensure that the results would be consistent across different subtypes of renal tumour and would be useful for researchers focusing on precision medicine.

These analytical approaches allowed for a comprehensive evaluation of the normalization methods in terms of their ability to accurately identify significant proteins and their impact on downstream analyses, such as differential expression analysis and pathway analysis. Enrichment pathways analysis was carried out using the software platform Cytoscape v3.9.1 [29] and the application StringApp [30] v2.0.0 for protein-protein analysis and biochemical pathway analysis. Different types of databases for enrichment pathway analysis (GO biological process [31], The Kyoto Encyclopedia of Genes and Genomes - KEGG [32], Reactome pathways [33]) were considered to access the impact of normalization on downstream analysis.

3. Results

The data pre-processing steps were standardized across all normalization methods. A detailed workflow of the pre-processing steps is illustrated in Fig. 1 and Supplementary material Fig. SM1.

3.1. Comparing each type of renal carcinoma against NAT

After comparing the datasets in pairs, we tested three normalization methods. Our findings revealed that both z-score normalization and quantile normalization produced similar results in terms of the number of statistically significant differentially expressed proteins throughout all comparisons, as shown in Fig. 2A–D. However, quantile

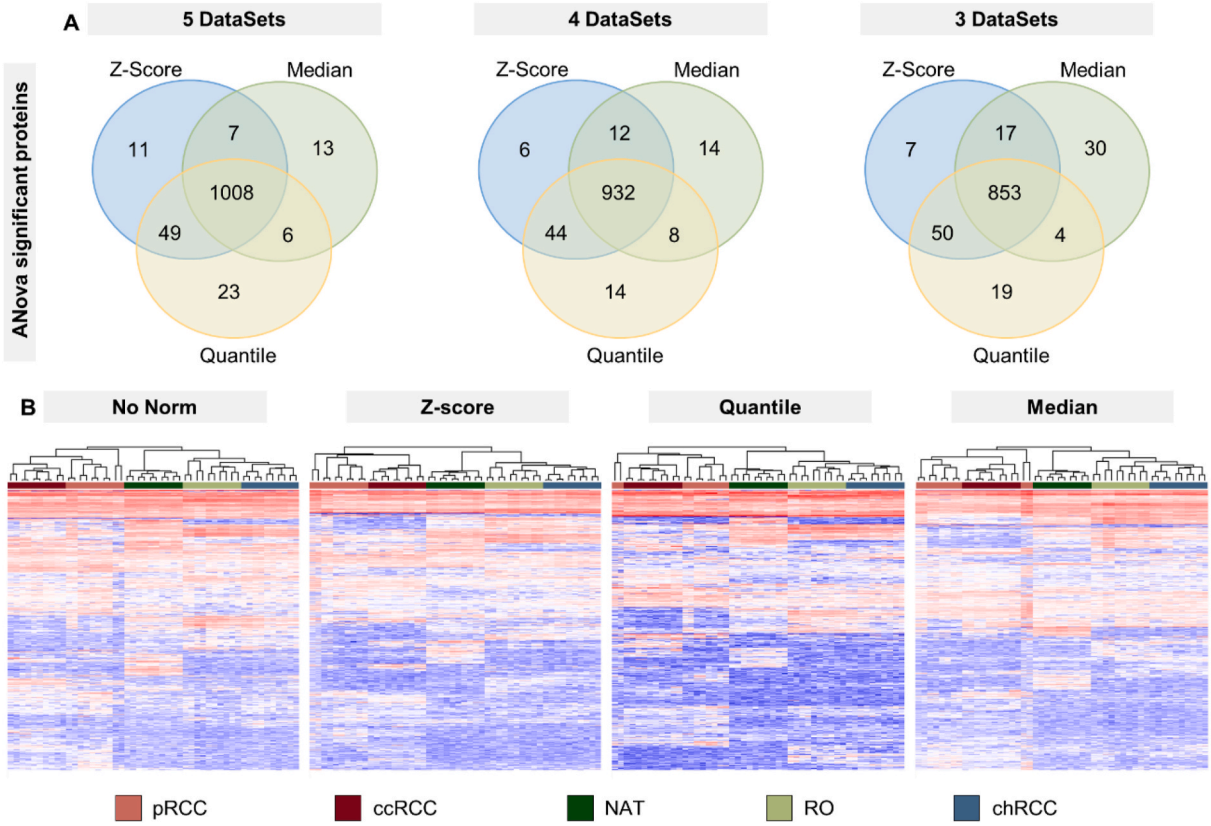


Fig. 3. Proteins found statistically differentially expressed using the comparison of: **A:** 5 datasets: All tumours against NAT; 4 datasets chRCC, ccRCC and RO tumours against NAT and 3 datasets: ccRCC and RO against NAT. B: Hierarchical clusters were performed by using proteins that were found to be statistically different using each normalization method when comparing all tumours against NAT (average linkage, no constraint, pre-processing with k-means, and Euclidean distance, n = 25).

normalization consistently identified a slightly higher number of proteins, depending on the carcinoma compared against the near adjacent tissue, NAT. On the other hand, the median divide method consistently produced about 37% fewer differentially expressed proteins for all the pairs compared, as depicted in Fig. 2. This trend was also observed for the down- and upregulated proteins as shown in Fig. SM2.

In a previous study utilizing the same raw data, our team employed mass spectrometry to identify four potential biomarkers: PLIN2, TUBB3, LAMP1, and HK1. The Z-score and Total Protein Approach (TPA) were utilized for this purpose, and these markers are depicted in Fig. SM2. TPA is a label-free mass spectrometry method developed by Wiśniewski et al. [34,35] that enables the measurement of absolute expression levels of numerous proteins without the need for standards.

Using TPA, we were able to highlight these potential biomarkers, and their validity was subsequently confirmed through immunohistochemical tissue arrays [18]. Interestingly, in our current study, we observed statistically significant differential expression for the same proteins across all three normalization methods employed, with similar p-values. This suggests that when a protein exhibits high significance, it will be consistently identified as significant regardless of the specific normalization method used.

These findings emphasize the robustness and reliability of the identified biomarkers and support the notion that they are biologically relevant. It further reinforces the notion that the choice of normalization method does not significantly impact the detection of highly significant proteins.

Furthermore, we compared the top 100 proteins with the highest log *P*-values across the three normalization methods in all two-group comparisons. We found that the z-score and quantile normalization methods had 94 proteins in common, while the median method shared 86 and 84 proteins with the z-score and the quantile methods, respectively. These results suggest that the z-score and quantile normalization methods may be more consistent with each other in identifying highly differentially expressed proteins, while the median method may identify a slightly different set of proteins.

3.2. Comparison of the two, three and four renal carcinomas at a time against NAT

Our next step was to compare the tumours datasets in groups of 2, 3 and 4 against NAT. Proteins for further processing were selected if they were found in any dataset in at least 70% of the replicates.

Focusing on the results obtained comparing the four neoplasia's datasets against NAT (Fig. 3A) it can be concluded that when the comparison is not made in pairs, the numbers of proteins found to be statistically different is almost the same (i.e., 1075, 1086 and 1034 for the Z-score, the quantile and the median divide, respectively). In addition, 1008 of all the proteins were identified with any normalization method. The same patterns were repeated comparing three or two tumours against NAT, Fig. 3A respectively.

These results suggest that the choice of normalization method did not have a substantial impact on the number of differentially expressed proteins identified when datasets are not compared in pairs, as explained in the previous section. This is due to the fact that the MaxQuant statistics varied differently, from comparing pairs to comparing three or more datasets, so the differences were not on the same scale as those observed when comparing datasets in pairs, as shown in the previous section.

To cluster the samples, we exploited the differentially expressed proteins identified by each normalization method using the five datasets (Supplementary Material 3). As noted in Fig. 3B, we found that three distinct clusters were formed regardless of the normalization method.

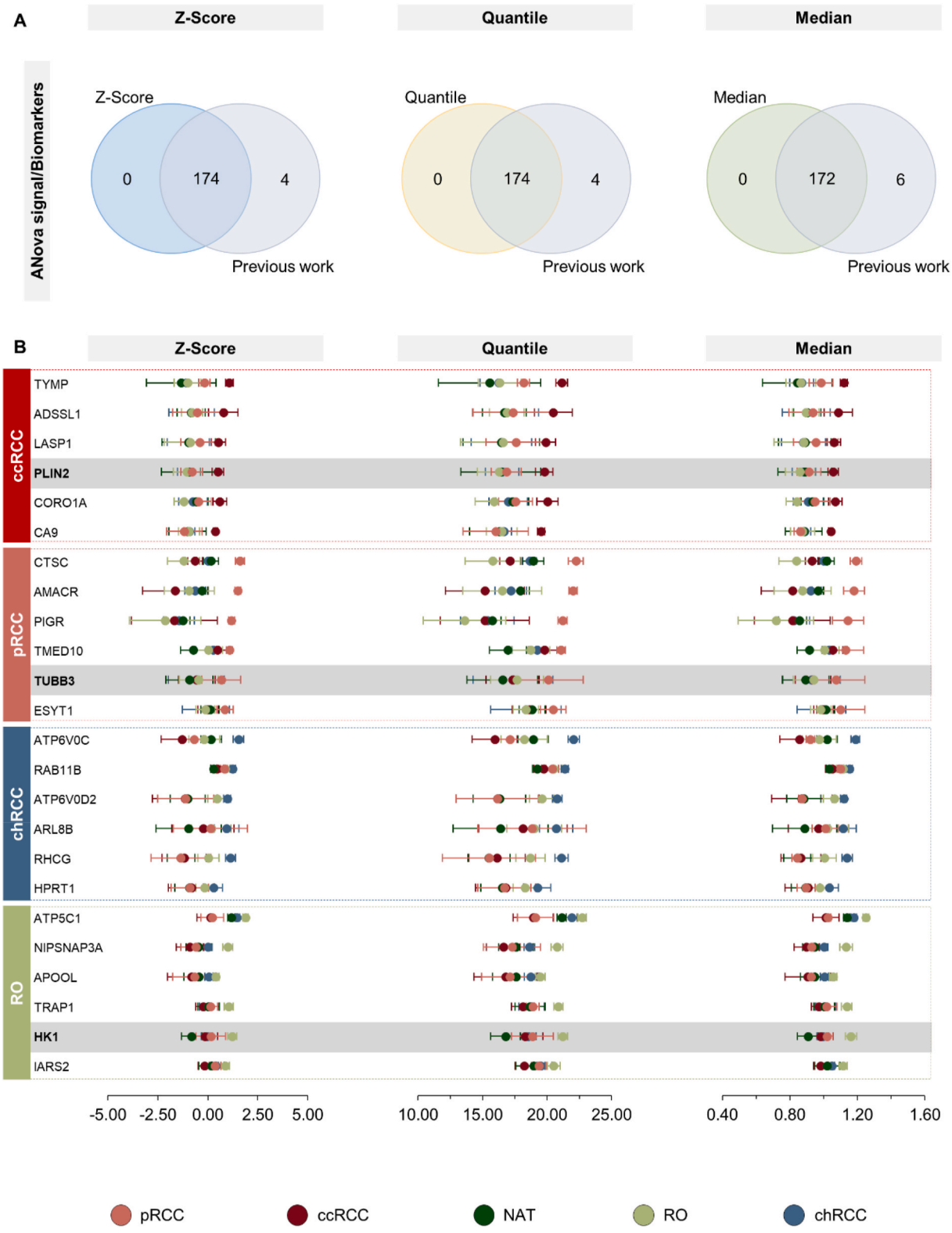
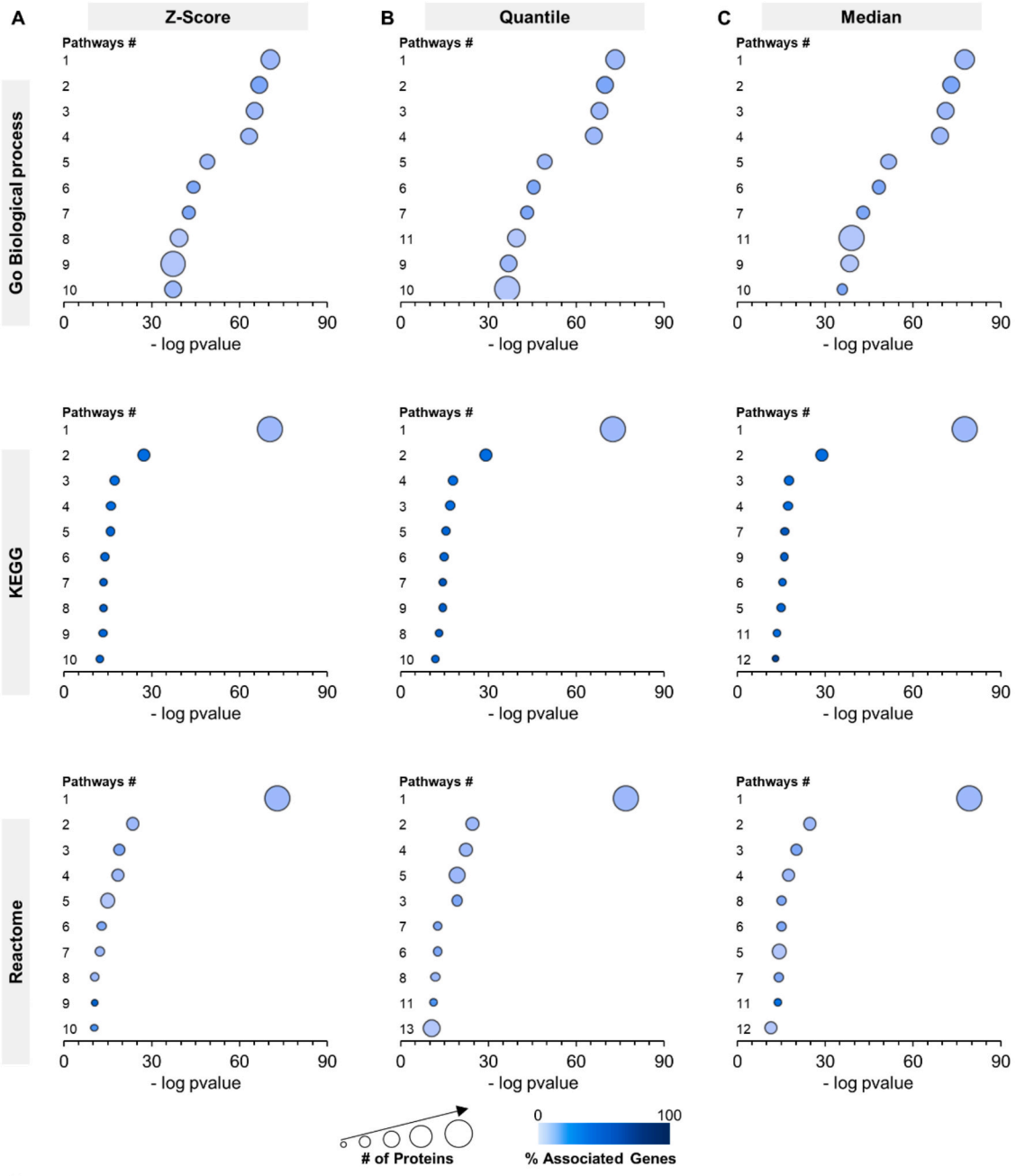


Fig. 4. Comparison of candidate biomarkers for each renal tumour subtype across all different type of data normalization. **A:** Biomarkers found in this work for each normalization method versus those found in our previous work [18]. **B:** Biomarkers proposed in our previous work (Fig. SM2) found with each normalization method in the present work. Note that for all of them the standard deviations of the normalized LFQ values overlap.

The tumours pCRR and ccRCC were grouped together in the first cluster, while RO and chRCC formed a distinct cluster in the second. The third one was NAT tissue. Notably, we observed that the proteomic expressions of RO and chRCC tumours were more similar to NAT than to pRCC and ccRCC.

When the differentially expressed proteins were analysed to extract those whose levels make them unique for just one type of carcinoma, the number of proteins was 178, regardless of the normalization process, as depicted in Fig. 4A. Furthermore, we also found that 174 out of these 178 biomarkers were present in our previous work [18], including the



D

Go Biological Pathway #		KEGG Pathway #		Reactome Pathway #	
#	Name	#	Name	#	Name
1	Small molecule metabolic process	1	Metabolic pathways	1	Metabolism
2	Carboxylic acid metabolic process	2	Carbon metabolism	2	Metabolism of amino acids and derivatives
3	Organic acid metabolic process	3	Biosynthesis of amino acids	3	Biological oxidations
4	Oxoacid metabolic process	4	Glycolysis / Gluconeogenesis	4	Neutrophil degranulation
5	Oxidation-reduction process	5	Valine, leucine and isoleucine degradation	5	Innate Immune System
6	Small molecule catabolic process	6	Fatty acid degradation	6	Protein localization
7	Monocarboxylic acid metabolic process	7	Citrate cycle (TCA cycle)	7	Fatty acid metabolism
8	Catabolic process	8	beta-Alanine metabolism	8	The citric acid (TCA) cycle and respiratory electron transport
9	Metabolic process	9	Pyruvate metabolism	9	Insulin receptor recycling
10	Organic substance catabolic process	10	Glycine, serine and threonine metabolism	10	Iron uptake and transport
11	Carboxylic acid catabolic process	11	Arginine and proline metabolism	11	Pyruvate metabolism and Citric Acid (TCA) cycle
		12	2-Oxocarboxylic acid metabolism	12	Metabolism of lipids
				13	Immune System

Fig. 5. Comparison of Top-10 pathways differentially expressed when chRCC is compared against NAT as per the proteins differentially expressed found with each normalization method. **A:** Z-score; **B:** Quantile; **C:** Median. Databases used for pathway enrichment analysis were Go biological processes, KEGG and Reactome.

four biomarkers validated by us using immunohistochemistry. These results suggest that the choice of normalization method has a negligible impact on the identification of specific biomarkers.

3.3. Comparison of the best biomarkers for immunohistochemistry

In a previous work we recommended several proteins as new potential biomarkers for immunohistochemistry (Supplementary material Fig. SM3). These proteins were first identified using mass spectrometry (LFQs) and then with TPA. Of these proteins, those represented in grey in Fig. 4B were selected in the previous work because the standard deviations did not overlap. These proteins were later validated against tissue arrays [18].

When we used the levels obtained through each normalization method for each protein to identify statistically significant proteins, we obtained two interesting results. Firstly, the three normalization methods delivered almost the same results, but they differed substantially from those obtained using TPA as described above. As shown in Fig. 4B, proteins PLIN2, TUBB3, and HK1 would have never been identified as the best potential biomarkers because the standard deviations overlap each other (Fig. 4). This unexpected result highlights the need to re-evaluate how data is treated and the importance of TPA in biomarker discovery.

3.4. Exploring the impact of normalization methods on identifying main pathways affected in renal tumours

To determine whether the identified dysregulated pathways changed as a function of the selected normalization method, we used the set of proteins found to be dysregulated when the five different datasets were compared. Then, through this data set, using an ANOVA test, we compared the proteins differentially expressed between NAT and chRCC.

Fig. 5 shows the 10 most important (as per p-value) dysregulated pathways for (i) GO biological process, (ii) KEGG, and (iii) Reactome pathways. The results show that for the GO biological processes the first 7 pathways affected are the same, with no variations nor in the p-value, neither in the number of proteins involved, thus showing no differences in the normalization methods. For the KEGG, the Z-score and the quantile normalizations were found more consistent between them than when compared to the median, as they have the same top-10 pathways. For the Reactome pathways, the three normalization methods were found to provide similar results.

4. Conclusions

When comparing two groups of datasets, such as healthy versus diseased samples, the Z-score normalization and the quantile methods tend to yield a higher number of differentially expressed proteins compared to the median normalization method. Thus, in these cases, we recommend using the Z-score normalization method. However, when comparing a larger number of datasets groups (3, 4, or 5), the differences in terms of proteins and biochemical pathways statistically different among the three normalization methods are negligible, and any can be selected. Interestingly, our findings show that four biomarkers, previously identified using the z-score normalization (LFQs) in combination with the TPA approach, and later validated with immunohistochemistry [18], could not have been easily highlighted using any of the normalization methods assessed in this study. This discovery opens new avenues of research in biomarker discovery as it emphasises the TPA approach as a powerful tool in biomarker discovery.

Author contributions

J.L.C., H.M.S., and L.B.C. designed the bioinformatics pipelines. L.B. C. and P.A.D.T.C. performed the bioinformatics analysis. J.L.C drafted the first manuscript. P.A.D.T.C., H. M. S., L.B.C., R.D., J.W., C.L., L.M. and M.P. read the final draft, corrected it and made valuable suggestions. R.D. provided the solid biopsies, and J.W. helped with the total protein approach. S.J. made the proteomics sample treatment and H.M. S. acquire the raw LC-MS/MS data.

Funding

PROTEOMASS Scientific Society, #PM001/2019, #PM003/2016. Fundacao para a Ciencia e a Tecnologia (FCT/MCTES), UIDB/50006/2020, UIDP/50006/2020, LA/P/0008/2020, SFRH/BD/144222/2019, SFRH/BD/120537/2016. University of Pittsburgh Hillman Cancer Center shared resource facilities (Cancer Genomics Facility and The Health Science Tissue Bank) supported in part by award P30CA047904.

Availability of data and materials

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [21] partner repository with the dataset identifier PXD023296.

Declaration of competing interest

All authors declare nor financial or *personal relationships that may be perceived as influencing this work*.

Data availability

Data is submitted n Pride. We providfe the link in the text.

Acknowledgements

PROTEOMASS Scientific Society is acknowledged by the funding provided to the Laboratory for Biological Mass Spectrometry Isabel Moura (#PM001/2019 and #PM003/2016). Authors acknowledge the funding provided by the Associate Laboratory for Green Chemistry - LAQV which is financed by national funds from FCT/ MCTES, *Fundação para a Ciência e a Tecnologia* and *Ministério da Ciência, Tecnologia e Ensino Superior, through* the projects UIDB/50006/2020 and UIDP/50006/202). H. M. S. acknowledges LAQV (LA/P/0008/2020) funded by FCT/MCTES for his research contract. L. B. C. thanks the FCT/MEC for the FCT PhD grant 2019 (SFRH/BD/144222/2019). S. J. thanks FCT/MEC (Portugal) for her PhD grant reference SFRH/BD/120537/2016. This project utilized the University of Pittsburgh Hillman Cancer Center shared resource facilities (Cancer Genomics Facility and The Health Science Tissue Bank) supported in part by award P30CA047904 (Dr.R. Dhir).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2023.124953.

References

[1] E. Dubois, A.N. Galindo, L. Dayon, O. Cominetti, Assessing normalization methods in mass spectrometry-based proteome profiling of clinical samples, Biosystems 215–216 (2022), 104661, https://doi.org/10.1016/j.biosystems.2022.104661.

[2] B.J.A. Mertens, Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies, in: Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry, Springer International Publishing, Cham, 2017, pp. 1–21, https://doi.org/10.1007/978-3-319-45809-0_1.

[3] A.M. De Livera, M. Sysi-Aho, L. Jacob, J.A. Gagnon-Bartsch, S. Castillo, J. A. Simpson, T.P. Speed, Statistical methods for handling unwanted variation in metabolomics data, Anal. Chem. 87 (2015) 3606–3615, https://doi.org/10.1021/ac502439y.

[4] A. Chawade, E. Alexandersson, F. Levander, Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets, J. Proteome Res. 13 (2014) 3114–3120, https://doi.org/10.1021/pr401264n.
[5] Y. V Karpievitch, A.R. Dabney, R.D. Smith, Normalization and missing value imputation for label-free LC-MS analysis, BMC Bioinf. 13 (2012) S5, https://doi.org/10.1186/1471-2105-13-S16-S5.
[6] B.M. Bolstad, R.A. Irizarry, M. Åstrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193, https://doi.org/10.1093/bioinformatics/19.2.185.
[7] S.J. Callister, R.C. Barry, J.N. Adkins, E.T. Johnson, W. Qian, B.-J.M. Webb-Robertson, R.D. Smith, M.S. Lipton, Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics, J. Proteome Res. 5 (2006) 277–286, https://doi.org/10.1021/pr050300l.
[8] K. Kultima, A. Nilsson, B. Scholz, U.L. Rossbach, M. Fälth, P.E. Andrén, Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides, Mol. Cell. Proteomics 8 (2009) 2285–2295, https://doi.org/10.1074/mcp.M800514-MCP200.
[9] T. Välikangas, T. Suomi, L.L. Elo, A Systematic Evaluation of Normalization Methods in Quantitative Label-free Proteomics, Brief Bioinform, 2016, p. bbw095, https://doi.org/10.1093/bib/bbw095.
[10] B.-J.M. Webb-Robertson, M.M. Matzke, J.M. Jacobs, J.G. Pounds, K.M. Waters, A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors, Proteomics 11 (2011) 4736–4741, https://doi.org/10.1002/pmic.201100078.
[11] A. Chawade, M. Sandin, J. Teleman, J. Malmström, F. Levander, Data processing has major impact on the outcome of quantitative label-free LC-MS analysis, J. Proteome Res. 14 (2015) 676–687, https://doi.org/10.1021/pr500665j.
[12] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, Mol. Cell. Proteomics 13 (2014) 2513–2526, https://doi.org/10.1074/mcp.M113.031591.
[13] B. Zhang, L. Käll, R.A. Zubarev, DeMix-Q: quantification-centered data processing workflow, Mol. Cell. Proteomics 15 (2016) 1467–1478, https://doi.org/10.1074/mcp.O115.055475.
[14] A. Pursiheimo, A.P. Vehmas, S. Afzal, T. Suomi, T. Chand, L. Strauss, M. Poutanen, A. Rokka, G.L. Corthals, L.L. Elo, Optimization of statistical methods impact on quantitative proteomics data, J. Proteome Res. 14 (2015) 4118–4126, https://doi.org/10.1021/acs.jproteome.5b00183.
[15] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.-J.L. Ham, D.M. Bunk, L.E. Kilpatrick, D.D. Billheimer, R.K. Blackman, H.L. Cardasis, S.A. Carr, K. R. Clauser, J.D. Jaffe, K.A. Kowalski, T.A. Neubert, F.E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.J. Fisher, B. W. Gibson, C.R. Kinsinger, M. Mesri, H. Rodriguez, S.E. Stein, P. Tempst, A. G. Paulovich, D.C. Liebler, C. Spiegelman, Repeatability and reproducibility in proteomic identifications by liquid chromatography−tandem mass spectrometry, J. Proteome Res. 9 (2010) 761–776, https://doi.org/10.1021/pr9006365.
[16] B. Bolstad, preprocessCore: A collection of pre-processing functions. R package version 1.62.1,, 2023, in: https://github.com/bmbolstad/preprocessCore. (Accessed 21 July 2023).
[17] Y. Zhao, L. Wong, W.W. Bin Goh, How to do quantile normalization correctly for gene expression data analyses, Sci. Rep. 10 (2020), 15534, https://doi.org/10.1038/s41598-020-72664-6.
[18] S. Jorge, J.L. Capelo, W. LaFramboise, S. Satturwar, D. Korentzelos, S. Bastacky, G. Quiroga-Garza, R. Dhir, J.R. Wiśniewski, C. Lodeiro, H.M. Santos, Absolute quantitative proteomics using the total protein approach to identify novel clinical immunohistochemical markers in renal neoplasms, BMC Med. 19 (2021) 196, https://doi.org/10.1186/s12916-021-02071-9.
[19] S. Jorge, J.L. Capelo, W. LaFramboise, R. Dhir, C. Lodeiro, H.M. Santos, Development of a robust ultrasonic-based sample treatment to unravel the proteome of OCT-embedded solid tumor biopsies, J. Proteome Res. 18 (2019) 2979–2986, https://doi.org/10.1021/acs.jproteome.9b00248.
[20] E.W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D. S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R.L. Moritz, J.J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, J.A. Vizcaíno, The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition, Nucleic Acids Res. 45 (2017), https://doi.org/10.1093/nar/gkw936. D1100–D1106.
[21] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D.J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, J.A. Vizcaíno, The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences, Nucleic Acids Res. 50 (2022), https://doi.org/10.1093/nar/gkab1038. D543–D552.
[22] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372, https://doi.org/10.1038/nbt.1511.
[23] S. Tyanova, T. Temu, A. Carlson, P. Sinitcyn, M. Mann, J. Cox, Visualization of LC-MS/MS proteomics data in MaxQuant, Proteomics 15 (2015) 1453–1456, https://doi.org/10.1002/pmic.201400449.
[24] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805, https://doi.org/10.1021/pr101065j.
[25] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M.Y. Hein, T. Geiger, M. Mann, J. Cox, The Perseus computational platform for comprehensive analysis of (prote) omics data, Nat. Methods 13 (2016) 731–740, https://doi.org/10.1038/nmeth.3901.
[26] S. Tyanova, J. Cox, Perseus, A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research, 2018, pp. 133–148, https://doi.org/10.1007/978-1-4939-7493-1_7.
[27] S. Yu, D. Ferretti, J.P. Schessner, J.D. Rudolph, G.H.H. Borner, J. Cox, Expanding the perseus software for omics data analysis with custom plugins, Curr Protoc Bioinformatics 71 (2020), https://doi.org/10.1002/cpbi.105.
[28] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47, https://doi.org/10.1093/nar/gkv007. –e47.
[29] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504, https://doi.org/10.1101/gr.1239303.
[30] N.T. Doncheva, J.H. Morris, J. Gorodkin, L.J. Jensen, Cytoscape StringApp: network analysis and visualization of proteomics data, J. Proteome Res. 18 (2019) 623–632, https://doi.org/10.1021/acs.jproteome.8b00702.
[31] The gene ontology resource: 20 years and still GOing strong, Nucleic Acids Res. 47 (2019) D330, https://doi.org/10.1093/nar/gky1055. –D338.
[32] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (2017) D353–D361, https://doi.org/10.1093/nar/gkw1092.
[33] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C.D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase, Nucleic Acids Res. 46 (2018) D649–D655, https://doi.org/10.1093/nar/gkx1132.
[34] J.R. Wiśniewski, M.Y. Hein, J. Cox, M. Mann, A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards, Mol. Cell. Proteomics 13 (2014) 3497–3506, https://doi.org/10.1074/mcp.M113.037309.
[35] J.R. Wiśniewski, Label-free and standard-free absolute quantitative proteomics using the "total protein" and "proteomic ruler" approaches, Methods Enzymol. 585 (2017) 49–60, https://doi.org/10.1016/bs.mie.2016.10.002.