Journal of proteome research

pubs.acs.org/jpr

Article

MetaProD: A Highly-Configurable Mass Spectrometry Analyzer for Multiplexed Proteomic and Metaproteomic Data

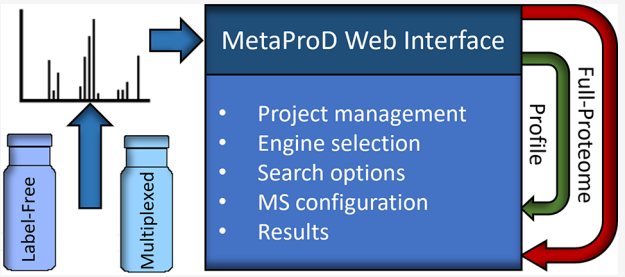
Jamie Canderan, Moses Stamboulian, and Yuzhen Ye*

Cite This: *J. Proteome Res.* 2023, 22, 442−453

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The microbiome has been shown to be important for human health because of its influence on disease and the immune response. Mass spectrometry is an important tool for evaluating protein expression and species composition in the microbiome but is technically challenging and time-consuming. Multiplexing has emerged as a way to make spectrometry workflows faster while improving results. Here, we present MetaProD (MetaProteomics in Django) as a highly configurable metaproteomic data analysis pipeline supporting label-free and multiplexed mass spectrometry. The pipeline is open-source, uses fully open-source tools, and is integrated with Django to offer a web-based interface for configuration and data access. Benchmarking of MetaProD using multiple metaproteomics data sets showed that MetaProD achieved fast and efficient identification of peptides and proteins. Application of MetaProD to a multiplexed cancer data set resulted in identification of more differentially expressed human proteins in cancer tissues versus healthy tissues as compared to previous studies; in addition, MetaProD identified bacterial proteins in those samples, some of which are differentially abundant.



KEYWORDS: *mass spectrometry, metaproteomics, proteomics, multiplexing, isobaric labeling, differentially expressed proteins, bacterial proteins*

INTRODUCTION

The human microbiome has been shown to be influential in human health and diseases, such as type 2 diabetes[1] and colorectal cancer.[2] Microbiota can be involved in important metabolic pathways in the host related to nutrition[3] and can be involved in the development of immunity and protection.[4,5] Microbiota can also act as an important biomarker for identifying a disease.[6] Metaprotemics has emerged as a field focused on the identification of quantification of bacterial proteins to allow for a determination of protein function,[7] microbiome species composition,[8] locality,[9] changes in bacterial gene expression in response to disease,[10] and information on future treatments that may be focused on bacterial proteins.[11]

Shotgun proteomics using mass spectrometry (MS) is a widely used technique in metaproteomics because of the ability to analyze complex samples containing thousands of proteins[12] that may be common in microbiome-based samples. These studies have been utilized in such cases as to provide data for carcinogenesis models[13] and identify target pathways for therapeutic treatments.[14]

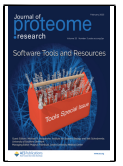
Many readily available pipelines exist utilizing mass spectrometry for protein identification and quantification. There are two general categories of the approaches: label-free proteomics and multiplex labeling.[15]

Label-free approaches have the advantages of generally requiring less sample and preparation complexity[15] but may have problems with high protein coefficients of variance (CVs) between different replicates[16] and missing peptides or proteins in a replicate that makes statistical analysis more challenging.[17] Another consideration is the length of time MS experiments may take. Label-free approaches generally run one sample on the instrument at a time, and this can cause experiments with dozens or hundreds of samples to be time-consuming, particularly as it has been shown that longer gradients on the liquid chromatography instrument have been shown to reduce CV values between replicates.[18]

Multiplexing approaches offer the advantage of allowing one to run multiple samples at the same time on the instrument, which has been shown to reduce CV values in replicates because the same peptide from different samples elutes together,[19] improve the ability for relative and absolute quantification,[20] allow for a higher number of peptide identifications, and ultimately allow for more overall

Special Issue: Software Tools and Resources 2023

Received: September 30, 2022
Published: January 23, 2023



https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

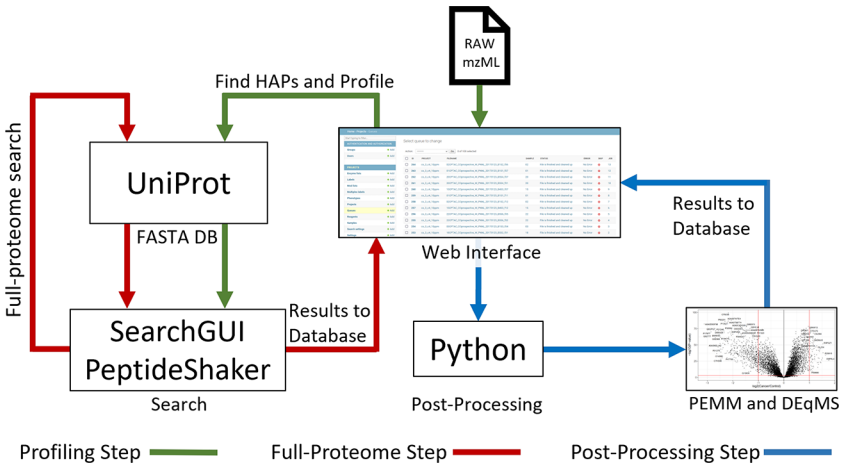


Figure 1. Overview of the MetaProD workflow showing the profiling step (green), full-proteome step (red), and postprocessing (blue).

quantification per instrument.[19,20] Utilizing labeling approaches, such as TMT or iTRAQ, therefore may be critical for large-scale studies focused on identifying changes in bacterial composition and protein expression based on the disease state of the individual.

Many metaproteomics pipelines exist, such as Metapro-IQ,[21] MetaLab,[22] and HAPiID[23] for label-free approaches and IsoProt[24] for labeled data, but these pipelines tend to be tied to a single-search algorithm when multiple algorithms together have shown improved results[25] or rely on Docker containers, which may not be suitable for high-performance computing (HPC) applications.

MetaProD is a highly configurable metaproteomics pipeline supporting both label-free and labeled experiments starting from raw mass spectrometry data all the way through generating tables of results in a Django-based web interface. MetaProD allows users to configure all projects and settings in the web-based interface (such as a choice of multiple search algorithms) and other typical options that may be changed during mass spectrometry experiments (such as enzyme specificity and missed cleavages). The web interface also allows users to view results, including protein expression in normalized spectral abundance factor[26] (NSAF) or number of peptide spectrum matches (PSM) and species expression, in a web browser and export desired results to downloadable files for further analysis elsewhere. Users can have the option to configure this interface to suit their own analysis needs using Django, but many of the typical use-case scenarios have been provided for. Tests of MetaProD using metaproteomic data of the known microbial community and other real metaproteomics data sets showed that MetaProD achieved fast and efficient identification of peptides and proteins from metaproteomic data of a mixture of species associated with different environments/hosts.

EXPERIMENTAL PROCEDURES

Pipeline Implementation

General Information. MetaProD is designed to work with label-free or multiplexed metaproteomic data but can support general proteomics data. For metaproteomic data analysis, MetaProD uses a two-step strategy including a profiling step and a full-proteome search for identification of peptides from metaproteomic data as with HAPiID.[23] We showed previously[23] that a two-step approach is important for efficient and fast peptide identification from metaproteomic data with unknown species composition. Figure 1 shows a general overflow of the MetaProD workflow.

The pipeline is developed using Python scripts to interface with the different proteomics software and generate results and is fully integrated with the Django web framework to allow for a web-based graphical user interface (GUI) for configuration and viewing of results. Users of the pipeline have the option to expand upon the web interface as desired using the extensive documentation provided by Django and examples provided in the code. The proteomics pipeline itself is run using simple command-line arguments on a Linux-based system but should be portable to any system where Django and Java are available. The system is designed to allow for implementation on HPC systems by dividing projects into jobs, which then can be fed into HPC workload managers, such as Slurm.

Generation of the Protein Database. A FASTA database containing protein sequences is generated by downloading the list of bacterial pan-proteomes[27] from Uniprot[28] (release 2021_03 was used for testing, but the software can download a newer version). The list of bacterial reference proteomes is also downloaded from Uniprot (release 2021_03), and the two lists are combined to a final FASTA file containing only proteins in nonduplicated proteomes by filtering any duplicate proteomes that appear in both lists. Supporting Information File 1 contains the specific list of Uniprot proteome IDs used for this paper. The list of proteins is filtered down to highly abundant proteins (HAPs) for the initial profiling step based on methods discussed by HAPiID[23] to limit the size of the initial FASTA database required. This is done by including any protein containing the term "ribosomal" in the protein name and reduces the initial FASTA file used from 42 125 535 sequences to 510 154. The full Uniprot human reference proteome (Uniprot ID UP000005640) can be added along with a list of common contaminants obtained from the Common Repository of Adventitious Proteins (CRAP)[29] depending on the configuration in the web interface to help reduce false-positives or otherwise include those proteins in the search results. The FASTA file is then processed using SearchGUI[30] (version 4.1.16) to append decoy sequences to be used for false-discovery rate (FDR) calculations.[31]

Proteomics Processing. MetaProD uses a two-step strategy including a profiling step and full-proteome search

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

for identification of peptides from metaproteomic data. Unlike HAPiID[23] and many other pipelines that can use only one search engine peptide identification, MetaProD can utilize multiple search engines. RAW files generated from the mass spectrometer are converted by ThermoRawFileParser[32] to convert the file from the RAW to mzML, as needed. SearchGUI is then used as an interface for the six mass spectrometry search engines with Comet,[33] MetaMorpheus,[9] MSGF+,[34] Myrimatch,[35] OMSSA,[36] and X!Tandem[37] being available as the search engines, which can be selected individually or together for both the profiling and full-proteome steps, and the previously mentioned FASTA file is used as a search database. The web interface allows the user to configure many common mass spectrometry settings, such as precursor and fragment tolerances in either parts per million (ppm) or daltons (Da), precursor charge ranges, peptide length filtering, proteases, post-translation modifications (PTMs) (with the option of adding new ones, as needed), and PSM, peptide, and protein-level FDR rates. PeptideShaker[38] (version 2.2.9) is used to combine the results from multiple search engines using the same settings as SearchGUI. For multiplexed data, Reporter[39] (version 0.9.8) is used to calculate the expression ratios from the multiplex labels for the PSMs. Custom Python scripts are implemented for protein inference to generate the minimum list of proteins to explain the peptides using a greedy approach and to filter the results to the configured FDR for spectra, peptides, and proteins.

The information from the initial profile step is used to generate a list of species covering a configurable amount of total protein NSAF (90% by default). This species list is used to regenerate a FASTA database containing all proteins from the proteomes for these species as described by HAPiID.[23] The SearchGUI, PeptideShaker, and Reporter steps are then repeated using the new FASTA database for this full-proteome step.

Additional Python scripts are used for multiplexed data to map the labels back to specific patients and phenotypes, to calculate the normalized protein expression ratios from the PSM ratios generated by Reporter, and to interface with the R packages PEMM[40] for imputation of missing protein values and DEqMS[41] for generation of differentially expressed proteins.

The final results are stored in a SQL database and then accessible to the Django web interface or user-generated SQL queries.

Web Interface. The Django-based web interface includes a configurable administration panel where projects can be created, the file queue can be managed, and specific search settings can be set for each project. Examples of the configuration interface are shown in Supporting Information Figures S1a and S1b. The interface includes example label-free and TMT-10 projects along with a predefined list of modifications and enzymes to use, but these can be expanded by the user in the future as long as they are supported by SearchGUI and PeptideShaker. The web interface also allows one to configure detailed sample information for multiplexed samples, such as phenotype, label, and patient identifier.

The web interface also includes a detailed results section for each project. This includes summary information for a project, which includes the number of peptides, PSM, and protein per file or project, NSAF information, and the top 10 species by both NSAF and PSM for the profile and proteome steps. Other information available per project includes the full file lists, full species lists, full protein list, full peptide list, and full PSM list. Each page also allows one to click for more details. The protein page, for example, allows one to view the accession number, the Uniprot proteome, the organism, and the protein description and to click a link to view all peptides and PSMs associated with a specific protein as shown in Supporting Information Figure S2. Much of this information can also be downloaded into tab-separated value (TSV) files for use elsewhere.

A full help page is available to explain the use of the Web site, and all aspects of the web interface can also be configured by the user by following the Django documentation if they wish to expand upon it or change aspects of the layout. The user can also change the stylesheet to quickly customize the appearance of the Web site.

Pipeline Tests

Settings Used Throughout. MetaProD was set to use the following settings unless otherwise described: HCD fragmentation, Q-Exactive as the instrument, 8−30 peptide length, 2−4 precursor charge range, 1% PSM/peptide/protein FDR, and 0−1 precursor isotope range.

MetaProD was set to use an enzyme for digestion, and trypsin was used as the protease in all cases with a maximum of two missed cleavages along with specific digestion except where described. Carbamidomethylation of C was set as a fixed modification and oxidation of M was set as a variable modification for all data sets.

Parts per million (PPM) error rates were set to either 5, 10, or 20 ppm for both precursor and fragment as described for individual data sets.

Finally, we tested multiple search engine configurations as described for each data set. The search engines tested include Comet, MetaMorpheus, MSGF+, MyriMatch, OMSSA, and X! Tandem individually for both the profile and full-proteome steps. We also used different configurations for the profile and full-proteome steps as summarized in Table 1.

Table 1. Summary of Search Engine Configurations Used Throughout This Paper[a]

identifier	search engines
C,3	Comet (Profile); Comet, MetaMorpheus, MSGF+ (Full)
C,4	Comet (Profile); Comet, MetaMorpheus, MSGF+, X!Tandem (Full)
3	Comet, MetaMorpheus, MSGF+ (Profile, Full)
4	Comet, MetaMorpheus, MSGF+, X!Tandem (Profile, Full)
6	Comet, MetaMorpheus, MSGF+, X!Tandem, MyriMatch, OMSSA (Profile, Full)

[a]Configurations that use the same single-search engine for profiling and full-proteome search are not shown in this table. C,3 and C,4 use different combinations of search engines in the two steps, whereas configurations 3, 4, and 6 use the same combination of search engines in both steps.

Label-Free CAMPI Data Sets. Raw mass spectrometry data was generated using a defined mixed culture from lab-scale bioreactors containing eight sequenced microbes: *Anaerostipes caccae* (DSMZ 14662), *Bacteroides thetaiotaomicron* (DSMZ 2079), *Bifidobacterium longum* (NCC 2705), *Blautia producta* (DSMZ 2950), *Clostridium butyricum* (DSMZ 10702), *Clostridium ramosum* (DSMZ 1402), *Escherichia coli* K-12 (MG1655), and *Lactobacillus plantarum* (DSMZ 20174) as part of the Critical Assessment of MetaProteome

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

Investigation (CAMPI).[42] The list of species and corresponding Uniprot proteome are reflected in Supporting Information Table S1, and full details of the sample preparation and processing are available in the original study.[42] The RAWs file used in this study (samples S07 and S11) were downloaded from ProteomeXchange with the data set identifier PXD023217.

The known species and strain composition of the sample allow one to test the pipeline performance of a multispecies FASTA database and determine potential configuration options for future studies. The single RAW file was run through the pipeline using individual and multiple search engines and multiple ppm-error settings, and the number of proteomes selected during profiling along with the PSM, peptide, and protein identifications during the full-proteome step were recorded along with the runtime of each case to benchmark the performance of the different search algorithms and the FASTA generation method.

We tested each search engine individually in both profile and full-proteome steps. We also tested using different combinations of search engines as shown in Table 1. Each combination was tested at 5, 10, and 20 ppm error rate for both precursor and fragment to benchmark error settings and the corresponding runtime and identifications.

We additionally investigated using 90% and 80% of NSAF for the profiling step to probe the number of proteomes selected during profiling and the resulting accuracy during the full-proteome step.

The CAMPI study also analyzed a fecal sample, which we included in our study. Raw mass spectrometry data was generated from a fecal sample obtained from a 33 year old omnivorous nonsmoking woman. The RAW files used in this study (sample F07) were downloaded from ProteomeXchange with the data set identifier PXD023217.

We tested the fecal data with the same parameters as the S07 and S11 samples previously mentioned.

Label-Free Human Mucosal−Luminal Interface Data Set. Human musosal−luminal interface (HMI) samples were collected during endoscopy from the ascending colon of eight children.[21] The bacteria in the samples were isolated, processed, and digested with trypsin and run with a 4 h gradient on a Q-Exactive mass spectrometer to produce a total of eight RAW files. Full details of the sample preparation and processing are available in the original study.[21] These eight files were downloaded from ProteomeXchange with the data set identifier PXD003528.

The eight samples were run through the pipeline using the six search engines as the only engine for both profiling and the full-proteome steps at 5 ppm error and the C,3 and C,4 combinations from Table 1 at both 5 and 10 ppm. The pipeline was set to select the top 90% of NSAF for species selection for the full-proteome step in all cases.

We additionally ran the C,3 and C,4 combinations at 5 and 10 ppm error using semispecific protease cleavage to benchmark the effect of using a semispecific setting versus fully specific.

Label-Free Wastewater SD6 Data Set. Microbial communities were sampled from the surface of the anoxic treatment phase at a biological wastewater treatment plant in Luxembourg.[43] The SD6 sample used in this paper was collected on October 12, 2011. The SD6 sample was prepared for mass spectrometry, split into six fractions, and run on a Q-Exactive mass spectrometer set to higher-energy collision dissociation to produce a total of six RAW files. Full details of the sample preparation are available in the original study.[43] The six RAW files were downloaded from PeptideAtlas with the data set identifier PASS00577.

The six fractions were run through the pipeline using the same search engine combinations as described for the HMI data set and grouped together as a single sample for data analysis.

Multiplexed Colon Cancer (CO) Data. Raw mass spectrometry data was generated by The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) (study ID PDC000117).[44] Tissue samples from 100 patients were collected from both a tumerous and a nontumerous site on the colon of each patient. The samples were prepared for mass spectrometry by using trypsin as a digestion reagent and TMT-10 as an isobaric labeling reagent. The 22 TMT-10 samples were fractionated using a reverse-phased HPLC with a C18 column to produce 12 fractions for each sample and run on a Thermo Scientific Q-Exactive Plus mass spectrometer set to high-energy collision dissociation (HCD) and the data-dependent acquisition (DDA) set to the top 12 spectra to generate a total of 264 RAW files that were used in the data analysis pipeline. Full details on the sample preparation are available in the original study,[14] and all data are available on the CPTAC Web site[44] with the data set identifier PDC000116.

The C,4 at 10 ppm search engine configuration (specific, 90% of NSAF) was used for this data set based on the results from the other data sets along with the settings used for all data sets described previously, but TMT 10-Plex of K and TMT 10-plex of the N-term were added as fixed modifications.

The PSM expression ratios obtained after the search were normalized relative to the reference channel(s). PSMs were grouped by sample, modified sequence, and protein and their median ratio was taken to be the expression ratio of the corresponding peptide. The peptides were then mapped to individuals and phenotypes. The peptides were then normalized by dividing each by the median for that channel. These peptides were then filtered to include only those that had values in at least 50% of the channels with a channel being a ratio for an individual for a given phenotype (cancer or normal).

Peptide ratios were $\log_2$ transformed. Missing values were imputed using PEMM[40] with a phi value of 0; PEMM uses a penalized EM algorithm to incorporate a missing data mechanism for imputation of missing values.

The median of the $\log_2$ peptide expressionratios for a protein was taken to be the protein expression ratio. These protein expression ratios were normalized to have equal medians (0) per phenotype of an individual and run through DEqMS[41] to identify differentially expressed proteins. The DEqMS results were filtered to only include proteins with a Benjamini and Hochberg (BH) adjusted p-value of less than 0.05.

Pipeline Availability

MetaProD, all source code, software, requirements, installation instructions, and full documentation are available on GitHub at https://github.com/mgtools/MetaProD. Information about a publicly available Amazon EC2 image with MetaProD, all required software, and example projects are available in the GitHub documentation.

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

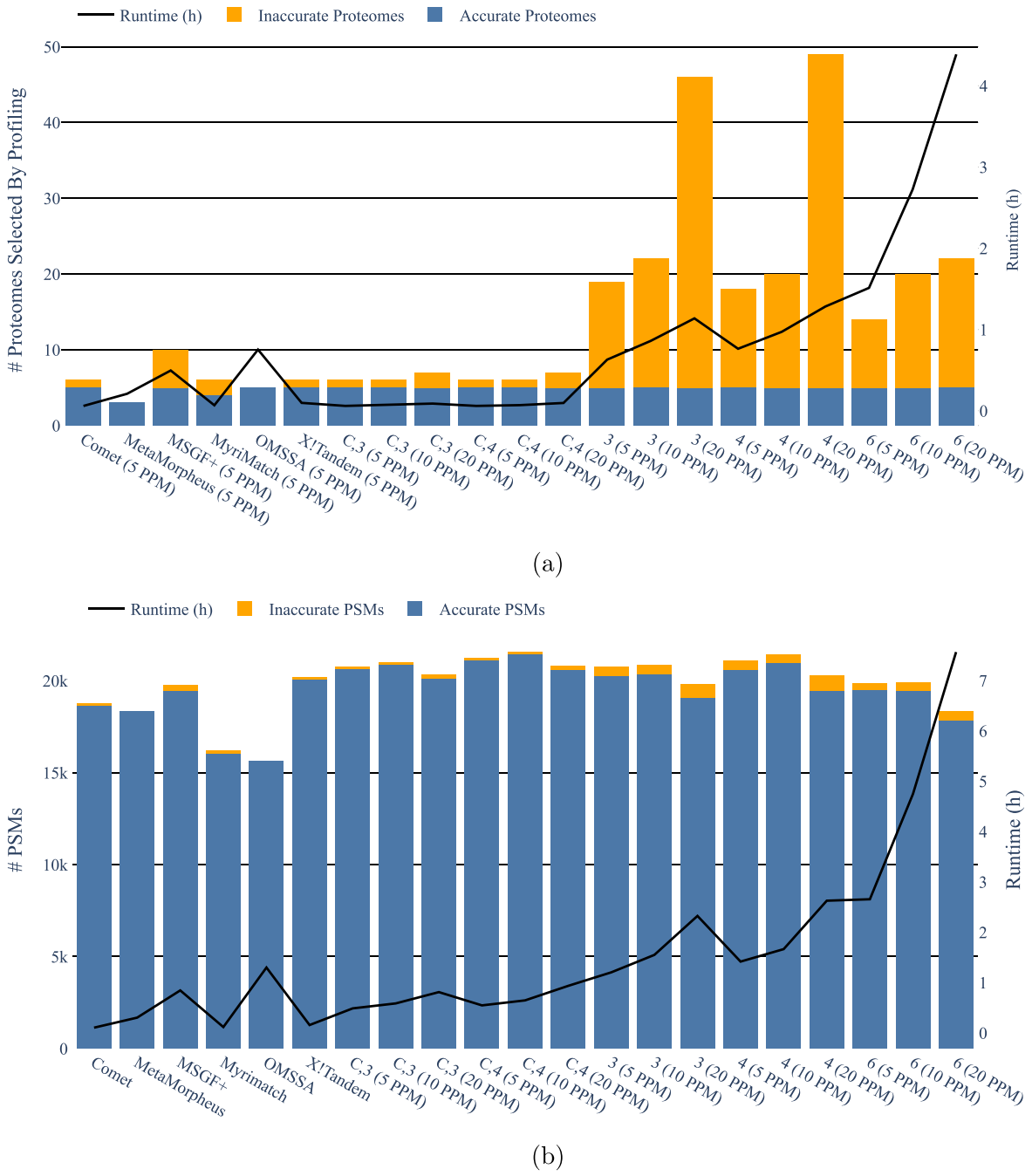


Figure 2. Profile and full-proteome performance of the different engine combinations for the SIHUMIx S07 data set at 90% of NSAF. (a) Profile performance of the search engine combinations showing the number of correct and inaccurate proteomes selected during the profile step along with the runtime of the SearchGUI step for the combination. (b) Full-proteome performance of the search engine combinations showing the total number of PSMs identified along with the number of PSMs believed to come from species not present in the sample. The runtime for both SearchGUI steps (profile + proteome) for each combination is also shown.

RESULTS AND DISCUSSION

Evaluation of MetaProD Using the CAMPI Data Sets

Because the SIHUMIx data sets (S07 and S11) have known bacterial composition, we can use them to evaluate the impacts of different parameters on the performance of the two steps of MetaProD (profiling and full-proteome search). The profiling performance of the various search engines and ppm settings for the S07 sample at 90% of NSAF are reflected in Figure 2a and both the S07 and S11 samples for both 80% and 90% of NSAF in Supporting Information Tables S2, S3, S5, and S6. The SIHUMIx data sets have a potential of eight proteomes corresponding to the eight species listed in Supporting Information Table S1. Supporting Information Table S2 indicates that most combinations identified three accurate and zero inaccurate proteomes at 80% NSAF for the S07 data set with a few of the combinations identifying inaccurate proteomes, particularly at 20 ppm error. Supporting Information Table S3 indicates that at 90% of NSAF for the S07 data set, four of the search engines (Comet, MSGF+, OMSSA, and X!Tandem) selected five accurate proteomes

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

during the profile step while MyriMatch selected four and MetaMorpheus only selected three with MSGF+ showing the highest number of incorrect proteomes (five) among the individual search engines with the rest showing only one or two incorrect. Comet (213 s), MyriMatch (243 s), and X! Tandem (349 s) showed the quickest profile runtime with MSGF+ (1784 s) and OMSSA (2710 s) showing the slowest. For the C,3 and C,4 combinations, the number of incorrect proteomes increased for a 20 ppm error rate versus 5 and 10 ppm error. Using multiple search engines for the profile step resulted in significantly increased numbers of incorrect proteomes along with significantly increased runtime. Increasing the ppm error rate for the profile step also resulted in an increase of runtime overall due to an increase in search complexity, particularly as more search engines were used.

For the S11 data set, as shown in Supporting Information Tables S5 and S6, most combinations identified four or five accurate proteomes with one or two inaccurate proteomes. Increasing the number of search engines used during the profile greatly increased the number of inaccurate proteomes selected with the 3 (5 ppm) combination resulting in 15 inaccurate proteomes selected and the 6 (20 ppm) combination resulting in 48 inaccurate proteomes. The trend shown with the S07 data set continues with Comet (1482 s) showing the quickest profile performance at 90% of NSAF and multiple search engines and higher ppm error drastically increasing the profiling time with the 6 search engine (20 ppm) combination being the slowest at 89 317 s.

Based on these results, we chose Comet by itself to select the species of interest during the profile step as other combinations either resulted in more inaccurate proteomes or significantly increased runtime. We also chose the 90% NSAF combination over the 80% combination due to the increased number of accurate proteomes selected, but note that 80% NSAF may result in fewer inaccurate proteomes and this setting may be adjusted to reflect the goal of a researcher.

The number of PSMs identified from the full-proteome search, accuracy (percent of PSMs from species believed to be in the sample), and runtimes for the various search engine and ppm combinations are show in Figure 2b for the S07 data set and Supporting Information Tables S4 and S7 for the S07 and S11 data sets. MyriMatch and OMSSA showed a clear decrease in the number of PSMs identified compared to the other individual search engines (Comet, MetaMorpheus, MSGF+, and X!Tandem) for both data sets. The C,4 combination at 10 ppm error resulted in the highest number of overall PSM identifications for the S07 data set and the third highest for the S11 data set with 4 (10 ppm) and 3 (10 ppm) being the highest. MSGF+ showed the lowest accuracy among the individual search engine combinations for the S07 data set and X!Tandem the lowest for the S11 data set followed by MSGF+. The number of inaccurate PSMs increased if multiple search engines were used during profiling. Five and 10 ppm error rates showed a relatively similar performance with slightly more PSMs identified at 10 ppm, but the 20 ppm error rate showed both a decrease in the number of PSMs identified and an increase in the percent of inaccurate PSMs.

We also generated species-level breakdown of the S07 and S11 data sets using both NSAF and peptide counts as shown in Supporting Information Figure S3. In all cases, *B. thetaiotaomicron*, *Blautia sp. YL58*, and *E. coli* were the top three species identified with the expression of the remaining species, such as *Erysipelatoclostridium ramosum*, varying depending on the sample and method used.

We compared the MetaProD identification results for the S07 data set using the combination resulting in the highest number of PSM identifications (C,4 10 ppm) as well as X! Tandem (5 ppm) with the reported CAMPI results.[42] CAMPI used two types of search database: a reference database and a multiomic database, where the reference database combined reference proteomes of the known strains (except for *B. producta*, for which the whole genus *Blautia* was used) and the multiomic database was generated from metagenomic and metatranscriptomic data sequenced from a matching sample. Table 2 shows that MetaProD identified slightly more PSMs compared to the CAMPI results, particularly for the C,4 combination (21 558 PSMs), but we could attribute this to different configurations of these two approaches (e.g., the number of search engines and the ppm parameter) and differences in the Uniprot databases.

Table 2. Comparison of the Number of PSMs and % of Spectra for MetaProD versus the Reported Results from the CAMPI Study

combination	no. of PSMs	% of total spectra in sample[a]
X!Tandem (5 ppm)	20 201	43.12
C,4 (10 ppm)	21 558	46.02
CAMPI Reference (5 ppm)	18 805	40.14
CAMPI Multi-Omic (5 ppm)	20 085	42.87

[a]% of total spectra is based the total spectra of 46 847 for the S07 sample reported by the CAMPI study.

Table 3 shows the species breakdown of the protein identifications from our study (C,4 10 ppm) and those reported by the CAMPI study. MetaProD identified 2871 total unique nonhuman/non-CRAP proteins. The top five most abundant species identified by MetaProD are all true species in the SHIUMIx data set. MetaProD showed significantly more resulting from *B. producta* and *E. coli*, whereas CAMPI showed a more significant percentage of its proteins resulting from *B. thetaiotaomicron*. It should be noted that the CAMPI study reported that they did not have a reference proteome available for *B. producta*, which may have resulted in less identifications for that particular species for their pipeline and thus affected the overall percentages of each species. The MetaProD results

Table 3. Species Breakdown of the Proteins Found by Our Pipeline vs That Reported in the CAMPI Study

species	% MetaProD proteins	% CAMPI proteins
B. thetaiotaomicron	53.95	82.77
B. producta	19.78	0.4
E. coli K-12	14.84	7.18
A. caccae	6.13	6.24
C. ramosum	3.69	0.55
Blautia argi[a]	**1.60**[a]	0
C. butyricum	0	1.19
B. longum	0	0.99
Bacteroides dorei[a]	0	**0.41**[a]
L. plantarum	0	0.21
Clostridiales bacterium[a]	0	**0.06**[a]

[a]Those in bold are species not among the eight species present in the sample.

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

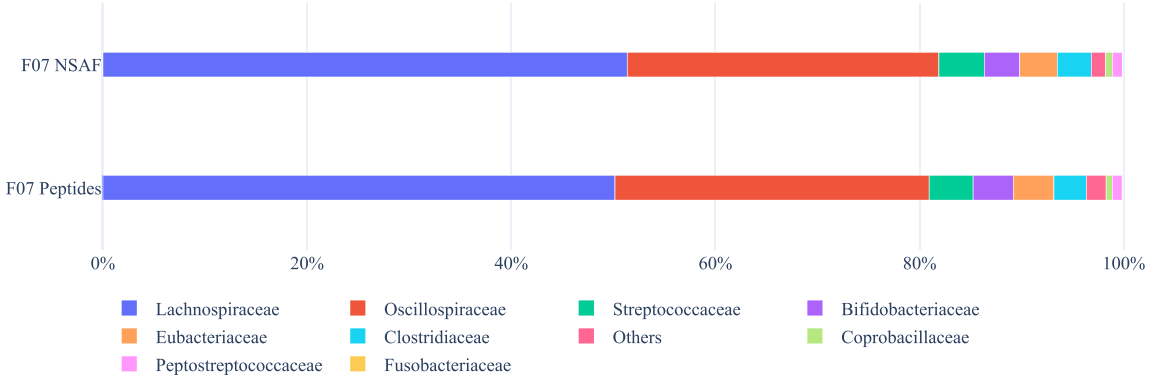


Figure 3. Breakdown of the family-level identifications for the C,4 10 ppm configuration on the CAMPI F07 data sets using NSAF and peptide counts.

Table 4. Number of Peptides Identified Compared to HAPiID-MSGF and MetaPro-IQ

	HM403	HM415	HM454	HM455	HM466	HM467	HM494	HM503
MetaProD (Full, All)[a]	22 053	17 195	20 746	21 372	19 031	19 039	22 158	24 782
MetaProD (Full, Bac)[b]	12 833	12 072	16 795	19 809	12 941	14 951	19 410	23 244
MetaProD (Bac, Bac)[c]	12 367	11 633	16 351	19 557	12 303	14 400	19 176	23 284
HAPiID-MSGF+	12 962	12 535	17 619	20 803	13 862	15 924	21 108	24 962
MetaPro-IQ	12 606	12 179	15 863	18 677	11 733	12 724	19 248	23 632

[a]Includes bacteria, humans, and contaminants in the search database (Full) and keeps all identified peptides (All). [b]Includes bacteria, humans, and contaminants in the search database (Full) and keeps only the bacterial peptides (Bac). [c]Includes only bacterial proteins in the search database (Bac) and therefore only identifies bacterial peptides (Bac).

were more similar to the reported CAMPI results using a metagenomic database along with their Unipept results, both of which would not have the issue with *Blautia*. MetaProD missed low-abundance species including *B. longum* (0.99%), *C. butyricum* (1.19%), and *L. plantarum* (0.21%). There are false positives among the low-abundance species, identified by either MetaProD or CAMPI. This suggests that it may be desirable to not include low-abundance ones to avoid introducing false identification of bacterial species when analyzing metaproteomic data sets with unknown species composition.

We also evaluated the full-proteome performance of the fecal F07 data set at both 80% and 90% NSAF for all the combinations as with S07 and S11 as shown in Supporting Information Tables S8 and S9. For most combinations, more PSMs, peptides, and proteins were identified using the 90% NSAF setting with the exceptions being using Myrimatch and OMSSA as the only search engines and the 4 and 6 search engine combinations. The highest number of PSMs (74 385) identified was by the C,4 (10 ppm) combination at 90% of NSAF and the 4 (10 ppm) combination (74 608) at 80% of NSAF. However, the runtime of this 4 (10 ppm) combination was 318 min compared to 182 min for the C,4 (10 ppm) combination.

For the F07 data set, we also generated a family-level breakdown of the results using both NSAF and peptide count based on the C,4 (10 ppm) combination as shown in Figure 3. The top five families identified were *Lachnospiraceae*, *Oscillospiraceae* (a heterotypic synonym of *Ruminococcaceae*), *Streptococcaceae*, *Bifidobacteriaceae*, and *Eubacteriaceae*, which is consistent with the CAMPI study.

Analysis of the Human Mucosal–Luminal Interface (HMI) Data Set

We used MetaProD to analyze the HMI data set and compared its performance with HAPiID (MSGF+ search engine) and MetaPro-IQ. We ran MetaProD in two different settings: one using the full-search database including human, contaminants, and bacteria, and the other one using a search database consisting of bacterial proteins only. Table 4 shows the MetaProD identification results using the C,4 configuration at 10 ppm and specific cleavage for the protease. We also used this data set with the full-search database to evaluate many of the search engine combinations and to evaluate changing from specific to semispecific cleavage as shown in Supporting Information Tables S10 and S11.

Results from HAPiID (which uses a curated search database of 3357 genomes based on human gut bacteria) and MetaPro-IQ (which uses the integrated gene catalog (IGC)[45] containing 9.9 million genes generated from 1267 human gut samples as the search database in a two-step approach) are also shown for comparison. Using the bacteria-only search database resulted in identification of fewer unique (unmodified sequence) bacterial peptides in MetaProD compared to using the full-search database also containing human proteins and contaminants for each sample except HM503. About 1.5% of the total peptides that were assigned to humans using the full-search database would have also been assigned to bacteria in the absence of human sequences. The higher number of total identifications and potential for misclassification for some peptides suggest that it may be desirable to include human proteins and contaminants in a database search to reduce false identification of bacterial peptides and proteins, especially when the samples come from humans or otherwise could contain human proteins.

The runtimes of the three pipelines were also compared as reflected in Supporting Information Table S12. MetaProD

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

shows improved performance in most cases as compared to HAPiID and especially compared to MetaPro-IQ. Both MetaProD and HAPiID use a similar two-step strategy with HAPs, but the main difference is that HAPiID uses MSGF+ for the first step and MetaProD uses Comet, by default. Supporting Information Table S10 shows runtimes of the individual search engines for MetaProD and shows that Comet (4163 s) for both steps is significantly faster than MSGF+ for both steps (35 042 s), and even though MetaProD used four search engines for the proteome step, the overall speed of Comet for profiling and then using four search engines for the second step (28 418 s) is faster than using MSGF+ for both steps while also providing more identifications overall for this data set (244 476 PSMs for C,4 versus 201 785 PSMs for MSGF+).

Analysis of the Wastewater MetaProteomics Data Set

The comparison between the search engines and specific versus semispecific enzymes is available in Supporting Information Tables S13 and S14. Similar to the HMI data set, the C,4 10 ppm combination with specific protease specificity showed the highest number of identifications (21 884 PSMs and 11 250 peptides). Of these 11 250 peptides, 9798 were identified as belonging to bacterial species.

We pooled the results from all fractions and took the unique unmodified sequences from the MetaProD results (8801 peptides) using the C,4 10 ppm (specific) combination and compared its performance with other previously published methods of protein identification that utilize a multiomic (metagenomic and metatranscriptomic) database from matching samples for metaproteomic data analysis. The two multiomics-based approaches are Contig2Pro,[46,47] which uses proteins predicted (using FragGeneScan[48]) from assembly contigs of matching metagenomic and metatranscriptomic data to build a search database for metaproteomic data analysis, and Graph2Pro,[46,47] which uses a de Bruijn assembly graph (from MegaHit[49]) to predict tryptic peptides to improve the use of matching multiomic data for metaproteomic data and can be used to approximate the upbounds of metaproteomic identification. We filtered the Contig2Pro/Graph2Pro results to match the 8−30 sequence length filtering used by MetaProD. Table 5 summarizes the results. MetaProD (8801 peptides) performed slightly worse than the Contig2Pro approach (9021 peptides) and worse than the Graph2Pro approach (12 764 peptides) . We include a Venn diagram of the overlap between the unique MetaProD results and the unique filtered Contig2Pro/Graph2Pro (Metahit) results as Supporting Information Figure S4, which shows roughly half of the MetaProD peptides overlap with both Contig2Pro and Graph2Pro, but 43% are unique to MetaProD. These results suggest that MetaProD performs reasonably well on nonhuman metaproteomics such as wastewater data sets without using matching metagenomic and/or metatranscriptomic data.

Table 5. Identification Results for MetaProD on the Wastewater Metaproteomic Data Set (SD6) Compared to Other Pipelines

	no. of PSMs	no. of peptides
MetaProD	21 884	8 801
Contig2Pro	19 199	9 021
Graph2Pro	27 495	12 764

Identification of Differentially Expressed Proteins from Multiplexed Colon Cancer Data

To demonstrate the utility of MetaProD for multiplexed metaproteomic data analysis and for exploration of bacterial proteins in human cancer tissues, we applied MetaProD to a collection of colon cancer metaproteomic data. Figure 4 shows that many peptides were only identified in a small fraction of channels (a specific phenotype for a specific individual). We first filtered out the peptides that were identified in less than 50% of the channels. For those retained, MetaProD applied PEMM[40] to imputate the missing abundance values for those channels before estimating the quantification of proteins and applying differential analysis. In total, MetaProD was able to identify 3 515 558 unique PSMs (39.9% of the total spectra), 233 368 unique peptides, 9169 unique proteins, and 7670 unique differentially expressed proteins (BH-adjusted p-value <0.05) in the colon cancer data set, and the full breakdown of human versus bacteria and a comparison to the reported CPTAC results[14] is available in Table 6. There were significantly more human PSMs (3 494 405) identified as compared to bacterial PSMs (21 153). Many of the identified PSMs were related to human blood, such as hemoglobin (68 041 PSMs) and albumin (75 279 PSMs).

Figure 5 shows the volcano plot of the distribution of MetaProD's differentially expressed proteins (shown as genes). In summary, MetaProD identified more differentially expressed (including up- and downregulated) human proteins than CPTAC (see Table 6). Figure 6 shows the overlap of the genes that are upregulated in the cancerous samples identified by MetaProD and CPTAC. Table 7 lists the top five up- and downregulated proteins identified by MetaProD. Among the top five upregulated results from MetaProD, the genes S100P, GPRC5A, and FAP were among the differentially expressed genes found by the CPTAC pipeline, but TRIM29 and CALD1 were not. We note TRIM29[50] and CALD1[51] have both been implicated in colorectal type cancers, so this suggests that the identification of these genes by our approach is not erroneous. The top five downregulated genes identified by MetaProD were found to be linked to cancer, including LPAR1,[52] CHGA,[53] SYN2,[54] GAP43,[55] and NCAM1.[56] CPTAC did not report a final list of downregulated genes, and therefore a more direct comparison to their downregulated results was not possible.

MetaProD was able to identify a significant number of bacterial peptides from the colon cancer data set. Figure 7 shows a breakdown of the species composition based on the full set of identified peptides. It shows that proteins associated with a wide variety of bacterial species could be identified from the cancer data sets, with the top 10 genuses explaining only about 23% of identified proteins (the top three genuses are *Steptomyces*, *Pseudomonas*, and *Candiddatus*). Supporting Information Figure S5 shows the breakdown of the top 25 microbial species identification. As with human proteins, a list of differentially abundant bacterial proteins was identified (BH-adjusted p-value <0.05), and there were 10 bacterial proteins that had an absolute value $\log_2 \geq 1$ in Table 8. Notably, all differentially abundant bacterial proteins meeting this criteria were more abundant in the healthy samples (negative fold-change values). It should also be noted that none of the 10 bacterial matches were identified by peptides that could have matched human proteins. Eight of the 10 proteins from Table 8 (A0A3M2LJ42 and R6WD61 had no information) had Gene Ontology (GO) annotations available

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

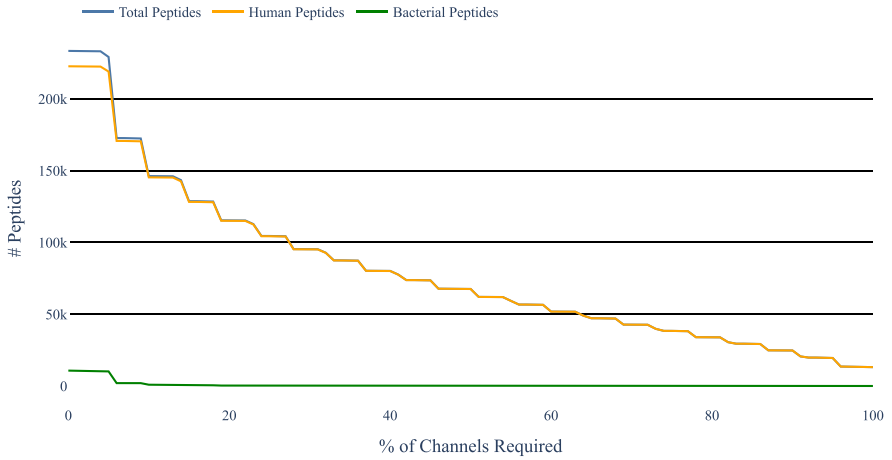


Figure 4. Total number of peptides selected when filtering results by the percent of missing channels prior to imputation for the colon cancer data set.

Table 6. Identification Breakdown for the Colon Cancer Data Set

	no. of pep. (no filter)	no. of pep. (filter)	no. of prot.	no. of diff[a]	no. of up[b]	no. of down[c]
MetaProD Human	222 657	64 988	9142	7646	77	522
MetaProD Bacteria	10 711	28	27	24	0	10
CPTAC Pipeline	N/A	N/A	6422	4744	31	417

[a]Shows the number of differentially expressed proteins. [b]Shows upregulated (more expressed in cancer). [c]Shows downregulated proteins (more expressed in healthy).

in Uniprot, and this is shown in Supporting Information Figure S6. Note that a single protein may have multiple annotations as three proteins had four annotations, one protein had three, and four proteins had one. Three of these eight proteins were annotated as integral components of the membrane, and three of them were related to translation, the ribosome, and rRNA binding. The other two proteins did not share function with any of the others.

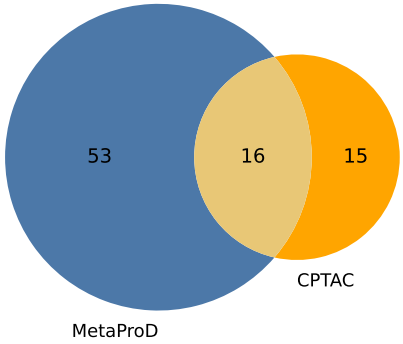


Figure 6. Comparison of the overlap of the differentially expressed upregulated genes with a fold change (FC) ≥ 2 and a *p*-value <0.05 found by the CPTAC pipeline and MetaProD.

CONCLUSION

MetaProD offers highly configurable options in a web-based interface to be suitable for many metaproteomics-based mass spectrometry applications (both label-free and multiplexed).

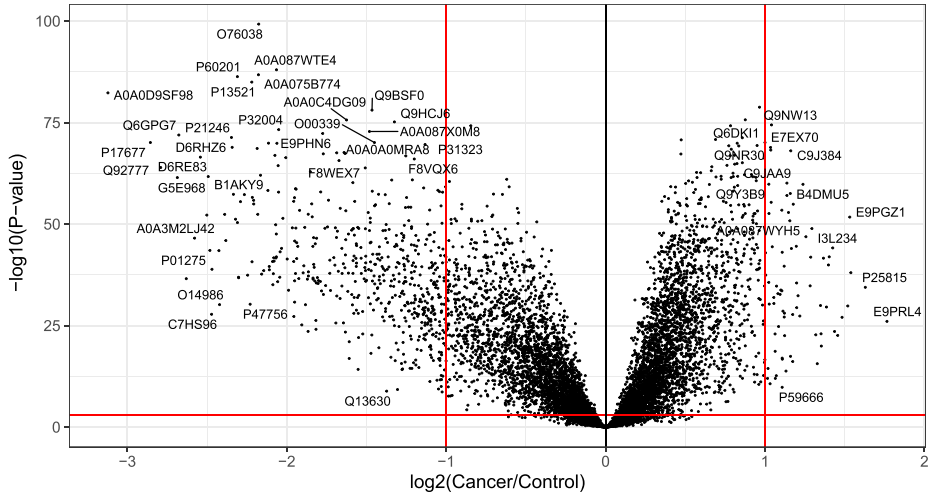


Figure 5. Volcano plot of the differentially expressed proteins (shown as genes) in the colon cancer data set showing significantly more downregulated genes.

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

Table 7. Top Five Human Proteins That Are More Expressed in Colon Cancer Samples and Healthy Samples, Respectively, Sorted by the Fold Change

accession	gene	description	$\log_2$(FC)
E9PRL4	TRIM29	tripartite motif-containing protein 29	1.77
P25815	S100P	protein S100-P	1.63
A0A3B3ITN8	GPRC5A	retinoic acid-induced protein 3	1.54
E9PGZ1	CALD1	caldesmon	1.53
A0A0D9SEN1	FAP	prolyl endopeptidase FAP	1.52
Q6GPG7	LPAR1	lysophosphatidic acid receptor 1	−2.68
G5E968	CHGA	chromogranin A	−2.69
Q92777	SYN2	synapsin-2	−2.79
P17677	GAP43	neuromodulin	−2.86
A0A0D9SF98	NCAM1	neural cell adhesion molecule 1	−3.12

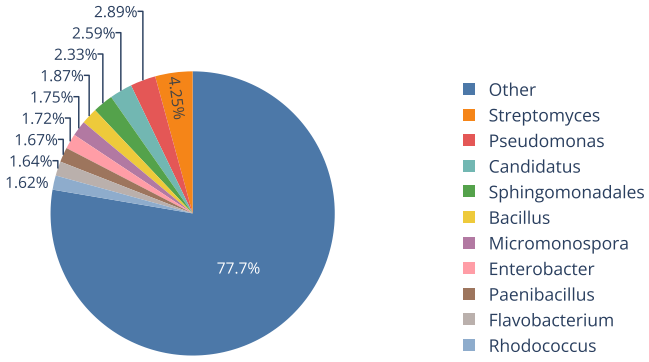


Figure 7. Genus-level breakdown of the colon cancer data showing the top 10 genuses by peptide count and the total percentage of any remaining genus. The full set of identified peptides (unfiltered) was used for this analysis.

Table 8. Top Differentially Abundant Bacterial Proteins Identified from the Colon Cancer Data Set

accession	PPID	description	$\log_2$(FC)[a]
A0A1E3W7U5	UP000094472	0S ribosomal protein L6	−2.63
A0A3M2LJ42	UP000261811	uncharacterized protein	−2.58
C7HS96	UP000003821	50S ribosomal protein L24	−2.47
A0A2I9DAV0	UP000236569	3-hydroxyacyl-CoA dehydrogenase	−1.95
A0A2 V2YZY8	UP000246635	Yip1-like protein	−1.55
C7HT74	UP000003821	uncharacterized protein	−1.40
A0A2 V2YQV4	UP000246635	uncharacterized protein	−1.39
R6WD61	UP000018231	peptidase U32 family	−1.26
C8XBG2	UP000002218	50S ribosomal protein L1	−1.21
A0A4R1UYP4	UP000295319	NAD(P)-dependent dehydrogenase	−1.10

[a]Positive value indicates higher abundance in cancer samples, and negative indicates higher abundance in healthy samples.

MetaProD performs well in terms of both number of identifications and runtime compared to many other previously published pipelines and has the potential to be expanded in the future to integrate new algorithms or techniques and updated as new versions of component software are deployed or as Uniprot updates their database.

The web interface allows for a simple and quick way to create new projects, modify settings, and administer many projects at once with the potential for user and password-based access. The resulting Web site allows for the potential for results to be viewable by many users at once using a web browser and can be deployed internally to keep results private or publicly to allow for broader access.

MetaProD was also able to identify differentially expressed human and bacterial proteins from a real-world CPTAC cancer data set, showing one of the many useful applications of the pipeline.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00614.

List of the Uniprot proteome IDs used (XLSX)

Figures S1−S6 and Tables T1−T14, including species, proteome selection and processing time, search algorithms, Web site configuration interface, and species identifications (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yuzhen Ye − *Computer Science Department, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, Indiana 47408, United States;* orcid.org/0000-0003-3707-3185; Email: yye@indiana.edu

Authors

Jamie Canderan − *Informatics Department, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, Indiana 47408, United States*

Moses Stamboulian − *Informatics Department, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, Indiana 47408, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00614

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Zhao, L. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **2018**, *359*, 1151−1156.

(2) Sánchez-Alcoholado, L.; Ramos-Molina, B.; Otero, A.; Laborda-Illanes, A.; Ordóñez, R.; Medina, J. A.; Gómez-Millán, J.; Queipo-Ortuño, M. I. The Role of the Gut Microbiome in Colorectal Cancer Development and Therapy Response. *Cancers* **2020**, *12*, 1406.

(3) Young, V. B. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ* **2017**, j831.

(4) Maslowski, K. M.; Mackay, C. R. Diet, gut microbiota and immune responses. *Nat. Immun.* **2011**, *12*, 5−9.

(5) Hooper, L. V.; Gordon, J. I. Commensal Host-Bacterial Relationships in the Gut. *Science* **2001**, *292*, 1115−1118.

(6) Manor, O.; Dai, C. L.; Kornilov, S. A.; Smith, B.; Price, N. D.; Lovejoy, J. C.; Gibbons, S. M.; Magis, A. T. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **2020**, *11*, 5206.

(7) Moon, C.; Stupp, G. S.; Su, A. I.; Wolan, D. W. Metaproteomics of Colonic Microbiota Unveils Discrete Protein Functions among Colitic Mice and Control Groups. *Proteomics* **2018**, *18*, 1700391.

(8) Kleiner, M.; Thorson, E.; Sharp, C. E.; Dong, X.; Liu, D.; Li, C.; Strous, M. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **2017**, *8*, 1558.

(9) Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems* **2019**, *4*, e00115−19.

(10) Mayers, M. D.; Moon, C.; Stupp, G. S.; Su, A. I.; Wolan, D. W. Quantitative Metaproteomics and Activity-Based Probe Enrichment Reveals Significant Alterations in Protein Expression from a Mouse Model of Inflammatory Bowel Disease. *J. Proteome Res.* **2017**, *16*, 1014−1026.
(11) Duong, M. T.-Q.; Qin, Y.; You, S.-H.; Min, J.-J. Bacteria-cancer interactions: bacteria-based cancer therapy. *Exp. Mol. Med.* **2019**, *51*, 1−15.
(12) Hettich, R. L.; Pan, C.; Chourey, K.; Giannone, R. J. Metaproteomics: Harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem.* **2013**, *85*, 4203−4214.
(13) Arteta, A. A.; Sánchez-Jiménez, M.; Dávila, D. F.; Palacios, O. G.; Cardona-Castro, N. Biliary Tract Carcinogenesis Model Based on Bile Metaproteomics. *Front. Oncol.* **2020**, *10*, 1032.
(14) Vasaikar, S. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **2019**, *177*, 1035−1049.e19.
(15) Pappireddi, N.; Martin, L.; Wühr, M. A Review on Quantitative Multiplexed Proteomics. *ChemBioChem* **2019**, *20*, 1210−1224.
(16) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ *. *Mol. Cell. Proteomics* **2014**, *13*, 2513−2526.
(17) Webb-Robertson, B.-J. M.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; Waters, K. M. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J. Proteome Res.* **2015**, *14*, 1993−2001.
(18) Wöhlbrand, L.; Rabus, R.; Blasius, B.; Feenders, C. Influence of NanoLC Column and Gradient Length as well as MS/MS Frequency and Sample Complexity on Shotgun Protein Identification of Marine Bacteria. *Microb. Physiol.* **2017**, *27*, 199−212.
(19) Sonnett, M.; Yeung, E.; Wühr, M. Accurate, Sensitive, and Precise Multiplexed Proteomics Using the Complement Reporter Ion Cluster. *Anal. Chem.* **2018**, *90*, 5032−5039.
(20) O'Connell, J. D.; Paulo, J. A.; O'Brien, J. J.; Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* **2018**, *17*, 1934−1942.
(21) 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. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **2016**, *4*, 1−12.
(22) Cheng, K.; Ning, Z.; Zhang, X.; Li, L.; Liao, B.; Mayne, J.; Stintzi, A.; Figeys, D. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **2017**, *5*, 1−10.
(23) Stamboulian, M.; Li, S.; Ye, Y. Using high-abundance proteins as guides for fast and effective peptide/protein identification from human gut metaproteomic data. *Microbiome* **2021**, *9*, 1−17.
(24) Griss, J.; Vinterhalter, G.; Schwämmle, V. IsoProt: A Complete and Reproducible Workflow To Analyze iTRAQ/TMT Experiments. *J. Proteome Res.* **2019**, *18*, 1751−1759.
(25) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining Results of Multiple Search Engines in Proteomics. *Mol. Cell. Proteomics* **2013**, *12*, 2383−2393.
(26) Zybailov, B. L.; Florens, L.; Washburn, M. P. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol. BioSyst.* **2007**, *3*, 354.
(27) Pan proteomes. https://www.uniprot.org/help/pan_proteomes.
(28) TheUniProtConsortium. UniProt: the universal protein knowledgebase. *Nucl. Acids Res.* **2017**, *45*, D158−D169.
(29) cRAP protein sequences. https://www.thegpm.org/crap/.
(30) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17*, 2552−2555.
(31) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. *Proteome Bioinform.* **2010**, *604*, 55−71.
(32) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRawFileParser: modular, scalable and cross-platform RAW file conversion. *J. Proteome Res.* **2020**, *19*, 537−542.
(33) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.
(34) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(35) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6*, 654−661.
(36) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **2004**, *3*, 958−964.
(37) Muth, T.; Vaudel, M.; Barsnes, H.; Martens, L.; Sickmann, A. XTandem Parser: an open-source library to parse and analyse X! Tandem MS/MS search results. *Proteomics* **2010**, *10*, 1522−1524.
(38) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33*, 22−24.
(39) Barsnes, H.; Vaudel, M. Reporter. http://compomics.github.io/projects/reporter.
(40) Chen, L. S.; Prentice, R. L.; Wang, P. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **2014**, *70*, 312−322.
(41) Zhu, Y.; Orre, L. M.; Tran, Y. Z.; Mermelekas, G.; Johansson, H. J.; Malyutina, A.; Anders, S.; Lehtiö, J. DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis *. *Mol. Cell. Proteomics* **2020**, *19*, 1047−1057.
(42) Van Den Bossche, T. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat. Commun.* **2021**, *12*, 7305.
(43) Muller, E. E. L. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* **2014**, *5*, 5603.
(44) Edwards, N. J.; Oberti, M.; Thangudu, R. R.; Cai, S.; McGarvey, P. B.; Jacob, S.; Madhavan, S.; Ketchum, K. A. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **2015**, *14*, 2707−2713.
(45) Li, J. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **2014**, *32*, 834−841.
(46) Tang, H.; Li, S.; Ye, Y. A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. *PLoS Comput. Biol.* **2016**, *12*, e1005224.
(47) Li, S.; Tang, H.; Ye, Y. A meta-proteogenomic approach to peptide identification incorporating assembly uncertainty and genomic variation. *Molecular & Cellular Proteomics* **2019**, *18*, S183−S192.
(48) Rho, M.; Tang, H.; Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucl. Acids Res.* **2010**, *38*, e191.
(49) Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **2015**, *31*, 1674−1676.
(50) Han, J.; Zuo, J.; Zhang, X.; Wang, L.; Li, D.; Wang, Y.; Liu, J.; Feng, L. TRIM29 is differentially expressed in colorectal cancers of different primary locations and affects survival by regulating tumor immunity based on retrospective study and bioinformatics analysis. *J. Gastrointest. Oncol.* **2022**, *13*, 1132−1151.
(51) Yokota, M.; Kojima, M.; Higuchi, Y.; Nishizawa, Y.; Kobayashi, A.; Ito, M.; Saito, N.; Ochiai, A. Gene expression profile in the

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453

activation of subperitoneal fibroblasts reflects prognosis of patients with colon cancer. *Int. J. Cancer* **2016**, *138*, 1422−1431.
(52) Leve, F.; Peres-Moreira, R. J.; Binato, R.; Abdelhay, E.; Morgado-Díaz, J. A. LPA Induces Colon Cancer Cell Proliferation through a Cooperation between the ROCK and STAT-3 Pathways. *PLoS One* **2015**, *10*, e0139094.
(53) Zhang, X.; Zhang, H.; Shen, B.; Sun, X–F. Chromogranin-A Expression as a Novel Biomarker for Early Diagnosis of Colon Cancer Patients. *IJMS* **2019**, *20*, 2919.
(54) Vicente, C. M.; da Silva, D. A.; Sartorio, P. V.; Silva, T. D.; Saad, S. S.; Nader, H. B.; Forones, N. M.; Toma, L. Heparan Sulfate Proteoglycans in Human Colorectal Cancer. *Anal. Cell. Pathol.* **2018**, *2018*, e8389595.
(55) Chen, X.; Wu, H.; Feng, J.; Li, Y.; Lv, J.; Shi, W.; Fan, W.; Xiao, L.; Sun, D.; Jiang, M.; Shi, M. Transcriptome profiling unveils GAP43 regulates ABC transporters and EIF2 signaling in colorectal cancer cells. *BMC Cancer* **2021**, *21*, 24.
(56) Fernández-Briera, A.; García-Parceiro, I.; Cuevas, E.; Gil-Martín, E. Effect of human colorectal carcinogenesis on the neural cell adhesion molecule expression and polysialylation. *Oncology* **2010**, *78*, 196−204.

https://doi.org/10.1021/acs.jproteome.2c00614
J. Proteome Res. 2023, 22, 442−453