QCactus: interactive quality control software for statistical analysis and reporting metrics of raw proteomics data

Zachary L Dwight, MS, MBA**; Nathan Hendricks,** PhD**; Santosh D Bhosale,** PhD**; Jonathan T Bui,** BS**; Monica Ghaly,** MS**; Susan M Mockus,** PhD, MBA
Precision Biomarker Laboratories, Cedars-Sinai Medical Center, Beverly Hills, CA

Precision Biomarker Laboratories

Cedars Sinai

Introduction

Mass spectrometry-based quantitative proteomics is a complex but rich source of valuable health insights and protein biomarkers. The potential of proteomics data to expand the health landscape for patients in multi-omics initiatives and complement genomic information is increasing at a rapid pace. However, proteomic investigation and multi-omics initiatives are becoming increasingly data-centric and highly dependent on the quality and assessment of the information being included or excluded at different points in laboratory and research workflows. Proteomics, compared to other 'omics' fields, is less standardized in context of data processing, file formatting, and structure of electronic information. Often, scientists are evaluating data quality far downstream of acquisition; focused on cohort analysis or visualization of differentially expressed proteins.

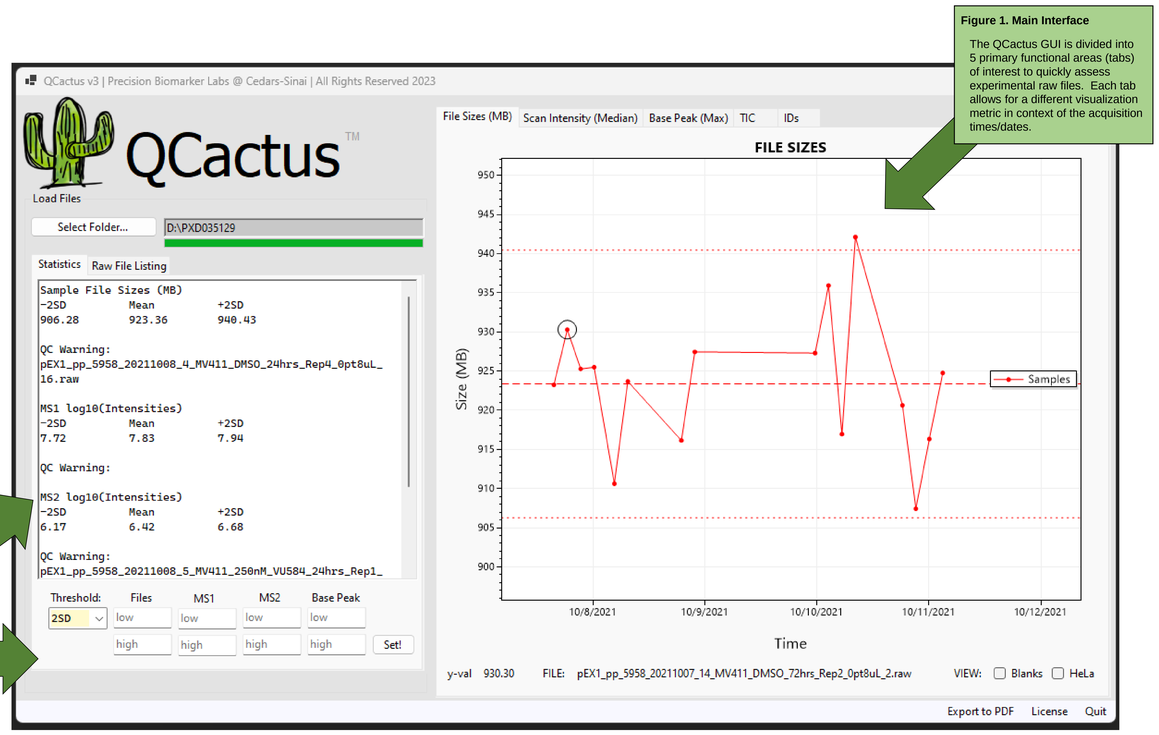


Figure 1. Main Interface

The QCactus GUI is divided into 5 primary functional areas (tabs) of interest to quickly assess experimental raw files. Each tab allows for a different visualization metric in context of the acquisition times/dates.

Figure 1. Main Interface

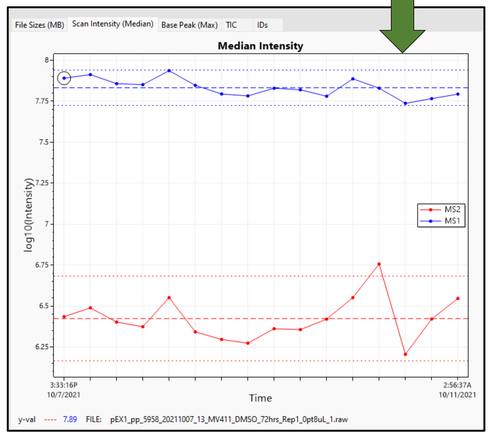
Threshold inputs allow the user to set specific QC standards. If a sample file fails to meet the thresholds, it is reported in the statistics window with a QC warning along with other summarizing statistics.

Methods

Though complex and non-uniform across instrumentation, proteomics data consolidated to derive answers to clinical questions requires computational curation. Focusing on assessment of raw data at acquisition, common mistakes can be avoided that impact laboratory performance and undermine quality of analytical conclusions. Here we introduce QCactus, a vendor independent tool to assess the quality of raw MS data acquired in data independent acquisition (DIA) mode. The areas of focus for assessment that are available via raw file interrogation include identity-free metrics and identity-based benchmarking. Identity-free metrics are focused on the technical characteristics of MS runs and allows for the evaluation of file sizes, scan intensities (MS1 and MS2), base peak intensity, and TIC information. Identity-based metrics include protein and peptide identifications.

Figure 2. Median Intensity View

The figures above and below were created with a public PRIDE data set: **PXD035129**. Below, we can visually identify one file on the MS2 series has fallen outside the 2SD threshold. This file is identified and listed in the statistics window with a warning as well. This file may need additional follow up.



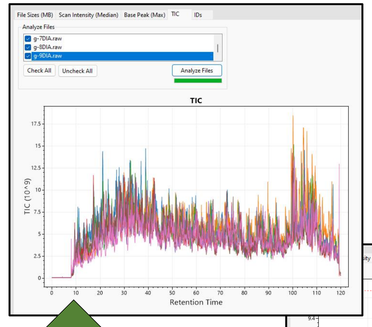
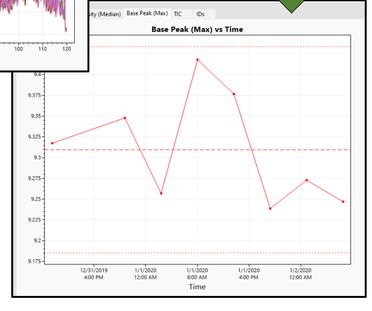


Figure 3a. Assessing a Reference Set

Manual visual inspection of the TIC plots also confirmed no issues and that the **PXD043389** set is concise and consistent enough to use as a benchmark.

Figure 3b. Assessing a Reference Set

The set included here, **PXD043389**, is described as a stable proteomics reference made available to help researchers assess multi-omics workflows for accuracy and reproducibility. This data set, 8 files in total, had no QC warnings for any of the non-ID-based metrics or plots



Preliminary Data

QCactus is a quick, statistics focused raw data analysis software to confirm data quality before downstream computational analysis and report standardized metrics and statistical or user-defined thresholds. Computational time is a resource that should be protected from waste like any laboratory process and assessing raw data from the mass spectrometer is critical. First, acquired raw data must pass an integrity check to avoid processing of corrupted files. If raw data pass the integrity check, files are processed and mined for a collection of metrics. Utilizing those metrics and consolidating into statistical measures, files (samples) that fall out of acceptable QC thresholds are highlighted and reported to the user. Additional projects (sets of raw files) may also be added as additional series to compare current runs to historical benchmarks and metrics. All metrics and plots may be exported to a formatted QC report. The software successfully achieves the goal of automating the identification of sample issues through raw file interrogation which is independent of a commercial vendor.

Conclusions

QCactus aids in removing any subjective decisions and allows scientists to quickly address problems before they utilize resources or waste time and effort downstream. QCactus was developed to alleviate the need for costly repeated experiments by incorporating quality control metrics into an easy-to-use interface at the acquisition phase. C# and the .NET framework were chosen to best build an interactive desktop application that could provide the desired processing, statistics, visualization, and reporting features. Identity-based metrics are calculated via an embedded version of MSFragger. QCactus is freely available per GitHub and without registration with links and materials found at **https://www.cs-pbl.com/software**.

Conflict of Interest: None to declare.

Contact Information: Zachary.Dwight@cshs.org