REVIEW

Computational quality control tools for mass spectrometry proteomics

Wout Bittremieux[1,2], Dirk Valkenborg[3,4,5], Lennart Martens[6,7,8] and Kris Laukens[1,2]

[1] Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[2] Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp/Antwerp, University Hospital, Edegem, Belgium
[3] Flemish Institute for Technological Research (VITO), Mol, Belgium
[4] CFP, University of Antwerp, Antwerp, Belgium
[5] I-BioStat, Hasselt University, Diepenbeek, Belgium
[6] Medical Biotechnology Center, VIB, Ghent, Belgium
[7] Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[8] Bioinformatics Institute Ghent, Ghent University, Zwijnaarde, Belgium

As mass-spectrometry-based proteomics has matured during the past decade, a growing emphasis has been placed on quality control. For this purpose, multiple computational quality control tools have been introduced. These tools generate a set of metrics that can be used to assess the quality of a mass spectrometry experiment. Here we review which types of quality control metrics can be generated, and how they can be used to monitor both intra- and inter-experiment performances. We discuss the principal computational tools for quality control and list their main characteristics and applicability. As most of these tools have specific use cases, it is not straightforward to compare their performances. For this survey, we used different sets of quality control metrics derived from information at various stages in a mass spectrometry process and evaluated their effectiveness at capturing qualitative information about an experiment using a supervised learning approach. Furthermore, we discuss currently available algorithmic solutions that enable the usage of these quality control metrics for decision-making.

Received: June 30, 2016
Revised: July 28, 2016
Accepted: August 19, 2016

Keywords:
Bioinformatics / Mass spectrometry / Quality control

1 Introduction

In the past decade, mass-spectrometry-based proteomics has evolved into an extremely powerful analytical technique to identify and quantify proteins in complex biological samples. This high-throughput approach can yield a considerable volume of complex data for each experiment. As it has matured, over the last few years a growing emphasis has been placed on quality assurance (QA). This attention on QA is of the utmost importance to safeguard confidence in the acquired results: in cases where this has been lacking mass spectrometry proteomics has sometimes suffered from exaggerated claims [1, 2]. To anticipate this evolution, a shift to "quality by design" is now taking place [3]. This means that the "designing and developing formulations and manufacturing processes ensure a predefined product quality." As such, QA

Correspondence: Prof. Kris Laukens, Campus Middelheim—M.G.111 Middelheimlaan 1, 2020 Antwerpen, Belgium
E-mail: kris.laukens@uantwerpen.be
Fax: +32 (0) 3 265 37 77

Abbreviations: **cRAP**, common repository of adventitious proteins; **DDA**, data-dependent acquisition; **DIA**, data-independent acquisition; **GUI**, graphical user interface; **ID**, identification; **iMonDB**, Instrument monitoring database; **MBR**, match-between-runs; **NIST**, National Institute of Standards and Technology; **PDF**, portable document format; **PNNL**, Pacific Northwest National Laboratory; **PSM**, peptide-spectrum-match; **PTXQC**, proteomics quality control; **QA**, quality assurance; **QC**, quality control; **SimpatiQCo**, simple automatic quality control; **SProCoP**, statistical process control in proteomics; **XML**, extensible markup language

Colour Online: See the article online to view Fig. 3 in colour.

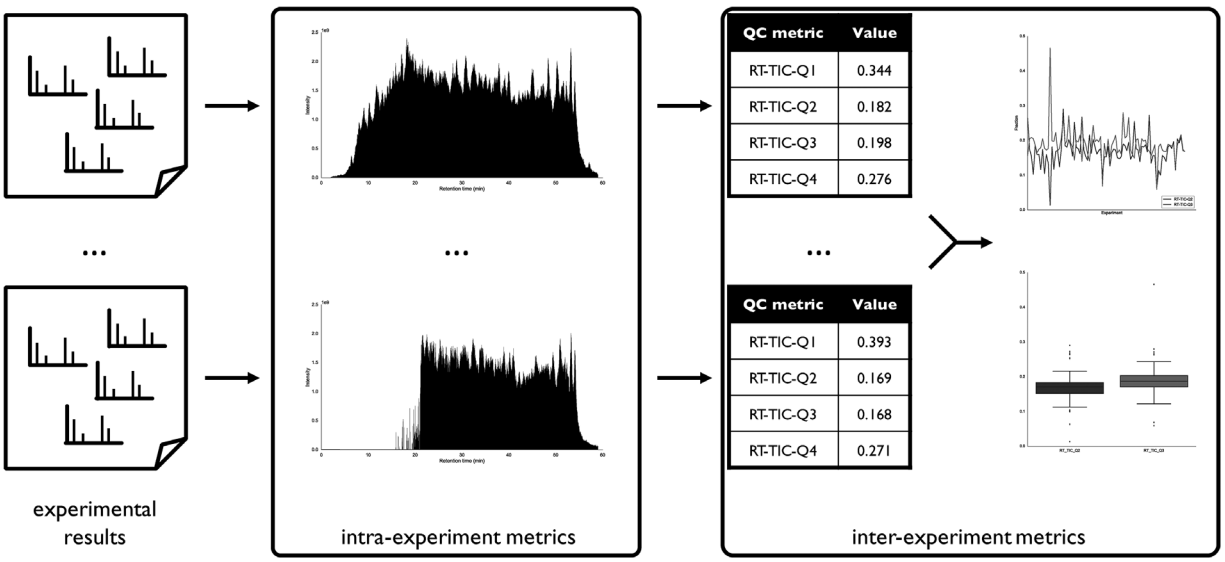

Figure 1. Intra-experiment metrics evaluate the quality of a single experiment, whereas inter-experiment metrics can be used to compare the quality of multiple experiments.

consists of multiple aspects of which quality control (QC) is an essential component, but other elements such as a careful experimental design [4–6] are equally vital.

Whereas the experimental design has to be established prior to the initiation of an experiment, QC takes place while or after the experimental results are obtained. Nonetheless, QC and experimental design should not be discussed in isolation, as they are interwoven. For example, a QC sample can consist of a single peptide, a single protein digest, or a complex lysate, and this decision influences the type of QC metric(s) that can be investigated [7–9]. Furthermore, one has to decide how many QC runs to include in the experiment and to what extent and in which order these QC runs are interleaved with the biological samples under consideration. The goal of QC is then to leverage the experimental set-up to comprehend how well an instrument performs and how confident the results from the experiments are.

Related to the experimental design and based on the type of performance we want to monitor, there are multiple approaches to QC. A typical example consists of the use of QC samples with a simple sample content interleaved between the biological samples. The interesting aspect of such QC samples is that they have a controlled, limited, and known sample content. They are typically measured on a frequent basis, which allows us to extract periodic information on the performance of the mass spectrometer. Of course, to understand this performance expressive QC metrics that provide information indicative of the quality of the experimental results need to be derived. Some straightforward and commonly used QC metrics include the number of identifications or the sequence coverage. Although these metrics give a global view of the performance, they do not allow us to pinpoint specific elements of the workflow where a failure might have arisen.

Instead, more granular QC metrics providing information on the chromatography, the ion signal, the spectrum acquisition, etc., might be used.

Over the years dozens of QC metrics have been proposed, generated by a range of bioinformatics tools. In this paper, we will list the main QC tools and explain their use cases and capabilities. Furthermore, we will provide an empirical assessment of which type of QC metrics is the most adequate in detecting low-quality experiments.

1.1 QC metrics

We can primarily distinguish QC metrics based on whether they represent information about a single experiment, or about multiple experiments, as illustrated in Fig. 1.

Intra-experiment metrics give information about a single experiment and are computed at the level of individual scans or identifications. These metrics show the evolution of a specific measure over the experiment run time, such as, for example a chromatogram of the total ion current (TIC) over the retention time, or the mass accuracy of the identified spectra.

Inter-experiment metrics, on the other hand, assess a specific part of the quality of an experiment using a single measurement for the whole experiment. These values can subsequently be compared for multiple experiments, for example through a longitudinal analysis to evaluate the performance over time. Often an intra-experiment metric can be converted to an inter-experiment metric through summarization. This is illustrated in Fig. 1, where a TIC chromatogram enables the assessment of the chromatographic performance by visualizing the intensity distribution over the retention time. Using summary statistics this continuous information can be converted to inter-experiment metrics detailing the fraction of

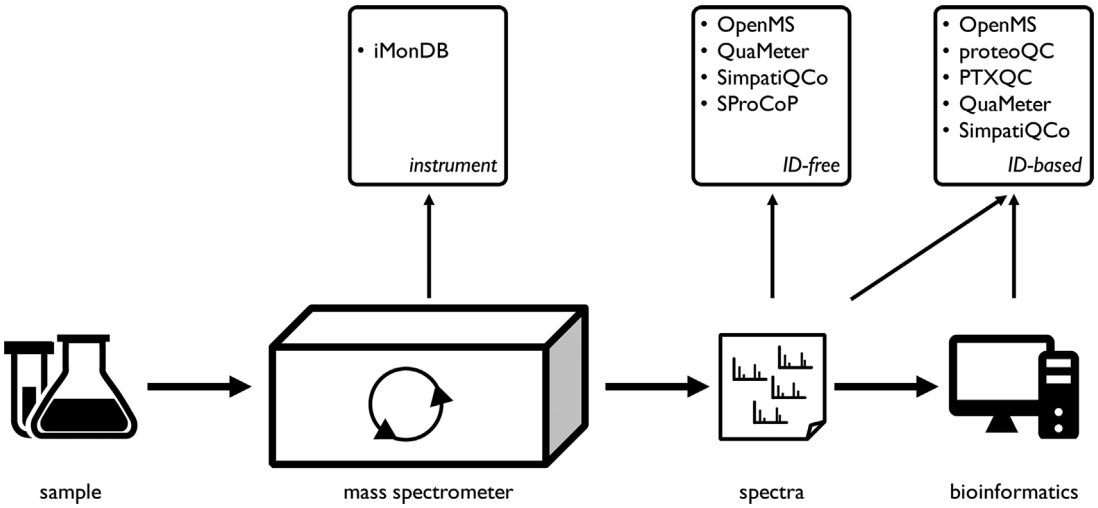

Figure 2. QC tools can capture qualitative information at different stages of a mass spectrometry experiment. For each type of QC metrics the representative tools are listed.

the total retention time that was required to accumulate a certain amount of the TIC, which gives a high-level assessment at the experiment level of the chromatographic stability.

To compare inter-experiment metrics, multiple observations for different experiments are required. Therefore, QC tools that analyze these metrics usually include a database back-end for the persistent storage of historical data. On the other hand, intra-experiment metrics can be computed from only a single experiment and there is no comparison with external data. As a result, QC tools that exclusively generate intra-experiment metrics are generally easier to set up, as no external data storage needs to be provided. Because the use cases and requirements differ between these two types of tools, we will further make a distinction between tools that generate metrics for individual experiments, tools that compare a limited group of experiments and do not necessarily require a complex back-end for data storage, and tools for longitudinal tracking that store QC data for a large number of experiments.

A second distinction between various metrics can be made based on from which stage in a mass spectrometry workflow they represent the quality of the system. As shown in Fig. 2, we can distinguish between instrument metrics, identification-free (ID-free) metrics, and identification-based (ID-based) metrics.

ID-free metrics and ID-based metrics are similar in the sense that they are both computed from the spectral results. *ID-free* metrics are derived solely from the spectral results, i.e. from the raw spectral data directly generated by the mass spectrometer. These metrics aim to capture information over the whole mass spectrometry workflow and include for example the shape of the peaks or the course of TIC detailing the chromatography, the number of MS1 and MS2 scans or the scan rate detailing the spectrum acquisition, or the charge state distribution detailing the ionization. The advantage of ID-free metrics is that they are generated directly from the raw spectral data, which makes it possible to instantly generate these metrics as soon as a mass spectrometry run has been completed.

ID-based metrics are derived from the spectral results as well, but they combine these data with subsequently obtained identification results. Examples include aforementioned metrics such as the number of identifications in terms of peptide-spectrum matches (PSMs), peptides, or proteins; or the sequence coverage for a known sample. Other detailed metrics can be computed as well, for example by comparing the difference in retention time for similar identifications to assess the chromatographic stability, the number of spectra identified as the same peptide to measure the dynamic sampling, or by linking information similar to the ID-free metrics with the identification results. Compared to ID-free metrics, the computation of ID-based metrics is somewhat more involved because it additionally requires the identifications results. Furthermore, the computation of ID-based metrics can be negatively influenced by suboptimal identification settings. However, in general the inclusion of identifications can provide a more detailed qualitative assessment of the experimental results.

Finally, *instrument* metrics do not look at the spectral data but derive information directly from instrument readouts. These are typically very sensitive, low-level metrics, such as the status of the ion source, the vacuum, or a turbo pump, depending on the type of instrument. An advantage of instrument metrics is that they directly indicate which part of the instrument is outside its normal range of operation. This facilitates troubleshooting and can be a driver for maintenance scheduling. On the other hand, these metrics cannot be directly related to the experimental results, instead they provide a secondary source of QC information. Furthermore, instrument metrics are instrument- and vendor-specific, and

Table 1. Overview of the discussed QC tools and their main characteristics

Tool	Interface	Operating system	Experiment type	Instrument	ID-free	ID-based	Website
QuaMeter [17, 18]	Command-line	Windows, Linux	discovery DDA	×	√	√	http://proteowizard.sourceforge.net/
OpenMS [23]	KNIME	Cross-platform	discovery DDA	×	√	√	http://www.openms.de/
proteoQC [28]	R	Cross-platform	discovery DDA	×	×	√	http://bioconductor.org/packages/proteoQC
PTXQC [31]	R	Windows, cross-platform	discovery & quantification DDA	×	×	√	https://github.com/cbielow/PTXQC
SProCoP [34]	Skyline	Windows	SRM & PRM	×	√	×	http://proteome.gs.washington.edu/software/skyline/tools/sprocop.html
SimpatiQCo [38]	web	Windows	discovery DDA	×	√	√	http://ms.imp.ac.at/?goto = simpatiqco
iMonDB [39]	GUI	Windows	any	√	×	×	https://bitbucket.org/proteinspector/imondb/

are typically not included in open file formats such as the mzML format [10].

Each distinct type of metric can give a different view on the quality of the data. However, not all metrics are always applicable; often metrics are especially relevant for a particular type of sample. For example, monitoring the sequence coverage is mostly applicable when using samples that contain a single protein digest, whereas the number of protein identifications is applicable to samples that consist of a complex lysate. Additionally, the type of experiment also plays an important role. For example, the number of identifications is very relevant for a discovery experiment, but less so for a targeted experiment. In contrast, instrument metrics are largely agnostic to the type of experiment and the sample content, but they can significantly vary between different instrument models and vendors.

2 QC tools

In recent years, QC has become a key focus of attention in academic, industrial, and governmental proteomics laboratories. This trend is exemplified (and possibly driven) by the numerous QC tools that have been developed over the past few years. Initial work by Rudnick et al. [11] described for the first time how computational QC metrics can be used to objectively assess the quality of a mass spectrometry proteomics experiment. Whereas previously QC was mostly performed manually by monitoring a few key measurements, this work showed how a comprehensive set of QC metrics can be used to thoroughly investigate the system performance. A set of 46 mainly ID-based metrics was defined and implemented in a pipeline of Perl programs by researchers at the National Institute of Standards and Technology (NIST), called NIST MSQC. This set of metrics has since then been reimplemented in several lab-specific data processing pipelines. Support for NIST MSQC itself has been discontinued in early 2016 and the original implementation is no longer available, but several of the reimplementations remain under active development.

It has been demonstrated that computational QC metrics provide objective criteria that can accurately capture the quality of a mass spectrometry experiment, and there has been a proliferation of tools that can compute such metrics. Here, we will detail the primary tools, their characteristics, and their usage. Table 1 provides an overview of the discussed tools.

2.1 Tools evaluating individual experiments

2.1.1 QuaMeter

QuaMeter was initially developed as a user-friendly and open-source alternative to NIST MSQC. NIST MSQC consisted of a graphical user interface (GUI) wrapper around multiple individual tools and scripts with various inter-dependencies, which resulted in a complex pipeline. Additionally, some elements of this pipeline could only be modified to a limited extent. NIST MSQC could exclusively compute metrics from Thermo Scientific raw files, and only supported three search engines to provide identifications: the NIST MSPepSearch or the SpectraST [12] spectral library search engines, or the OMSSA [13] sequence database search engine. These limitations restricted the applicability of NIST MSQC.

Instead, QuaMeter consists of a single multi-platform command-line application that is able to compute QC metrics from raw files originating from instruments produced by multiple vendors. Using the ProteoWizard [14] library it is able to read spectral data stored in a wide variety of vendor-specific raw files (restricted to the Windows platform) and open standard file formats, such as mzML [10]. Furthermore, it can utilize identification results produced by any search engine in the standard mzIdentML [15] or pepXML format through external processing using IDPicker [16].

The initial QuaMeter version [17] computed a set of 42 ID-based QC metrics equivalent to those defined by Rudnick et al. [11]. In a subsequent version QuaMeter improved upon this by also including functionality to compute a set of 45 ID-free QC metrics [18]. Both sets of metrics are inter-experiment summary metrics, although the output is exported to simple tab-delimited text files, so the visualization and analysis thereof has to be done using external software or code scripts. Without advanced visualization or analysis functionality QuaMeter focuses solely on computing QC metrics. Especially the set of ID-free metrics, which requires only the spectral data, can very easily be computed. For the set of ID-based metrics some prior processing of the identification results by IDPicker is required, which can make this process slightly more cumbersome. Only a limited configuration is required, and through the command-line functionality the computation can easily be automated. This makes QuaMeter a powerful tool that computes an extensive set of inter-experiment QC metrics.

2.1.2 OpenMS

OpenMS is a comprehensive open-source software library that offers a wide range of algorithms and tools for mass-spectrometry-based proteomics and metabolomics [19]. It consists of various small processing tools that can be used to construct complex analysis workflows [20, 21]. These workflows can be designed visually using the KNIME workflow engine [22], where each tool functions as an individual node in the workflow.

The various OpenMS nodes can be used to build complex QC pipelines [23]. The provided QC nodes can compute a set of intra-experiment metrics, consisting of both ID-free and ID-based metrics. OpenMS supports a range of search engines to generate identifications for the ID-based metrics, for which there exist specific nodes, including Mascot, MS-GF+ [24], Myrimatch [25], OMSSA [13], and X!Tandem [26]. Example QC metrics include the number of spectra (identified or otherwise), peptides, and proteins; mass accuracy statistics; and the mass over charge and retention time acquisition ranges. These metrics are complemented by various plots that provide further details, such as a TIC chromatogram, a histogram of the mass accuracy of the identified peptides, or a histogram of the charge distribution of the detected ion features. OpenMS exports this information to an Extensible Markup Language-based (XML) qcML file [23], which can be visualized in a web browser through an embedded stylesheet, or to a Portable Document Format (PDF) report.

Due to the wealth of algorithms and tools that are available in the OpenMS software library, the provided QC workflows can potentially be easily extended to compute additional metrics. Furthermore, there is no need to be restricted to algorithms natively provided by OpenMS, as the available functionality can easily be extended through custom nodes, for example by using the built-in support for the R statistical programming language [27]. This makes it possible to build granular workflows and achieve a very fine-grained control, although expert knowledge of the OpenMS ecosystem and the KNIME environment is recommended to do so. The constructed workflows can subsequently be exported and shared. Both OpenMS and KNIME are cross-platform tools, ensuring the universal applicability of these workflows.

2.2 Tools comparing groups of experiments

2.2.1 proteoQC

The proteoQC package [28] for the R programming language [27, 29] can be used to generate a HTML report detailing the experimental quality. Prior to executing proteoQC the experimental design has to be specified by configuring each spectral data file representing a sample as belonging to a specific fraction, technical replicate, and biological replicate. The generated QC report contains intra-experiment metrics for each individual sample, as well as aggregated information to compare samples at the level of their fractions, technical replicates, and biological replicates.

To generate a set of intra-experiment ID-based metrics for each sample, proteoQC uses the rTANDEM package [30] to interface the X!Tandem [26] sequence database search engine in R to provide identification results. For each sample some individual metrics and QC plots are generated, such as a breakdown of the precursor ion charge states, the mass accuracy, information on the number of spectra and peptides that were used to identify distinct proteins during protein inference, etc. Furthermore, when identifying the data proteoQC automatically adds the common Repository of Adventitious Proteins (cRAP, http://www.thegpm.org/crap/) database to the user-provided protein database. The cRAP database contains contaminants such as common laboratory proteins, like trypsin, or contaminants transfered through dust or contact, like keratin, and proteoQC reports which of these contaminants were detected in the samples. Additionally, proteoQC reports on the reproducibility of the results by comparing the number of identified spectra, peptides, and proteins per fraction, technical replicate, and biological replicate, and their overlap between the replicates.

By incorporating the experimental design proteoQC can make informed comparisons between individual samples, which provides QC information on an additional level. Furthermore, proteoQC is fully cross-platform within the popular R programming language.

However, as the QC pipeline has to be configured programmatically, some R experience is recommended to utilize proteoQC.

2.2.2 PTXQC

Proteomics Quality Control (PTXQC) [31] is an R-based quality control pipeline for MaxQuant [32], a highly popular

software suite for quantitative proteomics. Like MaxQuant, PTXQC supports a wide range of quantitative proteomics workflows, including stable isotope labeling with amino acids in cell culture (SILAC), tandem mass tags, and label-free quantification. After initial processing of the spectral data by MaxQuant, PTXQC uses the MaxQuant output results to compute various QC metrics. PTXQC requires as input the custom text files generated by MaxQuant and the MaxQuant configuration settings, and hence cannot be used to process any other type of data. As PTXQC is written in the R programming language, it is fully cross-platform. Additionally, easy drag-and-drop functionality to execute the QC analyses is provided for the Windows operating system.

PTXQC produces an extensive report that contains a set of 24 intra- and inter-experiment metrics. These metrics are divided into four categories corresponding to the specific MaxQuant output source the metrics are derived from: "ProteinGroups", "Evidence", "Msms", and "MsmsScans". The metrics cover a wide range of information, including the intensity of the detected features and peptides, the potential presence of contaminants, the mass accuracy of the identified peptides and fragments, the number of missed cleavages detailing the enzyme specificity, and the number of identified peptides and proteins. Other metrics are specifically related to the MaxQuant "match-between-runs" (MBR) [33] functionality. MBR aligns the retention times of multiple runs and transfers their identifications across features that have the same accurate mass and a similar retention time, providing more data for the downstream quantification of proteins. PTXQC assesses the MBR performance by evaluating the retention time alignment and by checking whether the identification transfer seems correct. All of these metrics are then visualized and compared between the different raw files that constitute the considered MaxQuant project using detailed figures. Furthermore, each of the metrics is converted to an individual score for each experiment using automated scoring functions. Most of these scores are absolute scores generated by comparing the observation to a threshold, for example such as whether the number of detected contaminants is too excessive, or generated by evaluating a specific characteristic of the observation, for example such as the extent to which the mass deviations are centered around zero. Other scores are computed for a single raw file using the other raw files as a reference, for example by comparing the number of missed cleavages in each individual raw file to the average number of missed cleavages. Finally, some other scores are evaluated relative to settings extracted from MaxQuant, such as the mass accuracy compared to the width of the precursor mass window. All these scoring functions generate inter-experiment metrics that are used to compare the quality of the different experiments. Usefully, PTXQC provides a heatmap overview of the inter-experiment metrics, which yields an assessment of the quality at a glance and facilitates pinpointing the low-performing experiments.

Although PTXQC can exclusively be used to analyze MaxQuant results, through this tight integration it is able to compute some highly relevant and specialized QC metrics. These metrics do not only assess the quality of the spectral data, but also provide information on the subsequent bioinformatics processing by MaxQuant. Furthermore, the addition of a high-level heatmap at the start of the report is very useful to get a quick overview of the quality, after which the more detailed visualizations can be employed to further investigate potential problems.

2.2.3 SProCoP

Statistical Process Control in Proteomics (SProCoP) [34] is a QC script written in R [27] that can be used as a plugin [35] for the popular Skyline [36] tool for targeted proteomics. SProCoP applies well-established statistical process control techniques such as the Shewhart control chart and the Pareto chart. The purpose of a Shewhart control chart is to track performance over time and identify outliers that deviate excessively from the expected behavior. Further, the Pareto chart is a combination of a bar and line graph, which displays the number of deviating measurements for each metric along with its cumulative percentage, and provides feedback on which metrics are more variable and may require attention.

Using these statistical process control techniques SProCoP monitors the performance of five inter-experiment QC metrics based on targeted peptides present in QC samples with a known sample content or spiked into real samples: signal intensity, mass measurement accuracy, retention time reproducibility, peak full width at half maximum, and peak symmetry. Measurement thresholds are defined empirically based on a reference set of samples with a known good quality, after which the performance of other samples in the Skyline project can be investigated.

Through its integration with Skyline SProCoP is vendor-independent and can be used for a wide range of targeted and discovery workflows. Additionally these statistical process control techniques are available online (http://www.qcmylcms.com/) and have been implemented in the Panorama [37] repository for targeted proteomics from Skyline. Panorama AutoQC is a utility application that monitors for new data files and automatically invokes Skyline to process the data. The QC metrics are stored in Panorama and the statistical process control charts similar to SProCoP can be visualized through the Panorama web application.

2.3 Tools for longitudinal tracking

2.3.1 SimpatiQCo

SIMPle AuTomatIc Quality COntrol (SimpatiQCo) [38] not only computes various QC metrics, it also stores and visualizes these metrics for longitudinal monitoring of the system performance. It uses a PostgreSQL database as back-end, and

an Apache webserver to provide a web-based front-end for configuration and visualization.

SimpatiQCo can compute QC metrics from a limited selection of Thermo Scientific and SCIEX instruments. Raw files from these instruments can be uploaded to the web server manually, or can be added automatically through a "hot folder" that is monitored continuously for new raw files. These raw files are then submitted to a linked Mascot server for peptide identifications. Next, SimpatiQCo calculates a range of ID-free and ID-based QC metrics such as the number of MS1 and MS2 scans, the number of identified PSMs and proteins, the TIC, and information on lock masses (if applicable). Further, specific peptides and proteins can be investigated in detail using metrics such as the peak area and width and the elution time of peptides of interest, and the protein sequence coverage. For each QC metric the range of acceptable values is learned based on the historical observations using robust statistical measures to take outlying values into account. This information is then displayed in the metric plots using a color-coded background band to highlight deviating system performance. Further, external messages can be entered manually, for example pertaining to instrument maintenance. These messages will be superimposed on the metric plots to relate the external events to the evolution of the metrics.

SimpatiQCo consists of a number of different components, such as the database, the web server, and various processing tools. These components need to be installed individually, and although a step-by-step installation guide is available online, this complicated process is not recommended for novice users. Furthermore, not all of the configuration can be done through the graphical web-based client. For example, to process raw files these must be able to be linked to a specific instrument. Unfortunately, an instrument definition can only be created by manually adding a record in the corresponding table of the PostgreSQL database.

SimpatiQCo is a powerful tool to track system performance over time, albeit with some technical limitations. Namely, SimpatiQCo is only able to process raw files generated on a limited number of instrument models and only supports the commercial Mascot search engine for peptide identifications.

2.3.2 iMonDB

Unlike the previous tools the [39] Instrument MONitoring DataBase (iMonDB) does not compute metrics from the spectral results, but extracts instrument metrics from the raw files. The iMonDB uses a MySQL database to store its information. This database acts as a server, with two separate standalone GUI applications that can connect to the database as clients, each with a specific task: the iMonDB Collector processes raw files and stores the instrument metrics in the database, whereas the iMonDB Viewer retrieves the information from the database and visualizes it.

The iMonDB supports a wide range of instruments manufactured by Thermo Scientific, although it does not support other instrument vendors. Prior to extracting instrument metrics from a raw file, a corresponding instrument definition has to be created. This can be done through the iMonDB Collector, which allows the full configuration through its graphical user interface. Further, extraction of the instrument metrics can be done manually through the GUI, or can be done through command-line functionality provided by the iMonDB Collector. This command-line functionality can be used to automatically run the iMonDB Collector using an external scheduling tool, such as the native operating system scheduler.

The behavior over time of the metrics for each instrument can be viewed using the iMonDB Viewer. Similar to functionality provided by SimpatiQCo it is possible to add additional information pertaining to external events and show this on the metric plots to link this to the evolution of the metrics. It is also possible to export a PDF file of the external events for reporting purposes.

A unique aspect of the iMonDB is that this is the only tool that is able to systematically analyze instrument metrics. The advantage of these instrument metrics, which provide information at the lowest level, is their high sensitivity, which makes it possible to detect emerging defects in a timely fashion. However, because these metrics are instrument-dependent they are usually not retained during conversion to open formats, such as mzML [10]. Due to this limitation the iMonDB needs to work with vendor-specific raw files directly, which is currently limited to Thermo Scientific raw files. Furthermore, there is a multitude of instrument metrics that are extracted, which makes it hard to comprehend which metrics are most useful to monitor systematically, even for expert users. Nevertheless, these instrument metrics can be very useful to detect malfunctioning instrument elements before these have a deleterious effect on the experimental results, preventing potential loss of valuable sample content.

2.4 Other tools

As mentioned previously, NIST MSQC [11] was the first tool that generated computational QC metrics, although it was recently retired in early 2016.

Metriculator [40] is a web-based tool for storing and visualizing QC metrics longitudinally. However, Metriculator does not compute QC metrics directly but critically depends upon an embedded version of NIST MSQC. Unfortunately, the installation process for Metriculator is not very straightforward; it has many Ruby dependencies whose installation might fail, and which are presently outdated or even no longer supported.

LogViewer [41] is a simple visualization tool that presents a set of 11 instrument metrics, such as MS1 and MS2 ions injection times, and ID-free metrics, such as the charge state and mass distributions. As input it uses log files from Thermo

instruments exported by RawXtract [42], which has been deprecated presently.

A different approach is used by SprayQc [43]. Whereas the other discussed tools compute QC metrics post-acquisition, SprayQc directly interfaces with peripheral equipment to continuously monitor its performance. SprayQc is able to automatically track the stability of the electrospray through computer vision, the status of the liquid chromatography pumps, the temperature of the column oven, and the continuity of the data acquisition. In case a malfunctioning is detected SprayQc can automatically take corrective actions and warn the instrument operator. This is a valuable approach to minimize the loss of precious sample content and provide early notifications, and it can complement the other QC tools that provide a post-acquisition quality assessment.

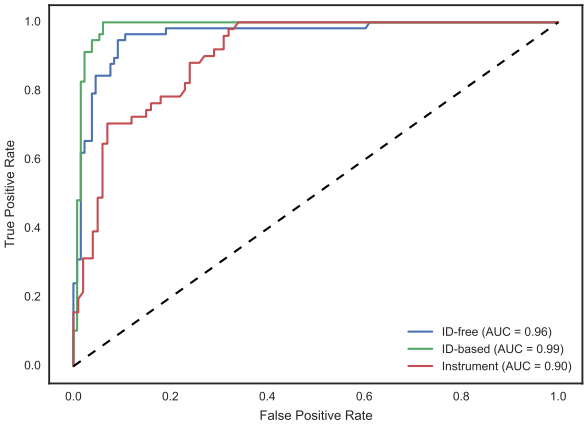

Figure 3. ROC curve showing the classification performance for the Velos Orbitrap instrument type. ROC curves for the other instrument types indicate similar results (data not shown).

3 Metrics evaluation

We compared various sets of metrics to assess their effectiveness in expressing the quality of a mass spectrometry proteomics experiment. Typically this is not a straightforward task because, as we have reviewed in the previous sections, each QC tool has its own characteristics and requirements, and use cases can vary as some tools are specific to certain experimental workflows and sample types. Meanwhile most tools also represent some of their QC information through visualizations. Although these quickly provide useful insights for human users, this data is not suitable for an objective, automatic comparison.

To compare different types of metrics we used the set of instrument metrics computed by the iMonDB [39], the set of ID-free metrics computed by QuaMeter [18], and the set of ID-based metrics as identified by Rudnick et al. [11]. These sets of metrics are very comprehensive and all of these inter-experiment metrics can readily be used to compare experiments to each other. To be able to determine whether or not these metrics can capture qualitative information about an experiment, we used a public dataset for which the quality of the experiments is known. The dataset consists of a number of complex QC LC-MS runs performed on several different instruments at the Pacific Northwest National Laboratory (PNNL) [44]. Each sample had an 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, being either "Exactive", "LTQ IonTrap", "LTQ Orbitrap", or "Velos Orbitrap", with each of these instrument groups consisting of multiple individual instruments. We refer to the original publication by Amidan et al. [44] for further information on the experimental procedures and the dataset details.

This public dataset already contains the precomputed set of ID-free metrics by QuaMeter and the set of ID-based metrics by SMAQC (the PNNL in-house reimplementation of the NIST MSQC metrics defined by Rudnick et al. [11]; https://github.com/PNNL-Comp-Mass-Spec/SMAQC). We further used the iMonDB to compute the set of instrument metrics. To this end all experimental raw files, precomputed QC metrics, and the expert quality annotations were retrieved from the PRIDE database [45].

To quantify the expressiveness of these three sets of metrics, each capturing a different type of QC information, we employed a binary classifier. As the quality of the experiments was manually assessed by expert instrument operators, this labeling can be used as the ground truth to train the classifier. We used the acceptable experiments, with their quality designated as either "good" or "ok", as the positive class, and the inferior experiments, with their quality designated as "poor", as the negative class. When given an experiment represented by its QC metrics, the classification task consists of correctly predicting the experiment's quality. Prior to training the classifier we removed redundant features that have a very low variance and we rescaled the features robust to outliers by centering by the median and scaling by the interquartile range. Next, for each separate instrument type we trained a random forest classifier, for which we split the data into 65–35% training and testing subsets that are equally stratified according to their quality labels. This classifier has been coded in Python and uses the random forest implementation from scikit-learn [46], along with functionality provided by NumPy [47] and pandas [48]. The code is available as open source at https://bitbucket.org/proteinspector/qc-evaluation/.

As illustrated by the ROC curve in Fig. 3 all three types of QC metrics are adept at discriminating high-quality experiments from low-quality experiments. This shows that all of the different tools can give us valuable insights into the quality of an experiment, and that information captured at various different stages of the mass spectrometry process should be investigated. ID-based metrics slightly outperform ID-free metrics, most likely because the ID-based metrics can employ additional information provided by the identifications. This difference is minimal however, which is perhaps not

surprising as both types of metrics take similar properties of the spectra into account. This reinforces previous research which showed that ID-based metrics are not significantly influenced by slight differences in the identifications, such as when using an alternative search engine [17]. This also shows the excellent efficacy of ID-free metrics in objectively evaluating the quality based solely on spectral information. Because ID-based metrics require additional computational steps to obtain the identifications, whereas ID-free metrics can be directly computed from the spectral results, ID-free metrics might be preferred if a speedy quality assessment is required. In contrast, instrument metrics perform a little worse at correctly identifying low-quality experiments. This is likely because they are only secondary results that are not always directly related to the data quality. Nevertheless, these metrics still have merit as they do not depend on a specific type of experiment or sample content, but are applicable on all occasions. Furthermore, by combining the individual classifiers for the various types of metrics in an ensemble classifier a further performance gain can be achieved because the different types of metrics each provide a complementary view on the quality.

4 Using QC metrics for decision-making

As tools for computational QC have proliferated in recent years, the challenge in this field is now shifting from the computation of QC metrics toward informed decision-making based on these metrics. However, interpreting these metrics is not trivial. First, considerable domain knowledge is required to understand what each metric signifies. Second, the metrics form a high-dimensional data space, which complicates their analysis. Different elements in a mass spectrometry workflow do not function in isolation but instead influence each other, which has to be taken into account while analyzing metrics representing information about these elements. Therefore, univariate approaches are generally insufficient; instead multivariate approaches that can deal with the high-dimensional data space should be preferred, while also taking the curse of dimensionality into account [49].

To this end Wang et al. [18] have developed a robust multivariate statistical toolkit to interpret QC metrics. They have used a PCA transformation to reduce the data to a low-dimensional approximation, in which they were able to successfully detect outlier low-quality experiments based on pairwise dissimilarities. Furthermore, they developed an ANOVA model which enabled them to identify whether the observed variability was attributable to lab-dependent factors, batch effects, or biological variability. Such work driving the understanding of QC metrics is highly valuable, and these analyses have been applied to great effect for multiple studies. For example, it was used to assess the quality of the experimental results for various studies conducted by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium [50–52].

Similar work was done by Bittremieux et al. [53], who applied unsupervised outlier detection to identify low-quality experiments. Subsequently they used a specialized outlier interpretation technique to determine which QC metrics mostly contributed to the decrease in quality. The advantage of this approach is that all QC metrics are used to identify low-quality experiments, unlike when using a dimensionality reduction, such as PCA, which discards some of the information. Meanwhile, the advanced outlier interpretation pinpointing the most relevant QC metrics can yield actionable information for domain experts to optimize their experimental set-up.

Whereas these previous analyses used unsupervised techniques, Amidan et al. [44] trained a supervised classifier to discriminate low-quality experiments from high-quality experiments. A supervised approach will generally perform better than an unsupervised approach but will require initial training. Furthermore, a supervised classifier might have to be retrained to adapt it to data generated by a different instrument or in a different laboratory. Amidan et al. [44] have expended significant effort in manually annotating the quality of over a thousand experiments to generate training data, which allowed them to build a highly performant logistic regression classifier.

These analyses are extremely valuable, as they allow us to achieve a deeper understanding of the mass spectrometry processes and the properties of what makes a high-quality experiment. These algorithmic approaches provide a thorough quality assessment of the spectral data, which enforces informed decision-making, and which has the potential to automatically drive the spectral acquisition in the future.

5 Conclusion

We have given an overview of the available computational tools to generate QC metrics for mass-spectrometry-based proteomics. These tools enable assessing the performance of the experimental set-up and detecting unreliable results. These are essential requirements to inspire confidence in the experimental results, which will prove to be a crucial step in the maturation of proteomics technologies, and which will allow us to for example routinely apply these technologies into a clinical setting [3, 54]. Another potential application where an accurate assessment of the data quality is paramount, is in the reuse of public data [55–58]. As public data repositories keep expanding and the potential for data reuse grows, we envision that data submissions to public repositories will soon have to be accompanied by QC parameters at the time of submission, or will have a standard set of QC metrics calculated automatically after submission [58].

Finally, most current QC tools are limited to the typical use case of bottom-up data-dependent acquisition (DDA) discovery experiments, and their QC metrics often cannot be directly translated to other types of experiments. Less research has been done on QC for other types of workflows, such as data-independent acquisition (DIA) [59] or top-down

proteomics [60], or even related mass-spectrometry-based domains, such as metabolomics [61]. In the next few years we will likely see the efforts on QC expanded to these types of workflows as well, which will further bolster the diverse and powerful mass spectrometry ecosystem.

W.B., D.V., L.M., and K.L. acknowledge the support of the VLAIO grant "InSPECtor" (IWT project 120025).

The authors have declared no conflict of interest.

6 References

[1] Baggerly, K. A., Morris, J. S., Edmonson, S. R. Coombes, K. R., Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst.* 2005, *97*, 307–309.

[2] Ezkurdia, I., Calvo, E., Del Pozo, A., Vázquez, J. et al., The potential clinical impact of the release of two drafts of the human proteome. *Expert Rev. Proteomics* 2015, *12*, 579–593.

[3] Tabb, D. L., Quality assessment for clinical proteomics. *Clin. Biochem.* 2013, *46*, 411–420.

[4] Hu, J., Coombes, K. R., Morris, J. S., Baggerly, K. A., The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct. Genomic Proteomic* 2005, *3*, 322–331.

[5] Cairns, D. A., Statistical issues in quality control of proteomic analyses: good experimental design and planning. *Proteomics* 2011, *11*, 1037–1048.

[6] Maes, E., Kelchtermans, P., Bittremieux, W., De Grave, K. et al., Designing biomedical proteomics experiments: state-of-the-art and future perspectives. *Expert Rev. Proteomics* 2016, *13*, 495–511.

[7] Paulovich, A. G., Billheimer, D., Ham, A.-J. L., Vega-Montoto, L. et al., Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* 2010, *9*, 242–254.

[8] Köcher, T., Pichler, P., Swart, R., Mechtler, K., Quality control in LC-MS/MS. *Proteomics* 2011, *11*, 1026–1030.

[9] Bereman, M. S., Tools for monitoring system suitability in LC MS/MS centric proteomic experiments. *Proteomics* 2015, *15*, 891–902.

[10] Martens, L., Chambers, M., Sturm, M., Kessner, D. et al., mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, 2011, *10*, R110.000133–R110.000133.

[11] Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V. et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* 2010, *9*, 225–241.

[12] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655–667.

[13] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[14] Chambers, M. C., Maclean, B., Burke, R., Amodei, D. et al., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 2012, *30*, 918–920.

[15] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O. et al., The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* 2012, *11*, M111.014381–M111.014381.

[16] Ma, Z.-Q., Dasari, S., Chambers, M. C., Litton, M. D. et al., IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* 2009, *8*, 3872–3881.

[17] Ma, Z.-Q., Polzin, K. O., Dasari, S., Chambers, M. C. et al., QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal. Chem.* 2012, *84*, 5845–5850.

[18] Wang, X., Chambers, M. C., Vega-Montoto, L. J., Bunk, D. M. et al., QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* 2014, *86*, 2497–2509.

[19] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A. et al., OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* 2008, *9*, 163.

[20] Junker, J., Bielow, C., Bertsch, A., Sturm, M. et al., TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *J. Proteome Res.* 2012, *11*, 3914–3920.

[21] Aiche, S., Sachsenberg, T., Kenar, E., Walzer, M. et al., Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics* 2015, *15*, 1443–1447.

[22] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R. et al., KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor.* 2009, *11*, 26–31.

[23] Walzer, M., Pernas, L. E., Nasso, S., Bittremieux, W. et al., qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Mol. Cell. Proteomics* 2014, *13*, 1905–1913.

[24] Kim, S., Pevzner, P. A., MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* 2014, *5*, 5277.

[25] Tabb, D. L., Fernando, C. G., Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* 2007, *6*, 654–661.

[26] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[27] R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www R-project.org/.

[28] Wen, B., Gatto, L., proteoQC: an R package for proteomics data quality control. 2016. R package version 1.6.0.

[29] Gatto, L., Christoforou, A., Using R and bioconductor for proteomics data analysis. *Biochim. Biophys. Acta* 2014, *1844*, 42–51.

[30] Fournier, F., Joly Beauparlant, C., Paradis, R., Droit, A., rTANDEM, an R/Bioconductor package for MS/MS protein identification. *Bioinformatics* 2014, *30*, 2233–2234.

[31] Bielow, C., Mastrobuoni, G., Kempa, S., Proteomics quality control: quality control software for MaxQuant results. *J. Proteome Res.* 2016, *15*, 777–787.

[32] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, *26*, 1367–1372.

[33] Cox, J., Hein, M. Y., Luber, C. A., Paron, I. et al., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 2014, *13*, 2513–2526.

[34] Bereman, M. S., Johnson, R., Bollinger, J., Boss, Y. et al., Implementation of statistical process control for proteomic experiments via LC MS/MS. *J. Am. Soc. Mass Spectrom.* 2014, *25*, 581–587.

[35] Broudy, D., Killeen, T., Choi, M., Shulman, N. et al., A framework for installable external tools in Skyline. *Bioinformatics* 2014, *30*, 2521–2523.

[36] MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M. et al., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010, *26*, 966–968.

[37] Sharma, V., Eckels, J., Taylor, G. K., Shulman, N. J. et al., Panorama: a targeted proteomics knowledge base. *J. Proteome Res.* 2014, *13*, 4205–4210.

[38] Pichler, P., Mazanek, M., Dusberger, F., Weilnböck, L. et al., 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.

[39] Bittremieux, W., Willems, H., Kelchtermans, P., Martens, L. et al., iMonDB: mass spectrometry quality control through instrument monitoring. *J. Proteome Res.* 2015, *14*, 2360–2366.

[40] Taylor, R. M., Dance, J., Taylor, R. J., Prince, J. T., Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics* 2013, *29*, 2948–2949.

[41] Sweredoski, M. J., Smith, G. T., Kalli, A., Graham, R. L. J., Hess, S., LogViewer: a software tool to visualize quality control parameters to optimize proteomics experiments using Orbitrap and LTQ-FT mass spectrometers. *J. Biomol. Tech.* 2011, *22*, 122–126.

[42] McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J. et al., MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* 2004, *18*, 2162–2168.

[43] 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.

[44] Amidan, B. G., Orton, D. J., LaMarche, B. L., Monroe, M. E. et al., Signatures for mass spectrometry data quality. *J. Proteome Res.* 2014, *13*, 2215–2222.

[45] Vizcaíno, J. A., Csordas, A., del Toro, N., Dianes, J. A. et al., 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 2016, *44*, D447–D456.

[46] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. et al., Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 2011, *12*, 2825–2830.

[47] 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.

[48] McKinney, W., Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, Austin, Texas, USA 2010, pp. 51–56.

[49] Aggarwal, C. C., Hinneburg, A., Keim, D. A., On the surprising behavior of distance metrics in high dimensional space. *Proceedings of the 8th International Conference on Database Theory - ICDT '01*, volume 1973 of *Lecture Notes in Computer Science*, London, England, Springer, Berlin Heidelberg 2001, pp. 420–434.

[50] Zhang, B., Wang, J., Wang, X., Zhu, J. et al., Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014, *513*, 382–387.

[51] Slebos, R. J. C., Wang, X., Wang, X., Zhang, B. et al., Proteomic analysis of colon and rectal carcinoma using standard and customized databases. *Sci. Data* 2015, *2*, 150022.

[52] Tabb, D. L., Wang, X., Carr, S. A., Clauser, K. R. et al., Reproducibility of differential proteomic technologies in CPTAC fractionated xenografts. *J. Proteome Res.* 2016, *15*, 691–706.

[53] Bittremieux, W., Meysman, P., Martens, L., Valkenborg, D., Laukens, K., Unsupervised quality assessment of mass spectrometry proteomics experiments by multivariate quality control metrics. *J. Proteome Res.* 2016, *15*, 1300–1307.

[54] Martens, L., Bringing proteomics into the clinic: the need for the field to finally take itself seriously. *Proteomics Clin. Appl.* 2013, *7*, 388–391.

[55] Eisenacher, M., Schnabel, A., Stephan, C., Quality meets quantity - quality control, data standards and repositories. *Proteomics* 2011, *11*, 1031–1036.

[56] Foster, J. M., Degroeve, S., Gatto, L., Visser, M. et al., A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* 2011, *11*, 2182–2194.

[57] Kinsinger, C. R., Apffel, J., Baker, M., Bian, X. et al., Recommendations for mass spectrometry data quality metrics for open access data (Corollary to the Amsterdam Principles). *Mol. Cell. Proteomics* 2012, *10*, O111.015446–O111.015446.

[58] Martens, L., Public proteomics data: how the field has evolved from sceptical inquiry to the promise of in silico proteomics. *EuPA Open Proteom.* 2016, *11*, 42–44.

[59] Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y., MacCoss, M. J., Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat. Protoc.* 2015, *10*, 887–903.

[60] Toby, T. K., Fornelli, L., Kelleher, N. L., Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* 2016, *9*, 499–519.

[61] 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.

