Bioinformatics, 38(7), 2022, 2072–2074
https://doi.org/10.1093/bioinformatics/btac040
Advance Access Publication Date: 26 January 2022
Applications Note

Systems biology

RawHummus: an R Shiny app for automated raw data quality control in metabolomics

Yonghui Dong [1], Yana Kazachkova[2], Meng Gou[3], Liat Morgan[1], Tal Wachsman[1], Ehud Gazit[1] and Rune Isak Dupont Birkler[1,]*

[1]Metabolite Medicine Division, BLAVATNIK CENTER for Drug Discovery, Tel Aviv University, Tel Aviv 69978, Israel, [2]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel and [3]College of Life Science, Liaoning Normal University, Dalian 116081, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 10, 2021; revised on January 16, 2022; editorial decision on January 17, 2022; accepted on January 24, 2022

Abstract

Motivation: Robust and reproducible data is essential to ensure high-quality analytical results and is particularly important for large-scale metabolomics studies where detector sensitivity drifts, retention time and mass accuracy shifts frequently occur. Therefore, raw data need to be inspected before data processing to detect measurement bias and verify system consistency.

Results: Here, we present RawHummus, an R Shiny app for an automated raw data quality control (QC) in metabolomics studies. It produces a comprehensive QC report, which contains interactive plots and tables, summary statistics and detailed explanations. The versatility and limitations of RawHummus are tested with 13 metabolomics/lipidomics datasets and 1 proteomics dataset obtained from 5 different liquid chromatography mass spectrometry platforms.

Availability and implementation: RawHummus is released on CRAN repository (https://cran.r-project.org/web/packages/RawHummus), with source code being available on GitHub (https://github.com/YonghuiDong/RawHummus). The web application can be executed locally from the R console using the command 'runGui()'. Alternatively, it can be freely accessed at https://bcdd.shinyapps.io/RawHummus/.

Contact: yonghui.dong@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolomics is a crucial technique in modern biological and biomedical research. Liquid chromatography mass spectrometry (LCMS) is currently the prime method for metabolomics studies owing to its high throughput, soft ionization and excellent metabolite coverage (Yang *et al.*, 2019; Zhou *et al.*, 2012). Unfortunately, detector sensitivity drifts, retention time (RT) and mass accuracy shifts within or across different measurement sequences frequently occur in LCMS, and therefore pose a negative impact on accurate biomarker discovery and system biology studies (Simader *et al.*, 2015). Like other omics techniques such as genomics and proteomics, metabolomics experiments often involve measurements of numerous samples and generate vast amounts of data (Shaham-Niv *et al.*, 2021). As such, rapid raw data quality check is essential to monitor the quality of the analytical run during sample analysis, or before time-consuming data processing and statistical analysis in order to detect any putative sources of error (Simader *et al.*, 2015).

The use of quality control (QC) is now routine to monitor, evaluate and correct system variations in metabolomics studies (Begou *et al.*, 2018). To date, however, only a few QC tools have been developed for metabolomics. Among them, QCScreen (Simader *et al.*, 2015) and MeTaQuaC (Kuhring *et al.*, 2020) are two excellent tools, which provide comprehensive system evaluation based on a set of predefined mass features. However, since system variation does not always occur in a linear fashion, they may fail to determine the overall system performance for untargeted analyses. In contrast, QC tools for proteomics are flourishing, such as PTXQC (Bielow *et al.*, 2016), Qcloud (Chiva *et al.*, 2018), QC-ART (Stanfill *et al.*, 2018) and RawBeans (Morgenstern *et al.*, 2021). In principle, metabolomics data can be also checked using proteomics QC software, however, they are limited to specific vendors and/or designed to evaluate some metrics which are irrelevant to metabolomics studies (e.g. charge distribution).

To this end, we have developed an R shiny application, RawHummus, for rapid and objective raw data evaluation.

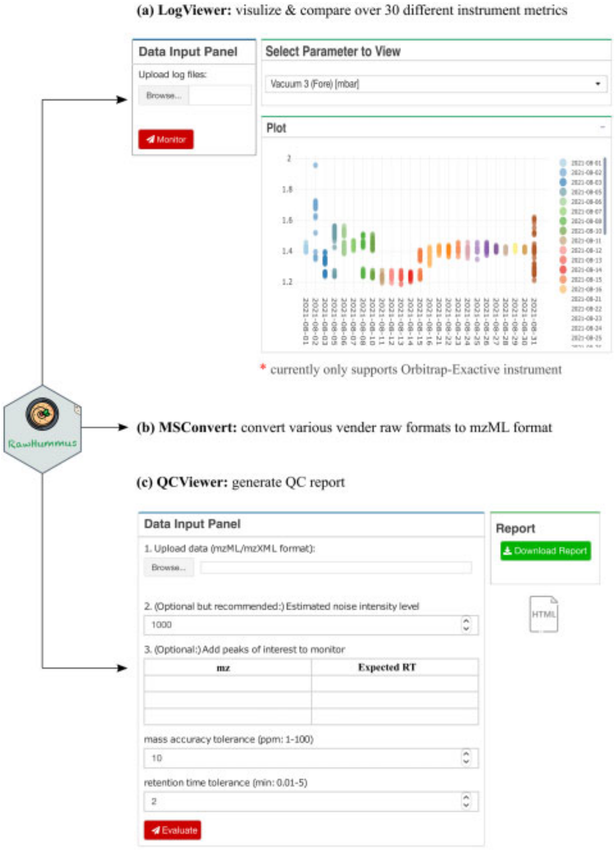
RawHummus accepts the generic mzXML or mzML format and is thus vendor independent. It creates a comprehensive HTML-based report containing interactive plots and tables, summary statistics and detailed explanations, which could assist metabolomics users to evaluate the quality of raw data.

2 Materials and methods

RawHummus is a Shiny App (Beeley and Sukhdeve, 2018) built in an R framework (R Core Team, 2020), which is released on both CRAN and GitHub. It includes three major functions, LogViewer, MSconvert and QCViewer (Fig. 1). To make RawHummus more user-friendly, detailed instructions are provided in each function tab in the web application.

Apart from technical and biological variations, metabolomics data quality is also vulnerable to instrumental fluctuation (e.g. temperature and humidity changes within the instrument). Such variation needs to be identified and minimized before sample injection. Thermo Q-Exactive series instrument, one of the most popular MS for metabolomics, can regularly monitor instrument status and save the information as daily log files. LogViewer was thus developed to interactively visualize the log files. Users can upload either a single log file to monitor instrument status of the day or multiple files to compare daily variations in instrumental parameters (Fig. 1a). LogViewer supports over 40 different instrument metrics, such as ambient temperature and ambient humidity. A representative log file with a complete metrics list is shown in Supplementary File S1.

The first step in many metabolomics data processing workflows is file conversion (Adusumilli *et al.*, 2017). Vendor-specific binary data need to be converted to open-format files (e.g. mzXML and mzML) for further manipulation by most software. Different tools are available for this purpose, among which MSConvert (part of ProteoWizard) is widely used (Chambers *et al.*, 2012). To facilitate file conversion, a command line version of MSConvert is incorporated in RawHummus. After installation of MSConvert, raw files can be easily converted to mzML format by RawHummus without any additional parameter settings (Fig. 1b). For further information regarding installation and usage of MSConvert, readers are kindly referred to elsewhere (Adusumilli *et al.*, 2017; Chambers *et al.*, 2012).



1. Upload data (mzML/mzXML format):

2. (Optional but recommended:) Estimated noise intensity level: 1000

3. (Optional:) Add peaks of interest to monitor

mz	Expected RT

mass accuracy tolerance (ppm: 1-100): 10

retention time tolerance (min: 0.01-5): 2

Fig. 1. Representative workflow and major functions of RawHummus

The converted files can be directly submitted to QCViewer for QC report generation (Fig. 1c). RawHummus adopts 12 quality metrics which are closely related to LC peak shape, RT, mass accuracy, detector sensitivity and fragmentation to evaluate chromatogram, MS1 and MS2 of the raw data. A description of each metric is given in Table 1. In addition to allowing users to define unlimited numbers of mass features for system evaluation, RawHummus also automatically selects six mass features evenly across the entire RT range for a more unbiased quality check. The report contains interactive plots and tables, summary statistics and detailed explanations, which can be used for data QC and instrumental error detection.

3 Results

Thirteen metabolomics/lipidomics datasets acquired from five different LCMS platforms were used to demonstrate the versatility and limitations of RawHummus (Supplementary File S2). Among them six were home datasets acquired from Thermo Q-Exactive Focus, nine were downloaded from Metabolights repository with kind permission from all the authors, including two datasets from Waters Xevo G2 QTOF (Dong *et al.*, 2020; Saw *et al.*, 2021), two from Bruker Maxis IITM QTOF (Dávila-Lara *et al.*, 2020), one from Agilent 6550 iFunnel QTOF (Meister *et al.*, 2021) and two from AB Sciex TripleTOF 6600 (Deng *et al.*, 2020). The total size of each dataset ranges from 0.34 to 12.15 GB and the corresponding data analysis time varies between 1.2 and 21 min on a PC with 32 GB memory and a 3.1 GHz Intel Core i7 processor (Supplementary File S2). The resulting QC reports are shown in Supplementary Files S3–S15. A list of demo metabolomics data files is deposited in GitHub (https://github.com/YonghuiDong/RawHummus_DemoData) for users to test the software.

Increasing the number of QC files and total file sizes lead to elevated data analysis time and memory usage. In this regard, RawHummus is more suitable for evaluating system performance within one project rather than long-term performance monitoring. RawHummus is deployed on ShinyApp.io (Rstudio) with a free plan and the bundle size is limited to 1 GB, therefore users are suggested to run RawHummus locally when large amounts of files need to be assessed. In addition, we have tentatively tested a proteomics dataset (Kazachkova *et al.*, 2021) and compared the reports with the ones obtained by RawBeans (Morgenstern *et al.*, 2021) (Supplementary Files S17 and S18). Although they share some similar metrics and show consistent results, due to the absence of some proteomics-specific metrics in RawHummus, proteomics users are suggested to use RawBeans or other proteomics dedicated tools for raw data QC. Likewise, test on the demo files shows that RawHummus is more suitable for metabolomics raw data QC compared with RawBeans (Supplementary Files S19 and S20).

4 Conclusion

RawHummus is a user-friendly tool providing a quick and effortless evaluation of an instrument performance and metabolomics data quality. The resulting QC report is presented in an HTML format which contains interactive plots and tables, summary statistics and detailed explanations. In addition, the QC report allows for early detection of instrumental errors and problematic samples prior to detailed data processing and statistical analysis, and it can be used as Supplementary Material in publications.

Table 1. Overview of quality metrics used in RawHummus report

Section	Metric	Explanation
Chromatogram	TIC plot	It is total ion current at each scan plotted as an intensity point for each raw file. Overlaid TIC plot is used for rapid inspection of RT and ion intensity fluctuations.
	Summed TIC bar plot	It is summed TIC of all scans in a raw file. It is used to check global ion intensity variations among raw files.
	TIC correlation analysis	Pairwise Pearson correlation analysis of raw files. It is used to evaluate chromatogram similarity, i.e. peak shape similarity and RT shift. Pearson correlation coefficient above 0.85 indicates that the two raw files are similar.
MS1	Max. mass difference (ppm)	It is maximum *m/z* variation among each selected mass feature across all the raw files. It is used to evaluate the mass accuracy. If the max. mass difference is over 5 ppm, this value will be highlighted in red.
	Max. RT difference (min)	It is maximum RT variation among each selected mass feature across all the raw files. It used to evaluate RT shifts. If the max. RT difference is over 1 min, this value will be highlighted in red.
	Max. intensity fold change	It is the maximum intensity fold change among each selected mass feature across all the raw files. It is used to evaluate ion intensity variation. If max. intensity fold change is over 1.5, this value will be highlighted in red.
	Intensity CV (%)	It is intensity coefficient of variance (or relative standard deviation, RSD) of each selected mass feature across all the raw files. It is used to evaluate intensity variation. If intensity CV is over 30%, this value will be highlighted in red.
MS2	No. of MS2 events	It is number of triggered MS/MS spectra per file. It is used to evaluate MS2 event.
	Precursor ion distribution across mass plot	It is density plot of the precursor ion across mass range based on the triggered MS/MS events.
	Precursor ion distribution across RT plot	It is density plot of the precursor ion across RT range based on the triggered MS/MS events.
	Cosine similarity of precursor ion distribution across mass	It measures the similarity of precursor ion distribution across mass. Cosine similarity score above 0.85 indicates that the precursor distributions across mass are similar between two files.
	Cosine similarity of precursor ion distribution across RT	It measures the similarity of precursor ion distribution across mass. Cosine similarity score above 0.85 indicates that the precursor distributions across RT are similar between two files.

Funding

This study was supported by the Metabolite Medicine Division, BLAVATNIK CENTER for Drug Discovery, Tel Aviv University, funded by Blavatnik Family Foundation.

Conflict of Interest: none declared.

References

Adusumilli,R. *et al.* (2017) Data conversion with ProteoWizard msConvert. In: Comai,L. *et al.*, eds. *Proteomics: Methods and Protocols, Methods in Molecular Biology*. Springer, New York, NY, pp. 339–368.

Beeley,C. and Sukhdeve,S.R. (2018) *Web Application Development with R Using Shiny: Build stunning graphics and interactive data visualizations to deliver cutting-edge analytics*. Packt Publishing Ltd.

Begou,O. *et al.* (2018) Quality control and validation issues in LC-MS metabolomics. In: Theodoridis,G.A. *et al.*, eds. *Metabolic Profiling: Methods and Protocols, Methods in Molecular Biology*. Springer, New York, NY, pp. 15–26.

Bielow,C. *et al.* (2016) Proteomics quality control: quality control software for MaxQuant results. *J. Proteome Res.*, **15**, 777–787.

Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Chiva,C. *et al.* (2018) QCloud: a cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS One*, **13**, e0189209.

Dávila-Lara,A. *et al.* (2020) Metabolomics analysis reveals tissue-specific metabolite compositions in leaf blade and traps of carnivorous nepenthes plants. *IJMS*, **21**, 4376.

Deng,W. *et al.* (2020) Metabolomics study of serum and urine samples reveals metabolic pathways and biomarkers associated with pelvic organ prolapse. *J. Chromatogr. B*, **1136**, 121882.

Dong,Y. *et al.* (2020) High mass resolution, spatial metabolite mapping enhances the current plant gene and pathway discovery toolbox. *N. Phytol.*, **228**, 1986–2002.

Kazachkova,Y. *et al.* (2021) The GORKY glycoalkaloid transporter is indispensable for preventing tomato bitterness. *Nat. Plants*, **7**, 468–480.

Kuhring,M. *et al.* (2020) Concepts and software package for efficient quality control in targeted metabolomics studies: MeTaQuaC. *Anal. Chem.*, **92**, 10241–10245.

Meister,I. *et al.* (2021) High-precision automated workflow for urinary untargeted metabolomic epidemiology. *Anal. Chem.*, **93**, 5248–5258.

Morgenstern,D. *et al.* (2021) RawBeans: a simple, vendor-independent, raw--data quality-control tool. *J. Proteome Res.*, **20**, 2098–2104.

R Core Team. (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saw,N.M.M.T. *et al.* (2021) Influence of extraction solvent on nontargeted metabolomics analysis of enrichment reactor cultures performing enhanced biological phosphorus removal (EBPR). *Metabolites*, **11**, 269.

Shaham-Niv,S. *et al.* (2021) Metabolite medicine offers a path beyond lists of metabolites. *Commun. Chem.*, **4**, 115.

Simader,A.M. *et al.* (2015) QCScreen: a software tool for data quality control in LC-HRMS based metabolomics. *BMC Bioinformatics*, **16**, 341.

Stanfill,B.A. *et al.*; TEDDY Study Group. (2018) Quality control analysis in real-time (QC-ART): a tool for real-time quality control assessment of mass spectrometry-based proteomics data. *Mol. Cell. Proteomics*, **17**, 1824–1836.

Yang,Q. *et al.* (2019) Metabolomics biotechnology, applications, and future trends: a systematic review. *RSC Adv.*, **9**, 37245–37257.

Zhou,B. *et al.* (2012) LC-MS-based metabolomics. *Mol. BioSyst.*, **8**, 470–481.