analytical chemistry

Technical Note

Cite This: Anal. Chem. 2018, 90, 685−689

pubs.acs.org/ac

MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go

Thilo Muth,*[,†] Fabian Kohrs,[‡] Robert Heyer,[‡] Dirk Benndorf,[‡,§] Erdmann Rapp,[§] Udo Reichl,[‡,§] Lennart Martens,[∥,⊥] and Bernhard Y. Renard[†]

[†]Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany

[‡]Bioprocess Engineering, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

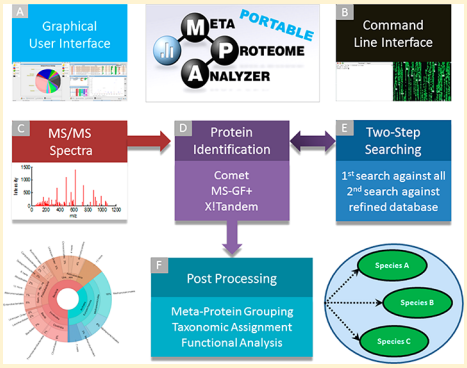
[§]Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, 39106 Magdeburg, Germany

[∥]Department of Biochemistry, Ghent University, 9000 Ghent, Belgium

[⊥]VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

Supporting Information

ABSTRACT: Metaproteomics, the mass spectrometry-based analysis of proteins from multispecies samples faces severe challenges concerning data analysis and results interpretation. To overcome these shortcomings, we here introduce the MetaProteomeAnalyzer (MPA) Portable software. In contrast to the original server-based MPA application, this newly developed tool no longer requires computational expertise for installation and is now independent of any relational database system. In addition, MPA Portable now supports state-of-the-art database search engines and a convenient command line interface for high-performance data processing tasks. While search engine results can easily be combined to increase the protein identification yield, an additional two-step workflow is implemented to provide sufficient analysis resolution for further postprocessing steps, such as protein grouping as well as taxonomic and functional annotation. Our new application has been developed with a focus on intuitive usability, adherence to data standards, and adaptation to Web-based workflow platforms. The open source software package can be found at https://github.com/compomics/meta-proteome-analyzer.



The key role of microbial consortia has recently gained increased attention due to promising findings on their functional repertoire in the human intestinal tract. Complex microbial communities fulfill essential host-related functions regarding nutrient uptake, digestion, and immune response.[1] Importantly, the human gut microbiome has also been correlated with pathological states such as type-2 diabetes,[2] cardiovascular disease,[3] Crohn's disease,[4] inflammatory bowel disease,[5] and obesity.[6,7] In more general terms, the importance of microbial communities is related to the well-known fact that microbes are critical to the niche system (e.g., human host) in which they reside. One of the most common approaches for studying microbial communities presents genome analysis, using either 16S rRNA gene sequencing or shotgun whole metagenome sequencing.[8] While these techniques are highly useful tools for gaining insights into the composition and functional potential of a microbial community, these do lack the ability to capture the actual functional profile of such a community at a given time point and under specific conditions. However, such profiling is essential to demonstrate that predicted biological processes are actually present and active in a given sample and can be only gained from the functionally active snapshot of microbial communities.[9] Metaproteomics, the mass spectrometry-based analysis of multispecies proteins from microbial samples, aims to elucidate the functional expression and taxonomic origin of such microbial consortia.[10−12] This proteomic technique is also employed for rapidly detecting pathogens and studying their host-adaptation mechanisms.[13] The application of metaproteomics has led to promising findings in recent studies for which disease-associated protein markers could be identified, e.g., when analyzing samples from bovine blood serum[14] or human oral saliva.[15] While throughput and resolution of instrumentation have evolved dramatically within the past decade, the analysis and interpretation of the upcoming data still remains a challenge. This can mainly be attributed to the complexity and heterogeneity of microbiome samples, which can contain proteins from hundreds or thousands of different species.[16] Despite the increase in popularity of metaproteomics, existing proteome bioinformatics methods have not yet been sufficiently

Received: August 30, 2017
Accepted: December 7, 2017
Published: December 7, 2017

DOI: 10.1021/acs.analchem.7b03544
Anal. Chem. 2018, 90, 685−689

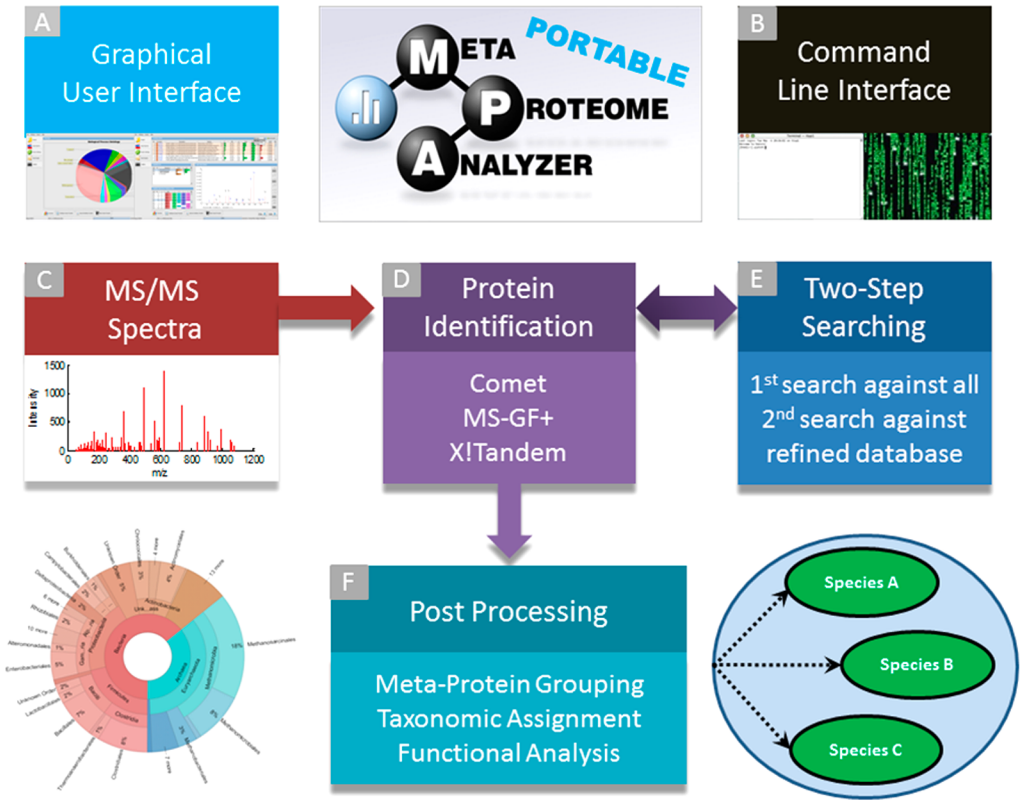


Figure 1. Overview on the MPA Portable workflow. The software can be accessed using either the graphical user interface (A) or the command line interface (B). User-provided MS/MS spectra (C) are processed within the application for matching against a FASTA database by up to three different search algorithms (X!Tandem, MS-GF+, and Comet) (D). As an alternative to conventional searching, a two-step search (E) can be applied to iteratively reduce the search space. Further postprocessing steps (F) include grouping of homologous proteins to meta-proteins and the fully automated assignment of peptides and protein to taxonomic levels and functional annotations, as described for the original MPA software package.[21]

adapted to adequately address these challenges,[17] and tailored solutions for metaproteomics remain rare.[18−21]

In this article, we present the MetaProteomeAnalyzer (MPA) Portable software and demonstrate all novel features and improvements which have been developed since the original MPA publication.[21] MPA Portable is a lightweight and freely available application which serves as a one-stop solution for processing and analyzing metaproteomics data. In contrast to the original server-based MPA software,[21] the MPA Portable tool requires no further installation steps and is independent of any relational database system. In addition to the graphical user interface (GUI), which can be used for in-depth data exploration, a command line interface (CLI) has also been added to MPA Portable. This allows the program to be executed as part of a larger, scripted workflow, for instance, on a high-performance cluster environment. While a standalone version (including a guided tutorial) is available for download on the GitHub Web site (https://github.com/compomics/meta-proteome-analyzer), the whole MPA workflow has also been included within the community-accepted multiomics informatics platform of Galaxy-P.[22] In addition to the previously supported database search engine X!Tandem,[23] the newly developed software now also integrates the SEQUEST-derivative Comet[24] and MS-GF+[25] as search algorithms. Similar to the original development, MPA Portable also allows the results of multiple search engines to be combined to increase the overall peptide and protein identification yield, but it adds the ability to perform an optional two-step search workflow.[26] In the two-step searching approach, the spectra are first matched against a wide search space (e.g., the whole UniProt database) without applying any FDR filtering. On the basis of the results of this search, the proteins with at least one PSM are retained by which a new sequence database is created. In a second round, a typical search against the reduced search space is applied with stringent FDR filtering. The objective of such an iterative search procedure is to increase the number of highly confident peptide spectrum matches and, consequently, to improve overall protein identification yield. This strategy is particularly useful for metaproteomics data analysis which usually suffers from a decreased identification rate when searching against large protein sequence databases, which is in turn caused by the higher chance of retrieving high-scoring false positive identifications.[27] Moreover, to improve compatibility with existing proteomic software tools, the import of files stored in the mzIdentML standard data format (version 1.2)[28] has been also implemented. This latter feature is particularly useful for the reprocessing of identification results that have been generated using external tools or elsewhere, as is for instance the case for data obtained from the public domain PRIDE database.[29] Figure 1 provides an overview of an MPA Portable workflow, comprising all typical steps of data processing, ranging from the input of MS/MS spectra, over protein identification, to the MPA-specific postprocessing features such as protein grouping and automated sequence annotation at the taxonomic and functional level.[21]

We tested our proposed software workflow on two experimental data sets from samples with known composition. The first benchmarking data set was established by mixing the

DOI: 10.1021/acs.analchem.7b03544
Anal. Chem. 2018, 90, 685−689

bacterial strains (5BCT) *Bacillus subtilis*, *Escherichia coli*, *Pseudomonas fluorescens*, *Micrococcus luteus*, and *Desulfovibrio vulgaris* with a protein ratio of 1:1:1:1:1. The corresponding sample was prepared in-house and sample specifications can be found in the Supporting Information. The mass spectrometry data have been deposited to the ProteomeXchange Consortium via PRIDE,[29] with data set identifier PXD007681. In addition, the second data set which was used for evaluation derived from a lab-assembled mixture of nine microbial organisms (9MM) published by Tanca et al.[30]

The first evaluation concerned the performance of the newly integrated search algorithms in MPA Portable against UniProtKB/Swiss-Prot (2016/12/13) with regard to the accuracy when assigning peptides to given reference taxa at the species level. In this analysis, the assignments were classified at the peptide level as follows (i) "correct unique" when a peptide was matched unambiguously to a protein from the correct taxon, (ii) "correct" when a peptide was matched to the correct taxon but was also shared with proteins from incorrect taxa, and (iii) "incorrect" when a peptide was assigned to species which were not contained in the original sample. In addition, we also applied a taxonomy filter after combining the search results within the MPA application (as previously described by Tanca et al.[30]) corresponding to 5% of the total number of taxon-specific assignments. Thus, only taxa with a higher number of peptide assignments than the specified filter threshold were taken into consideration in that case.

The results show that the combination of hits from multiple search engines within the MPA Portable workflow significantly increases the number of correct unique and correct taxon-specific peptides (Figure 2). In addition, the results show that the number of incorrect assignments could be decreased by filtering for the most abundant organisms using a relatively high threshold of 5%. However, for more complex samples, it is not recommended to use such a stringent taxonomic filtering, as it may considerably reduce the proportion of sparsely identified but essential organisms.

The second evaluation of our software concerned the effect of the newly implemented two-step search strategy. In general, the composition and size of the protein sequence database has a strong impact on the results in any proteomics data analysis workflow, and it has been recommended to focus on relevant sequences for better identification yields.[31] In metaproteomics, however, the actual microbial composition in the sample is commonly unknown and the identification of relevant sequences is therefore highly problematic. Indeed, it would be particularly damaging to introduce selection bias by the mistaken removal of relevant taxa and their reference proteomes. To automate the process of building a sample-optimized search database with sufficient taxonomic coverage and depth, we therefore implemented the previously described two-step search approach.[26] To evaluate the performance of two-step in comparison to conventional searching, we first searched the 9MM data set of known species composition against a tailored database which only contained the protein sequence sets from the nine expected organisms. Next, we matched the same data set against UniProtKB/Swiss-Prot using (i) conventional and (ii) two-step searching. The two-step method was used by searching against the whole database without applying any FDR threshold in the first round. The protein identifications obtained from this first round then serve as the reduced database in a second round search on the same data. In this second round, a stringent FDR threshold of 1% is

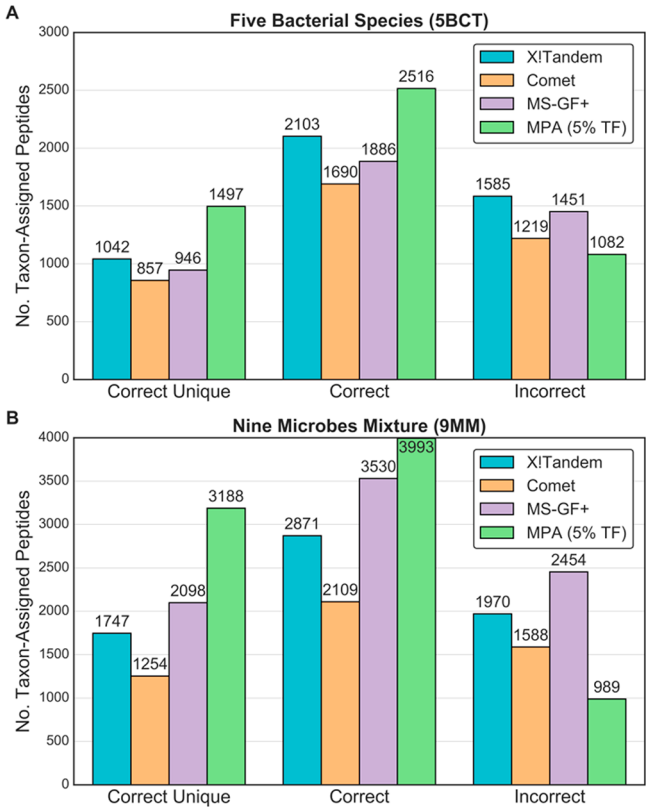


Figure 2. Taxon-specific peptide assignment performance for 5BCT and 9MM reference data. The numbers of correct unique, correct (i.e., unique and shared) and incorrect taxon-specific peptide identifications are shown as bar charts for data sets 5BCT (A) and 9MM (B) when using X!Tandem (blue), Comet (orange), MS-GF+ (violet), and MPA Portable with an applied taxon filter (TF) of 5% (green). For the latter, the results of all three database search algorithms were combined by taking the union of all hits. The data sets were searched against UniProtKB/Swiss-Prot and filtered by an FDR threshold of 1%.

applied to reduce the number of false positive hits. For each search setting, the peptide identifications were classified into correct and incorrect taxon assignments as described above. We also applied a taxonomy filter with a threshold ranging from 0% to 10% to test the influence of this parameter. In this analysis, the database search results from all three search algorithms X! Tandem, Comet, and MS-GF+ were combined by taking the union of the respective hits.

The results of the search method evaluation can be found in Figure 3. As expected, the most correct taxon assignments at the peptide level could be obtained when searching the data against the 9MM reference database (Figure 3A). When searching the same data against the whole UniProtKB/Swiss-Prot database, on average, only around 69% of the original hits could be correctly assigned. This can be explained by the increased search space and the peptide sequence ambiguities among homologous species. When applied to the UniProtKB/Swiss-Prot search, the two-step searching approach recovered around 80% of the peptides originally identified against only the 9MM reference database, thus showing better performance than the standard search. However, the proportion of incorrect species assignments is also higher for two-step approach when compared to the standard search (Figure 3B). Fortunately, this effect can be minimized when increasing the taxon filter

DOI: 10.1021/acs.analchem.7b03544
Anal. Chem. 2018, 90, 685−689

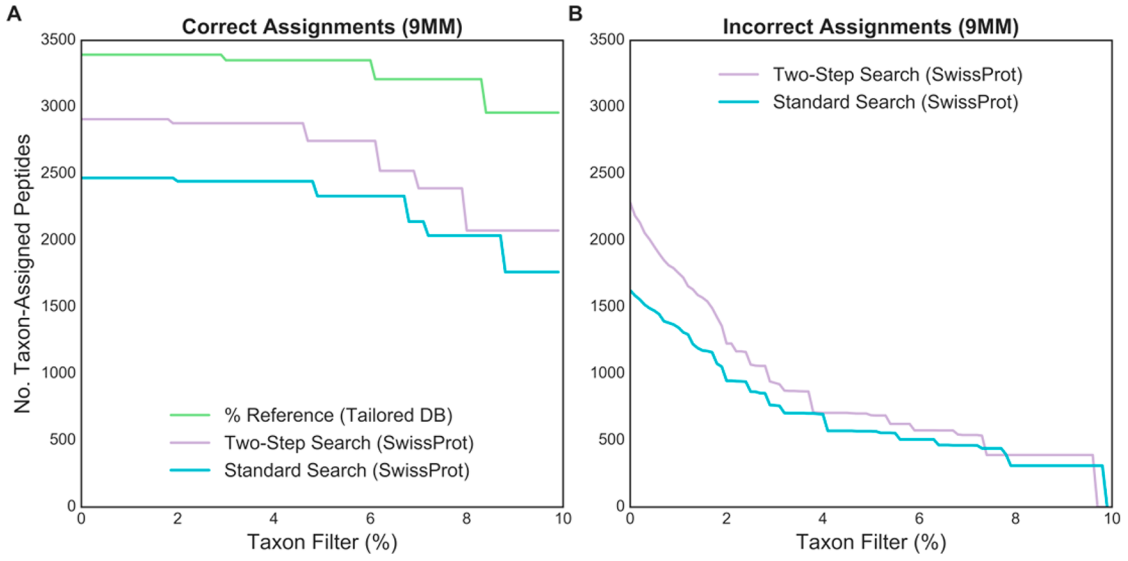


Figure 3. Performance evaluation of standard *versus* two-step searching for 9MM reference data. The numbers of correct (A) and incorrect (B) taxon-specific peptides at 1% FDR are shown as line charts for the 9MM data set, which was searched against a tailored reference database (green) and against UniProtKB/Swiss-Prot using standard search (blue) and two-step search approach (violet). The taxonomy filter shown on the *x*-axis was applied for values between 0% and 10%. Note that incorrect taxon-specific peptides were not found for the tailored reference database, since only known microbial species were included in the search database.

threshold, demonstrating that the two-step approach is optimally beneficial when used in combination with taxonomic filtering. It should be noted, however, that the results from the described two-step approach should be treated carefully as the actual FDR of the overall process is likely to be higher than the 1% FDR set for the second round.[27]

The availability of the MPA Portable software marks another step toward fulfilling the needs of the metaproteomics community, which requires reliable and easily accessible solutions for analyzing its valuable high-throughput data. Moreover, the addition of a command-line interface to MPA Portable enables analyses using high-performance cluster hardware.[32] This is particularly important in the context of metaproteomics, because searches are often performed against very large protein sequence databases to cover a wide taxonomic range (e.g., UniProtKB/TrEMBL[33] with over 88 million entries as of July 2017). In our evaluation on the performance of the taxonomic assignment, a rather limited number of fewer than 10 species from microbial mixture samples was used. However, data from more complex samples are required to improve the performance of our pipeline. Overall, the research community of metaproteomics would strongly benefit from such well-defined reference data for benchmarking and optimization of analytical workflows and software tools. Moreover, the thorough assessment of protein FDR estimation represents an important next step in the field which would highlight demand that tools still need to become more reliable at the statistical side for making results better reproducible. It should be also noted that the output of the command-line execution of MPA Portable can be fully imported into the GUI, thus allowing large data sets (e.g., hundreds of thousands to millions of MS/MS spectra) to be analyzed on a cluster environment, while being visualized in the application on a local desktop computer. Consequently, MPA Portable now offers users the combination of usability with computational power in a single package. Finally, to make the developed software even more hardware-independent and sustainable, it has also been integrated into the Galaxy-based workflow for metaproteomics analysis.[34] Similarly, we plan to distribute it within the system-agnostic BioContainers framework which allows software to be installed and executed under an isolated and controllable environment.[35]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b03544.

Details on the preparation and measurement of the microbial sample (5BCT) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mutht@rki.de. Phone: +49 30 18754 2990. Fax: +49 30 18754 2328.

ORCID

Thilo Muth: 0000-0001-8304-2684

Lennart Martens: 0000-0003-4277-658X

Notes

The authors declare no competing financial interest.

The MS proteomics data in this paper have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository:[29] dataset identifier PXD007681.

■ ACKNOWLEDGMENTS

B.Y.R. acknowledges financial support by Deutsche Forschungsgemeinschaft (DFG), Grant Number RE3474/2-1. L.M. acknowledges support from NCI-ITCR Grant 1U24CA199347.

■ REFERENCES

(1) Round, J. L.; Mazmanian, S. K. *Nat. Rev. Immunol.* **2009**, *9*, 313–323.

DOI: 10.1021/acs.analchem.7b03544
Anal. Chem. 2018, 90, 685–689

(2) Larsen, N.; Vogensen, F. K.; van den Berg, F. W.; Nielsen, D. S.; Andreasen, A. S.; Pedersen, B. K.; Al-Soud, W. A.; Sorensen, S. J.; Hansen, L. H.; Jakobsen, M. *PLoS One* **2010**, *5*, e9085.
(3) Howitt, M. R.; Garrett, W. S. *Nat. Med.* **2012**, *18*, 1188−1189.
(4) Haberman, Y.; Tickle, T. L.; Dexheimer, P. J.; Kim, M. O.; Tang, D.; Karns, R.; Baldassano, R. N.; Noe, J. D.; Rosh, J.; Markowitz, J.; Heyman, M. B.; Griffiths, A. M.; Crandall, W. V.; Mack, D. R.; Baker, S. S.; Huttenhower, C.; Keljo, D. J.; Hyams, J. S.; Kugathasan, S.; Walters, T. D.; et al. *J. Clin. Invest.* **2014**, *124*, 3617−3633.
(5) Morgan, X. C.; Tickle, T. L.; Sokol, H.; Gevers, D.; Devaney, K. L.; Ward, D. V.; Reyes, J. A.; Shah, S. A.; LeLeiko, N.; Snapper, S. B.; Bousvaros, A.; Korzenik, J.; Sands, B. E.; Xavier, R. J.; Huttenhower, C. *Genome Biol.* **2012**, *13*, R79.
(6) Ley, R. E.; Turnbaugh, P. J.; Klein, S.; Gordon, J. I. *Nature* **2006**, *444*, 1022−1023.
(7) Kolmeder, C. A.; Ritari, J.; Verdam, F. J.; Muth, T.; Keskitalo, S.; Varjosalo, M.; Fuentes, S.; Greve, J. W.; Buurman, W. A.; Reichl, U.; Rapp, E.; Martens, L.; Palva, A.; Salonen, A.; Rensen, S. S.; de Vos, W. M. *Proteomics* **2015**, *15*, 3544−3552.
(8) Jovel, J.; Patterson, J.; Wang, W.; Hotte, N.; O'Keefe, S.; Mitchel, T.; Perry, T.; Kao, D.; Mason, A. L.; Madsen, K. L.; Wong, G. K. S. *Front. Microbiol.* **2016**, *7*, 459.
(9) Tanca, A.; Abbondio, M.; Palomba, A.; Fraumene, C.; Manghina, V.; Cucca, F.; Fiorillo, E.; Uzzau, S. *Microbiome* **2017**, *5*, 79.
(10) Hettich, R. L.; Pan, C.; Chourey, K.; Giannone, R. J. *Anal. Chem.* **2013**, *85*, 4203−4214.
(11) Wilmes, P.; Bond, P. L. *Trends Microbiol.* **2006**, *14*, 92−97.
(12) VerBerkmoes, N. C.; Denef, V. J.; Hettich, R. L.; Banfield, J. F. *Nat. Rev. Microbiol.* **2009**, *7*, 196−205.
(13) Welker, M.; Moore, E. R. *Syst. Appl. Microbiol.* **2011**, *34*, 2−11.
(14) Lamont, E. A.; Janagama, H. K.; Ribeiro-Lima, J.; Vulchanova, L.; Seth, M.; Yang, M.; Kurmi, K.; Waters, W. R.; Thacker, T.; Sreevatsan, S. *J. Clin Microbiol* **2014**, *52*, 536−543.
(15) Belstrom, D.; Jersie-Christensen, R. R.; Lyon, D.; Damgaard, C.; Jensen, L. J.; Holmstrup, P.; Olsen, J. V. *PeerJ* **2016**, *4*, e2433.
(16) Zhang, X.; Chen, W.; Ning, Z.; Mayne, J.; Mack, D.; Stintzi, A.; Tian, R.; Figeys, D. *Anal. Chem.* **2017**, *89*, 9407−9415.
(17) Muth, T.; Renard, B. Y.; Martens, L. *Expert Rev. Proteomics* **2016**, *13*, 757−769.
(18) Mesuere, B.; Willems, T.; Van der Jeugt, F.; Devreese, B.; Vandamme, P.; Dawyndt, P. *Bioinformatics* **2016**, *32*, 1746−1748.
(19) Penzlin, A.; Lindner, M. S.; Doellinger, J.; Dabrowski, P. W.; Nitsche, A.; Renard, B. Y. *Bioinformatics* **2014**, *30*, i149−156.
(20) Zhang, X.; Ning, Z.; Mayne, J.; Moore, J. I.; Li, J.; Butcher, J.; Deeke, S. A.; Chen, R.; Chiang, C. K.; Wen, M.; Mack, D.; Stintzi, A.; Figeys, D. *Microbiome* **2016**, *4*, 31.
(21) Muth, T.; Behne, A.; Heyer, R.; Kohrs, F.; Benndorf, D.; Hoffmann, M.; Lehteva, M.; Reichl, U.; Martens, L.; Rapp, E. *J. Proteome Res.* **2015**, *14*, 1557−1565.
(22) Boekel, J.; Chilton, J. M.; Cooke, I. R.; Horvatovich, P. L.; Jagtap, P. D.; Kall, L.; Lehtio, J.; Lukasse, P.; Moerland, P. D.; Griffin, T. J. *Nat. Biotechnol.* **2015**, *33*, 137−139.
(23) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.
(24) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. *Proteomics* **2013**, *13*, 22−24.
(25) Kim, S.; Pevzner, P. A. *Nat. Commun.* **2014**, *5*, 5277.
(26) Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J. *Proteomics* **2013**, *13*, 1352−1357.
(27) Muth, T.; Kolmeder, C. A.; Salojarvi, J.; Keskitalo, S.; Varjosalo, M.; Verdam, F. J.; Rensen, S. S.; Reichl, U.; de Vos, W. M.; Rapp, E.; Martens, L. *Proteomics* **2015**, *15*, 3439−3453.
(28) Vizcaino, J. A.; Mayer, G.; Perkins, S.; Barsnes, H.; Vaudel, M.; Perez-Riverol, Y.; Ternent, T.; Uszkoreit, J.; Eisenacher, M.; Fischer, L.; Rappsilber, J.; Netz, E.; Walzer, M.; Kohlbacher, O.; Leitner, A.; Chalkley, R. J.; Ghali, F.; Martinez-Bartolome, S.; Deutsch, E. W.; Jones, A. R. *Mol. Cell. Proteomics* **2017**, *16*, 1275−1285.
(29) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. *Proteomics* **2005**, *5*, 3537−3545.
(30) Tanca, A.; Palomba, A.; Deligios, M.; Cubeddu, T.; Fraumene, C.; Biosa, G.; Pagnozzi, D.; Addis, M. F.; Uzzau, S. *PLoS One* **2013**, *8*, e82981.
(31) Sticker, A.; Martens, L.; Clement, L. *Nat. Methods* **2017**, *14*, 643−644.
(32) Verheggen, K.; Barsnes, H.; Martens, L. *Proteomics* **2014**, *14*, 367−377.
(33) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. *Nucleic Acids Res.* **2004**, *32*, D115−D119.
(34) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. *Proteomics* **2015**, *15*, 3553−3565.
(35) da Veiga Leprevost, F.; Gruning, B. A.; Alves Aflitos, S.; Rost, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; Bai, M.; Jimenez, R. C.; Sachsenberg, T.; Pfeuffer, J.; Vera Alvarez, R.; Griss, J.; Nesvizhskii, A. I.; Perez-Riverol, Y. *Bioinformatics* **2017**, *33*, 2580−2582.

DOI: 10.1021/acs.analchem.7b03544
Anal. Chem. 2018, 90, 685−689