Journal of proteome research

Article

pubs.acs.org/jpr

Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics

Wout Bittremieux,[†,‡] Pieter Meysman,[†,‡] Lennart Martens,[¶,§,∥] Dirk Valkenborg,[⊥,#,▽] and Kris Laukens*[,†,‡]

[†]Department of Mathematics and Computer Science, University of Antwerp, 2020 Antwerp, Belgium

[‡]Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp/Antwerp University Hospital, 2650 Edegem, Belgium

[¶]Department of Medical Protein Research, VIB, 9000 Ghent, Belgium

[§]Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium

[∥]Bioinformatics Institute Ghent, Ghent University, 9000 Ghent, Belgium

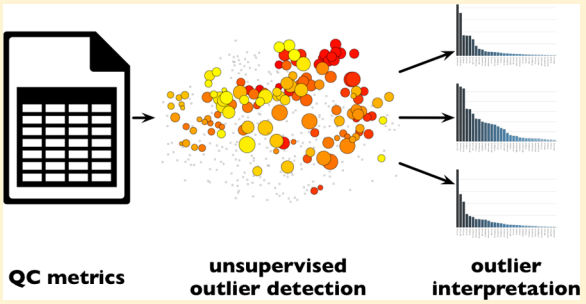
[⊥]Flemish Institute for Technological Research (VITO), 2400 Mol, Belgium

[#]CFP, University of Antwerp, 2020 Antwerp, Belgium

[▽]I-BioStat, Hasselt University, 3590 Diepenbeek, Belgium

Supporting Information

ABSTRACT: Despite many technological and computational advances, the results of a mass spectrometry proteomics experiment are still subject to a large variability. For the understanding and evaluation of how technical variability affects the results of an experiment, several computationally derived quality control metrics have been introduced. However, despite the availability of these metrics, a systematic approach to quality control is often still lacking because the metrics are not fully understood and are hard to interpret. Here, we present a toolkit of powerful techniques to analyze and interpret multivariate quality control metrics to assess the quality of mass spectrometry proteomics experiments. We show how unsupervised techniques applied to these quality control metrics can provide an initial discrimination between low-quality experiments and high-quality experiments prior to manual investigation. Furthermore, we provide a technique to obtain detailed information on the quality control metrics that are related to the decreased performance, which can be used as actionable information to improve the experimental setup. Our toolkit is released as open-source and can be downloaded from https://bitbucket.org/proteinspector/qc_analysis/.



KEYWORDS: *mass spectrometry, proteomics, quality control, quality assessment, outlier detection, outlier interpretation*

INTRODUCTION

Mass-spectrometry-based proteomics forms a powerful analytical technique that can be used to identify and quantify complex protein samples. Yet, despite the many technological and computational advances, performing a mass spectrometry experiment is still a highly complex activity, and its results are subject to a large variability.[1] Consequently, applying these techniques routinely in, for example, a clinical setting is often still not possible.[2] As a remedy, for mass spectrometry proteomics to mature into an analytical, transparent, and reproducible discipline, an additional focus on quality assurance and "quality by design" are required.[3]

This variability exhibited by mass spectrometry experiments can originate from different sources, such as steps in the experimental setup that are not (yet) fully understood, stochastic processes, or the bioinformatics processing workflow, and it impedes achieving reproducible results across multiple experiments.[1,4] To understand and evaluate how technical variability affects the results of an experiment, we have been introduced to several quality control (QC) and performance metrics.[5−13] These metrics are computationally derived from the output of a mass spectrometry experiment and aim to capture the important operational characteristics of a mass spectrometer to provide an objective evaluation of the quality of the experiment.

Initial work by Rudnick et al.[5] defined a set of QC metrics that are mostly linked to peptide and protein identifications to assess the quality of an experiment. These metrics were further reimplemented in QuaMeter,[6] which was later on extended to provide a set of identification-free metrics as well.[9] Other tools focus explicitly on the behavior of the metrics over time to

Received: January 13, 2016

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

study them in a longitudinal fashion,[7,8,13] use statistical process control,[10] or derive the metrics from the instrument settings rather than from the spectral data.[13,14]

However, despite the availability of QC metrics covering a wide range of qualitative information, a systematic approach to quality control is still lacking. Instead, quality control practices frequently involve only monitoring a few simplified performance measures in a spreadsheet or sometimes are even limited to only checking QC metrics retrospectively in an ad hoc fashion when a malfunction is suspected.

One of the barriers impeding the adoption of a systematic quality control workflow is the lack of knowledge about what specific QC metrics signify. It is often unclear which metrics can be applied to detect which problems, and the acceptable variability within metrics is ill-defined.[15] Most aforementioned tools illustrate their applicability by highlighting a few particular use cases where they were able to detect inferior experiments; however, this often does not translate to a more general setting. Specific metrics might only be relevant for highly exceptional situations, or the metrics might be hard to interpret and link to actionable solutions even for a domain expert.

Another issue that is generally overlooked during the interpretation of QC metrics is that interdependencies have to be taken into account. During a mass spectrometry experiment, the different steps do not function in isolation; instead, they influence each other. Because most tools only look at a single metric at the same time in a univariate fashion, these dependencies are ignored, which can lead to erroneous results.[15] To accommodate for these dependencies, one can employ a multivariate approach, as has been done by Wang et al.,[9] who analyzed the variability present in samples originating from multiple National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) studies using multivariate statistics, such as principal component analysis (PCA). Whereas Wang et al.[9] mainly applied unsupervised techniques, a different approach was employed by Amidan et al.[16] Here, first the quality of a multitude of experiments was manually reviewed by expert instrument operators, after which this labeling was used to train a supervised classifier to distinguish good experiments from poor experiments. In general, such multivariate approaches can protect against false positives in the event of a high number of variables and can allow us to detect patterns that are invisible when evaluating each metric individually. An additional advantage is that certain multivariate techniques provide insight on how some of the variables are related to each other, which can improve the understanding of the QC metrics and their interpretation.

Here, we will illustrate how computationally derived QC metrics can be used to provide an initial discrimination between low-quality and high-quality experiments in an unsupervised fashion prior to manual investigation. The presented techniques will take into account the multidimensional feature space exhibited by the QC metrics while also prioritizing the most relevant QC metrics. A special emphasis will be laid on the interpretability of the obtained results, as to integrate a systematic approach to quality assurance into existing mass spectrometry proteomics workflows, the interpretability of the qualitative information, even to nonexpert users, is paramount. Finally, we will unify these different steps to present an open-source toolkit of powerful techniques for the analysis and interpretation of mass spectrometry proteomics QC metrics.

EXPERIMENTAL SECTION

Quality Control Metrics

Our goal is to discriminate the low-quality experiments from the high-quality experiments when considering multiple experiments. Suitable data for this kind of analysis are for example standard quality control samples, such as the simple BSA samples that are used by several laboratories. The advantage of such QC runs is that they are measured on a very frequent basis and that their low sample complexity and controlled sample content allow us to easily asses the operation of a mass spectrometer. Although the computation of QC metrics is not restricted to such samples, a controlled and consistent sample content and operating procedure facilitate interpreting the analysis results. If LC−MS runs vary over time due to biological or technical changes in the sample itself, such as a different wet lab protocol or a different biological condition, this may complicate discerning the origin of potentially detected anomalies.

Experimental Data

We used two public data sets to investigate the experiment quality. The first data set consists of a number of standard quality control LC−MS runs performed on several different instruments at the Pacific Northwest National Laboratory (PNNL).[16] Each sample had identical content (whole-cell lysate of *Shewanella oneidensis*), and the quality of the various runs has been manually annotated by expert instrument operators as being either "good", "ok", or "poor". We split up the various runs depending on the instrument type, with each instrument group consisting of multiple individual instruments. Both the experimental raw files and the expert annotations have been retrieved from the PRoteomics IDEntifications (PRIDE) database.[17] Please see the original publication for further information on the experimental procedures.[16]

The second data set was generated as part of The Cancer Genome Atlas (TCGA) project, in which the aim was to perform a proteogenomic characterization of human colon and rectal cancer.[18] For this study, 95 samples from 90 patients were obtained, with each sample fractionated into 15 concatenated peptide fractions before being subjected to an LC−MS analysis. This resulted in 1425 raw files, which were retrieved from the CPTAC data portal.[19] For a more detailed exposition of the sample content and preparation, please refer to the original publications.[18,20]

Table 1 shows an overview of how the data from these two sources has been split into four data sets. The PNNL data sets consist of a unique public data resource, as they not only contain high-quality measurements, as is common, but explicitly also include low-quality measurements. Furthermore, the expert annotations provide crucial

Table 1. Overview of the Characteristics of the Various Datasets[a]

reference	denomination	instrument model (accession)	number of raw files	expert annotation
Amidan et al.[16]	PNNL LTQ-IonTrap	Thermo LTQ MS (MS:1000447)	225	yes
Amidan et al.[16]	PNNL LTQ-Orbitrap	Thermo LTQ Orbitrap (MS:1000449), Thermo LTQ Orbitrap XL (MS:1000556)	379	yes
Amidan et al.[16]	PNNL Velos-Orbitrap	Thermo LTQ Orbitrap Velos (MS:1001742)	538	yes
Zhang et al.[18]	TCGA	Thermo LTQ Orbitrap Velos (MS:1001742)	1425	no

[a]The instrument accession numbers refer to their identifiers in the Proteomics Standards Initiative - Mass Spectrometry (PSI-MS) controlled vocabulary.[21]

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

information to validate the detection of low-quality experiments. The TCGA data set can be used to highlight how instrument performance evolves over time when working with complex sample contents because all raw files have been obtained on a single Orbitrap Velos mass spectrometer using the same operating procedure over an extended time period.

Metrics Generation

Over the past few years, multiple sets of QC metrics have been defined.[5−13] Here, we focus on so-called identification-free (ID-free) metrics, which have the advantage that they are directly derived from the raw data and that they do not depend on identification results. This enables the generation of the metrics as soon as the experimental raw data is available instead of having to analyze that raw data in a potentially computation-heavy peptide and protein identification workflow. Furthermore, the lack of dependency on the identification results eliminates possible sources of computational variability and prevents suboptimal settings in the various bioinformatics steps from influencing the quality assessment.[22]

Specifically, we used the ID-free metrics computed by QuaMeter,[9] which are listed in Supplementary Table S1. These metrics are derived from the raw spectral data and provide information on various stages of a mass spectrometry experiment: they include information on the chromatography, the MS and MS/MS performance, and the charge distribution. For the PNNL data, several different sets of QC metrics were already computed, and we restricted the metrics under consideration to the QuaMeter metrics. For the TCGA data, we used QuaMeter version 1.1.91 to produce the QC metrics.

To prepare the QC metrics for analysis, one must perform several preprocessing steps. First, invariant metrics with a low information content are removed, as they have the same value for each experiment and needlessly increase the dimensionality without adding additional information. Furthermore, mutually dependent metrics can be derived from each other, so the duplicated metrics can be omitted without any information loss. Removing the low-variance and correlated metrics decreases the dimensionality while retaining all embedded information, and prevents irrelevant metrics from deteriorating the subsequent analyses. See Supplementary Section 1 for more details on the various preprocessing steps.

Furthermore, an appropriate visualization can often provide crucial insights in the nature of the data.[23] See Supplementary Section 2 for an aside on various optimized techniques that can be used to visualize high-dimensional data such as those under consideration here.

Quality Analysis

Because most quality control tools are able to generate (at least) several dozens of metrics, any single experiment can be characterized by multiple QC metrics. Therefore, it is often not clear which metrics are most interesting in general or even which metrics are relevant in a specific situation. The numerous metrics form a multidimensional data space, which results in several challenges during the analysis. For example, the measurement for a single metric might slightly deviate while all other metrics are firmly within the normal range of operation. In such cases, the deviating measurement might simply be due to random fluctuations and not actually due to an abnormal performance. When the metrics are evaluated individually, a multiple test correction is an often overlooked necessity to avoid spurious results. Furthermore, as the different stages of a mass spectrometry experiment do not function in isolation but instead influence each other, likewise some metrics will be correlated. For example, a problem during ionization will lead to different charge-state proportions, might influence the number of MS/MS scans, and will have an impact on the number of successfully identified spectra. Again, it is inadequate to only look at a single metric, which might lead to incomplete or even false conclusions. These simple examples illustrate that analyzing each metric individually is often insufficient, and instead, multivariate techniques that take into account all metrics simultaneously should be used.

However, not all multivariate techniques are always applicable. For example, when a multivariate approach using a dimensionality reduction, such as PCA, is applied, part of the data is lost, which can likewise lead to faulty or incomplete results. Furthermore, an additional disadvantage of using PCA-based and related techniques is that the principal components are formed by linear combinations of the original features, which complicates their interpretation. Nevertheless, applying a dimensionality reduction can still be useful in specific situations (for example, when combining multiple sources of QC metrics). Merging multiple sets of QC metrics will drastically increase the data dimensionality, which can have a profound effect on the subsequent analysis. Namely, for high-dimensional data, various detrimental effects commonly subsumed under the term "curse of dimensionality" pose challenges for algorithms that make use of the full feature space.[24] Therefore, in some cases, applying a prior dimensionality reduction, or using algorithms that are optimized for a high-dimensional search space, might achieve a superior performance. Because in our analysis we have limited the QC metrics to only the ID-free metrics that are computed by QuaMeter, the dimensionality of the data set is not overly large to impede taking all features into account. Furthermore, using the full dimensionality has certain advantages, such as the availability of more information compared to when a dimensionality reduction technique would be applied.

We will next show how optimized techniques that take into account the multidimensional feature space can be used to assess the quality of mass spectrometry experiments and to discriminate low-quality from high-quality experiments.

Outlier Detection

Outlier detection can be used to detect deviating experiments with a low performance or a high level of (unexplained) variability. These outlying experiments can subsequently be analyzed to discover the source of the reduced performance to enhance the quality of future experiments. Additionally, outlier detection can be a vital step to remove invalid measurements ahead of further processing, such as, for example, sample identification, to ensure that these low-quality experiments do not unduly influence the output results.

Local Outlier Detection. We have used the Local Outlier Probability (LoOP)[25] algorithm to detect outlying experiments, as it has a few beneficial properties. Most importantly, LoOP identifies outliers on the basis of their local neighborhood.[26] This approach is more sensitive than global outlier detection methods, as outliers are identified on the basis of the density of the neighbors in their immediate vicinity, as opposed to the global data distribution. For example, when analyzing a sizable number of experiments performed over an extended time period, as we have done, it is conceivable that there will occur small environmental changes over time, which will have an influence on the experimental results. Because these effects might be more or less pronounced at certain times, this prohibits the use of a single global outlier measure. Instead, when the outlier measure is restricted to the local neighborhood, outliers will be identified on the basis of (excessive) differences with their closest matching experiments. Another advantage is that the LoOP outlier scores are normalized and can be expressed as a probability, whereas most other outlier detection algorithms report scores with an arbitrary scale, with scores often incomparable between different data sets or different parameter values, even when using the same algorithm. LoOP, however, consistently uses probabilities, which ensures that these outlier scores can readily be compared and straightforwardly be interpreted.[25]

Figure 1 shows a histogram of the outlier scores that have been assigned to all of the experiments in the PNNL LTQ-IonTrap data set. As can be seen, most experiments have a (relatively) low outlier score, with the bulk of the experiments having a score close to 0%. Other experiments have a higher outlier score, with some of them being marked as clear outliers. As each experiment has been assigned a numeric score, this enables the ranking of the various experiments by their gradation of being an outlier. Furthermore, generally a score threshold is set to distinguish outliers from nonoutliers. Although setting such a score threshold can sometimes be quite subjective, there are a few considerations that can be taken into account. First, we expect that most of the experiments are nonoutliers with a score close to zero, while outliers have a higher score over a wider range. Second,

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

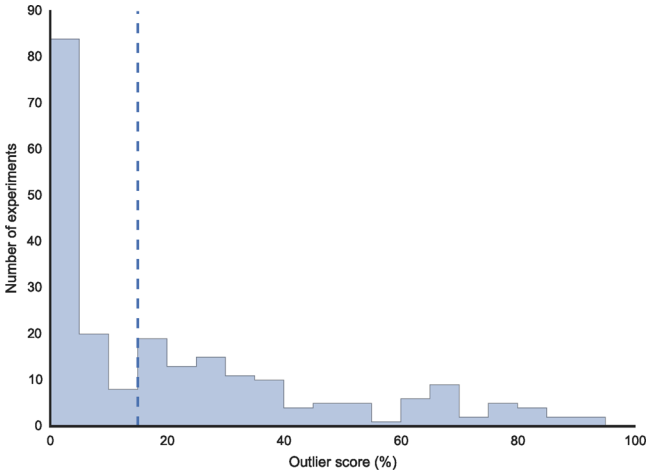


Figure 1. Histogram of the LoOP outlier scores for the PNNL LTQ-IonTrap data set. The score threshold of 15% is indicated by the dashed line.

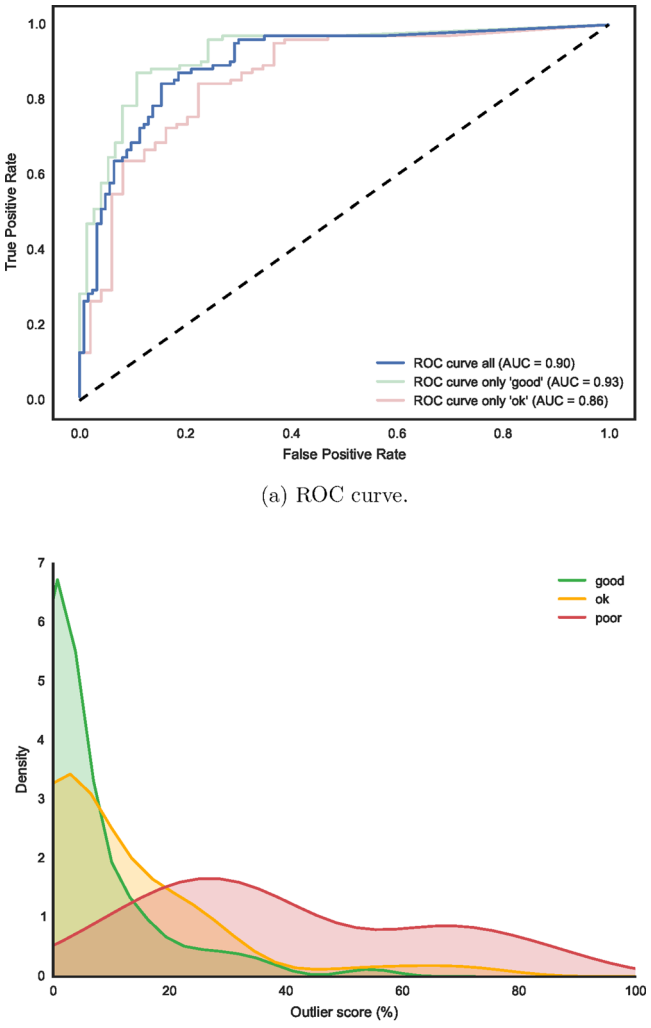
we aim to have a high sensitivity to detect most if not all of the low-quality experiments, so the threshold should be set quite conservatively to ensure that all of the low-quality experiments are detected, at the expense of some false positives. As the outlier detection strategy is an unsupervised method, some number of false positives is unavoidable, which should be filtered out in a subsequent manual evaluation step. Keeping in mind these considerations (for example, for the outlier histogram in Figure 1), a good choice for the score threshold would be 15%.

Outlier Validation. Because outlier detection is an unsupervised method, it is not straightforward to validate the results when real-life data sets are used because the ground truth is often unknown. However, for the PNNL data, the quality of the experiments was assessed by expert instrument operators, whose labeling can be used as the ground truth for validating the detected outliers, as each experiment was assessed as having either "good", "ok", or "poor" quality.

Using these quality assignments, the obtained outlier scores can be validated. Figure 2a shows the receiver operator characteristic (ROC) curve for the PNNL LTQ-IonTrap data set, while Supplementary Figure S3 shows the comparable ROC curves for the PNNL LTQ-Orbitrap and the PNNL Velos-Orbitrap data sets. The blue ROC curve shows the outlier detection performance when the "good" and "ok" experiments are considered as the positive class (i.e., the acceptable experiments) and the "poor" experiments as the negative class (i.e., the unacceptable, low-quality experiments). Additionally, the green curve shows the performance when only the "good" experiments are considered as the positive class, and the red curve shows the performance when only the "ok" experiments are considered as the positive class. These ROC curves clearly indicate that the outlier detection technique successfully manages to discriminate high-quality experiments from low-quality outlying experiments. Furthermore, they show that there is an optimal distinction between the experiments with the highest quality (labeled "good") and the low-quality experiments (labeled "poor"). Meanwhile, experiments with a slightly diminished but still sufficient quality (labeled "ok") can still be discriminated quite successfully from the low-quality experiments.

This is also indicated by Figure 2b and Supplementary Figure S4, which show the density of the outlier scores for the various quality labels. The figures show that the high-quality experiments indeed have a very low outlier score, and the low-quality experiments have a higher outlier score. Therefore, our previous assumptions stated to determine a score threshold when no validation information is available hold true, and the choice of 15% as threshold provides an adequate separation of high-quality and low-quality experiments.

Note that, although the previous results indicate that a very good performance detecting the low-quality experiments is achieved, we are



(a) ROC curve.

(b) Outlier score density for the various quality levels.

Figure 2. Outlier validation for the PNNL LTQ-IonTrap data set based on the expert quality assignments.

still outperformed by the original results by Amidan et al.[16] However, contrary to their approach, which entails a supervised classifier, our approach is fully unsupervised. For a supervised approach, training data is required, which in this case means that the quality of a considerable number of experiments needs to be annotated manually to provide the ground truth. However, our approach does not require a training phase but can be applied directly on a set of experiments of unknown quality. Therefore, the time-consuming manual curation can be forgone while still being able to successfully identify low-quality experiments. Furthermore, a supervised classifier has to be retrained for different instruments or even for different operating procedures on the same instrument, with each situation again incurring the need for manual curation to provide training data. Contrary to this, our outlier detection strategy can directly be applied to diverse data sets with widely varying characteristics, as evidenced by the consistent performance across data generated on different instrument types. However, because an unsupervised technique will inherently result in more false positives than a supervised technique, our outlier detection strategy is mostly suited as a filtering step to quickly provide an initial discrimination between high-quality and low-quality experiments. By conservatively setting the outlier score threshold, we determined that the experiments marked as outliers can subsequently be inspected manually in full detail to exclude any false positives. Because our outlier detection strategy is already quite sensitive despite being an unsupervised technique, the effort required for the manual evaluation will be significantly reduced. Furthermore, by using the local

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

neighborhood to determine whether an experiment is an outlier, optionally dissimilar data sources can even be combined while still achieving a similar performance (data not shown). Finally, our approach requires only a few simple and intuitive parameters, which can easily be understood and set, as is detailed in Supplementary Section 3.

Outlier Interpretation. Although we have shown that we can successfully differentiate low-quality from high-quality experiments, it is insufficient to only know that a specific experiment is an outlier; it is also of vital importance to know why the experiment is an outlier. For this purpose, the outlier score is only useful to a limited extent: it indicates how significantly an experiment is an outlier, but it does not provide an explanation as to what causes the experiment to be an outlier. For the provision of an explanation why an experiment is an outlier, the subspace in which the outlying experiment can be differentiated from the other experiments can be used.[27] Here, a subspace is formed by one or more attributes, which correspond to the various QC metrics. Thus, this subspace can be used to interpret the outlier by indicating which QC metrics best explain the outlying behavior. The relevant subspaces for an outlier can be used by domain experts to increase interpretability and investigate the performance of the experiment. Supplementary Section 4 provides the full details on how the relevant subspaces are identified for the various outliers.

Figure 3 shows an example of interpreting a specific outlying experiment from the TCGA data set. As detailed in Supplementary Section 4, to retrieve the relevant subspace for an outlier, first the feature importances for the various QC metrics are computed, as is shown in Figure 3a. Next, the subspace formed by the relevant QC metrics is extracted, as is shown in Figure 3b. The feature subspace explaining the outlier can be interpreted by domain experts and can provide insights in relationships between various QC metrics. For example, Figure 3b shows that this particular outlier exhibited an exceptionally high variance in peak heights, which may indicate problems with the chromatography.

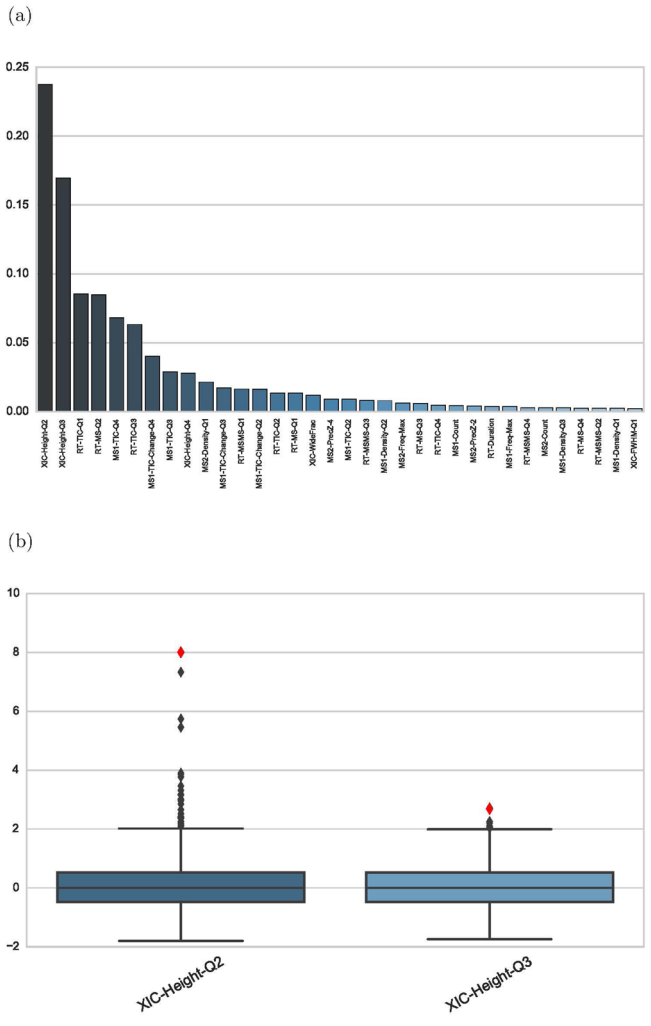


Figure 3. Feature importances and subspace for experiment "TCGA-AA-A02O-01A-23_W_VU_20130206_A0218_10A_R_FR07", which has an outlier score of 98.06%.

This example also shows the advantage of using a multivariate approach instead of only looking at single QC metrics individually. In Figure 3, this advantage is somewhat less pronounced, as the values in the outlier's subspace are individually both already significant outliers. However, the interplay of various metrics can be crucial to detect outliers, as is shown in Supplementary Figure S7. Here, the values in the outlier's subspace are well within 1.5 times the interquartile range, as denoted by the whiskers of the box plot, so this outlier cannot be detected using univariate techniques. However, by comparing against the local neighborhood and by analyzing all metrics simultaneously, we determined that the aberrant proportion of the metrics still results in a high outlier score. Meanwhile, prominent outliers for a single metric will still be detected as well, with explanatory subspaces consisting of only this single metric, as is shown in Supplementary Figure S8.

Furthermore, it is worth highlighting that taking the full set of metrics into account while detecting outliers has advantages over multivariate approaches in which a dimensionality reduction technique is used as well because all metrics are taken into account instead of only a lower-dimensional approximation. For example, when applying PCA, only the most significant principal components are retained to achieve a dimensionality reduction. In this case, only the metrics with the highest variance contribute significantly to the first few principal components. Therefore, using such an approach, a prominent outlier such as shown in Figure 3 cannot be detected successfully because the metrics describing the peak heights only have a limited contribution to the first two principal components, as can be verified in Supplementary Table S2.

Frequent-Outlier Subspaces. Next, by combining the explanatory subspaces for all individual outliers, it is possible to get a general view on which QC metrics are most relevant during the detection of deviating experiments. Most outliers can be interpreted by a limited number of QC metrics: for the TCGA data set, when setting the outlier score threshold at a conservative 25%, which results in 508 outliers (Supplementary Figure S9), on average, each subspace consists of about 2.3 QC metrics, with a minimum of one metric and a maximum of seven metrics. By considering each outlying experiment as a transaction and the QC metrics that form the subspace of the experiment as items in the transaction, we determined that frequent itemset mining[28] can be applied to detect QC metrics that often co-occur in the outliers' subspaces. Table 2 shows the QC metrics that form frequent itemsets with a minimum support of 5% (i.e., the QC metrics are present in the subspace of at least 5% of all outliers). Here, the higher the support, the more often QC metrics co-occur as important explanatory variables in the outliers' subspaces. Table 2 indicates that some QC metrics are more useful than others to detect outliers, as they are present more often. Furthermore, some sets of multiple metrics occur often as well, such as in the case of the itemset consisting of MS1-TIC-Q4 and MS1-TIC-Change-Q4, which has a total support of 6%. Indeed, it seems logical that these two metrics are related, as excessive changes in the TIC between the third and the fourth quartile (MS1-TIC-Change-Q4) will influence the total amount of TIC near the end of the experiment (MS1-TIC-Q4). Other pairs of co-occurring metrics were observed as well, such as the combination of metrics XIC-Height-Q2 and XIC-Height-Q3, concerning the chromatographic peak height, and metrics MS2-PrecZ-2 and MS2-PrecZ-4, concerning the precursor charge-states, although these and other combinations have a slightly lower support value and are not included in Table 2.

Spectral Identification Performance. We were able to show the accuracy of the outlier detection method using manually curated data sets. However, in general such an approach is not possible because it takes too much effort to manually assess the quality of a large

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 2. Overview of the QC Metrics that Frequently Occur in the Outliers' Explanatory Subspaces for the TCGA Dataset[a]

outlier subspace QC metric(s)	support (%)
MS1-TIC-Change-Q4	16
XIC-Height-Q4	15
RT-Duration	14
RT-TIC-Q3	14
XIC-Height-Q2	12
XIC-Height-Q3	11
XIC-WideFrac	11
MS1-TIC-Q4	10
MS1-TIC-Change-Q2	9
RT-TIC-Q2	8
MS2-PrecZ-4	8
MS1-TIC-Q2	8
RT-TIC-Q4	7
MS2-PrecZ-2	7
RT-MSMS-Q4	7
MS1-TIC-Q4, MS1-TIC-Change-Q4	6
MS1-Freq-Max	6
RT-MS-Q3	6
RT-MSMS-Q3	6
RT-MSMS-Q1	5
MS1-Density-Q3	5

[a]Exact support values can differ slightly between separate executions due to some variable effects while computing the explanatory subspaces.

collection of experiments. Therefore, the number of peptide-spectrum matches (PSMs) is often used as a stand-in quality measure, as one would expect that the low-quality experiments result in a lower number of identified spectra due to a diminished performance. When comparing the number of valid PSMs between the outlying experiments and the nonoutlying experiments for the TCGA data set, the latter result in a slightly higher number of PSMs, as is shown in Figure 4a, although the difference is not very pronounced. The outliers seem to contain both experiments that have a lower number of PSMs and experiments that have an average or even an above-average number of PSMs. This observation confirms prior findings that outliers for this data set do not necessarily arise due to sources impeding successful spectrum identifications but can possibly be attributed to a significant biological diversity between the various samples.[20]

However, when the explanatory subspaces for the outliers are taken into account, a distinction between several of the outliers can be made. As can be seen in Figure 4b, for some specific QC metrics, the number of PSMs for the outliers is notably lower than for the nonoutlying experiments. Conversely, for a few other QC metrics the number of PSMs for the outliers is very similar to those of the nonoutlying experiments. Supplementary Table S3 confirms the difference in terms of PSMs between the nonoutlying experiments and the outlying experiments for each of the frequently occurring QC metrics by computing a *t*-test with null hypothesis that they have identical expected values. The reported *p*-values show that this null hypothesis can be safely rejected in some instances, with significantly lower numbers of PSMs for some sets of outliers, which have also been highlighted in Figure 4b. We can hypothesize that the QC metrics for which the number of PSMs is comparable to that of the nonoutlying experiments mainly capture sources of variability that do not necessarily impede spectral identifications, such as, for example, biological differences. Meanwhile, the QC metrics that show a clear discrepancy in terms of valid PSMs compared to the nonoutlying experiments warrant a closer look, as they might indicate potential problems during the experiment resulting in a diminished performance. Because monitoring a large number of QC metrics on a regular basis is often unpractical, it might be more convenient to limit the

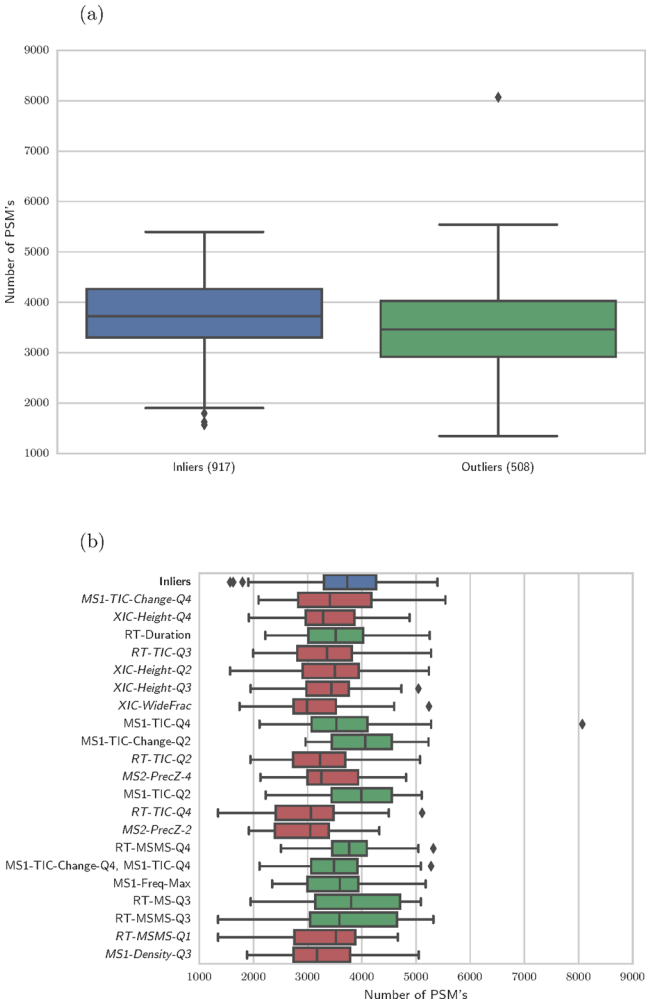


Figure 4. Comparison of the number of PSMs between the nonoutlying and the outlying experiments for the TCGA data set.

general analysis to a small number of user-friendly, well-understood, and discriminating metrics.[4] From Table 2 and Figure 4b it can be deduced that, for example, metrics detailing the chromatographic performance, the TIC accumulation, and the precursor ionization are suitable candidates. Because these metrics occur frequently in the explanatory subspaces, they can be used to detect a wide range of outliers. Furthermore, they seem to indicate a significant decrease in identification performance, highlighting the outliers that are most likely to have a negative influence on the eventual output results. The efficacy of these QC metrics has independently been noted before as well, with the chromatographic peak width and the electrospray ionization selected to monitor the experimental quality during a previous multicenter performance study.[4]

Software Availability

To aid users in the quality exploration of their own experiments, we have bundled the presented analysis techniques into a software toolkit tuned for mass spectrometry quality control. This software, which has been fully coded in Python (making use of scientific and machine learning libraries such as NumPy,[29] pandas,[30] and scikit-learn[31]) allows users to run the presented workflow to detect and interpret outlying experiments. The software can be used as a command-line application and exports the outlier analysis as a qcML file,[12] which can be viewed in any web browser.

All code is released as open source and is available at https://bitbucket.org/proteinspector/qc_analysis/.

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

CONCLUSIONS

A recent informal poll conducted by The Netherlands Proteomic Center (NPC) highlighted that proteomics researchers regard quality control as a crucial component during data analysis, but it also revealed that the majority of researchers do not incorporate systematic quality control approaches in their day-to-day workflows because applying currently existing quality control tools is perceived to be too hard. This clearly shows that there is still significant room for improvement to make quality control an ubiquitous step during mass spectrometry proteomics experiments, something that is vital for mass-spectrometry-based methods to mature into analytical, transparent, and reproducible disciplines.[2]

In this paper, we have presented a powerful technique to perform an initial filtering of low-quality mass spectrometry experiments based on computationally derived QC metrics, with a strong focus on providing easily interpretable results. After all, when identifying low-performance experiments, it is insufficient to just know that an experiment has failed; it is also crucial to understand why the experiment in question exhibits a decreased performance to be able to remedy the problems that caused the failure. Furthermore, we have shown that our approach can be successfully applied across different instruments and instrument types and for sample contents with varying complexity. As such, the methodology we have presented can play an important role in investigating the performance of experiments, as it is able to detect outlying experiments, as well as provide an explanation about the outlying behavior, which can be interpreted by domain experts. A potential disadvantage, however, is that to able to successfully detect low-quality experiments, it must be possible to establish a positive baseline, which requires that there are a sufficient number of experiments available and that the number of low-quality experiments does not exceed the number of high-quality experiments. Because in normal conditions, only a few low-quality experiments are present, and because our approach does not require manual setting of this positive baseline but instead is able to automatically infer it, we believe that this is not a limiting factor. Additionally, because we employ an unsupervised technique, some false positives are unavoidable. However, to determine which experiments exhibit a decreased performance the presented technique can be used as an initial filtering step prior to a detailed manual inspection, which will drastically decrease the number of experiments that need to be checked manually, resulting in significant time savings.

However, the quality control analyses presented here should not stand on their own; instead, they need to be integrated with the experimental results, and they should be closely linked to all operational information and relevant events pertaining to the mass spectrometry instruments. For example, environmental conditions, instrument maintenance schedules, etc., all can have a significant influence on the experimental results.[13,32] This information should ideally be structurally recorded in a sort of electronic lab notebook and should then be related to the experimental results and the quality analysis.

Furthermore, novel analyses or algorithmic approaches are insufficient on their own. Ideally there should be a consolidation of the developed quality control methods to date, and these methods should be made available to a wide audience in intuitive and user-friendly tools, something that is still severely lacking in the community at large.

Finally, although here we explicitly focused on quality control for proteomics, our approach has potential applications in other mass-spectrometry-based domains, such as metabolomics, as well. Current quality control practices in metabolomics mainly involve the direct comparison of features measured in specialized QC samples to subject samples,[33−35] whereas the QC metrics that were considered here form an additional level of abstraction as they are derived from the experimental results. However, because we employed identification-free metrics, which do not directly depend on domain-specific identification procedures but instead capture the general properties of a mass spectrometry experiment, most of these metrics can be straightforwardly applied outside of the proteomics setting as well. Therefore, to conclude, extending such computational QC approaches to related fields such as metabolomics can be a very interesting avenue of future research.

ASSOCIATED CONTENT

Supporting Information

All data used for the analyses presented here is available from the project Web site (https://bitbucket.org/proteinspector/qc_analysis/) along with the functionality to easily recreate all analyses. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00028.

Details on preprocessing, visualization, parameter configuration, and oulier interpretability. Figures showing metrics correlation matrices; multidimensional visualizations; outlier detection ROC curves, score densities, detection AUC versus the size of the local neighborhood, score threshold versus sensitivity and specificity; outlying experiment interpretation, and TCGA outlier score histogram. Tables showing QuaMeter identification-free quality control metrics, TCGA PCA loadings, and outlier subspace identification p-values. (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: +32(0)32653310; fax: +32(0)32653777; e-mail: kris.laukens@uantwerpen.be.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by SBO grant "InSPECtor" (120025) of the Flemish agency for Innovation by Science and Technology (IWT). Data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH).

REFERENCES

(1) Tabb, D. L.; et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2010**, *9*, 761−776.

(2) Martens, L. Bringing proteomics into the clinic: The need for the field to finally take itself seriously. *Proteomics: Clin. Appl.* **2013**, *7*, 388−391.

(3) Tabb, D. L. Quality assessment for clinical proteomics. *Clin. Biochem.* **2013**, *46*, 411−420.

(4) Campos, A.; et al. Multicenter experiment for quality control of peptide-centric LC-MS/MS analysis − A longitudinal performance

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX

assessment with nLC coupled to orbitrap MS analyzers. *J. Proteomics* **2015**, *127* (B), 264−274.
(5) Rudnick, P. A.; et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* **2010**, *9*, 225−241.
(6) Ma, Z.-Q.; Polzin, K. O.; Dasari, S.; Chambers, M. C.; Schilling, B.; Gibson, B. W.; Tran, B. Q.; Vega-Montoto, L.; Liebler, D. C.; Tabb, D. L. QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal. Chem.* **2012**, *84*, 5845−5850.
(7) Pichler, P.; Mazanek, M.; Dusberger, F.; Weilnböck, L.; Huber, C. G.; Stingl, C.; Luider, T. M.; Straube, W. L.; Köcher, T.; Mechtler, K. SIMPATIQCO: A server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *J. Proteome Res.* **2012**, *11*, 5540−5547.
(8) Taylor, R. M.; Dance, J.; Taylor, R. J.; Prince, J. T. Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics* **2013**, *29*, 2948−2949.
(9) Wang, X.; Chambers, M. C.; Vega-Montoto, L. J.; Bunk, D. M.; Stein, S. E.; Tabb, D. L. QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* **2014**, *86*, 2497−2509.
(10) Bereman, M. S.; Johnson, R.; Bollinger, J.; Boss, Y.; Shulman, N.; MacLean, B.; Hoofnagle, A. N.; MacCoss, M. J. Implementation of statistical process control for proteomic experiments via LC MS/MS. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 581−587.
(11) Bittremieux, W.; Kelchtermans, P.; Valkenborg, D.; Martens, L.; Laukens, K. jqcML: An open-source Java API for mass spectrometry quality control data in the qcML format. *J. Proteome Res.* **2014**, *13*, 3484−3487.
(12) Walzer, M.; et al. qcML: An exchange format for quality control metrics from mass spectrometry experiments. *Mol. Cell. Proteomics* **2014**, *13*, 1905−1913.
(13) Bittremieux, W.; Willems, H.; Kelchtermans, P.; Martens, L.; Laukens, K.; Valkenborg, D. iMonDB: Mass spectrometry quality control through instrument monitoring. *J. Proteome Res.* **2015**, *14*, 2360−2366.
(14) Scheltema, R. A.; Mann, M. SprayQc: A real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **2012**, *11*, 3458−3466.
(15) Bereman, M. S. Tools for monitoring system suitability in LC MS/MS centric proteomic experiments. *Proteomics* **2015**, *15*, 891−902.
(16) Amidan, B. G.; Orton, D. J.; LaMarche, B. L.; Monroe, M. E.; Moore, R. J.; Venzin, A. M.; Smith, R. D.; Sego, L. H.; Tardiff, M. F.; Payne, S. H. Signatures for mass spectrometry data quality. *J. Proteome Res.* **2014**, *13*, 2215−2222.
(17) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44*, D447−D456.
(18) Zhang, B.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513*, 382−387.
(19) Edwards, N. J.; Oberti, M.; Thangudu, R. R.; Cai, S.; McGarvey, P. B.; Jacob, S.; Madhavan, S.; Ketchum, K. A. The CPTAC Data Portal: A resource for cancer proteomics research. *J. Proteome Res.* **2015**, *14*, 2707−2713.
(20) Slebos, R. J. C.; Wang, X.; Wang, X.; Zhang, B.; Tabb, D. L.; Liebler, D. C. Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Sci. Data* **2015**, *2*, 150022.
(21) Mayer, G.; et al. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. *Database* **2013**, *2013*, bat009−bat009.
(22) Bell, A. W.; et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **2009**, *6*, 423−430.
(23) Oveland, E.; Muth, T.; Rapp, E.; Martens, L.; Berven, F. S.; Barsnes, H. Viewing the proteome: How to visualize proteomics data? *Proteomics* **2015**, *15*, 1341−1355.
(24) Zimek, A.; Schubert, E.; Kriegel, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining* **2012**, *5*, 363−387.
(25) Kriegel, H.-P.; Kröger, P.; Schubert, E.; Zimek, A. LoOP: Local outlier probabilities. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, November 2−6, 2009; pp 1649−1652.
(26) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, May 15−18, 2000; pp 93−104.
(27) Micenková, B.; Dang, X.-H.; Assent, I.; Ng, R. T. Explaining outliers by subspace separability. *Proceedings of the 13th IEEE International Conference on Data Mining*, Dallas, Texas, December 7−10, 2013; pp 518−527.
(28) Naulaerts, S.; Meysman, P.; Bittremieux, W.; Vu, T. N.; Vanden Berghe, W.; Goethals, B.; Laukens, K. A primer to frequent itemset mining for bioinformatics. *Briefings Bioinf.* **2015**, *16*, 216−231.
(29) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22−30.
(30) McKinney, W. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, Austin, TX, June 28−July 3, 2010; pp 51−56.
(31) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.
(32) Bennett, K. L.; Wang, X.; Bystrom, C. E.; Chambers, M. C.; Andacht, T. M.; Dangott, L. J.; Elortza, F.; Leszyk, J.; Molina, H.; Moritz, R. L.; Phinney, B. S.; Thompson, J. W.; Bunger, M. K.; Tabb, D. L. The 2012/2013 ABRF Proteomic Research Group Study: Assessing longitudinal intralaboratory variability in routine peptide liquid chromatography tandem mass spectrometry analyses. *Mol. Cell. Proteomics* **2015**, *14*, 3299−3309.
(33) Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249−2264.
(34) Gika, H. G.; Theodoridis, G. A.; Earll, M.; Wilson, I. D. A QC approach to the determination of day-to-day reproducibility and robustness of LC-MS methods for global metabolite profiling in metabonomics/metabolomics. *Bioanalysis* **2012**, *4*, 2239−2247.
(35) Godzien, J.; Alonso-Herranz, V.; Barbas, C.; Armitage, E. G. Controlling the quality of metabolomics data: new strategies to get the best out of the QC sample. *Metabolomics* **2015**, *11*, 518−528.

DOI: 10.1021/acs.jproteome.6b00028
J. Proteome Res. XXXX, XXX, XXX−XXX