# Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data*[S]

Bryan A. Stanfill‡, Ernesto S. Nakayasu§, Lisa M. Bramer‡, Allison M. Thompson¶, Charles K. Ansong§, Therese R. Clauss§, Marina A. Gritsenko§, Matthew E. Monroe§, Ronald J. Moore§, Daniel J. Orton§, Paul D. Piehowski§, Athena A. Schepmoes§, Richard D. Smith§, Bobbie-Jo M. Webb-Robertson‡‖, Thomas O. Metz§‖, and TEDDY Study Group

**Liquid chromatography-mass spectrometry (LC-MS)-based proteomics studies of large sample cohorts can easily require from months to years to complete. Acquiring consistent, high-quality data in such large-scale studies is challenging because of normal variations in instrumentation performance over time, as well as artifacts introduced by the samples themselves, such as those because of collection, storage and processing. Existing quality control methods for proteomics data primarily focus on post-hoc analysis to remove low-quality data that would degrade downstream statistics; they are not designed to evaluate the data in near real-time, which would allow for interventions as soon as deviations in data quality are detected. In addition to flagging analyses that demonstrate outlier behavior, evaluating how the data structure changes over time can aide in understanding typical instrument performance or identify issues such as a degradation in data quality because of the need for instrument cleaning and/or re-calibration. To address this gap for proteomics, we developed Quality Control Analysis in Real-Time (QC-ART), a tool for evaluating data as they are acquired to dynamically flag potential issues with instrument performance or sample quality. QC-ART has similar accuracy as standard post-hoc analysis methods with the additional benefit of real-time analysis. We demonstrate the utility and performance of QC-ART in identifying deviations in data quality because of both instrument and sample issues in near real-time for LC-MS-based plasma proteomics analyses of a sample subset of The Environmental Determinants of Diabetes in the Young cohort. We also present a case where QC-ART facilitated the identification of oxidative modifications, which are often underappreciated in proteomic experiments.** ***Molecular & Cellular Proteomics 17: 1824–1836, 2018. DOI: 10.1074/mcp.RA118.000648.***

Control of data quality is a fundamental need for facilitating scientific reproducibility, and it is also essential for the translation of experimental discoveries to clinical, industrial or environmental applications (1–3). In liquid chromatography-mass spectrometry (LC-MS)[1]-based proteomics studies, it has been demonstrated that performing robust quality control (QC) can improve overall protein quantification and subsequently yield more accurate statistical estimates of differential abundance by detecting outlier data points (4). To date, only a few tools have been developed to assess LC-MS-based proteomics data quality in the context of an entire study (reviewed in (5)), and most of these are implemented as post-hoc analyses to be utilized at the end of the experiment. However, because of the complexity of proteomic studies, especially those involving large sample sets or cohorts, performing QC assessment of proteomics data in real-time would offer significant advantages. The sources of variability in a proteomics experiment that are addressed by QC can be categorized into two groups: biological and technical. The goal of QC is to not remove normal biological variability; however, there are circumstances where an LC-MS analysis displays outlier behavior and should be flagged and evaluated. For example, a sample may display outlier behavior and further examination may find that the subject had a cofounder, such as a medical drug exposure. The biological profile is likely no longer normal in the context of the experimental design, and thus either the sample would need to be removed or the cofounder would need to be dealt with statistically - either way the issue could be addressed. Thus, there is a clear advantage when performing further analysis of the data, either post-hoc or in near real-time, to identify analyses or

From the ‡Computational and Statistical Analytics Division, §Biological Sciences Division, ¶Environmental and Molecular Sciences Laboratory, 902 Battelle Blvd, Pacific Northwest National Laboratory, Richland, Washington

Received January 31, 2018, and in revised form, March 13, 2018
Published, MCP Papers in Press, April 16, 2018, DOI 10.1074/mcp.RA118.000648

samples for further investigation. Technical variability is derived from sample collection, transportation, storage, preparation, and/or instrument performance. Teasing out the cause of outliers in the category of technical variability can be extremely challenging, but evaluation of some data parameters, such as peak intensities and peptide sequences identified, which can vary depending on the mass spectrometer, LC column (particularly important for multi-column platforms), time since last instrument cleaning, length of proteolytic digestion, and sample cleanup, can improve downstream analysis (6). These issues are further complicated when the proteomics study requires months or years to complete, as other parameters in the instrument can drift over time. Thus, QC evaluation that differentiates normal change over time from outlier behavior could dramatically improve overall data quality by notifying investigators of the need for instrument maintenance, thus minimizing instrument-related artifacts.

The need for reliable QC approaches for LC-MS-based proteomics studies can be measured by the increasing number of publications on the topic (4, 7–15). Initial research in this area resulted in several web-based applications that track individual QC metrics on the fly with varying levels of sophistication (7–10). In Wang *et al.*, (11) QC metrics are tracked, then quantified and the uncertainty is partitioned into sources such as lab and instrument type, but the evaluation is focused on the entire experiment rather than individual MS analyses. Amidan *et al.*, (12) proposed a method to identify poor-quality datasets using a supervised learning approach. Because of the dynamic nature of mass spectrometry data, the method performs well post-hoc, but the supervised algorithm is overly specific to the training data and therefore cannot accurately track the quality of experiments in real-time. Bielow *et al.*, (13) developed a software tool (PTXQC) that summarizes QC metrics to allow an expert to curate individual datasets more quickly. However, their method currently is tailored to the QuaMeter metrics (16) only and the extension to other types of metrics is not immediate. Finally, Bittremieux *et al.*, (14) proposed a powerful tool that is unsupervised in nature and can handle generic QC data of high-dimension.

We have developed a method, QC Analysis in Real Time (QC-ART), that identifies local and global deviations in data quality because of either biological or technical sources of variability. The procedure is similar to that of Matzke *et al.* (4) in the context of the statistical outlier algorithm employed but adds a dynamic modeling component to analyze the data in a streaming LC-MS environment. We demonstrate the accuracy of QC-ART on data from both label-free and isobarically-labeled (*i.e.* iTRAQ (17)) proteomics studies. QC-ART is general enough to be applied to any LC-MS-based study where appropriate QC metrics can be collected over time, such as metabolomics and lipidomics. Using hand-curated data, QC-ART was validated to achieve similar accuracy to state-of-the-art post-hoc analyses (4) but in real-time. Lastly, the capabilities and benefits of QC-ART for large-scale proteomics studies is demonstrated for data collected from analyses of a sample subset of The Environmental Determinants of Diabetes in the Young (TEDDY) cohort (18). Using QC-ART in this study, we identified multiple types of both naturally occurring issues and those because of deliberate, yet subtle manipulations of instrument operating parameters, and the results were compared with current state-of-the-art methods (14). We also demonstrated the utility of QC-ART in identification of oxidations of tryptophan, tyrosine and cysteine residues, which are often overlooked in peptide identification. QC-ART is available as an online application (https://ascm.shinyapps.io/BAS_QCART), where researchers can perform analyses by uploading their own data. Additionally, the source code necessary to implement QC-ART as a standalone application is freely available as the R package *QCART* on GitHub (https://github.com/stanfill/QC-ART), and the source code for a corresponding web interface is freely available at https://github.com/stanfill/QC-ART-Web-App.

## EXPERIMENTAL PROCEDURES

As discussed above, existing algorithms for QC of LC-MS-based untargeted proteomics data are not designed for streaming applications (14). However, for purposes of comparison to the state-of-the-art, QC-ART is compared with several existing algorithms and results are compared in a post-hoc fashion to determine if QC-ART performs as well as or better than existing approaches in a dynamic manner. The two key algorithms used to compare with QC-ART were Robust Mahalanobis Distance on Peptide Abundance Vectors (RMD-PAV) and an unsupervised QC method, InSPECTor. These two algorithms are described at a high level for comparative purposes, and details of these methods are available in (4) and (14), respectively. For both algorithms, data matrices are represented at the sample level, *i.e.* in the case where samples are fractionated before LC-MS analysis, the data for a single sample is the sum of all fractions.

*Existing Post-Hoc QC Algorithms*—RMD-PAV is a post-hoc analysis technique used to identify outlier LC-MS data based on all of the quantified peptide peak intensities for a sample. The observed peptide peak intensities are summarized as an abundance distribution represented by a set of statistical metrics, such as median and skewness. This set of metrics is reduced using robust principal components analysis (rPCA), and then the robust Mahalanobis distance between the metrics transformed using the rPCA coordinate system is measured. The distance computed for each sample is compared against percentiles of the appropriate chi-square distribution to determine how extreme an instrument run is relative to the rest of the

[1] The abbreviations used are: LC, Liquid Chromatography; AUC, Area Under the Curve; BPMZ, Base peak m/z; iTRAQ, Isobaric Tags for Relative and Absolute Quantitation; LTQ, Linear Ion Trap; MASIC, MS/MS Automated Selected Ion Chromatogram Generator; *m/z*, mass-to-charge ratio; NIST, National Institute of Standards and Technology; PCA, Principal Component Analysis; PIMZ, Parent ion m/z; QC, Quality control; QC-ART, Quality Control Analysis in Real-Time; RMD-PAV, Robust Mahalanobis Distance on Peptide Abundance Vectors; ROC, Receiver Operating Characteristic; rPCA, Robust Principal Component Analysis; SARS-CoV, Severe Acute Respiratory Syndrome Coronavirus; TEDDY, The Environmental Determinants of Diabetes in the Young.

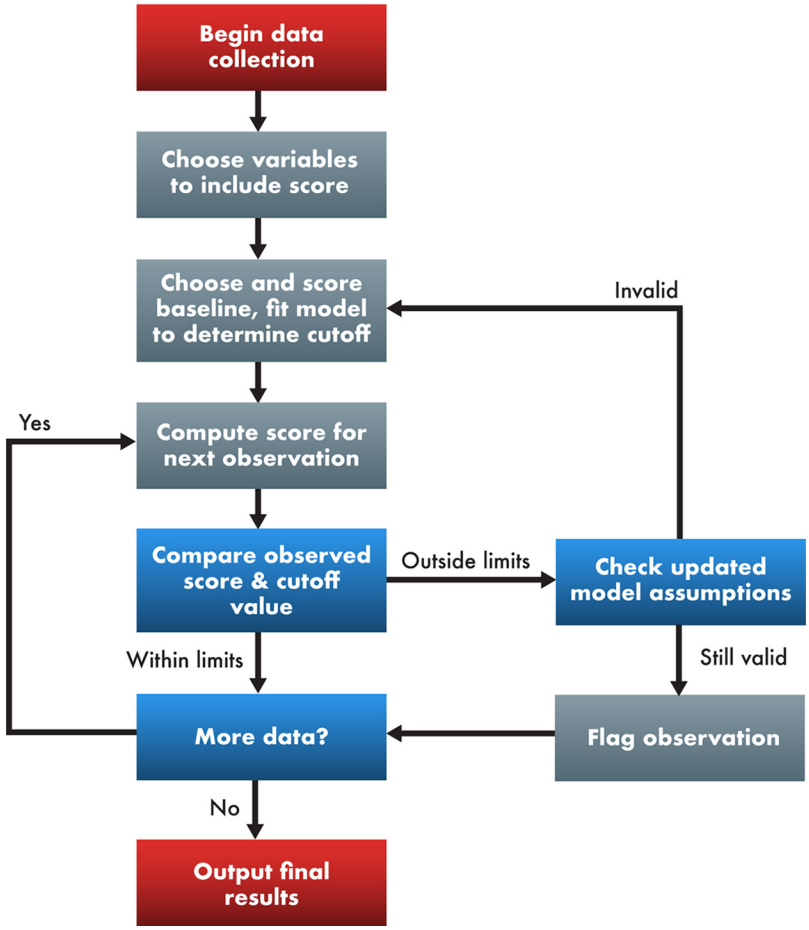


FIG. 1. **The workflow of the QC-ART algorithm expressed as a flow chart.** The steps in the flow chart are either input/output nodes (colored red), process nodes (colored gray) or decision nodes (colored blue).

data set. A score akin to a *p* value is used to identify samples that may be outliers.

InSPECTor is a recently developed method that represents the cutting edge in unsupervised outlier detection for large LC-MS experiments. It is based on a local outlier probability distance metric, which identifies potentially outlying instrument runs by finding a group of *k* instrument runs most like the analysis in question. InSPECTor then uses the standard normal density kernel to estimate the probability of each run being an outlier (15). The probability threshold used to identify outlying instrument runs is chosen by the user and varies from experiment to experiment.

*QC-ART*—QC-ART is an algorithm that uses a dynamic linear model to flag anomalies while accounting for typical instrument change over time. Furthermore, to perform this task in near real time, only metrics that can be computed in a rapid fashion are used, such as those defined by NIST (19) and proposed along with QuaMeter (16). Fig. 1 illustrates the generic workflow for QC-ART. Once data collection has started, variables are computed for each instrument analysis in near-real time. The model is fit using a baseline set of data, and as each new sample is analyzed by the instrument, it is immediately scored. If the data do not show any anomalous behavior in the context of the baseline set, the process continues with the results from analysis of the next sample. However, if the data appear anomalous in the context of the baseline, then the sample is flagged for follow up by a technician. At this point the user may evaluate the model assumptions and determine if the existing baseline is still appropriate or if a new one should be established. If the user believes that the instrument performance has changed or that the instrument needs to be cleaned, then the process begins again.

*QC-ART Variables*—It has been shown that summary statistics derived from reporter ion distributions from isobarically labeled proteomics data are beneficial in addition to NIST and QuaMeter QC metrics when assessing LC-MS-based proteomics data quality (20). Inspired by these metrics, we considered a large list of potential variables that could be generated rapidly for inclusion in QC-ART. However, to increase the rate at which QC-ART can process data, it was prudent to reduce the number of variables to just those that demonstrate predictive qualities for identifying low-quality data. PCA was used to identify a subset of the initial variables in conjunction with domain expertise associated with common sources of altered LC-MS data quality, such as nanoelectrospray instability. For label-free proteomics data, only the NIST QC metrics are used.

*Setting the Baseline*—A baseline data set comprised from good quality instrument runs is critical and driven by the researcher's goal(s). For real-time analysis, for example to track instrument performance over time, then a set of data from the beginning of the experiment from analyses that were performed under ideal instrumental conditions should be chosen. In this way, successive instrument runs that have scores significantly far away from the quality threshold will signify a shift in data quality that should be evaluated. The dynamic nature of QC-ART allows the baseline to change over time as needed. When using QC-ART to perform post-hoc QC, a baseline that is evenly distributed thorough out the course of the experiment in chronological order, which accounts for uncertainty because of variability in instrument performance over time, is selected. We investigate the impact of baseline size and quality on the accuracy of QC-ART in supplemental File S1.

*Scoring New MS Data*—An rPCA method is used to transform the data, and then the robust Mahalanobis distance between the reduced set of variables is computed to assess similarity, a modification of the existing Sign2 metric (21). Given that the metrics can vary dramatically in scale and the underlying statistical distributions of the metrics are unknown, rPCA is more accurate than traditional PCA methods at identifying outlying observations (22). Similarly, the robust Mahalanobis distance is used to score the instrument runs transformed to the rPCA space because it has been shown to be the preferred multivariate distance in the presence of extreme observations (4). The resultant scores, called QC-ART scores, are not guaranteed to follow a distribution; therefore, cutoff values based on percentiles of a common distribution are not appropriate. Thus, the baseline scores are used to build a model against which all future scores are compared. Previous QC methods assume a static linear model to the QC metrics, which implicitly assumes the mechanism generating the data is unchanged throughout the course of the experiment (7, 10, 13). Because instrument behavior is likely to change over time, we additionally implemented dynamic linear models, whose parameter estimates are continually updated when additional experiments are performed and observed to be high quality (23). Experiments that are identified to be of poor-quality should not be used to update the parameter estimates as doing so could decrease the chances of identifying poor-quality instrument runs performed later in the study.

For both the static and dynamic models, the assumptions associated with the model must be checked as new instrument runs are added to the dataset. See supplemental File S1 for a further discussion about model assumptions and how to verify them. Both the static and dynamic linear models are used to define threshold values for the QC-ART scores. The threshold values are chosen to control the probability of a false positive, *i.e.* a good quality instrument run was erroneously flagged as being of poor quality. The static model threshold is used to identify changes in instrument quality relative to the chosen baseline set only, whereas the dynamic model threshold is used to identify changes in instrument behavior relative to the baseline set after controlling for the recent behavior of the instrument. Because QC-ART scores are distances, they cannot be less than zero and a large score indicates that a given instrument run is different compared with the baseline set. Therefore, isolated large scores are interpreted as a single instrument run that warrants further investigation, *e.g.* because of an occluded electrospray ionization emitter or inappropriate database search parameters, but do not signify a systematic change in instrument quality. Several successive scores above the threshold value or deviations from the model assumptions represent a potential systematic change in data. Finally, the QC-ART scores on their own are not easily translated from one study to another. However, QC-ART scores can be translated to probabilities by using the probability distribution function implied by the static and dynamic models. Probabilities derived from the static model are interpreted as the probability that an extreme LC-MS instrument run occurred given the baseline data only. Alternatively, probabilities derived from the dynamic model are interpreted as the probability that an extreme LC-MS analysis occurred given the baseline data and recent changes in instrument behavior.

## RESULTS

We assessed the ability of QC-ART to identify outlying instrument runs using both a previously published label-free proteomics data set that has been expertly curated, and a new isobarically-labeled proteomics data set from analysis of a large sample cohort. QC-ART is compared against RMD-PAV and InSPECtor for both data sets.

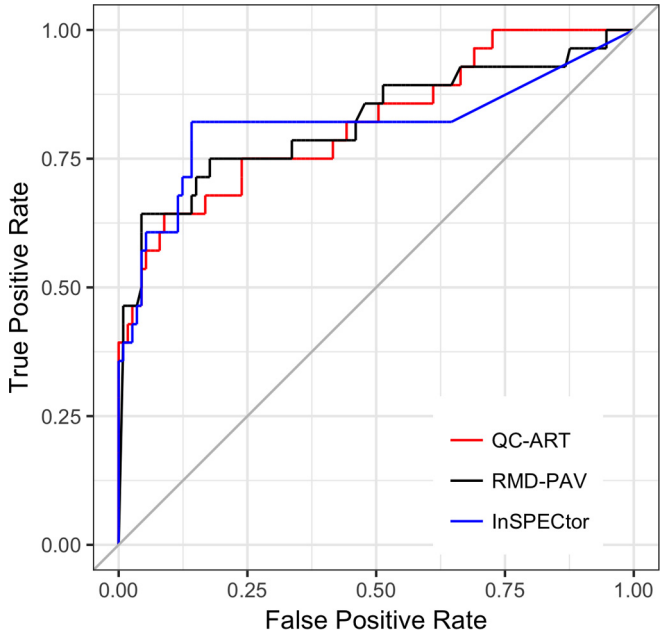


FIG. 2. **Comparison of the accuracy of InSPECTor, QC-ART and RMD-PAV to identify extreme peptide abundance distributions for the Calu-3 data via ROC curve analysis.** The area under the curve for each of the methods is 0.820 for QC-ART, 0.823 for RMD-PAV and 0.816 for InSPECtor, which indicates that QC-ART is able to achieve comparable levels of accuracy as both post-hoc techniques but in real-time.

*Real Data Benchmark - Expert Identified Outlier Runs*—The label-free proteomics data are comprised from analyses of a human lung-derived cell line, Calu-3, infected with Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV). A total of 141 LC-MS runs were performed using an LTQ-Orbitrap instrument, and the data were expertly curated, and 28 LC-MS analyses identified as potentially outlying (4). The same five statistics as previously published (4) were used to summarize each LC-MS run for the RMD-PAV and InSPECtor methods: (1) the fraction of missing peptides, (2) a group-wide correlation coefficient, (3) median, (4) skew, and (5) kurtosis of the abundance distribution. QC-ART scores were computed using the same five statistics except for the group-wide correlation coefficient because it could not be computed until all peptide quantifications were completed. The first ten LC-MS runs that were not suspected of being outliers were used as the baseline for QC-ART, which was applied to each sample in the order in which it was quantitated without using any information from samples that had not yet been quantitated. The InSPECtor method was implemented with a neighborhood size of ten to parallel the baseline choice for QC-ART. Results for other baseline sizes are given in supplemental File S1.

A receiver operating characteristic (ROC) curve analysis was used to compare the ability of QC-ART to identify extreme peptide abundance distributions relative to the RMD-PAV and InSPECtor methods (Fig. 2). QC-ART and RMD-PAV

TABLE I
*Variables used by QC-ART to identify instrument runs of poor quality*
The variables derived from the reporter ion file (top section) are used only when available, i.e., iTRAQ instrument runs.

| Source | Variable | Definition |
|---|---|---|
| | PIMZ_skew | Skewness of the parent ion *m/z* |
| | BPMZ_skew | Skewness of the base peak *m/z* |
| Reporter | I119_skew | Skewness of the blank reporter ion channel |
| Ion | I121_median | Median of the reference reporter ion channel |
| File | I121_skew | Skewness of the reference reporter ion channel |
| | WAPIC_skew | Skewness of the weighted average percent intensity correction |
| | Missingness | Percentage of missing data over all ions |
| | P_2C | Number of tryptic peptides; unique peptide count |
| | MS1_2B | Median TIC value for identified peptides from run start through middle 50% of separation |
| NIST | RT_MS_Q1 | The interval for the first 25% of all MS events divided by RT-Duration (RT-Duration is defined as the highest scan time observed minus the lowest scan time observed) |
| | RT_MS_Q4 | The interval for the fourth 25% of all MS events divided by RT-Duration |
| | RT_MSMS_Q1 | The interval for the first 25% of all MS/MS events divided by RT-Duration |
| | RT_MSMS_Q4 | The interval for the fourth 25% of all MS/MS events divided by RT-Duration |

perform comparably across the range of possible cut-off values and have extremely close area under the curve (AUC) values. Compared with QC-ART and RMD-PAV, the InSPECtor method achieved larger true-positive rates for the same false-positive rate over a limited range of cutoff values, but had lower accuracy overall, as indicated by the lower AUC score.

*Case Study - Longitudinal Cohort Study*—To test the utility of QC-ART for monitoring the quality of LC-MS data in real time, we used it to supervise data from plasma proteomics analyses of a subset of samples from TEDDY (18), a large prospective study with the goal of discovering factors that initiate the autoimmune response and destruction of the pancreatic beta cells, leading to the development of type 1 diabetes. To fulfill this goal, we are performing comprehensive plasma proteomic analyses of TEDDY samples to better understand progression of the disease. A total of 2252 plasma samples from 368 donors were pooled together by donor, depleted of the 14 most abundant proteins, then digested with trypsin and labeled with 8-plex iTRAQ reagent according to the manufacturer recommendations. Each 8-plex iTRAQ set was multiplexed by including 6 TEDDY samples plus one common reference sample (channel 121) that was generated by pooling aliquots from all donors, whereas the remaining iTRAQ channel (119) was not used. Each of the resulting sixty-two 8-plex iTRAQ sets was fractionated into 24 fractions, resulting in 1488 individual samples for LC-MS/MS analysis that in total required 14 months to complete, together with analysis of one independent QC sample (tryptic digest of the bacterium *Shewanella oneidensis*) and one blank every 24 fractions, which were used to assess instrument performance. The variables used by QC-ART to monitor these instrument runs are described in Table I. The NIST variables were calculated using the PNNL-developed software SMAQC, which is freely available on GitHub https://github.com/PNNL-Comp-Mass-Spec.

To monitor instrument performance using QC-ART, a set of instrument runs that were collected during peak instrument performance were chosen. In this instance, the first ten fraction sets of data (a total of 240 LC-MS/MS runs) collected after instrument cleaning, calibration and running an independent QC sample before each iTRAQ set were treated as the baseline for all future instrument runs (Fig. 3*A*). Additionally, because the fractions might contain completely different peptides, each of the 24 fractions of the iTRAQ sets was treated separately. For example, to assess the quality of the data collected in Fraction 1 of Set 11, the corresponding variables for Fraction 1 of Set 11 were compared against those same variables for Fraction 1 of Sets 1 through 10.

If an instrument run was known or later judged to be of poor-quality, its value was not used to update the dynamic threshold model. In practice, the samples that were flagged by QC-ART were reanalyzed at the end of the study because this methodology was under-development when the samples were being processed. For future studies employing QC-ART, flagged samples will be reanalyzed immediately. Five events of interest that occurred during the study are labeled in Fig. 3*A* (and all subsequent figures); (1) iTRAQ sets 16 and 17, (2) test runs with deliberately mistuned instrument parameters, (3) drop in instrument performance, (4) gap in analysis and (5) diluted samples. In July 2015, a still underdevelopment version of QC-ART flagged several runs of iTRAQ sets 16 and 17 as poor-quality datasets, at which point the instrument operator stopped the data collection and cleaned the mass spectrometry ion source. These samples were rerun in early August 2015, and the corresponding QC-ART scores returned to normal levels. We also collected 10 samples using deliberately mistuned mass spectrometry parameters and liquid chromatography gradient to assess QC-ART's ability to identify poor-quality data. The mistuned parameters were: (1) changing the mass spectrometer front lens voltage from −8 to −6, which affects the number of ions entering the ion trap; (2)

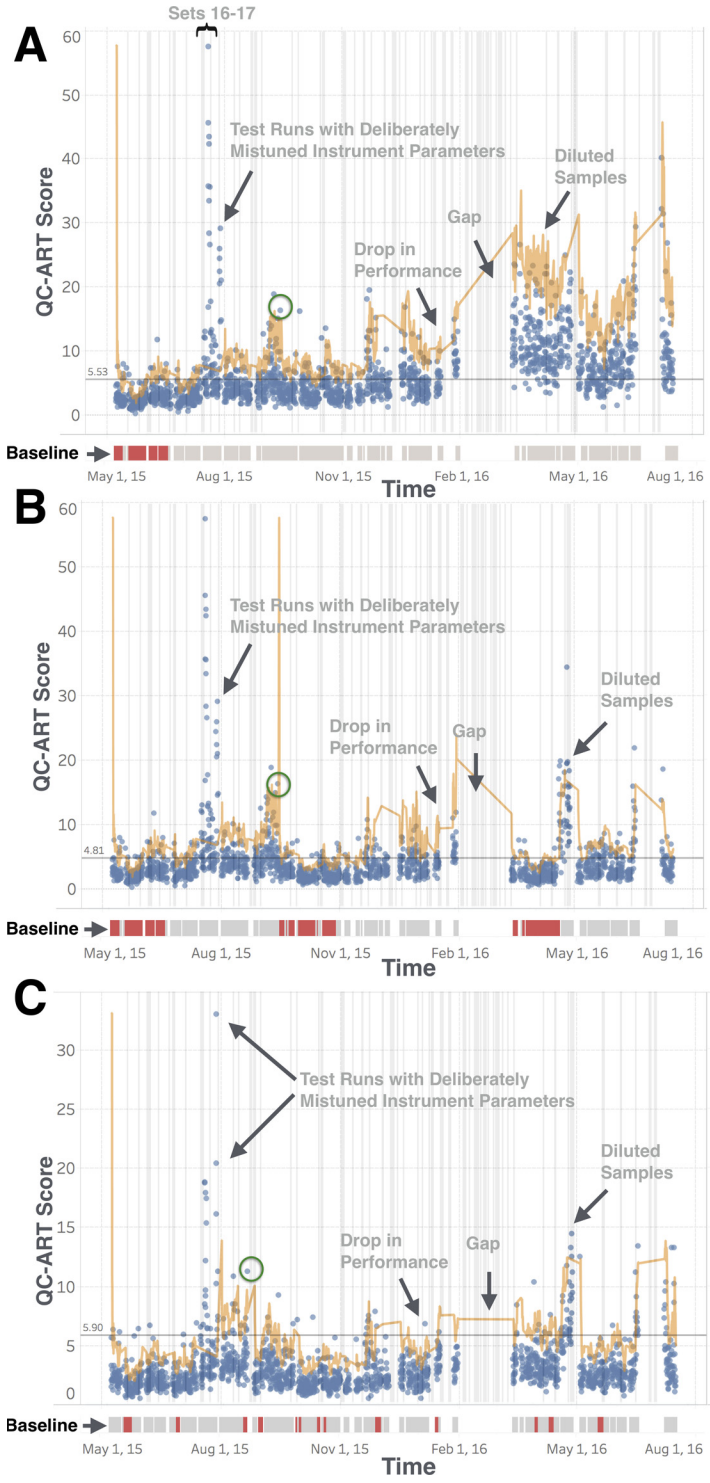


FIG. 3. **QC-ART scores of the iTRAQ data with static and dynamic thresholds when different sets of instrument runs were used as the baseline (red areas in timelines below figures).** In all figures, the point circled in green represents a sample that exhibited suspicious oxidation patterns, the gray vertical lines indicate instrument cleaning and recalibration events, and the horizontal lines represent static (gray) and dynamic (yellow) threshold values. *A*, When the first ten sets are used as the baseline then the poor quality test runs are correctly identified, but the fundamental shift in instrument performance over time causes the method to flag too many data points later in the experiment. *B*, Checking model assumptions throughout the course of the cohort study tells the researcher when a new baseline is needed, which allows the method to correctly identify instrument runs of poor quality without negatively impacting the false alarm rate. *C*, Distributing the baseline throughout the course of the study reduces some of the noise in the QC-ART scores, but they cannot be computed in real-time because some of the baseline samples occur late in the study.

inserting a dead volume in the LC tubing, which affects the chromatographic performance; (3) changing the mass spectrometer S-lens from 69 to 25%, which generates a bias toward lower *m/z* ions being transmitted to the ion trap; and (4) increasing the number of microscans, which reduces the overall number of collected spectra. In January 2016, the instrument behavior dropped significantly (labeled "Drop in instrument performance") and led to a period of instrument recalibration and cleaning (labeled "Gap"). Finally, in April 2016, a full set of samples seemed to be too concentrated and was then diluted before reanalysis, which led to differences in data quality (labeled "Diluted samples").

Using the first ten sets as the baseline, QC-ART correctly flagged the poor-quality test runs in July 2015. Similarly, the

QC-ART scores correctly increased in response to the decrease in instrument performance in January 2016. QC-ART did not, however, identify the diluted samples in late April 2016. QC-ART's inability to detect the difference in sample preparation is because of the difference in instrument behavior before and after the recalibration and cleaning in January 2016. To properly account for the change in instrument performance after the long maintenance period, a new baseline must be chosen. The need for a new baseline is also apparent when assessing model assumptions (supplemental File S1).

Fig. 3*B* illustrates a QC-ART update when model assumptions were violated, that is, a systematic change in instrument performance was identified. From Fig. 3*A* the distribution of the QC-ART scores changes throughout the course of the study, though it is not obvious exactly when that change is significant enough to warrant a new baseline. Tracking the model assumptions through time in conjunction with the QC-ART scores indicated that significant changes in instrument performance occurred in October 2015 and March 2016 (supplemental File S1). To account for the changed instrument behavior, two new baselines were chosen in October 2015 and again after the gap in experimental runs in early 2016 (red areas in timeline below Fig. 3*B*). Both retraining periods coincide with significant instrument maintenance that were initiated by the technicians and are immediately preceded by large shifts in QC-ART scores. This illustrates that when QC-ART is appropriately trained, it can identify all areas of interest. The poor-quality test runs in July 2015 have very large scores, with the scores leading up to the gap in January 2016 increasing as the instrument performance drops, and the diluted samples in July 2016 being identified successfully.

QC-ART can also be used as a post-hoc data quality tool by selecting a baseline comprised of data from analyses that are distributed throughout the course of the study (Fig. 3*C*). Compared with the results from when the baseline is comprised of data collected at the beginning of the study (Fig. 3*A*), spreading out the baseline greatly reduced the variance in scores as indicated by, *e.g.* reduced QC-ART scores associated with the poor-quality test runs in July 2015. Also, the variability in instrument performance through time was partially accounted for by the alternative baseline selection. Used as a post-hoc data analysis tool, QC-ART could identify the poor-quality test runs in July 2015, and the diluted samples in April 2016. QC-ART could not, however, identify the last group of instrument runs before the gap in January 2016. This is likely because data acquired just before it was included in the baseline set. This illustrates the importance of choosing an appropriate baseline set when using QC-ART.

To validate the QC-ART results, data from each LC-MS/MS analysis of the TEDDY cohort study was analyzed by the InSPECtor and RMD-PAV methods in a post-hoc fashion, *i.e.* the respective methods are applied once all the samples were analyzed at least once. The InSPECtor method was applied to all instrument runs using the same variables that informed QC-ART as reported in Table I (Fig. 4*A*). Unlike the QC-ART scores in Fig. 3*A*, the change in instrument performance over time cannot be detected with the InSPECtor method, which was expected given the local focus of its distance metric. The instrument runs starting in March 2016 do not appear to be different from those that occurred much earlier in the study even though they are quite different when compared directly. The upward trend in outlier scores starting in January 2016 indicates that InSPECtor could detect the degradation in instrument performance before the gap in early 2016, however. Based on the chosen 95% threshold, InSPECtor also identified the poor-quality test runs in July 2015 and data from some of the diluted samples in May 2016.

To apply the RMD-PAV method to the cohort study, the peptide reporter ion intensity data for each sample within each fraction set was extracted using MASIC (24). Because of the complexity of the cohort study, the collected data had to be manipulated to implement RMD-PAV. First, the full data set was reduced such that only the initial analysis of each iTRAQ set was retained. For example, samples from fraction set four were analyzed in May 2015 and March 2016, but only the results from May 2015 were used to compute RMD-PAV scores. Second, the iTRAQ 8-plex configuration used for this study creates results for six individual samples and a pooled reference sample. The RMD-PAV scores were derived from the pooled reference sample data because the data from the individual samples exhibited unwanted variation because of biological differences between the donors. Finally, because the study groups were spread across fraction sets, the suggested group average correlation variable typically used by RMD-PAV was replaced with a global average correlation. That is, the average pairwise correlation between the peptide abundance vectors for each sample is used in place of the average group correlation defined in Equation (1) of Matzke *et al.* (4). The RMD-PAV scores on the log base two scale that resulted from this implementation of RMD-PAV along with a 99.9% threshold are plotted in Fig. 4*B*.

The large group of samples with scores above the threshold starting in May 2016 suggests that RMD-PAV could recognize that data processed at the end of the study differ systematically from data collected at the beginning of the study. However, RMD-PAV was not able to differentiate the diluted samples from any of the other samples that occur after the break. Additionally, the data collected just before the gap does not look to be of poor-quality based on RMD-PAV scores. Note that the interesting data point plotted previously in green and the poor-quality test runs flagged by QC-ART and InSPECtor were not analyzed using RMD-PAV because they were reruns of samples analyzed earlier in the experiment.

The instrument run represented by the green point in Figs. 3 and 4*A* was flagged by QC-ART based on all three-baseline choices using both the dynamic and static thresholds. When the individual metrics corresponding to that run were investigated, no individual outlying values were noticed. To further

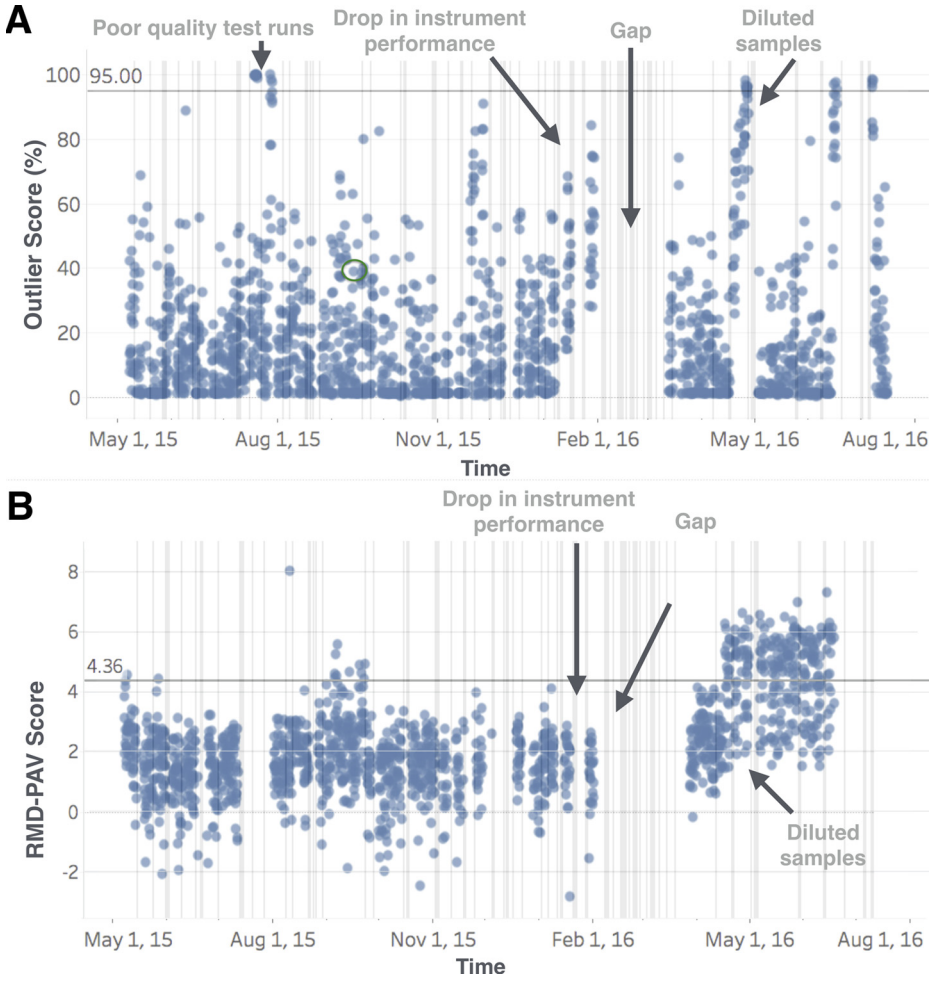


FIG. 4. **Results of the InSPECtor and RMD-PAV methods for the TEDDY iTRAQ data with labeled time periods of interest.** *A*, The InSPECtor outlier scores (%) were computed using a neighborhood size of 10 are plotted for each instrument run in chronological order. The horizontal line corresponds to a 95% threshold and the green point represents a sample that exhibited suspicious oxidation patterns. The mistuned sample runs and some of the diluted samples were correctly flagged to be of poor quality, but the instrument drift was not detected. *B*, The RMD-PAV scores were computed for each fraction set once. The solid horizontal line corresponds to a 99.9% threshold. Some of the diluted samples were correctly flagged to be of poor quality, but the large shift in instrument performance at the end of the cohort prevents RMD-PAV from identifying many of the other phenomena of interest during this study.

investigate this issue, a thorough manual inspection was carried out for several LC-MS runs with similar behavior. Although only small differences were observed in peak intensities and overall shape of the total-ion chromatogram, some regions of the chromatogram revealed different peak distributions compared with the corresponding fractions in different iTRAQ sets (Fig. 5*A*). By examining one of those regions (elution time 32–37 min) we observed extensive mass shifts of 15.99 Da, which corresponds to the addition of one oxygen atom (Fig. 5*B*). The sequence of a peptide in this region was determined to be GQYCYELDEK, which corresponded to amino acid residues 177–186 of human vitronectin (Uniprot ID: P04004). Surprisingly, this peptide lacks methionine residues, which are easily oxidized and usually set as a possible modification location in proteomic data analysis. Note that InSPECtor gave this instrument run a score below 50/100 (Fig. 4*A*), which could imply it is incapable of capturing subtle changes in data quality as reliably as QC-ART.

To determine possible oxidation sites, database searches were performed again but with MSGF+ (25) now considering potential oxidation in any amino acid residue. Those peptide identifications that were reported by MSGF as having oxidized residues were then analyzed by Ascore (26), a tool that calculates probabilities for specific localization of modifications, to assure the localization of the oxidation. The final oxidation counts were normalized by the total number for each amino acid (Fig. 5*C*), and the results showed enrichment in oxidations of cysteine, methionine, tryptophan and tyrosine residues (Fig. 5*D*). Although methionine oxidation was expected, cysteine was not because the samples were reduced with dithiothreitol during sample preparation. Tryptophan and tyrosine oxidations have been previously described in specific

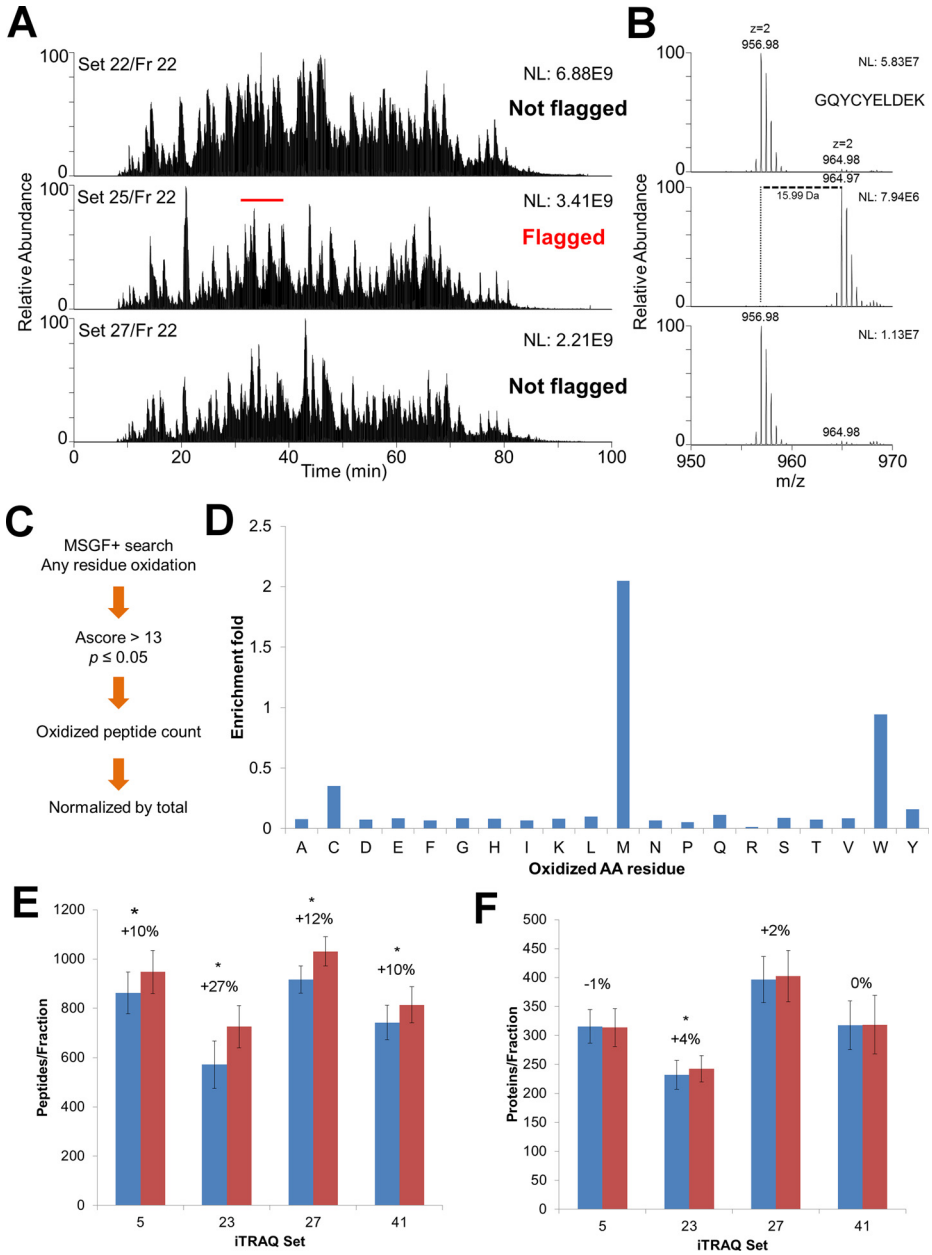


FIG. 5. **QC-ART leads to the identification of unexpected oxidations.** *A*, Total-ion chromatogram of a LC-MS/MS run and the corresponding high pH reversed-phase fractions of three different iTRAQ sets. *B*, A selected *m/z* range of the region highlighted in *A*. *C*, Workflow of the MSGF+ database searches to identify new oxidized residues. *D*, Normalized counts of oxidized amino acid residues. *E-F*, Average number of peptide (*E*) and protein (*F*) identifications per fraction. The blue bars represent the database search performed considering methionine oxidation as the only possible modification, whereas the red bars represent searches performed considering methionine, cysteine, tryptophan and tyrosine oxidations. The asterisks represent $p \leq 0.05$ by *t* test.

oxidative stress conditions (27, 28), but they are not usually considered during informatics processing for peptide identification. Moreover, oxidation on cysteine, tryptophan, tyrosine and proline residues were recently identified by an unbiased database search analysis of HeLa and HEK293 proteome (29). We then reanalyzed the data by performing the protein database searches considering these oxidations, which led to an increase of up to 27% in the number of identified peptides (Fig. 5*E*), but which was not reflected in a significant increase in the identification of proteins (Fig. 5*F*).

*Development of a User-friendly Interface*—To make QC-ART more accessible to instrument operators and core facility personnel, we developed a user-friendly, online application (https://ascm.shinyapps.io/BAS_QCART), where researchers can perform analyses on QC metrics of their own samples (Fig. 6*A*). The researcher simply uploads their data to the

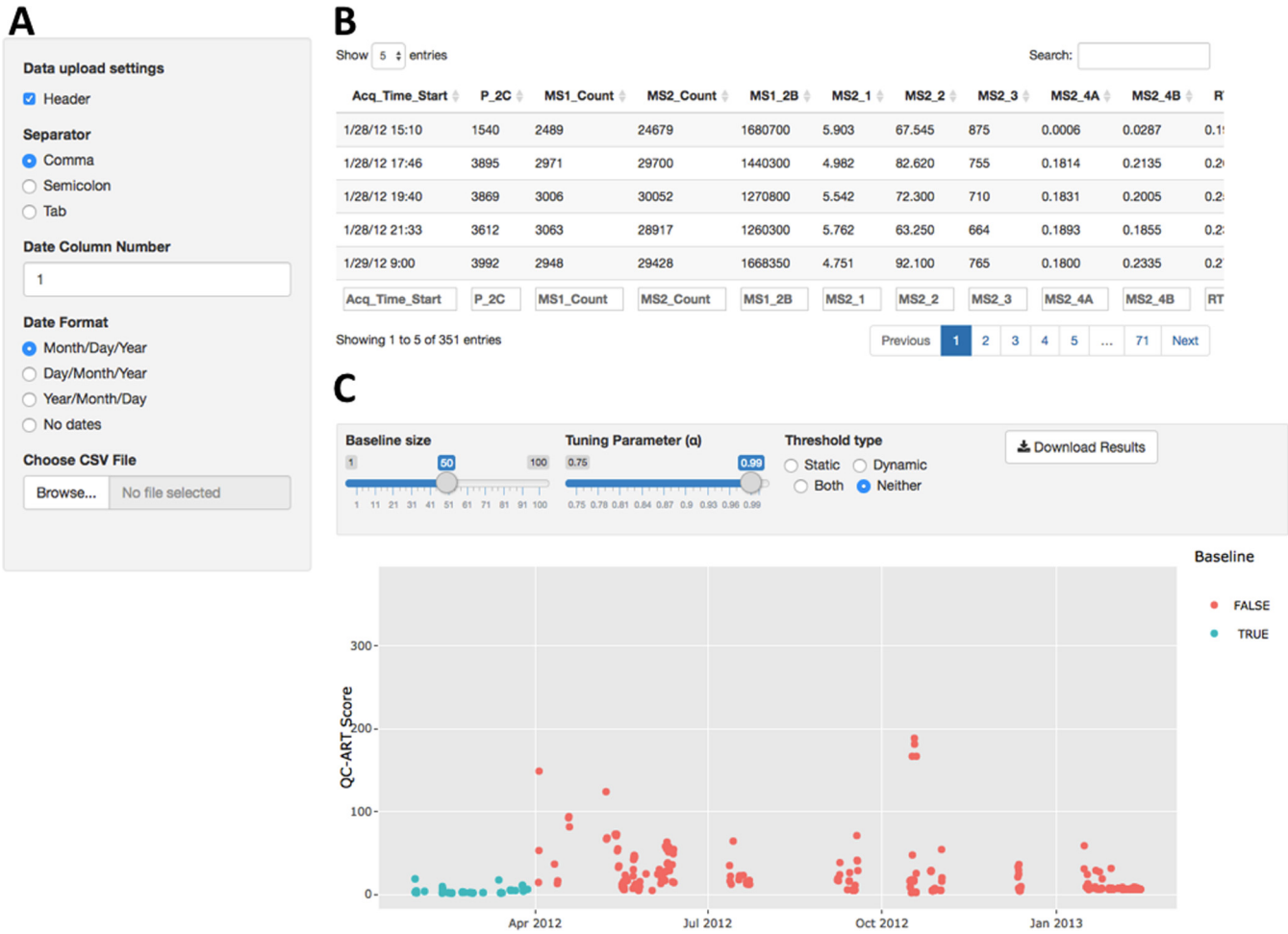


FIG. 6. **QC-ART interface.** *A*, Data upload and settings. *B*, Verification of the uploaded data. Once the data are uploaded they can be verified and searched. *C*, Results and output. The results can be visualized in the interactive graph and downloaded as a comma separated file. Tuning parameters, including baseline size, can be adjusted by experienced users to better discriminate poor-quality runs.

application, then QC-ART scores and thresholds are computed using default parameter values (Fig. 6*A*–6*C*). We have pre-set reliable threshold values based on our extensive training. However, advanced users can manipulate tuning parameters such as the baseline size and the proportion of variability explained by the principal components (Fig. 6*C*). The QC-ART scores are plotted in an interactive dot plot along with dynamic and static thresholds to identify instrument runs that may require further evaluation. The source code used by the online application to implement the QC-ART method is freely available as the R package *QCART* on GitHub (https://github.com/stanfill/QC-ART).

## DISCUSSION

QC-ART is a novel and powerful real-time QC tool. Its flexibility sets it apart from existing QC methods, with the added cost of more oversight by the researcher. The researcher must choose the baseline data sets, appropriate variables and a model to fit to the scores, but the insights derived from QC-ART are deeper than those currently attainable and offer informative metrics to allow researchers to actively steer data collection. Existing QC tools either automatically chose a baseline set using a machine-learning algorithm, *e.g.* InSPECtor, or use all instrument runs available, *e.g.* RMD-PAV. By selecting a baseline of instrument runs at the beginning of a long study when instrument performance is likely optimal, QC-ART can reveal previously unexplored sources of uncertainty, such as normal *m/z* instrument drift. Using a data set that was expertly curated to be of good or bad quality, we showed that QC-ART is equally or more accurate than RMD-PAV and InSPECtor, but the QC scores for each sample were available immediately after peptide identification and other data processing was completed, rather than after all samples were analyzed. In the context of a long running cohort study, neither InSPECtor nor RMD-PAV could identify the gradual change in instrument performance that was obvious in Fig. 3. The benefits of QC-ART relative to existing post-hoc tools like RMD-PAV is derived from the baseline flexibility, but also QC-ART's ability to fuse multiple sources of data. The inclusion of NIST variables such as

BPMZ skew, MS1 2B and P 2C, allowed QC-ART to identify samples that were prepared incorrectly, but had peptide abundance vectors similar to other runs. Because RMD-PAV is defined solely based on peptide abundance data, it was not able to identify the improperly prepared samples. Further, using all other instrument runs as a baseline makes it easy for RMD-PAV to identify large changes in instrument performance, but those large shifts in performance often mask the subtle changes such as the slow degradation of instrument performance in January 2016. Finally, by fitting a model to the QC-ART scores and continually checking the assumptions associated with those models, QC-ART was able to pinpoint exactly when the instrument needed service or cleaning. None of the existing methods for LC-MS-based proteomics analysis QC can identify change points in instrument performance with this level of rigor. In addition, QC-ART scores can be modeled either statically or dynamically, allowing QC-ART to identify both global and local changes in instrument behavior. Static thresholds are used exclusively in the literature to identify global outlying observations (7, 10, 13).

QC-ART is an important addition to the existing proteomics QC toolkit and can substantially reduce the amount of time required to identify and re-run samples that may have been subject to unwanted sources of variability.

*Acknowledgments*—We thank PNNL Graphic Designer Michael Perkins and Staff Scientist Aritra Dasgupta for assistance in preparing the figures. TEDDY families are warmly acknowledged for their participation in this study. The TEDDY Study Group is acknowledged for excellent contributions.

**Colorado Clinical Center:** Marian Rewers, M.D., Ph.D., PI[1,4,5,6,10,11], Kimberly Bautista[12], Judith Baxter[9,10,12,15], Ruth Bedoy[2], Danie[1] Felipe-Morales, Kimberly Driscoll, Ph.D.[9], Brigitte I. Frohnert, M.D.[2,14], Marisa Gallant, M.D.[13], Patricia Gesualdo[2,6,12,14,15], Michelle Hoffman[12,13,14], Rachel Karban[12], Edwin Liu, M.D.[13], Jill Norris, Ph.D.[2,3,12], Adela Samper-Imaz, Andrea Steck, M.D.[3,14], Kathleen Waugh[6,7,12,15], Hali Wright[12]. University of Colorado, Anschutz Medical Campus, Barbara Davis Center for Childhood Diabetes.

**Finland Clinical Center:** Jorma Toppari, M.D., Ph.D., PI[¥ˆ1,4,11,14], Olli G. Simell, M.D., Ph.D., Annika Adamsson, Ph.D.[ˆ12], Suvi Ahonen[*±§], Heikki Hyöty, M.D., Ph.D.[*±6], Jorma Ilonen, M.D., Ph.D.[¥¶3], Sanna Jokipuu[ˆ], Tiina Kallio[ˆ], Leena Karlsson[ˆ], Miia Kähönen[µ◇], Mikael Knip, M.D., Ph.D.[*±5], Lea Kovanen[*±§], Mirva Koreasalo[*±§2], Kalle Kurppa, M.D., Ph.D.[*±13], Tiina Latva-aho[µ◇], Maria Lönnrot, M.D., Ph.D.[*±6], Elina Mäntymäki[ˆ], Katja Multasuo[µ◇], Juha Mykkänen, Ph.D.[¥3], Tiina Niininen[±*12], Sari Niinistö[±§2], Mia Nyblom[*±], Petra Rajala[ˆ], Jenna Rautanen[±§], Anne Riikonen[*±§], Mika Riikonen[ˆ], Minna Romo[ˆ], Juulia Rönkä[µ◇], Jenni Rouhiainen[ˆ], Tuula Simell, Ph.D., Ville Simell[ˆ¥13], Maija Sjöberg[¥ˆ12,14], Aino Stenius[µ◇12], Maria Leppänen[ˆ], Sini Vainionpää[ˆ], Eeva Varjonen[¥ˆ12], Riitta Veijola, M.D., Ph.D.[µ◇14], Suvi M. Virtanen, M.D., Ph.D.[*±§2], Mari Vähä-Mäkilä[ˆ], Mari Åkerlund[*±§], Katri Lindfors, Ph.D.[*13] [¥]University of Turku, [*]University of Tampere, [µ]University of Oulu, [ˆ]Turku University Hospital, Hospital District of Southwest Finland, [±]Tampere University Hospital, [◇]Oulu University Hospital, [§]National Institute for Health and Welfare, Finland, [¶]University of Kuopio.

**Georgia/Florida Clinical Center:** Jin-Xiong She, Ph.D., PI[1,3,4,11], Desmond Schatz, M.D.[*4,5,7,8], Diane Hopkins[12], Leigh Steed[12,13,14,15], Jamie Thomas[*6,12], Janey Adams[*12], Katherine Silvis[2], Michael Haller, M.D.[*14], Melissa Gardiner, Richard McIndoe, Ph.D., Ashok Sharma, Joshua Williams, Gabriela Young, Stephen W. Anderson, M.D.[ˆ], Laura Jacobsen, M.D.[*14] Center for Biotechnology and Genomic Medicine, Augusta University. [*]University of Florida, [ˆ]Pediatric Endocrine Associates, Atlanta.

**Germany Clinical Center:** Anette G. Ziegler, M.D., PI[1,3,4,11], Andreas Beyerlein, Ph.D.[2], Ezio Bonifacio, Ph.D.[*5], Anja Heublein, Michael Hummel, M.D.[13], Sandra Hummel, Ph.D.[2], Annette Knopff[7], Charlotte Koch, Sibylle Koletzko, M.D.[¶13], Claudia Ramminger, Roswith Roth, Ph.D.[9], Marlon Scholz, Laura Schulzik[2], Joanna Stock[9,12,14], Katharina Warncke, M.D.[14], Lorena Wendel, Christiane Winkler, Ph.D.[2,12,15]. Forschergruppe Diabetes e.V. and Institute of Diabetes Research, Helmholtz Zentrum München, Forschergruppe Diabetes, and Klinikum rechts der Isar, Technische Universität München. [*]Center for Regenerative Therapies, TU Dresden, [¶]Dr. von Hauner Children's Hospital, Department of Gastroenterology, Ludwig Maximillians University Munich.

**Sweden Clinical Center:** Åke Lernmark, Ph.D., PI[1,3,4,5,6,8,10,11,15], Daniel Agardh, M.D., Ph.D.[13], Carin Andrén Aronsson, Ph.D.[2,12,13], Maria Ask, Jenny Bremer, Ulla-Marie Carlsson, Corrado Cilio, Ph.D., M.D.[5], Emelie Ericson-Hallström, Annika Fors, Lina Fransson, Thomas Gard, Rasmus Bennet, Carina Hansson, Susanne Hyberg, Hanna Jisser, Fredrik Johansen, Berglind Jonsdottir, M.D., Silvija Jovic, Helena Elding Larsson, M.D., Ph.D.[6,14], Marielle Lindström, Markus Lundgren, M.D.[14], Maria MÅnsson-Martinez, Maria Markan, Jessica Melin[12], Zeliha Mestan, Caroline Nilsson, Karin Ottosson, Kobra Rahmati, Anita Ramelius, Falastin Salami, Sara Sibthorpe, Anette Sjöberg, Birgitta Sjöberg, Evelyn Tekum Amboh, Carina Törn, Ph.D.[3,15], Anne Wallin, Åsa Wimar[14], Sofie Åberg. Lund University.

**Washington Clinical Center:** William A. Hagopian, M.D., Ph.D., PI[1,3,4,5,6,7,11,13,14], Michael Killian[6,7,12,13], Claire Cowen Crouch[12,14,15], Jennifer Skidmore[2], Ashley Akramoff, Jana Banjanin, Masumeh Chavoshi, Kayleen Dunson, Rachel Hervey, Shana Levenson, Rachel Lyons, Arlene Meyer, Denise Mulenga, Davey Schmitt, Julie Schwabe. Pacific Northwest Research Institute.

**Pennsylvania Satellite Center:** Dorothy Becker, M.D., Margaret Franciscus, MaryEllen Dalmagro-Elias Smith2, Ashi Daftary, M.D., Mary Beth Klein, Chrystal Yates. Children's Hospital of Pittsburgh of UPMC.

**Data Coordinating Center:** Jeffrey P. Krischer, Ph.D., PI[1,4,5,10,11], Sarah Austin-Gonzalez, Maryouri Avendano, Sandra Baethke, Rasheedah Brown[12,15], Brant Burkhardt, Ph.D.[5,6], Martha Butterworth[2], Joanna Clasen, David Cuthbertson, Christopher Eberhard, Steven Fiske[9], Dena Garcia, Jennifer Garmeson, Veena Gowda, Kathleen Heyman, Belinda Hsiao, Francisco Perez Laras, Hye-Seung Lee, Ph.D.[1,2,13,15], Shu Liu, Xiang Liu, Ph.D.[2,3,9,14], Kristian Lynch, Ph.D.[5,6,9,15], Colleen Maguire, Jamie Malloy, Cristina McCarthy[12,15], Aubrie Merrell, Steven Meulemans, Hemang Parikh, Ph.D.[3], Ryan Quigley, Cassandra Remedios, Chris Shaffer, Laura Smith, Ph.D.[9,12], Susan Smith[12,15], Noah Sulman, Ph.D., Roy Tamura, Ph.D.[1,2,13], Ulla Uusitalo, Ph.D.[2,15], Kendra Vehik, Ph.D.[4,5,6,14,15], Ponni Vijayakandipan, Keith Wood, Jimin Yang, Ph.D., R.D.[2,15]. Past staff: Michael Abbondondolo, Lori Ballard, David Hadley, Ph.D., Wendy McLeod. University of South Florida.

**Project scientist:** Beena Akolkar, Ph.D.[1,3,4,5,6,7,10,11]. National Institutes of Diabetes and Digestive and Kidney Diseases.

**Proteomics Laboratory:** Richard D. Smith, Ph.D., Thomas O. Metz, Ph.D., Charles Ansong, Ph.D., Bobbie-Jo Webb-Robertson, Ph.D., Hugh D. Mitchell, Ph.D., Ernesto S. Nakayasu, Ph.D., and Wei-Jun Qian, Ph.D. Pacific Northwest National Laboratory.

**Repository:** Sandra Ke, Niveen Mulholland, Ph.D. NIDDK Biosample Repository at Fisher BioServices.

**Other contributors:** Kasia Bourcier, Ph.D.[5], National Institutes of Allergy and Infectious Diseases. Thomas Briese, Ph.D.[6,15], Columbia

University. Suzanne Bennett Johnson, Ph.D.[9,12], Florida State University. Eric Triplett, Ph.D.[6], University of Florida.

**Committees:** [1]Ancillary Studies, [2]Diet, [3]Genetics, [4]Human Subjects/Publicity/Publications, [5]Immune Markers, [6]Infectious Agents, [7]Laboratory Implementation, [8]Maternal Studies, [9]Psychosocial, [10]Quality Assurance, [11]Steering, [12]Study Coordinators, [13]Celiac Disease, [14]Clinical Implementation, [15]Quality Assurance Subcommittee on Data Quality.

* The TEDDY Study Group is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, and UC4 DK100238 and by Contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work is supported in part by the National Institutes of Health/National Center for Advancing Translational Sciences Clinical and Translational Science Awards UL1 TR000064 (University of Florida) and the University of Colorado (UL1 TR001082), and TEDDY grant UC4 DK100238. Proteomics measurements were obtained using capabilities developed partially under National Institutes of Health grant P41GM103493 and were performed in the Environmental Molecular Sciences Laboratory, a U.S. Department Of Energy (DOE) sponsored national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA. Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830.

S This article contains supplemental material.

‖ To whom correspondence should be addressed: Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999, MSIN: K8-78, Richland, WA 99352. E-mail: Thomas.metz@pnnl.gov or bobbie-jo.webb-robertson@pnnl.gov.

Author contributions: B.A.S., E.S.N., L.M.B., B.-J.W.-R., and T.O.M. designed research; B.A.S., E.S.N., L.M.B., A.M.T., C.K.A., T.C., M.A.G., M.E.M., R.J.M., D.J.O., P.D.P., and A.A.S. performed research; B.A.S., E.S.N., L.M.B., A.M.T., C.K.A., M.E.M., D.J.O., P.D.P., B.-J.W.-R., and T.O.M. analyzed data; B.A.S., E.S.N., B.-J.W.-R., and T.O.M. wrote the paper; R.D.S. and T.S.G. contributed new reagents/analytic tools.

REFERENCES

1. Batt, A. L., Furlong, E. T., Mash, H. E., Glassmeyer, S. T., and Kolpin, D. W. (2017) The importance of quality control in validating concentrations of contaminants of emerging concern in source and treated drinking water samples. *Sci. Total Environ.* **579,** 1618–1628
2. Kocher, T., Pichler, P., Swart, R., and Mechtler, K. (2011) Quality control in LC-MS/MS. *Proteomics* **11,** 1026–1030
3. Mischak, H., Apweiler, R., Banks, R. E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J. H., Fliser, D., Girolami, M., Hermjakob, H., Hochstrasser, D., Jankowski, J., Julian, B. A., Kolch, W., Massy, Z. A., Neusuess, C., Novak, J., Peter, K., Rossing, K., Schanstra, J., Semmes, O. J., Theodorescu, D., Thongboonkerd, V., Weissinger, E. M., Van Eyk, J. E., and Yamamoto, T. (2007) Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics Clin. Appl.* **1,** 148–156
4. Matzke, M. M., Waters, K. M., Metz, T. O., Jacobs, J. M., Sims, A. C., Baric, R. S., Pounds, J. G., and Webb-Robertson, B. J. (2011) Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics* **27,** 2866–2872
5. Bittremieux, W., Valkenborg, D., Martens, L., and Laukens, K. (2017) Computational quality control tools for mass spectrometry proteomics. *Proteomics* **17,** 1–11
6. Piehowski, P. D., Petyuk, V. A., Orton, D. J., Xie, F., Moore, R. J., Ramirez-Restrepo, M., Engel, A., Lieberman, A. P., Albin, R. L., Camp, D. G., Smith, R. D., and Myers, A. J. (2013) Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *J. Proteome Res.* **12,** 2128–2137
7. Pichler, P., Mazanek, M., Dusberger, F., Weilnbock, L., Huber, C. G., Stingl, C., Luider, T. M., Straube, W. L., Kocher, T., and Mechtler, K. (2012) SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *J. Proteome Res.* **11,** 5540–5547
8. Scheltema, R. A., and Mann, M. (2012) SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11,** 3458–3466
9. Taylor, R. M., Dance, J., Taylor, R. J., and Prince, J. T. (2013) Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics* **29,** 2948–2949
10. Bereman, M. S., Johnson, R., Bollinger, J., Boss, Y., Shulman, N., MacLean, B., Hoofnagle, A. N., and MacCoss, M. J. (2014) Implementation of statistical process control for proteomic experiments via LC MS/MS. *J. Am. Soc. Mass Spectrom.* **25,** 581–587
11. Wang, X., Chambers, M. C., Vega-Montoto, L. J., Bunk, D. M., Stein, S. E., and Tabb, D. L. (2014) QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* **86,** 2497–2509
12. 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., and Payne, S. H. (2014) Signatures for mass spectrometry data quality. *J. Proteome Res.* **13,** 2215–2222
13. Bielow, C., Mastrobuoni, G., and Kempa, S. (2016) Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **15,** 777–787
14. Bittremieux, W., Meysman, P., Martens, L., Valkenborg, D., and Laukens, K. (2016) Unsupervised quality assessment of mass spectrometry proteomics experiments by multivariate quality control metrics. *J. Proteome Res.* **15,** 1300–1307
15. Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009) LoOP: local outlier probabilities. *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1649–1652, ACM
16. 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., and Tabb, D. L. (2012) QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal. Chem.* **84,** 5845–5850
17. Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
18. Hagopian, W. A., Lernmark, A., Rewers, M. J., Simell, O. G., She, J. X., Ziegler, A. G., Krischer, J. P., and Akolkar, B. (2006) TEDDY–The Environmental Determinants of Diabetes in the Young: an observational clinical trial. *Ann. N.Y. Acad. Sci.* **1079,** 320–326
19. Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V., Neta, P., Blonder, N., Billheimer, D. D., Blackman, R. K., Bunk, D. M., Cardasis, H. L., Ham, A. J., Jaffe, J. D., Kinsinger, C. R., Mesri, M., Neubert, T. A., Schilling, B., Tabb, D. L., Tegeler, T. J., Vega-Montoto, L., Variyath, A. M., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Carr, S. A., Fisher, S. J., Gibson, B. W., Paulovich, A. G., Regnier, F. E., Rodriguez, H., Spiegelman, C., Tempst, P., Liebler, D. C., and Stein, S. E. (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* **9,** 225–241
20. Ow, S. Y., Salim, M., Noirel, J., Evans, C., Rehman, I., and Wright, P. C. (2009) iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.* **8,** 5347–5355
21. Filzmoser, P., Maronna, R., and Werner, M. (2008) Outlier identification in high dimensions. *Comput. Stat. Data An.* **52,** 1694–1711
22. Li, G. Y., and Chen, Z. L. (1985) Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components - Primary Theory and Monte-Carlo. *J. Am. Stat. Assoc.* **80,** 759–766
23. West, M., and Harrison, J. (1997) *Bayesian forecasting and dynamic models*, 2 Ed., Springer-Verlag New York
24. Monroe, M. E., Shaw, J. L., Daly, D. S., Adkins, J. N., and Smith, R. D. (2008) MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput. Biol. Chem.* **32,** 215–217

25. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5,** 5277
26. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–1292
27. Houee-Levin, C., Bobrowski, K., Horakova, L., Karademir, B., Schoneich, C., Davies, M. J., and Spickett, C. M. (2015) Exploring oxidative modifications of tyrosine: An update on mechanisms of formation, advances in analysis and biological consequences. *Free Radical Res.* **49,** 347–373
28. Ehrenshaft, M., Deterding, L. J., and Mason, R. P. (2015) Tripping up Trp: Modification of protein tryptophan residues by reactive oxygen species, modes of detection, and biological consequences. *Free Radical Bio. Med.* **89,** 220–228
29. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14,** 513–520

# Supplement to "QC-ART: A tool for real-time quality control assessment of mass spectrometry-based proteomics data"

## Details on Scoring Algorithm

The algorithm we use to perform rPCA is as follows. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ represent the $i$th observation of $p$ variables used to represent the $i$th instrument run, then the reduced set of $q < p$ latent variables is given by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})$, which is computed as follows.

1. Robustify the data by subtracting the median and dividing by the median absolute deviation (MAD) for each variable $j = 1, \ldots, p$: $x_{ij}^{*} = \frac{x_{ij} - \text{median}(x_{1j}, \ldots, x_{nj})}{\text{MAD}(x_{1j}, \ldots, x_{nj})}$.

2. Project the data onto the unit sphere by dividing by the length of each observation:

   $$x_{ij}^{**} = \frac{x_{ij}^{*}}{\| \boldsymbol{x}_i^{*} \|}.$$

3. Find the right singular vectors of the robustly sphered data using standard matrix decomposition methodology:

   $$\boldsymbol{X}^{**} = \begin{bmatrix} \boldsymbol{x}_1^{**} \\ \vdots \\ \boldsymbol{x_n^{**}} \end{bmatrix} = \boldsymbol{UDV}^{\top}.$$

4. Retain the $q < p$ latent variables that account for $\alpha$ % of the total uncertainty and project the data into that lower dimension:

   $$\boldsymbol{Z} = \boldsymbol{X}^{**}\boldsymbol{V}^{*} \text{ where } \boldsymbol{V}^{*} = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix}, q \text{ is the smallest } k \text{ such that } \sum_{i=1}^{k} \frac{d_i}{\sum_{j=1}^{p} d_j} \geq \alpha/100.$$

The quantity $\alpha$ determines the number of roughly independent latent variables $q$ used in place of the initial set of $p$ potentially correlated variables. The value $\alpha = 95$ was used in the manuscript; other common values of $\alpha$ are 85, 90 and 99.

Once the raw metrics are transformed via rPCA, a multivariate distance metric is used to assess how similar each instrument run is to those in the baseline. Continuing with the notation above, the distance for the $i$th observation from the baseline set is given by $s_i$, which is computed as follows.

1. Robustly sphere the data in the lower dimension PC space:

$$z_{ij}^{*} = \frac{z_{ij} - \text{median}(z_{1j}, \ldots, z_{nj})}{\text{MAD}(z_{1j}, \ldots, z_{nj})} \text{ and } z_{ij}^{**} = \frac{z_{ij}^{*}}{\| \boldsymbol{z}_i^{*} \|},\ j = 1, \ldots, q$$

2. The QC-ART score for observation $i$ is the norm of the robustly sphere data in the PC space:

$$s_i = \sqrt{\sum_{j=1}^{q} (z_{ij}^{**})^2}$$

## Static and dynamic threshold models

The static and dynamic thresholds used to identify potentially outlying instrument runs are derived from appropriate statistical models, each of which are described next.

### Static threshold

The static threshold is derived from a mean plus noise model fit to the baseline instrument runs. That is, let $s_i$ for $i = 1, \ldots, n_b$ denote the QC-ART for instrument run $i$ then the threshold value is derived from the following linear model

$$\log(s_i) = \mu + \epsilon_i$$

where $\mu$ is the log-scale population mean of all scores and $\epsilon_i$ is the residual noise that is assumed to follow some known distribution. If the error terms $\epsilon_i$ are independent and approximately follow a normal distribution with mean zero and standard deviation $\sigma$ then the threshold value can be set to user chosen percentiles of the log normal distribution with mean $\bar{s} = \sum_{i=1}^{n_b} \log(s_i)/n_b$ and standard deviation $\hat{\sigma}(1 + \frac{1}{n_b})$ where $\hat{\sigma} = \sqrt{\sum_{i=1}^{n_b} [\log(s_i) - \bar{s}]^2 / n_b}$. In this manuscript we used 95$^{\text{th}}$ percentiles of the appropriate log normal distribution.

The prediction uncertainty $\hat{\sigma}(1 + \frac{1}{n_b})$ was used as the distribution standard deviation instead of the standard error $\hat{\sigma}/n_b$ because this model will be applied to new instrument runs that were not used to estimate the mean or standard deviation, thus the additional uncertainty associated with the unseen data must be added.

If appropriate, the static linear model can account for a linear trend in time by adding a term like $\beta t_i$ where $t_i$ is the time since the cohort study began that instrument run $i$ was processed. The regression coefficient $\beta$ is estimated once using only instrument runs in the baseline set.

### Dynamic threshold

To identify local variations in instrument performance, a dynamic model that adapts to local trends must be fit to the data. To do this we used the dynamic linear model defined by the following set of equations

$$\log[s(t_i)] = \mu(t_i) + v(t_i)$$

$$\mu(t_i) = \mu(t_{i-1}) + w(t_i)$$

for $i \geq 1$where $\mu(t_i)$ is the true log-scale process mean at time $t_i$, $v(t_i)$ is an error term in the observation equation and $w(t_i)$ is the error term in the process model. The error terms $v(t_i)$ and $w(t_i)$ are assumed to be independent and follow the log-normal distribution.

In most cases, maximum likelihood can be used to estimate the variance parameters associated with the error terms and the Kalman filter can be used to update estimates of the mean parameter $\mu(t_i)$ as more data are collected and provide one-step-ahead predictions for future observations along with an estimate of uncertainty. In this manuscript, we used 95$^{th}$ percentiles of the distribution of the one-step-ahead predictions to create the dynamic thresholds.

## The Impact of the Chosen Baseline

In the manuscript, we used the SARS-CoV data set as benchmark to compare the accuracy of the proposed QC-ART method relative to the existing methods InSPECtor and RMD-PAV. Using a baseline of size 10 for QC-ART and neighborhood size of 10 for InSPECtor, QC-ART performed comparably in real-time to both post-hoc methods. In this section, we extend this comparison to a much larger set of possible baseline sizes and quantify how the quality of the baseline affects the accuracy as well. In particular, we computed the accuracy of QC-ART and InSPECtor for all baseline sizes from 5 to 70 (half of the full dataset size). We additionally considered a second version of QC-ART where low-quality instrument runs were allowed into the baseline.

In Figure S1, the accuracy of both versions of QC-ART, InSPECtor and RMD-PAV are plotted as a function of baseline size (referred to as neighborhood size for InSPCEtor). A smooth curve was fit to the observed AUC values in order to make comparisons easier, and a 95% confidence interval (gray band around each curve) is plotted to illustrate the uncertainty in the smoothed estimates. The RMD-PAV method does not use a neighborhood size argument and therefore is constant across all values.

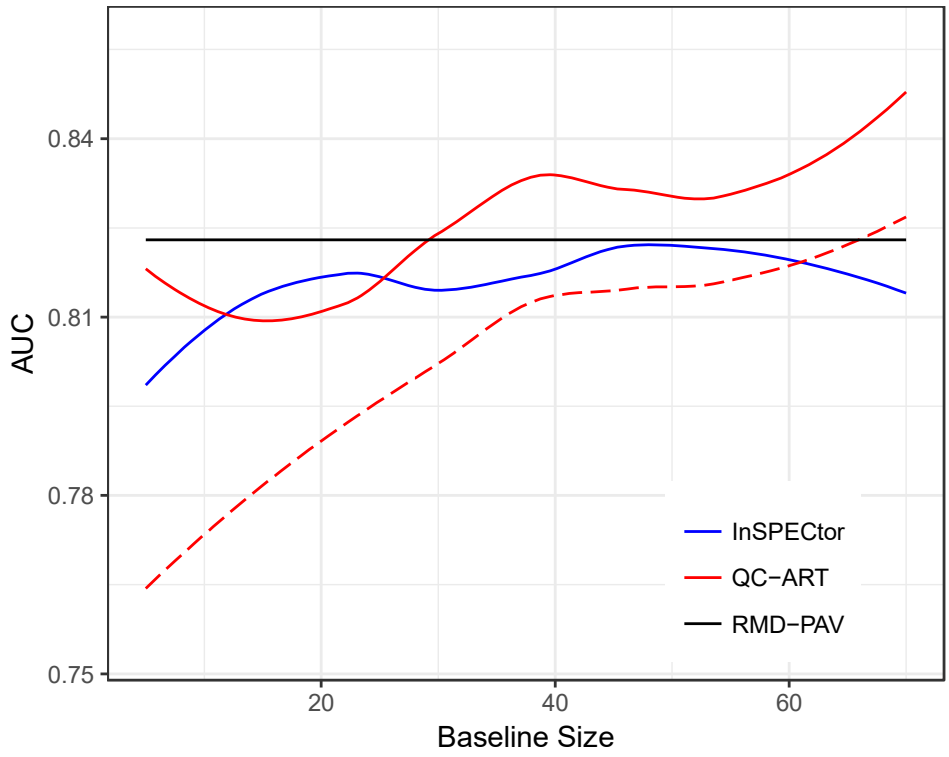


***Fig. S1 Smoothed AUC values with 95% confidence intervals based on the SARS-CoV data set for every baseline size from 5 to 70.*** *The two QC-ART curves are differentiated by the composition of the baseline: only good quality instrument runs (solid red line) or a mix of good and poor quality instrument run (dashed red line).*

For small baselines and baselines larger than 30, the confidence intervals for QC-ART (solid red line) and InSPECtor (blue line) do not overlap, indicating that QC-ART is significantly more accurate than InSPECtor. At no point is InSPECtor significantly more accurate than QC-ART. For baselines greater than 35, QC-ART is significantly more accurate than RMD-PAV (black line), which has an effective baseline size of 140 (the full dataset).

The flatness of the curve corresponding to the QC-ART method when only good quality runs are included in the baseline (solid red line), indicates that small changes in baseline size do not impact the accuracy of QC-ART. However, the increasing trend in accuracy indicates that larger baseline sizes are preferred to smaller ones. As expected, the accuracy of QC-ART is significantly worse when poor quality instrument runs are allowed in the baseline (dashed red line). We can therefore conclude that the quality of the baseline chosen to train QC-ART is more important than the size of the baseline.

## Checking model assumptions

After the static or dynamic model has been fit to the QC-ART baseline scores, the ability of the model to represent the QC-ART scores in the baseline set needs to be assessed. In particular, if a

model is appropriate then the standardized residuals should be independent and resemble random draws from a standard normal distribution. In Fig. S1, we assessed these assumptions for the dynamic linear models fit to the different baseline sets chosen for the iTRAQ data.

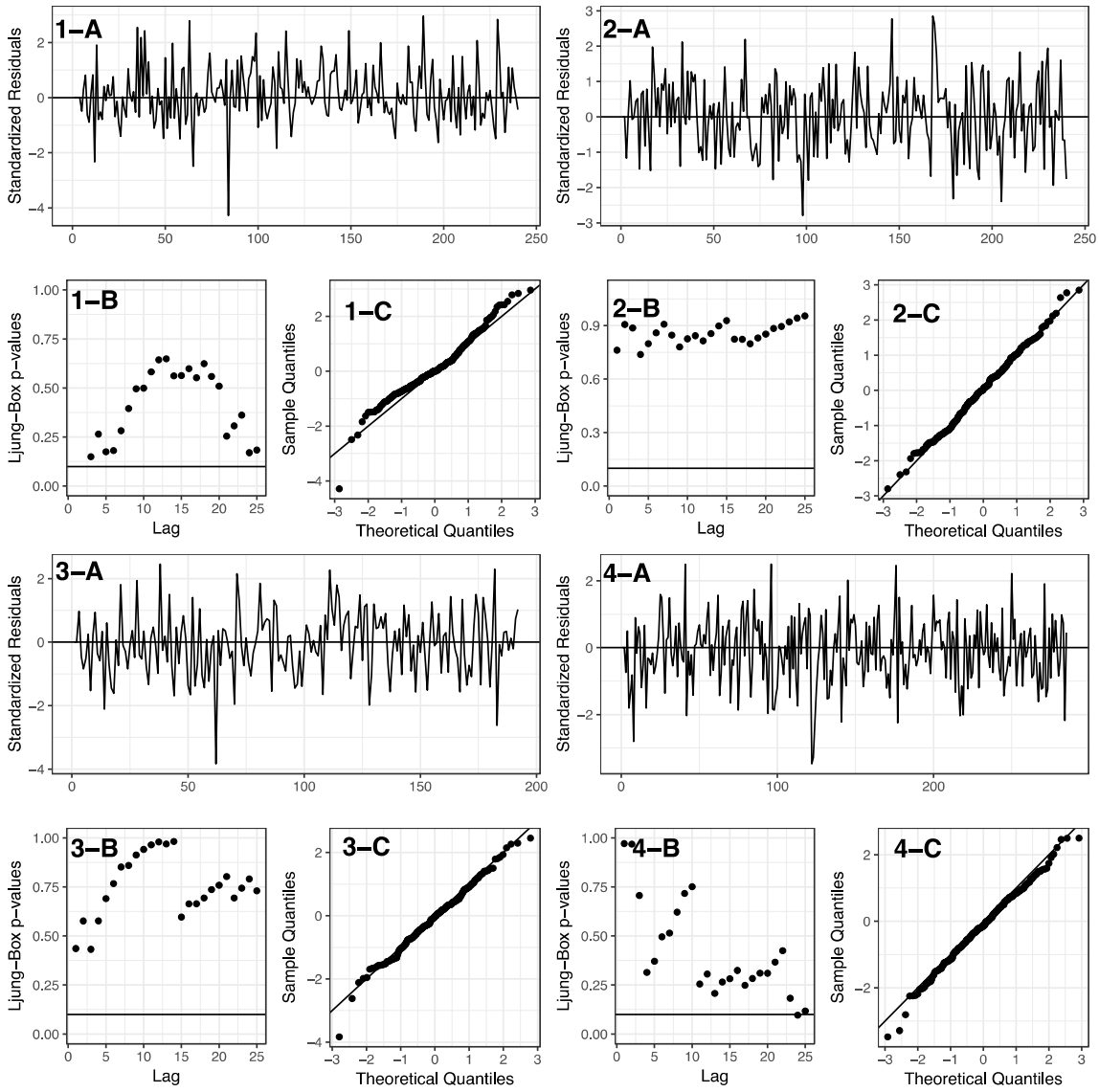


**Fig. S2 Plots to assess dynamic model assumptions for the four different baselines used to analyze the iTRAQ data.** A group of figures was created for each of the four iTRAQ baseline sets: the three groups colored red in Fig. 3B (labelled 1, 2, 3) as well as the baseline set in Fig. 3C. Each group of figures is comprised of a plot of the standardized residuals over time (labelled 'A'), a plot of the Ljung-Box p-values for different lag values (labelled 'B') and a normal quantile-quantile plot (labelled 'C'). The standardized residuals should not show any discernable pattern. The Ljung-Box test assess the degree to which the residuals are correlated at different lags (1); large *p*-values indicate the residuals are not significantly correlated at that or any lower lag. The normal quantile-quantile plot is used to assess the normality of the residuals, the points

should follow the diagonal line closely. See Chapter 2.8 of (2) for a more detailed discussion of model assessment for dynamic linear models.

The assumptions appear to be satisfied for each of the four dynamic linear models (**Fig. S1**). There is a possible outlier in the baseline composed of the first ten sets (see the figure labeled 1-C), but it was retained in the baseline as it was a valid instrument run.

For the SLC003 data, a static linear model was used to determine the threshold. The same assumptions were assessed for the static linear model that was used to determine the threshold for the SLC003 data (**Fig. S2**). Again, the model assumptions appear to be satisfied though it is very difficult to assess the normality of errors given the small sample size.

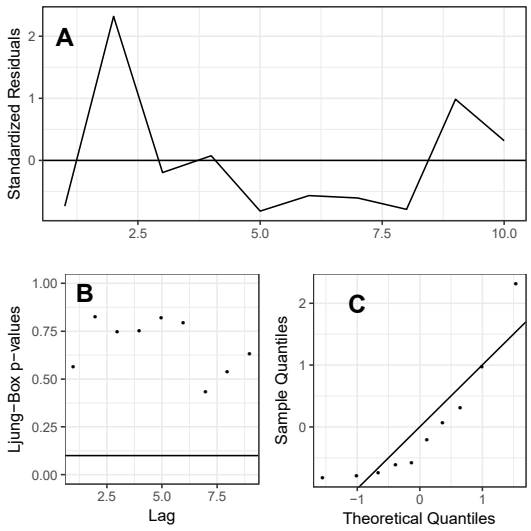


**Fig. S2 Plots to assess linear model assumptions for the baseline set of instrument runs used to analyze the SLC003 data.** All of the linear model assumptions appear to be satisfied .

## References

1. Ljung, G. M., and Box, G. E. (1978) On a measure of lack of fit in time series models. *Biometrika* 65, 297–303

2. Petris, G., Petrone, S., and Campagnoli, P. (2009) *Dynamic linear models with R* (Springer)