

Technical Brief

QCMAP: an interactive web-tool for performance diagnosis and prediction of LC-MS systems

Taiyun Kim^{1,5}, Irene Rui Chen^{1,5}, Benjamin L. Parker², Sean J. Humphrey², Ben Crossett³, Stuart J. Cordwell^{2,3}, Pengyi Yang^{1,4*}, Jean Yee Hwa Yang^{1,5*}

¹School of Mathematics and Statistics, ²School of Life and Environmental Sciences, ³School of Mass Spectrometry, ⁴Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, and ⁵Judith and David Coffey Life Lab, Charles Perkins Centre, University of Sydney, NSW 2006, Australia

Corresponding author:

Jean Yee Hwa Yang and Pengyi Yang

School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia

Email: pengyi.yang@sydney.edu.au or jean.yang@sydney.edu.au

Phone: 61-93513012

Fax: 61-93514534

Keywords:

Proteomics / LC-MS / Performance diagnosis / Interactive web-tool / Prediction

Abstract

The increasing role played by liquid chromatography-mass spectrometry (LC-MS)-based proteomics in biological discovery has led to a growing need for quality control on the LC-MS systems. While numerous quality control tools have been developed to track the performance of LC-MS systems based on a pre-defined set of performance factors (e.g. mass error, retention time), the precise influence and contribution of the performance factors and their generalization property to different biological samples are not as well characterized. Here, we developed a web-based application (QCMAP) for interactive diagnosis and prediction of the performance of LC-MS systems across different biological sample types. Leveraging on a standardized HeLa cell sample run as QC within a multi-user facility, we trained predictive models on a panel of commonly used performance factors to pinpoint the precise conditions to a (un)satisfactory performance in three LC-MS systems. We demonstrated that the learned model can be applied to predict LC-MS system performance for brain samples generated from an independent study. By compiling these predictive models into our web-application, QCMAP allows users to benchmark the performance of their LC-MS systems using their own samples and identify key factors for instrument optimization. QCMAP is freely available from: <http://shiny.maths.usyd.edu.au/QCMAP/>.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/pmhc.201900068](https://doi.org/10.1002/pmhc.201900068).

This article is protected by copyright. All rights reserved.

Liquid chromatography-mass spectrometry (LC-MS) is a maturing technology for high-throughput proteomics. Given the increasing role of LC-MS in biological discovery and clinical applications, quality control (QC) of LC-MS systems has become a prerequisite for MS-based proteomics experiments [1]. Numerous publicly available LC-MS QC tools have been developed [2] and many multi-user, core facilities also implement QC measures as part of their routine workflows [3]. Examples of popular public QC tools include MSQC [4], QuaMeter [5], SIMPATIQCO [6], SPROCoP [7], qcML for openMS [8], and PTXQC [9]. While these tools excel in assessment of overall performance of a LC-MS system, few of them offer an interactive and diagnostic view of individual performance factors such as mass error, percentage of MS/MS identified, and retention time. To optimize the instrument for best performance, it is particularly useful to analyze the performance factors leading to (un)satisfactory performance of the LC-MS system. Indeed, several recent LC-MS QC pipelines have taken into account the analysis and visualization of individual performance factors [10], [11]. However, these tools often restrict the analysis to a set of control samples such as those from CPTAC [12]. The precise characterization of factors that influence the performance of multiple LC-MS systems and their generalization properties across different biological sample types and organisms is still lacking.

Here, we developed an interactive web-based application for quality control of mass spectrometry-based proteomics systems (QCMAP). In particular, we trained predictive models on a standardized HeLa cell protein sample to characterize a panel of commonly utilized performance factors and their contribution to the performance of LC-MS systems. We demonstrated that the model trained on the HeLa sample can be applied to predict LC-MS system performance in analyzing brain samples obtained from mice in an independent study. Taken together, the proposed web-application provides (i) overall performance benchmark, (ii) interactive analysis of performance factors for multiple LC-MS systems, and also (iii) predictive models that are generalizable across different species and can be applied to assess LC-MS systems using diverse biological sample types.

The workflow of QCMAP is shown in Figure 1. Key components in QCMAP are described in detail below.

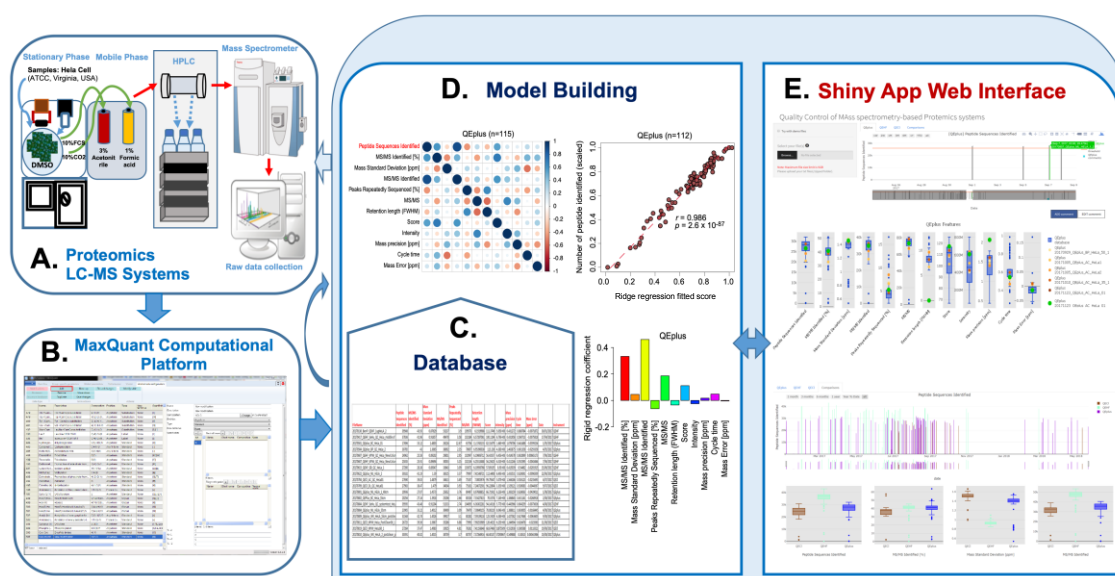


Figure 1. Schematic illustration of QCMAP web-application workflow.

Firstly, to train a set of models for performance assessment of multiple LC-MS systems, we compiled a total of 223 datasets generated from a standardized sample (1 μ g HeLa cell protein followed by tryptic digestion) by Sydney Mass Spectrometry core facility using three state-of-the-art LC-MS instruments – Q Exactive HF (QEHF), QEplus, and QEClassic (QECI) (Table S2). Raw LC-MS data (Figure. 1A) were first processed by the MaxQuant software [13] (Figure. 1B) and the text outputs (13 plain text files) for each dataset were collected to form a performance benchmark database (Figure. 1C). To pinpoint the performance of LC-MS systems, 11 commonly utilized performance factors [14] were extracted from each dataset in the database (Figure. S1). The 11 performance factors are listed in Table S1 and their values were standardized as described in the model building section of the supplementary file. Next, by using the number of unique peptides identified in the standardised HeLa sample as a surrogate of performance of each LC-MS system, a set of multivariate linear models (ridge regression) were trained to learn the relative importance of these 11 performance factors in contribution to the overall performance of the three LC-MS systems (Figure. 1D; Figure. S2 and S3A). Note that high peptide identification rates require optimal performance across the entire platform including chromatography, ionization, ion transfer, precursor detection and selection, fragmentation and MS/MS detection etc. Hence, a simplistic measure of unique peptides is beneficial for rapid assessment of the entire platform by all users. Ridge regression is a highly interpretable statistical model and an ideal choice for assessing the importance of the 11 performance factors provided they are scaled to a similar numeric range.

Secondly, for an interactive diagnosis of instrument and LC-MS systems, QCMAP allows users to simultaneously visualize and interpret data by providing an overall historic view and a set of boxplots for individual performance factors (Figure. 1E). While the overall historic view enables multiscale to summarize instrumentation performance over time, boxplots of performance factors illustrate how changes of other factors could affect a LC-MS system. We have also generated a ridge regression model for each instrument for accurate diagnostics of the instrument and assessed the contribution of each feature. The contributing features for each model were assessed and showed high correlation between instruments (Figure. S3B).

To utilise the predictive models for performance assessment, a user can upload the MaxQuant output files (either as a zip file or individual files) and a newly input dataset will instantly appear as a red bar compared to the existing database in the histogram plots (Figure. 1E). The top panel of Figure 1E shows the historical plot from weekly, monthly to yearly. This provides a quick summary view of the number of 'Peptide Sequences Identified'. The horizontal line in the bar plot indicates the different thresholds for different instruments. The selected bar in the plot is highlighted green when the red bar is clicked. The user can observe the value of features in the boxplot for the corresponding file selected from the bar plot (as shown in the middle panel of Figure. 1E).

On the 11 boxplots, the five latest datasets by date are shown in color gradient (yellow to brown) to help the user to identify the reason for any failure if a particular dot is outside the normal range of the boxplot. For example, if 'Mass precision [ppm]' shows a decreasing value along with the color gradient for the last 5 readings, then that indicates a suboptimal performance of the instrument on mass detection. The bottom panel of Figure. 1E is the comparison of different instruments for each parameter.

Thirdly, the predictive models trained from QCMAP are generalizable to different biological samples. To demonstrate this, we generated 10 datasets on a QECI instrument from mouse brain samples with different levels of quality (PRIDE: PXD010307). Four datasets were acquired directly after the instrument was cleaned and are high quality while six datasets were acquired on a dirty instrument producing different degrees of lower quality data. By applying our model that we trained for the QECI using HeLa samples, we accurately predicted the performance of the instrument with the four datasets acquired on a clean instrument having higher predicted scores using only the 11 performance factors in each brain sample (Figure. 2A). Furthermore, the lower quality data correspond to lower percentage of MS/MS identified while the FWHM remains largely unchanged for all datasets (Figure. 2B),

indicating the drop of performance is due to poorer mass spectral quality and not reduced chromatography performance. These results demonstrate that the trained models from QCMAP are predictive to diverse biological samples and is not locked or limited to standard quality control sample such as the HeLa samples. As such, this enables users to provide their own data generated from different biological samples to assess the performance of their LC-MS system. The generalization property of the trained predictive models greatly increases the applicability of QCMAP.

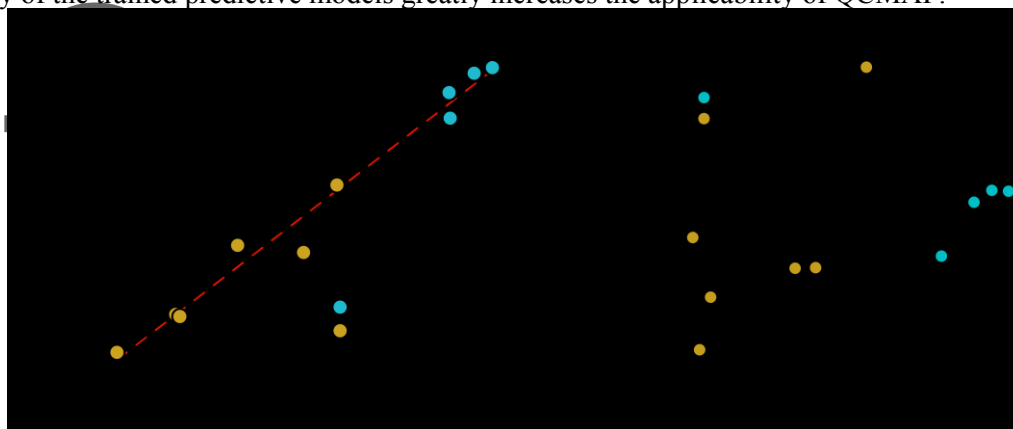


Figure 2: Prediction analysis of the 10 mouse brain samples generated from LC-MS using QECI instrument. **A.** Rigid regression prediction scores of mouse brain samples (x-axis) and the number of peptides identified from each sample (y-axis). Correlation and p-value are calculated from a linear fit of the Rigid regression prediction scores and the actual number of peptides identified. **B.** Scatter plot of percentage of MS/MS identified against FWHM.

Current models in QCMAP were trained for datasets from data-dependent acquisition (DDA) method. However, additional models can be trained and included for datasets acquired from other methods (e.g. data-independent acquisition (DIA)) on any instrument platform supported by MaxQuant. Peptide load is an important consideration when assessing instrument performance on proteome depth. Although QCMAP does not restrict the peptide load used for LC-MS system performance assessment, a consistent peptide load is required to assess instrument performance over longer periods. We also note that QCMAP is dedicated to rapid QC and monitoring of instrument performance and not the sample. Hence, the identical sample needs to be acquired over time. Lastly, while we have demonstrated that power of QCMAP in isolating performance factors on the LC and the MS system using the mouse brain dataset in this study, in our future work, we aim to generate additional synthetic datasets that capture unique performance factors associated with LC and MS systems; and generalizing the tool for samples obtained from affinity-purification (AP)-MS.

Acknowledgements

We thank members from the School of Mathematics and Statistics, School of Life and Environmental Sciences, and Sydney Mass Spectrometry for their feedback on this work. Judith & David Coffey Life Lab provided financial support.

Funding: This work is supported by ARC/DP (DP170100654) to JYHY and PY; ARC/DECRA (DE170100759) to PY; NHMRC/ECF (1072129) to BLP and NHMRC/CDF (1111338) to JYHY.

Conflict of Interest: none declared.

References

- [1] W. Bittremieux, D. L. Tabb, F. Impens, A. Staes, E. Timmerman, L. Martens, and K. Laukens, "Quality control in mass spectrometry-based proteomics," *Mass Spectrometry Reviews*. 2018.
- [2] W. Bittremieux, D. Valkenburg, L. Martens, and K. Laukens, "Computational quality control tools for mass spectrometry proteomics," *Proteomics*. 2017.
- [3] T. Köcher, P. Pichler, R. Swart, and K. Mechtler, "Quality control in LC-MS/MS," *Proteomics*. 2011.
- [4] P. A. Rudnick, K. R. Clauser, L. E. Kilpatrick, D. V. Tchekhovskoi, P. Neta, N. Blonder, D. D. Billheimer, R. K. Blackman, D. M. Bunk, H. L. Cardasis, A.-J. L. Ham, J. D. Jaffe, C. R. Kinsinger, M. Mesri, T. A. Neubert, B. Schilling, D. L. Tabb, T. J. Tegeler, L. Vega-Montoto, A. M. Variyath, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. A. Carr, S. J. Fisher, B. W. Gibson, A. G. Paulovich, F. E. Regnier, H. Rodriguez, C. Spiegelman, P. Tempst, D. C. Liebler, S. E. Stein, "Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses," *Mol. Cell. Proteomics*, 2010.
- [5] Z. Q. Ma, K. O. Polzin, S. Dasari, M. C. Chambers, B. Schilling, B. W. Gibson, B. Q. Tran, L. Vega-Montoto, D. C. Liebler, and D. L. Tabb, "QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation," *Anal. Chem.*, 2012.
- [6] P. Pichler, M. Mazanek, F. Dusberger, L. Weilnböck, C. G. Huber, C. Stingl, T. M. Luider, W. L. Straube, T. Köcher, and K. Mechtler, "SIMPATIQCO: A server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on orbitrap instruments," *J. Proteome Res.*, 2012.
- [7] M. S. Bereman, R. Johnson, J. Bollinger, Y. Boss, N. Shulman, B. MacLean, A. N. Hoofnagle, and M. J. MacCoss, "Implementation of statistical process control for proteomic experiments via LC MS/MS," *J. Am. Soc. Mass Spectrom.*, 2014.
- [8] M. Walzer, L. E. Pernas, S. Nasso, W. Bittremieux, S. Nahnsen, P. Kelchtermans, P. Pichler, H. W. P. van den Toorn, A. Staes, J. Vandenbussche, M. Mazanek, T. Taus, R. A. Scheltema, C. D. Kelstrup, L. Gatto, B. van Breukelen, S. Aiche, D. Valkenburg, K. Laukens, K. S. Lilley, J. V. Olsen, A. J. Heck, K. Mechtler, R. Aebersold, K. Gevaert, J. A. Vizcaíno, H. Hermjakob, O. Kohlbacher, L. Martens, "qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments," *Mol. Cell. Proteomics*, 2014.
- [9] C. Bielow, G. Mastrobuoni, and S. Kempa, "Proteomics Quality Control: Quality Control Software for MaxQuant Results," *J. Proteome Res.*, 2016.
- [10] C. Chiva, R. Olivella, E. Borràs, G. Espadas, O. Pastor, A. Solé, and E. Sabidó, "QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories," *PLoS One*, 2018.
- [11] X. Wang, M. C. Chambers, L. J. Vega-Montoto, D. M. Bunk, S. E. Stein, and D. L. Tabb, "QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics," *Anal. Chem.*, 2014.
- [12] N. J. Edwards, M. Oberti, R. R. Thangudu, S. Cai, P. B. McGarvey, S. Jacob, S. Madhavan, and K. A. Ketchum, "The CPTAC data portal: A resource for cancer proteomics research," *J. Proteome Res.*, 2015.
- [13] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat. Biotechnol.*, 2008.
- [14] B. G. Amidan, D. J. Orton, B. L. Lamarche, M. E. Monroe, R. J. Moore, A. M. Venzin, R. D.

Smith, L. H. Sego, M. F. Tardiff, and S. H. Payne, "Signatures for mass spectrometry data quality," *J. Proteome Res.*, 2014.

Author Manuscript

This article is protected by copyright. All rights reserved.