QCQuan: a web tool for the automated assessment of protein expression and data quality of labeled mass spectrometry experiments.

DOI: 10.1021/acs.jproteome.9b00072
Handle: http://hdl.handle.net/1942/28364

QCQuan: a web tool for the automated assessment of protein expression and data quality of labeled mass spectrometry experiments.

Joris Van Houtven,[†,‡,¶] Annelies Agten,[‡] Kurt Boonen,[†,¶] Geert Baggerman,[†,¶] Jef Hooybergs,[†,§] Kris Laukens,[‖,⊥] and Dirk Valkenborg[*,‡,¶,†]

†*VITO NV, Applied Bio & molecular Systems, Boeretang 200, Mol, BE 2400*

‡*Universiteit Hasselt, Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Agoralaan, Diepenbeek, BE 3590*

¶*Universiteit Antwerpen, Centre for Proteomics, Groenenborgerlaan 171, Antwerpen, BE 2020*

§*Universiteit Hasselt, Theoretical Physics, Agoralaan, Diepenbeek, BE 3590*

‖*Universiteit Antwerpen, Biomedical Informatics Research Center Antwerp (Biomina), Middelheimlaan 1, Antwerpen, BE 2020*

⊥*Universiteit Antwerpen, Advanced Database Research and Modelling (ADReM), Department of Mathematics & Computer Sciences, Middelheimlaan 1, Antwerpen, BE 2020*

E-mail: dirk.valkenborg@uhasselt.be

Abstract

In the context of omics disciplines and especially proteomics and biomarker discovery, the analysis of a clinical sample using label-based tandem mass spectrometry (MS) can be af-

fected by sample preparation effects or by the measurement process itself, resulting in an incorrect outcome. Detection and correction of these mistakes using state-of-the-art methods based on, f.i., mixed models can take a lot of (computing) time. MS-based proteomics labs are high-throughput and need to avoid a bottleneck in their quantitative pipeline by quickly discriminating between high- and low-quality data. To this end we developed an easy-to-use web-tool called QCQuan (available at qcquan.net) which is built around the CONSTANd normalization algorithm. It automatically provides the user with exploratory and quality control information as well as a differential expression analysis based on conservative, simple statistics. In this document we describe in detail the scientifically relevant steps that constitute the workflow, and assess its qualitative and quantitative performance on three reference data sets. We find that QCQuan provides clear and accurate indications about the scientific value of both a high- and a low-quality data set. Moreover, it performed quantitatively better on a third data set than a comparable workflow assembled using established, reliable software.

Keywords: label-based, tandem mass spectrometry, quantitative proteomics, data-driven, normalization, workflow, quality control

Introduction

In MS-based proteomics, labeled tandem-MS with multiplexing capabilities is chosen over label-free methods whenever accuracy, precision and instrument time are more important than proteome coverage[1]. As proteomics labs that use multiplexing/labeling have a high data throughput, their subsequent normalization and data analysis can become a bottleneck, especially when doing larger experiments possibly covering multiple tandem-MS runs.

On one hand it is time-consuming for a researcher to do the non-automated parts of the analysis, like verifying that the experiment was successfully conducted. This includes

doing exploratory analyses to verify that the data are consistent and suffice to answer the putative research question. There are some software packages available like Scaffold[2], Proteome Discoverer[3] (PD) and MaxQuant[4] which may alleviate this task, but none of them provide a single report file that summarizes both the outcome and quality of an experiment. Moreover, they each have their own drawbacks (f.i., the former being non-free, the middle using a time-consuming normalization procedure, and the latter lacking statistical support for combining samples from multiple MS runs), which leaves room for improvement.

On the other hand, even some automated steps can take up a lot of computing time, like the (accurate) normalization of quantification data. There are many normalization methods available, either data-driven or based on more complex statistical models as f.i. proposed by Oberg[5] and Hill[6] et al. The latter are in theory able to accurately normalize large multi-run experiments but in practice become computationally infeasible as the number of peptides or proteins exceeds about 2500 or 1000, respectively[5]. Hence, when aiming for a higher number of identifications one has to resort to data-driven methods, but most of these methods are limited in accuracy and become unsuitable when processing data from multiple runs.

These two bottleneck issues are important, because in a high data throughput environment there is a great opportunity cost coupled to spending too much time on one data set, or even worse, spending any time at all on less useful data or on false discoveries. Therefore, the proteomics community needs a workflow/tool that can rapidly determine the scientific value of large amounts of data in a short period of time. Such a tool should automate the bridging of the gap between data on the level of annotated PSMs (peptide-spectrum matches) and the enrichment analysis, as depicted in Figure 1. The tool would thus provide a reasonable part of the quality control (QC) and provide a fast but reliable quantification and differential expression analysis (DEA), also for large experiments with data across multiple tandem-MS runs.

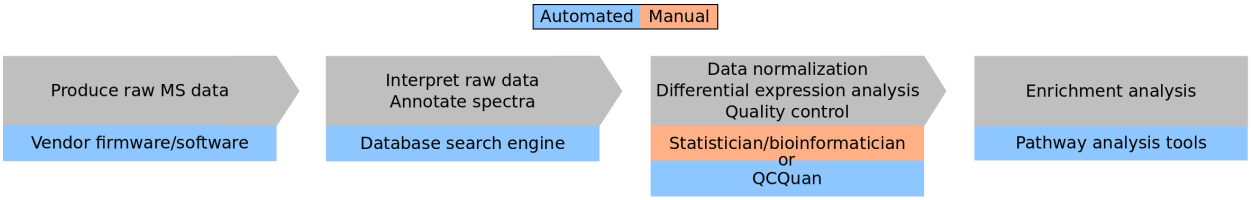


Figure 1: QCQuan bridges the automation gap in the enrichment analysis pipeline.

To this end, we built a web tool for quick assessment of labeled MS-based proteomics experiments called QCQuan (Quality Controlled Quantification) and present it in this paper. It is a workflow which takes PSM-level input data (in PD these are the `_PSMs.txt` files) and returns output on the protein level - in data files and an automated report - as well as a normalization result on the level of non-redundant peptides (i.e., aggregated across all retention times, charges and modifications). We designed QCQuan to be flexible, easy to use and transparent, so that it may be used by anyone, with any experimental setup (including not only TMT labeling, but also i-TRAQ, ICAT, SILAC and others), and could become a benchmark for future innovations in MS data analysis. The normalization issue is addressed by employing the data-driven CONSTANd algorithm, which has shown promising results[7], especially when handling multi-run experiments. Furthermore, QCQuan automatically generates a report containing a variety of QC plots and statistics, as well as crude differential expression results. The latter includes conservative estimates of fold changes and p-values, based on well-established, simple statistical practices. The report is intended to provide researchers with a quick test to triage their data sets: to check for each one whether it is a priority for further investigation, and whether or not something went obviously wrong in the corresponding experimental procedure, like the use of improper instrument settings or pipetting errors during sample preparation. After a positive assessment using QCQuan, one can subject the data set to a more thorough analysis, for example by using the statistical methods by Oberg[5] and Hill[6].

Experimental section

QCQuan workflow

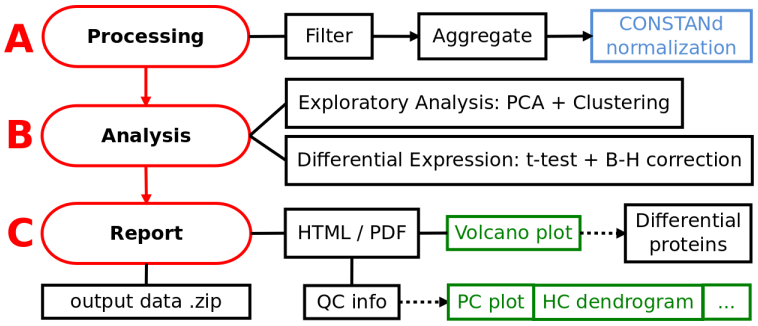


Figure 2: Simplified schematic of the QCQuan workflow. The three main steps are shown in red. The processing step is performed separately for each tandem-MS run. Graphics are shown in green.

The scientific workflow consists of three distinct steps, schematically shown in red in Figure 2 and summarized as follows:

A. **Processing**: clean the input data while gathering some QC information, and aggregate the PSMs (possibly from multiple search engines) to the non-redundant peptide level, then perform the CONSTANd normalization.

B. **Analysis**: while gathering additional QC information, transform the data from the peptide to the protein level and perform a differential expression analysis (DEA). Also perform an exploratory analysis (EA) for QC purposes, including a principal component analysis (PCA) and hierarchical clustering (HC) on the peptide-level data matrix.

C. **Report**: produce a PCA plot (first 2 components), HC dendrogram and for each non-reference condition a volcano plot with a list of the top differential proteins. From the gathered QC info, produce some relevant statistics as well as MS-1 calibration and intensity plots. Lastly, summarize all visualizations, statistics and other relevant information (including meta-data) into a PDF report.

Each step is further detailed in the sections below, and additional user experience details like the input and output file descriptions are outlined in the user experience section. Note that the variable names used in this paper correspond to those used in the PD software, but are user-specifiable.

A. Processing

Each tandem-MS run corresponds to one PSM file which is processed as follows:

1. **Filter.** Remove information from unnecessary columns/variables. Only keep information about the most likely associated proteins, namely the *'Master' Protein Accessions* and corresponding *Descriptions*, and remove information about the others. Remove PSMs that have missing values for columns/variables that the workflow strictly requires, which are: *Sequence*, *Master Protein Accessions*, and *First Scan*. If a PSM has only missing values for all samples of any one experimental condition, it is also removed. Next, remove PSMs with *Confidence* level (if available) worse than 'Medium', or with *Isolation Interference [%]* level (if available) higher than 30. Lastly, for each PSM we remove the labels (if present) from its list of modifications, and keep only information about the identity of modifications, not their location.
2. **Aggregate.** Remove redundancy due to the possible use of multiple search engines, which are specifiable by the user. Only the PSM of the engine/algorithm with the highest priority is kept and the others are removed. Then, aggregate on retention time *RT [min]* (RT) and on *Charge* (see aggregation section). At this point, the only possible leftover redundancy is due to modifications.
3. **Normalize.** Perform CONSTANd normalization[7] (see CONSTANd section).

The data is now normalized and on the level of non-redundant peptides, but still separated into one data frame per tandem-MS run.

B. Analysis

After the processing steps have been performed for each MS run individually, the data is ready for a collective analysis, which consists of two sub-procedures:

I. *The **exploratory analysis** assesses the (dis)similarity between samples, conditions and MS runs*, and consists of a PCA and a HC (with Euclidian distance metric and UPGMA mean linkage criterium). Both are performed on the data frame that results from an *inner join* (see Figure S1) on the peptide sequences of all processing step output data frames (amount equal to the number of different MS runs). This way, only peptides observed in at least one sample of each condition and in each MS run will be present, and each peptide occurs only once. The data matrix is then transposed so the samples take on the role of observations, and the peptides take on the role of variables or dimensions.

Missing values are imputed to be zero. The missing values from peptides not detected at all in a certain MS run or condition had already been removed. However, multiple samples can belong to the same condition or MS run (f.i. see Table S1), so a peptide may have been detected in one or more of those samples but not in the other(s). We need to impute or remove those remaining missing values, because PCA and HC cannot handle them. They are very probably missing not at random, because we know that the same corresponding peptide with only a different label was detected in that same MS run. We therefore reason that these quantification values are missing because they are in fact zero – or at least lower than the detection threshold – and thus impute them to be zero. A better strategy would be to consider more advanced imputation models for left-censored data, but this is out of the scope of this research.

II. *The **differential expression analysis** consists of a log_2 fold change calculation and t-test of protein quantifications between the different biological conditions and the*

reference condition. The protein quantifications used to compute the fold changes, are the averages of the corresponding peptide quantifications, which are pooled together per condition from all corresponding samples and all tandem-MS runs. In contrast, the t-test for each protein is performed on the same corresponding peptide quantifications but after averaging them within each sample (effect on statistics shown in Figure S3, Figure S5 and Figure S6). The latter is necessary because some peptide quantifications of the same protein within the same sample can be considered as repeated measurements of this protein and we cannot treat them as independent observations. Due to this reduction of information the statistical test is not as powerful as it could be, but that is justified since the idea behind QCQuan is to be simple and conservative. Together with the Benjamini-Hochberg correction we apply to the protein-level data, this will further keep the FDR under control.

Shared peptides are excluded by default. We calculate from the processed data frames the mapping between peptides and proteins, which is not a bijection since some 'shared' peptides correspond to multiple proteins. By default, the DEA excludes shared peptides and presents this as the 'minimal calculation', but the user may also instruct QCQuan via the web interface to do a second calculation which includes shared peptides. In this case, each protein gets the full contribution of each associated peptide. Enabling this 'full calculation' option hence provides a crude sensitivity analysis through comparison with the minimal calculation results.

C. Report

The report addresses what we assert to be *the three aspects that determine the scientific value of a data set: experimental quality, biological quality, and added value.*

The experimental quality of the data can be assessed by looking for anomalies in a variety of basic statistics. During the processing and analysis step, many useful quantities (e.g. the amount of isolation interference in each tandem-MS run, or the MS2 intensity

mean, maximum and standard deviation per reporter) are calculated which allow for the detection and pinpointing of anomalies in the data. Two figures may also be produced: the MS1 calibration plot showing the search engine score against the mass offset of each PSM in each tandem-MS run, and the MS1 intensity histogram showing how many PSMs are observed at which intensity, and how many of them were actually used (i.e. not discarded before the aggregation steps).

The biological quality of the data can be assessed using a PCA plot and HC dendrogram of the samples. Using the PCA and HC results from the analysis step, a PCA plot (scatter plot of the two first principal components for each sample) and a HC dendrogram are produced. Identical colors are used for samples belonging to the same biological condition, and identical markers are used for samples from the same tandem-MS run. With this exploratory information, one can easily assess the (dis)similarity between samples and/or conditions and/or experiments, in order to draw conclusions about the biological quality of the data.

The added value of the data can be assessed using a volcano plot and summary of top differential proteins. The protein-level data frames with differential expression results are now used to generate for each condition-reference pair a volcano plot, which colors the protein data points according to whether they exceed a log_2 fold change of 1, and whether their (adjusted) t-test p-values fall below 0.05 or not. The top 10 (by default, but this number is specifiable by the end-user) differential proteins ranked according to p-value are summarized in a table, along with their log_2 fold change, description, and total amount (before averaging repeated measurements) of supporting peptides observed. Although QCQuan should not be used for making high-precision estimates, this information allows the researcher to assess whether the data set may contain valuable results or not.

Reference data sets and methodology

Qualitative performance

To check whether QCQuan can properly aid in distinguishing data sets with regard to their biological quality and/or added value, we have two data sets ('Organs' and 'Failed') at our disposal.

The **Organs** data set (size: 99515 PSMs) is used by Bailey[8] et al., where biological samples were gathered from 8 organs of each 4 mice in TMT 8-plex experiments. This translates to 32 samples spread across 8 biological conditions in 4 tandem-MS runs, as can be inferred from the experimental design in Figure S2.

The **Failed** data set (size: 8446 PSMs) corresponds to an in-house experiment concerning three couples of biological conditions, evenly represented in 24 samples across four TMT 6-plex tandem-MS runs. As can be seen from Table S1, this means that for each combination of conditions there are 3 replicates.

Both of these data sets had their PSM files generated by PD2.1 using both Mascot and Sequest as search engines. The PSM files were then analyzed by QCQuan using the default settings, as well as the following specifiable parameters: Mascot as the Master PSM algorithm; only compute the minimal DEA (thus excluding shared peptides); 'muscle' and 'b' as the reference conditions for the Organs and Failed data sets, respectively.

Quantitative performance

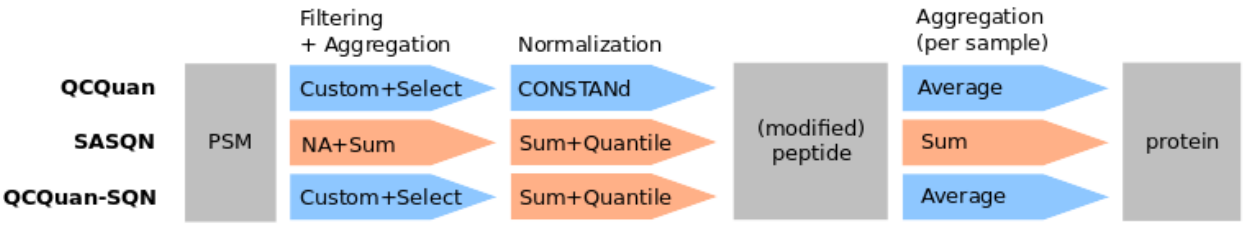


Figure 3: Main differences between the three workflows which are compared by their performance on the Spike-in data set.

To verify whether QCQuan's quantification values used in the DEA adequately represent reality, we use a technical data set which we call '**Spike-in**' (size: 2382 PSMs). It was made available by Gatto et al.[9] via ProteomeXchange (ID PXD000001) and is used in their accompanying `RforProteomics` (RfP) R package[10]. Its mzML and mzID files contain data from one 6-plex tandem-MS experiment, where "*four exogenous proteins were spiked into an equimolar Erwinia carotovora lysate with varying proportions in each channel of quantitation; yeast enolase (ENO) at 10:5:2.5:1:2.5:10, bovine serum albumin (BSA) at 1:2.5:5:10:5:1, rabbit glycogen phosphorylase (PHO) at 2:2:2:2:1:1 and bovine cytochrome C (CYT) at 1:1:1:1:1:2*". Entries were filtered out from the mzID file, which either corresponded to decoys, were not ranked first by the MS-GF+ search engine, or corresponded to shared peptides. The remaining entries were then combined with their corresponding mzID quantification and feature data and then saved as text files, which served as our PSM input files.

We treated each sample as a separate biological condition and ran QCQuan with the default settings, Mascot as the search engine, the sample of TMT reporter 129 as the reference condition. To assess the added value of QCQuan we also re-ran the analysis using two additional workflows we call 'SASQN'and 'QCQuan-SQN'. They are described below, and summarized in Figure 3.

SASQN was chosen as a transparent and trustworthy reference workflow, built using the functionalities provided by the RfP package to filter, aggregate and normalize the data in a way alternative to QCQuan's (see supplementary information). We first filtered out NA values and then used 'sum-wise' aggregation (i.e. summing quantification values whenever combining observations) to obtain quantifications on both the peptide and protein levels, after applying both a 'sum' (In RfP, sum normalization consists of re-scaling each observation's quantification values so that their sum is equal to one) and a quantile normalization, in that order, on the peptide level. This way, we can compare the spike-in protein and peptide fold changes ultimately obtained from both workflows, as well as the

theoretically expected fold changes.

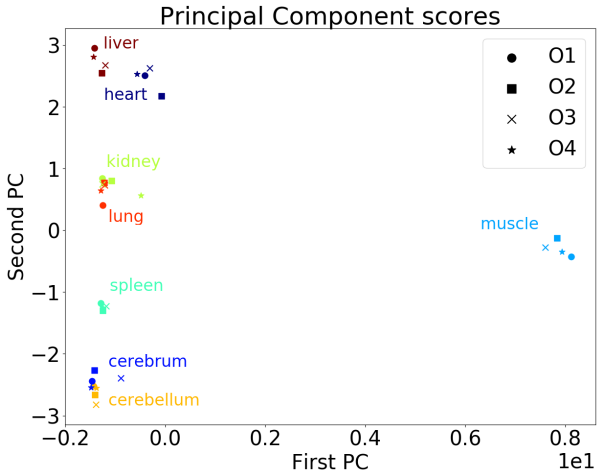
QCQuan-SQN is a hybrid between QCQuan and SASQN, where only QCQuan's CONSTANd normalization has been replaced by the sum and quantile normalization from SASQN. This way, one can assess the relative contributions of the normalization and the other parts of the workflow to the total of differences between the QCQuan and SASQN results.

Results and Discussion

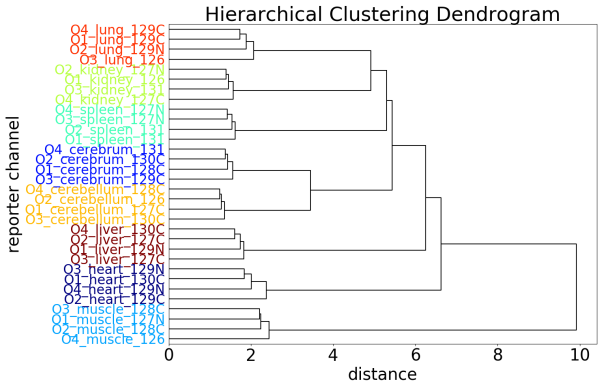
Qualitative

For the **Organs** data set (runtime: $146s$), as we had anticipated, both the PCA and HC plots in Figure 4 suggest that there are clear proteomic differences between samples from different types of organs, but not so much between samples from the same type of organ. The CONSTANd algorithm successfully normalizes the data from multiple independent runs, highlighting biological rather than experimental differences.

For the **Failed** data set (runtime: $25s$), we anticipated to also see significant proteomic differences between the conditions A and a, but the PCA and HC plots in Figure 5 suggest the opposite. This is also confirmed by the volcano plot in Figure S7, which shows only 5 significant differential proteins (using a 0.05 significance level) and still those have log_2 fold changes with absolute values smaller than 0.5. Surprised by this result, we inquired about the experimental procedures and discovered a human mistake was made in the wet lab. In this case, one could immediately infer from QCQuan's QC plots that something was awry.

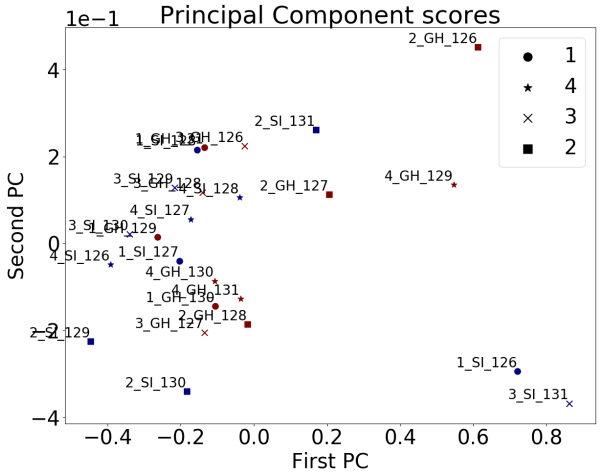


(a) PCA plot. Most of the organs seem to have their corresponding samples form a cluster. Note: labels have been adapted to avoid cluttering.

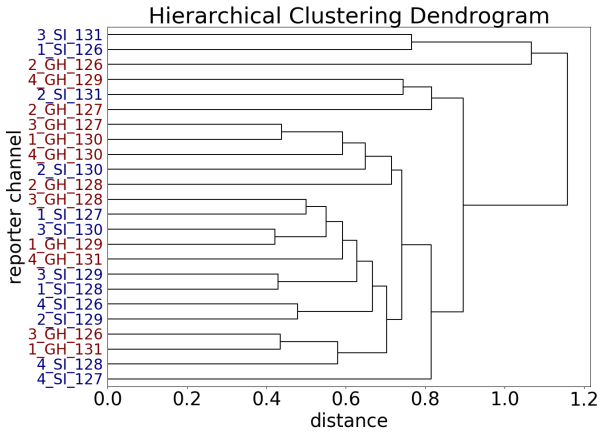


(b) HC dendrogram. Samples corresponding to identical organs are always clustered with each other before they are clustered with samples from other tandem-MS runs.

Figure 4: QCQuan uncovers biological similarities and dissimilarities between samples (color corresponds to organ type) in the Organs data set.



(a) PCA plot. The samples seem randomly scattered across the plane, regardless of the biological condition they represent.



(b) HC dendrogram. The samples seem randomly clustered together, regardless of the biological condition they represent.

Figure 5: QCQuan suggests there is no distinction between the biological conditions in the Failed data set.

Quantitative

The results of the three DEA approaches involving the Spike-in data set (QCQuan runtime: $31s$) are summarized in Figure 6.

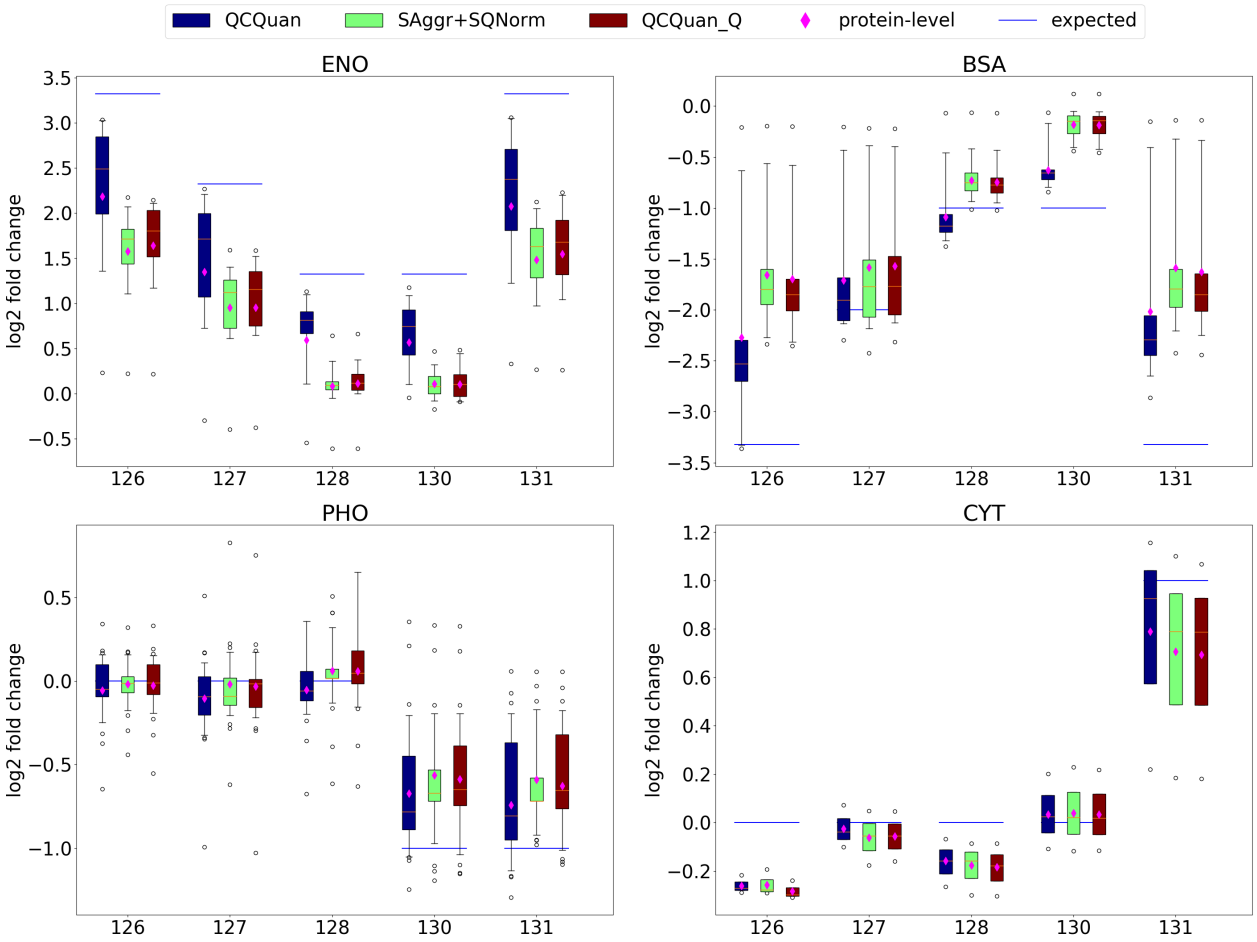


Figure 6: Boxplots of the Spike-in peptide (red line is the median) estimates and corresponding protein estimates (pink diamond) of the fold change, obtained using three different workflows. Estimates of QCQuan with CONSTANd normalization are consistently closer to the expected values (blue lines) than the other two methods, except when the expected fold changes are zero.

Firstly, they show that fold change estimates by QCQuan (using either normalization method) on both the peptide and protein level are in agreement with those provided by the SASQN workflow. All three approaches clearly exhibit the ratio compression phenomenon[11], but the QCQuan (with CONSTANd) approach is in most cases (except for very small fold changes) better at estimating the expected fold changes. However, one

has to keep in mind that the SASQN workflow was not chosen for its high accuracy, but rather for its reliability and transparency.

Secondly, the strong similarity between the QCQuan-SQN and SASQN estimates confirm that the difference in normalization method accounts for practically all of the difference with respect to the QCQuan estimates. Combined with the first finding, this clearly indicates that CONSTANd performed better at normalizing the data than did quantile normalization, and that the aggregation methods - although quite dissimilar - have but a relatively small effect on the results.

Another remarkable observation is that the protein estimates are usually less close to the expected values than the peptide (median) estimates. This is an artifact due to the negative skewness of the absolute fold changes of these particular peptide data. That skewness may in turn be a manifestation of the limits of the linear dynamic range[12] of the mass spectrometer, but this requires further investigation.

Conclusions

We find that the proposed workflow provided us with helpful indications about the scientific value of both a high- and a low-quality data set. Also, it performed quantitatively better on the Spike-in data set than a comparable workflow assembled using functionality from the RforProteomics R-package.

QCQuan is relatively fast, as it produces a QC and DEA report for a commonly sized (10^5 PSMs) tandem-MS experiment in under 3 minutes. It can handle data from multiple tandem-MS runs, a feat which is otherwise only possible through a computationally costly statistical approach like using mixed models.

One can use QCQuan either solely for normalization and data preprocessing purposes, or also for its built-in QC and DEA capabilities. One can do both a minimal and full protein inference, which also makes for a simplistic sensitivity analysis. There is built-in

compatibility for the use of multiple and custom PSM algorithms, as well as alternative kinds of quantification values like PD's S/N-ratio. Whenever information required by any non-essential sub-step like filtering or gathering of QC info is unavailable, the sub-step is skipped without error and the occurrence is logged in the report.

QCQuan does not require the user to have programming skills or other specific training, and a more extensive documentation is available at `https://qcquan.net`. Being a version-controlled web tool, it requires no installations or updates. The only steps to be taken are the collection of PSM files (optionally creating a simple variable name wrapper text file), selecting the desired settings on the website and uploading the data.

The drawbacks of the workflow are for instance a possible lack of statistical power due to the (intentionally) conservative nature of the approach, and the fact that it is currently only available via a web server. A possible risk is that due to the data-driven nature of the CONSTANd algorithm, poor experimental design may lead to unintended biases in the normalization step (see CONSTANd section in Supporting Information). All in all though, we believe that these are not drawbacks which would amplify any existing problems with the data, and they can be easily and naturally avoided by the everyday, informed researcher.

In summary, the proposed workflow allows researchers to quickly assess the scientific value (i.e. experimental and biological quality as well as added value) of a data set through its quality control (QC) and differential expression analysis (DEA) report, and this with reasonable and justifiable drawbacks. *QCQuan thus constitutes a solution to the data analysis bottleneck issues mentioned in the introduction.*

We hope that due to its transparency, simplicity in design and compatibility with any label-based tandem-MS technique, QCQuan may evolve to be a standardized proteomics DEA workflow for comparison with other, more specialized tools.

Supporting Information

The following supporting information is available free of charge at ACS website `http://pubs.acs.org`

1. Figure S1. Inner join result when combining two data frames.
2. Figure S2. Experimental design corresponding to the Organs data set.
3. Table S1. Experimental design corresponding to the Failed data set.
4. Figure S3. Comparison of DEA statistics for the Organs data set when treating repeated measurements as independent, and when averaging them.
5. Section User experience. Notes on user experience.
6. Section Aggregation. Notes on aggregation.
7. Figure S4. Elution profile of a peptide.
8. Section CONSTANd. Notes on CONSTANd normalization.
9. Figure S5. Comparison of p-values from the Organs data set when treating repeated measurements as independent, and when averaging them.
10. Figure S6. MA plots of the p-values from the Organs data set when treating repeated measurements as independent, and when averaging them.
11. Figure S7. Volcano plot of the Failed data set for condition b versus B.
12. Figure S8. Detailed schematic of the QCQuan workflow.
13. R_scripts.zip. R-code for generating PSMs (generate_PSMs.R) and R-code for SASQN workflow (SASQN_workflow.R).
14. Organs_input.zip. Example input data of the Organs data set: Organs_input.zip

15. Organs_full_output.zip. Example output data of the minimal and full expression analysis of the Organs data set: Organs_full_output.zip

16. Organs_full_Report.pdf Example report file of the minimal and full expression analysis of the Organs data set.

References

(1) Megger, D. A.; Pott, L. L.; Ahrens, M.; Padden, J.; Bracht, T.; Kuhlmann, K.; Eisenacher, M.; Meyer, H. E.; Sitek, B. Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2014**, *1844*, 967–976.

(2) Proteome Software, Scaffold. `http://www.proteomesoftware.com/products/scaffold/`.

(3) Thermo Fischer Scientific, Proteome Discoverer. `https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFAKJMAUH`.

(4) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **2008**, *26*, 1367.

(5) Oberg, A. L.; Mahoney, D. W.; Eckel-Passow, J. E.; Malone, C. J.; Wolfinger, R. D.; Hill, E. G.; Cooper, L. T.; Onuma, O. K.; Spiro, C.; Therneau, T. M.; et al., Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of proteome research* **2008**, *7*, 225–233.

(6) Hill, E. G.; Schwacke, J. H.; Comte-Walters, S.; Slate, E. H.; Oberg, A. L.; Eckel-Passow, J. E.; Therneau, T. M.; Schey, K. L. A statistical model for iTRAQ data analysis. *Journal of proteome research* **2008**, *7*, 3091–3101.

(7) Maes, E.; Hadiwikarta, W. W.; Mertens, I.; Baggerman, G.; Hooyberghs, J.; Valkenborg, D. CONSTANd : A Normalization Method for Isobaric Labeled Spectra by Constrained Optimization. *Molecular & Cellular Proteomics* **2016**, *15*, 2779–2790.

(8) Bailey, D. J.; McDevitt, M. T.; Westphall, M. S.; Pagliarini, D. J.; Coon, J. J. Intelli-

gent data acquisition blends targeted and discovery methods. *Journal of proteome research* **2014**, *13*, 2152–2161.

(9) Gatto, L.; Christoforou, A. Using R and Bioconductor for proteomics data analysis. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2014**, *1844*, 42–51.

(10) RforProteomics. `https://doi.org/doi:10.18129/B9.bioc.RforProteomics`.

(11) Savitski, M. M.; Mathieson, T.; Zinn, N.; Sweetman, G.; Doce, C.; Becher, I.; Pachl, F.; Kuster, B.; Bantscheff, M. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of proteome research* **2013**, *12*, 3586–3598.

(12) Tang, K.; Page, J. S.; Smith, R. D. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2004**, *15*, 1416–1423.

For TOC Only