Received: 24 July 2023 | Revised: 22 August 2023 | Accepted: 24 August 2023

DOI: 10.1002/pmic.202300134

Proteomics
Proteomics and Systems Biology

TECHNICAL BRIEF

HowDirty: An R package to evaluate molecular contaminants in LC-MS experiments

David Gomez-Zepeda[1,2,3] | Thomas Michna[3] | Tanja Ziesmann[3] | Ute Distler[3,4] | Stefan Tenzer[1,2,3,4]

[1]Helmholtz-Institute for Translational Oncology Mainz (HI-TRON), Mainz, Rheinland-Pfalz, Germany

[2]German Cancer Research Center (DKFZ), Heidelberg, Germany

[3]Institute for Immunology, University Medical Center of the Johannes-Gutenberg University, Mainz, Rheinland-Pfalz, Germany

[4]Research Center for Immunotherapy (FZI), University Medical Center of the Johannes-Gutenberg University, Mainz, Rheinland-Pfalz, Germany

Correspondence
David Gomez-Zepeda, Helmholtz-Institute for Translational Oncology Mainz (HI-TRON), Mainz, Rheinland-Pfalz, Germany.
Email: david.gomez-zepeda@dkfz-heidelberg.de

Ute Distler, Institute for Immunology, University Medical Center of the Johannes-Gutenberg University, Mainz, Rheinland-Pfalz, Germany.
Email: ute.distler@uni-mainz.de

Stefan Tenzer, Research Center for Immunotherapy (FZI), University Medical Center of the Johannes-Gutenberg University, Mainz, Rheinland-Pfalz, Germany.
Email: tenzer@uni-mainz.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: TE599/9-1, DI 2471/1-1, SFB1292/2 TP-B11, SFB1292/2 TP-Q1; Bundesministerium für Bildung und Forschung, Grant/Award Numbers: 031L0217A/B, 16LW0241K

Abstract

Contaminants derived from consumables, reagents, and sample handling often negatively affect LC-MS data acquisition. In proteomics experiments, they can markedly reduce identification performance, reproducibility, and quantitative robustness. Here, we introduce a data analysis workflow combining MS1 feature extraction in Skyline with HowDirty, an R-markdown-based tool, that automatically generates an interactive report on the molecular contaminant level in LC-MS data sets. To facilitate the interpretation of the results, the HTML report is self-contained and self-explanatory, including plots that can be easily interpreted. The R package HowDirty is available from https://github.com/DavidGZ1/HowDirty. To demonstrate a showcase scenario for the application of HowDirty, we assessed the impact of ultrafiltration units from different providers on sample purity after filter-assisted sample preparation (FASP) digestion. This allowed us to select the filter units with the lowest contamination risk. Notably, the filter units with the lowest contaminant levels showed higher reproducibility regarding the number of peptides and proteins identified. Overall, HowDirty enables the efficient evaluation of sample quality covering a wide range of common contaminant groups that typically impair LC-MS analyses, facilitating corrective or preventive actions to minimize instrument downtime.

KEYWORDS
contamination, LC-MS, sample preparation, software

Abbreviations: DDA, data-dependent acquisition; DIA, data-independent acquisition; FASP, filter-assisted sample preparation; MWCO, molecular weight cut-off; PEG, polyethylene glycol; PPG, polypropylene glycol; SP3, single-pot solid-phase-enhanced sample preparation.

1 | INTRODUCTION

1.1 | Tools for rapidly assessing molecular contaminants are needed to ensure the quality of LC-MS results

MS laboratories and facilities usually aim to optimize their sample preparation procedures to provide optimal results and to minimize contaminants that can be detrimental to the LC-MS instruments. This requires thoroughly evaluating the consumables and reagents to prevent sample contamination. For instance, polymers and plasticizers can leak from plastic consumables, and hand-care products often contain substances interfering with LC-MS analyses [1]. These molecules often elute in reversed-phase LC within the retention time range of peptides and may induce ion suppression [2], harming the overall performance and reproducibility of MS results [3]. Thus, there is a need for tools to evaluate sample quality and levels of molecular contaminants efficiently. In addition, the output of such tools needs to be readily interpretable by the final users, which often include collaborators from different backgrounds. This will facilitate communication and thus assist in rapidly adapting protocols or laboratory practices to improve sample quality and ensure optimal results.

1.2 | Feature detection using Skyline enables identifying molecular contaminant features across multiple LC-MS platforms

Previously, Rardin [3] published a strategy to evaluate the presence of common molecular contaminants in LC-MS experiments using the software Skyline to detect contaminant-associated MS1 features. To perform the targeted data extraction, Rardin compiled an extensive molecular transition list including 64 parent molecules and 800 molecular species, hereby called contaminant groups and contaminants, respectively. One of the major advantages of this strategy is that Skyline can directly process raw data from most major MS vendors. Moreover, Skyline is widely used by the LC-MS community [4, 5]. The approach of Rardin allows MS experts to assess possible sample contamination. However, the output can be challenging to analyze and interpret by non-experts.

1.3 | HowDirty closes this gap by generating an interactive HTML report that users with different backgrounds can easily interpret

To facilitate the evaluation of small-molecule and polymer contaminants in proteomics and peptidomics samples, we developed a workflow using an R-markdown [6, 7], tidyverse-based [8] code to evaluate and plot the degree of contamination from the Skyline output [3] and compile the results in an interactive HTML document. The report is self-contained and can be archived as part of the analysis documentation or shared with collaborators.

2 | IMPLEMENTATION

After installing the HowDirty package in R, the contaminant evaluation workflow can be completed following the steps summarized in Figure 1 and detailed in the tutorial published on Github. This approach is compatible with data-dependent acquisition (DDA) and data-independent acquisition (DIA). After configuring Skyline using the molecular contaminant template compiled by Rardin [3], raw files are loaded, and the MS1 features are extracted. At this stage, it is recommended to evaluate the feature identification since it is solely based on the precursor m/z (see section Limitations). Then, the user exports the results to a CSV file. This file is loaded into the HowDirty template with a sample annotation file containing final file names and groups (i.e., conditions), which will be used for statistics and plots. In addition, it is possible to load threshold files representing the normal status of an instrument (see Section 4, Contamination Thresholds). Finally, when the HowDirty template is compiled (i.e., "knitted" in R markdown), it generates a self-contained HTML with interactive plots and tables as well as a result summary compiled in an Excel file. In addition, since the package is open-source, the template and other functions provided in HowDirty can be incorporated in other data analysis pipelines.

3 | ABUNDANCE CALCULATION

The list of common contaminants comprises a large variety of substances, such as polymers (polyethylene glycol, PEG; polypropylene glycol, PPG; etc.), polysiloxanes, etc. These molecules have different ionization efficiencies, which may be further affected by ionization competition. Therefore, their peak areas cannot be directly compared within a sample or between samples of different origins and batches. To consider the proportion of a contaminant within a sample, we calculated a normalized abundance (NormAbundance) as denoted in Equation (1).

$$\text{Abundance} = \frac{\text{Area}_i}{\text{Area}_{TIC}} \quad (1)$$

Area_i = peak area of analyte i; Area_{TIC} = total ion count area

To summarize the abundance of contaminants across contaminant groups and samples, HowDirty reports quantiles at 25%, 50% (median), 75%, and 90%, and the summed abundance. This simple strategy avoids biases introduced by low or missing values. To consider the possible heterogeneity of the samples, the median is used for plots and comparisons at the contaminant group level. To summarize the potential contamination by diverse types of molecules, the summed contaminant

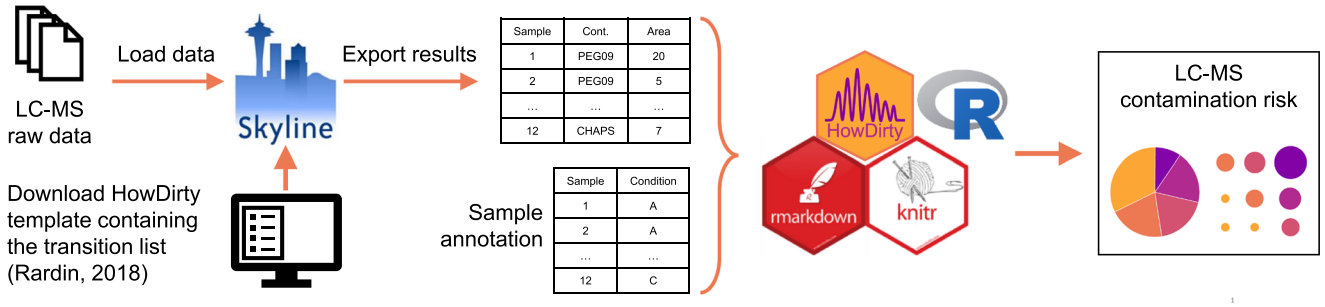

FIGURE 1 Workflow for the evaluation of LC-MS sample contamination using Skyline [4, 5], the molecular contaminant transition list [3], and the R-based package HowDirty.

TABLE 1 Quantile segments used by HowDirty to assign the contamination risk level based on the reference dataset.

From (≥)	To (<)	Contamination risk level
0%	25%	1) Very low
25%	50%	2) Low
50%	75%	3) Medium
75%	90%	4) High
90%		5) Very high
ND		6) No threshold in reference

Abbreviation: ND, not detected in the reference dataset.

abundance is used at the sample level. Using such simple metrics facilitates the interpretation by both MS experts and non-experts.

4 | CONTAMINATION THRESHOLDS

HowDirty can be used to generate lab- and instrument-specific contamination thresholds. To establish thresholds that represent the normal status of sample chemical background on a LC-MS instrument platform, it is recommended to generate a reference dataset. This can be done by processing files acquired over an extended time period (e.g., spanning multiple months), including samples of diverse origins. The result Excel file can be used as reference input in future analyses. Most samples in such a long period should be within acceptable contamination levels. Thus, HowDirty uses the reference results to calculate thresholds based on the Abundance quantiles for each contaminant (Table 1), the contaminant group (median abundance), and the total per sample (sum of all contaminants). For instance, if a sample in a new dataset has a PEG19 abundance above 90% as compared to the samples in the reference dataset, it will be tagged as "5) Very High" to indicate the high degree of contamination and the risk of further contamination if these or some similar samples are injected in the LC-MS platform.

To illustrate the reference dataset generation, we processed 2 months of analyses performed in an Ultimate 3000 – Exploris 480 LC-MS (Thermo Fisher Scientific). The samples originated from multiple laboratories and included, among others, cell lysates and mouse tissue extracts. The protein digest was performed in our laboratory following our standard operation procedures using either filter-assisted sample preparation (FASP) or single-pot solid-phase-enhanced sample preparation (SP3) digestion protocols [9–11] (details in Supporting Information S1). We included the resulting reports in Supporting Information (S2, HTML report; S3, Excel report).

5 | HOWDIRTY ENABLES THE EFFICIENT EVALUATION OF SAMPLE PREPARATION PROCEDURES AND CONSUMABLES

To demonstrate the application of HowDirty in evaluating sample preparation consumables, we tested the effect of using molecular weight cut-off (MWCO) ultrafiltration units from three different providers for the FASP digestion (n = 5 each) of HeLa lysate aliquots (n = 5 each). The filter devices tested were: Vivacon (Sartorius), Microcon (Merck), and Nanosep (Cytiva). We analyzed the resulting peptides by LC-MS (n = 3 each) and processed the data using the HowDirty workflow (Figure 2A). The HowDirty report can be downloaded from the Supporting Information (S4, HTML report; S5, Excel report). This paragraph will reference the plots by their subsection within the HTML file (S3). The summary pie-chart immediately shows some contaminated samples (Figures 2B and S4.4.1). The condition-grouped plots indicated that using filters from Nanosep resulted in a higher degree of sample contamination (Figures 2C and S4.4.2). The contaminant group plots showed that the correspondent samples contained high levels of PEG (Figure S4.5.1). The expected signal intensity patterns were observed in the contaminant-specific plot, showing a bell-shaped pattern with an apex at PEG09_NH4 (Figures 2D and S4.6.1.1). The pseudo-chromatograms allowed us to confirm that the PEG molecules eluted across the retention time from the lowest to the highest degree of polymerization, providing additional confirmation of the identity of this contaminant (Figures 2E and S4.6.2.1).

To evaluate the impact of the degree of contamination on peptide and protein identification, we processed the DDA files in MSFragger [12] (Figure 2F). There was no significant difference in the number of peptides identified across the three types of filter units. However, the Nanosep filters resulted in significantly fewer protein groups

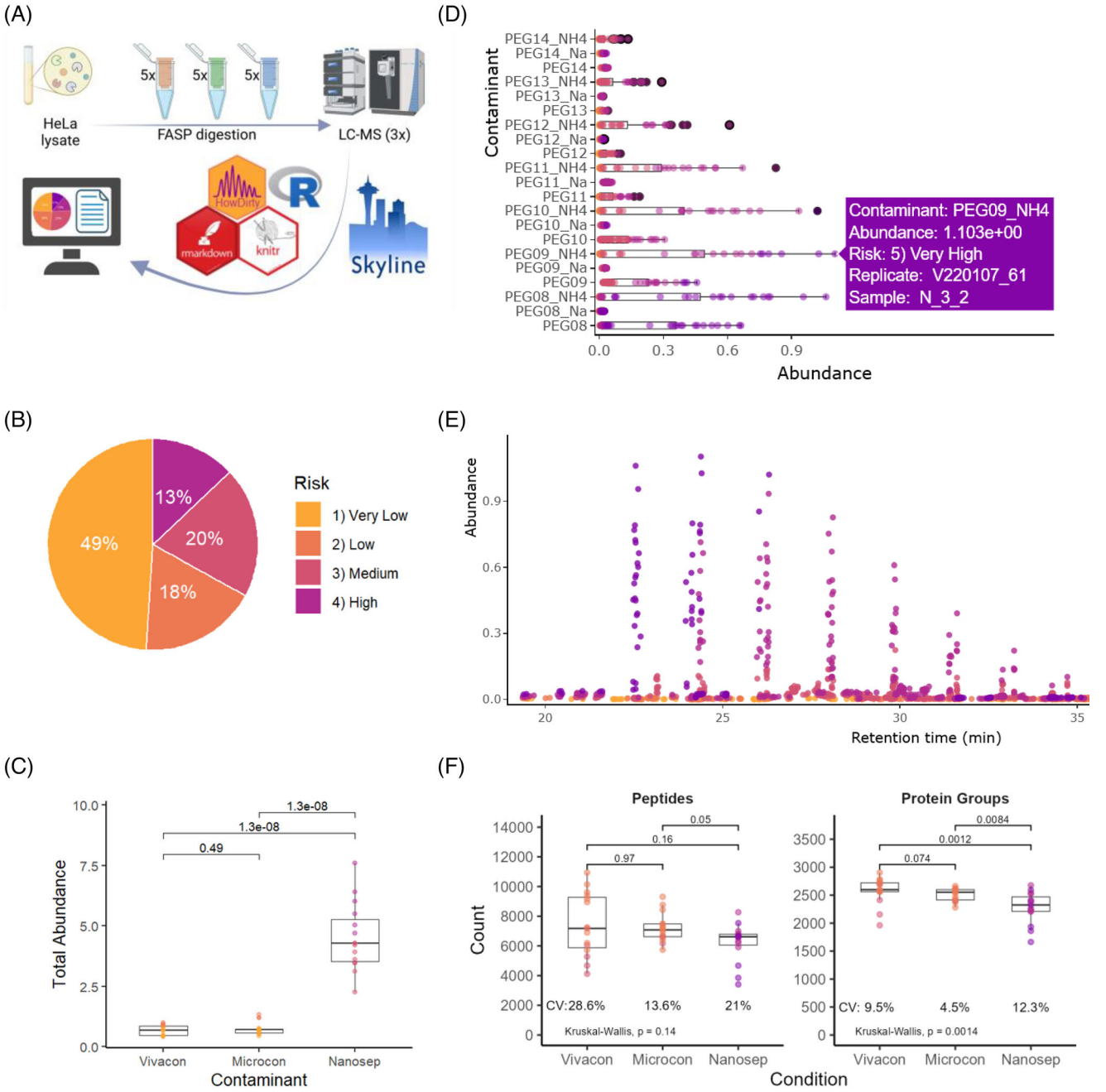

FIGURE 2 Evaluation of ultrafiltration units from different providers using HowDirty. (A) Experimental design (Created with BioRender.com). (B) Global summary of the contamination risk evaluation and color legend for all the plots. (C) Condition boxplot with Wilcoxon signed-rank test. (D) PEG contaminant-specific abundance plot; cropped at PEG 14 for visualization. (E) Pseudochromatogram plot showing the abundance of possible PEG molecules across the retention time. (F) Count of unique Peptides and Protein Groups by DDA; differences were assessed by a Kruskal–Wallis test, followed by a Wilcoxon signed-rank test.

(median = 2325) compared Microcon (median = 2553, $p = 0.0084$), and to Vivacon (2599, $p = 0.0012$). Notably, the variability in identified protein groups was the highest for the Nanosep filters, compared to Vivacon and Microcon (CV = 12.3%, 9.5%, and 4.5%, respectively). Altogether, the Microcon filters provided the lowest degree of contamination and also the lowest peptide and protein group variability. Thus, we decided to use Microcon filters for our standard operation procedures. In summary, this example shows that the HowDirty workflow is an efficient tool for evaluating sample preparation procedures and the consumables implicated.

6 | LIMITATIONS

The workflow requires using two specialized programs, Skyline and R. However, Skyline is already widely used by the LC-MS community.

In addition, users that are not familiar with Skyline can access the multiple tutorials provided by the developers (https://skyline.ms). Similarly, to minimize the requirement of R programming knowledge to use HowDirty we provide a ready-to-use template and a step-by-step tutorial. A future iteration of the workflow could be incorporated as a Skyline plugin to facilitate its implementation.

The contaminant feature identification in Skyline is based only on the MS1 m/z, which can lead to some false identifications. Thus, we recommend to further evaluate the data if contaminations are reported by HowDirty. For instance, by verifying the expected chromatographic pattern from lower to higher degree of polymerization in the case of PEG and PPG (Figure 2E), and CHAPS usually elutes at late stages in reversed phase (RP)-LC but it may be miss assigned to earlier eluting peptide peaks.

7 | CONCLUSIONS

We introduced a workflow to evaluate the degree of contamination with diverse substances in LC-MS samples. This approach builds on the use of Skyline [4, 5] to extract contaminant MS1 features, previously presented by Rardin [3], and our R package HowDirty to generate a self-contained interactive HTML report. The summary statistics and plots provided by HowDirty will enable the LC-MS community to efficiently assess sample quality and rapidly take corrective or preventive action to minimize instrument downtime. Although using Skyline requires a moderate expertise, the HowDirty report can be easily interpreted by users without MS background with minimal clarification, facilitating inter-lab communication and potential troubleshooting of upstream sample handling.

8 | METHODS

Methods are further described in Supporting Information S1.

8.1 | Materials and substances

All reagents used were analytical or LC-MS grade, and LoBind tubes (Eppendorf) were employed to minimize sample loss. Ultrafiltration units (MWCO 30 kDa) from three different providers were evaluated for FASP digestion in the example experiment: Vivacon 500, 30,000 MWCO Hydrosart (Sartorius, ref. VN01H23); Microcon 30 kDa Centrifugal Filter Unit with Ultracel-30 membrane (Merck, ref. MRCF0R030); Nanosep Omega-membrane centrifugal filter 30 kDa (Cytiva, formerly Pall Lab, ref. OD030C35).

8.2 | Proteomics sample preparation

Whole-cell lysates were obtained from different origins, including HeLa, *E. coli*, Saccharomyces bayanus, other human cell lines, or mouse tissue. Samples were digested with trypsin using modified versions of FASP [10] or SP3 [11] protocols as described by [9].

8.3 | LC-MS analyses

The digests were injected in an Ultimate 3000 - Exploris 480 LC-MS (Thermo Fisher Scientific), and the peptides were resolved on a reversed-phase C18 column (HSS-T3, 100Å, 1.8 μm, 75 μm × 250 mm; Waters Corporation) at 55°C in a 44 min gradient from 2 to 35% at a flow rate of 300 nL/min. Eluted molecules were ionized in positive mode. For the reference dataset, MS/MS data were acquired in DDA or DIA mode. The example dataset evaluating ultrafiltration units was acquired in DDA mode.

8.4 | LC-MS data analysis

Contaminant MS1 features were extracted using Skyline (v21.2.0.568) [4, 5], and the exported results were processed using the R package HowDirty described in this manuscript. Peptide and protein identification was performed using MSFragger (v3.2) [12].

AUTHOR CONTRIBUTIONS

Following CREdiT classification. Conceptualization: D.G.Z. and U.D. Methodology: D.G.Z., U.D., T.M., and S.T. Software: D.G.Z. and T.Z. Validation: D.G.Z., U.D., T.M., and T.Z. Formal analysis: D.G.Z. and T.M. Investigation: D.G.Z., U.D., and T.M. Resources: S.T. Data curation: D.G.Z., U.D., T.M., and T.Z. Writing—original draft: D.G.Z. Writing—review & editing: D.G.Z., U.D., T.M., T.Z., and S.T. Visualization: D.G.Z. and T.M. Supervision: D.G.Z., U.D., and S.T. Project administration: U.D. and S.T. Funding acquisition: S.T. and U.D.

ACKNOWLEDGMENTS

The authors acknowledge Christina Jung, Claudia Darmstadt, and Lucas Kleinort for their technical assistance on sample and instrument preparation for LC-MS analyses; and Malte Sielaff for testing the workflow. D.G.Z. and S.T. acknowledge funding from Bundesministerium für Bildung und Forschung, (BMBF) as part of the National Research Node "Mass spectrometry in Systems Medicine" (MSCoreSys) [031L0217A/B, 16LW0241K]. S.T., T.Z. were supported by DFG priority program SPP 2225 (Grant No TE599/9-1 to S.T.). The work was further supported by the German Research Foundation (DFG; Project Number 318346496, SFB1292/2 TP-Q1 to S.T., TP-B11 to U.D. and DI 2471/1-1 to U.D).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The HowDirty R package can be downloaded and installed from https://github.com/DavidGZ1/HowDirty. The Skyline contamination template file compiled by Rardin [3] can be downloaded from the Panorama

Public data repository: https://panoramaweb.org/labkey/contaminants.url. The raw and result files of the experiment evaluating ultrafiltration units for FASP digestion have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) [13] via the jPOSTrepo partner repository [14] with the dataset identifiers PXD044032 and JPST002256, respectively.

ORCID

Stefan Tenzer https://orcid.org/0000-0003-3034-0017

REFERENCES

1. Keller, B. O., Sui, J., Young, A. B., & Whittal, R. M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta*, *627*, 71–81. https://doi.org/10.1016/j.aca.2008.04.043
2. Annesley, T. M. (2003). Ion suppression in mass spectrometry. *Clinical Chemistry*, *49*, 1041–1044. https://doi.org/10.1373/49.7.1041
3. Rardin, M. J. (2018). Rapid assessment of contaminants and interferences in mass spectrometry data using Skyline. *Journal of the American Society for Mass Spectrometry*, *29*, 1327–1330. https://doi.org/10.1007/s13361-018-1940-z
4. Pino, L. K., Searle, B. C., Bollinger, J. G., Nunn, B., Maclean, B., & Maccoss, M. J. (2020). The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrometry Reviews*, *39*, 229–244. https://doi.org/10.1002/mas.21540
5. Maclean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., & Maccoss, M. J. (2010). Skyline: Aan open source document editor for creating and analyzing targeted proteomics experiments, *Bioinformatics*, *26*, 966–968. https://doi.org/10.1093/bioinformatics/btq054
6. Xie, Y., Allaire, J. J., Grolemund, G., & Markdown, R. (2018). *The definitive guide*. Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown
7. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/
8. Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*, 1686. https://doi.org/10.21105/joss.01686
9. Sielaff, M., Kuharev, J., Bohn, T., Hahlbrock, J., Bopp, T., Tenzer, S., & Distler, U. (2017). Evaluation of FASP, SP3, and iST protocols for proteomic sample preparation in the low microgram range. *Journal of Proteome Research*, *16*, 4060–4072. https://doi.org/10.1021/acs.jproteome.7b00433
10. Wisniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nature Methods*, *6*, 359–362. https://doi.org/10.1038/nmeth.1322
11. Hughes, C. S., Foehr, S., Garfield, D. A., Furlong, E. E., Steinmetz, L. M., & Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular Systems Biology*, *10*, 757. https://doi.org/10.15252/msb.20145625
12. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics, *Nature Methods*, *14*, 513–520. https://doi.org/10.1038/nmeth.4256
13. Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., & Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology*, *32*, 223–226. https://doi.org/10.1038/nbt.2839
14. Okuda, S., Watanabe, Y., Moriya, Y., Kawano, S., Yamamoto, T., Matsumoto, M., Takami, T., Kobayashi, D., Araki, N., Yoshizawa, A. C., Tabata, T., Sugiyama, N., Goto, S., & Ishihama, Y. (2017). jPOSTrepo: An international standard data repository for proteomes. *Nucleic Acids Research*, *45*, D1107–D1111. https://doi.org/10.1093/nar/gkw1080

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202300134 in the Supporting Information section at the end of the article.

How to cite this article: Gomez-Zepeda, D., Michna, T., Ziesmann, T., Distler, U., & Tenzer, S. (2023). HowDirty: An R package to evaluate molecular contaminants in LC-MS experiments. *Proteomics*, e2300134. https://doi.org/10.1002/pmic.202300134