QCMAP: An Interactive Web-Tool for Performance Diagnosis and Prediction of LC-MS Systems

Taiyun Kim, Irene Rui Chen, Benjamin L. Parker, Sean J. Humphrey, Ben Crossett, Stuart J. Cordwell, Pengyi Yang, and Jean Yee Hwa Yang**

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 (QC) 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, a web-based application (QCMAP) is developed 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, predictive models are trained on a panel of commonly used performance factors to pinpoint the precise conditions to a (un)satisfactory performance in three LC-MS systems. It is 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/.

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 clinical proteomic tumor analysis consortium (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

T. Kim, I. R. Chen, Dr. P. Yang, Prof. J. Y. H. Yang
School of Mathematics and Statistics
University of Sydney
NSW 2006, Australia
E-mail: pengyi.yang@sydney.edu.au; jean.yang@sydney.edu.au

Dr. B. L. Parker, Dr. S. J. Humphrey, Prof. S. J. Cordwell
School of Life and Environmental Sciences
University of Sydney
NSW 2006, Australia

Dr. B. Crossett, Prof. S. J. Cordwell
Sydney Mass Spectrometry
University of Sydney
NSW 2006, Australia

Dr. P. Yang
Computational Systems Biology Group
Children's Medical Research Institute
Faculty of Medicine and Health
University of Sydney
Westmead, NSW 2145, Australia

T. Kim, I. R. Chen, Prof. J. Y. H. Yang
Judith and David Coffey Life Lab
Charles Perkins Centre
University of Sydney
NSW 2006, Australia

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/pmic.201900068

DOI: 10.1002/pmic.201900068

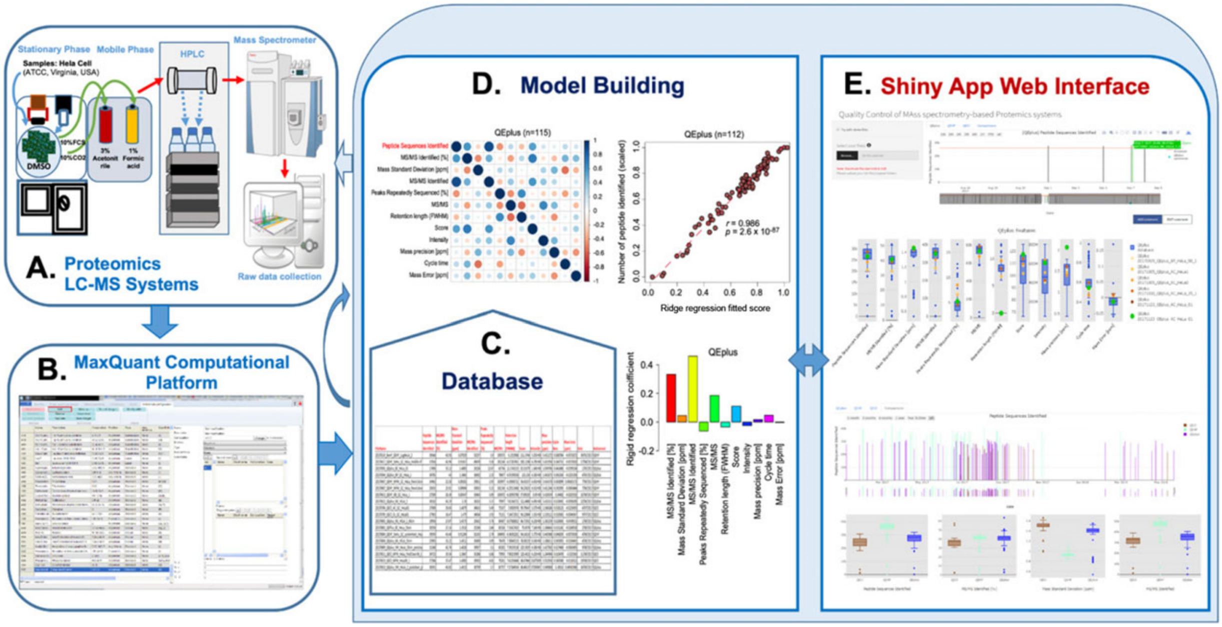
Stationary Phase
Mobile Phase
HPLC
Mass Spectrometer
Samples: Hela Cell (ATCC, Virginia, USA)
10%FCS
10%CO2
DMSO
3% Acetonitrile
1% Formic acid
Raw data collection
A. Proteomics LC-MS Systems
B. MaxQuant Computational Platform
C. Database
D. Model Building
QEplus (n=115)
Peptide Sequences Identified
MS/MS Identified [%]
Mass Standard Deviation [ppm]
MS/MS Identified
Peaks Repeatedly Sequenced [%]
MS/MS
Retention length (FWHM)
Score
Intensity
Mass precision [ppm]
Cycle time
Mass Error [ppm]
QEplus (n=112)
Number of peptide identified (scaled)
Ridge regression fitted score
r = 0.986
p = 2.6 x 10^-87
QEplus
Rigid regression coefficient
E. Shiny App Web Interface
Quality Control of MAss spectrometry-based Proteomics systems
QEplus features
Peptide Sequences Identified

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

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.

First, 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 μg 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 (QECl) (Table S2, Supporting Information). 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, Supporting Information). The 11 performance factors are listed in Table S1, Supporting Information, 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 standardized 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, Supporting Information). 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.

Second, 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, Supporting Information).

To utilize 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

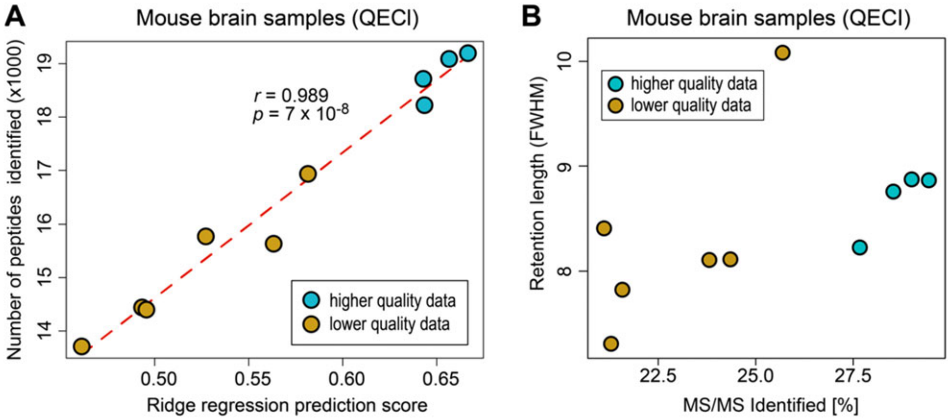

Figure 2. Prediction analysis of the ten mouse brain samples generated from LC-MS using QECl instrument. A) Rigid regression prediction scores of mouse brain samples (x-axis) and the number of peptides identified from each sample (y-axis). Correlation and p-value are calculated from a linear fit of the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

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 five 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 ten datasets on a QECl 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 QECl 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.

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. Last, 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.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors 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. This work was supported by ARC/DP (DP170100654) to J.Y.H.Y. and P.Y.; ARC/DECRA (DE170100759) to P.Y.; NHMRC/ECF (APP1072129) to B.L.P. and NHMRC/CDF (APP1111338) to J.Y.H.Y.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

interactive web-tools, liquid chromatography-mass spectrometry (LC-MS), performance diagnosis, prediction, proteomics

Received: February 18, 2019
Revised: May 7, 2019
Published online: June 13, 2019

[1] W. Bittremieux, D. L. Tabb, F. Impens, A. Staes, E. Timmerman, L. Martens, K. Laukens, *Mass Spectrom. Rev.* **2018**, *37*, 697.

[2] W. Bittremieux, D. Valkenborg, L. Martens, K. Laukens, *Proteomics* **2017**, *17*, 1600159.

[3] T. Köcher, P. Pichler, R. Swart, K. Mechtler, *Proteomics* **2011**, *11*, 1026.

[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 et al., *Mol. Cell. Proteomics* **2010**, *9*, 225.

[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, D. L. Tabb, *Anal. Chem.* **2012**, *84*, 5845.

[6] P. Pichler, M. Mazanek, F. Dusberger, L. Weilnböck, C. G. Huber, C. Stingl, T. M. Luider, W. L. Straube, T. Köcher, K. Mechtler, *J. Proteome Res.* **2012**, *11*, 5540.

[7] M. S. Bereman, R. Johnson, J. Bollinger, Y. Boss, N. Shulman, B. MacLean, A. N. Hoofnagle, M. J. MacCoss, *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 581.

[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. Valkenborg, 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, *Mol. Cell. Proteomics* **2014**, *13*, 1905.

[9] C. Bielow, G. Mastrobuoni, S. Kempa, *J. Proteome Res.* **2016**, *15*, 777.

[10] C. Chiva, R. Olivella, E. Borràs, G. Espadas, O. Pastor, A. Solé, E. Sabidó, *PLoS One* **2018**, *13*, e0189209.

[11] X. Wang, M. C. Chambers, L. J. Vega-Montoto, D. M. Bunk, S. E. Stein, D. L. Tabb, *Anal. Chem.* **2014**, *86*, 2497.

[12] N. J. Edwards, M. Oberti, R. R. Thangudu, S. Cai, P. B. McGarvey, S. Jacob, S. Madhavan, K. A. Ketchum, *J. Proteome Res.* **2015**, *14*, 2707.

[13] J. Cox, M. Mann, *Nat. Biotechnol.* **2008**, *26*, 1367.

[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, S. H. Payne, *J. Proteome Res.* **2014**, *13*, 2215.