analytical chemistry

pubs.acs.org/ac Article

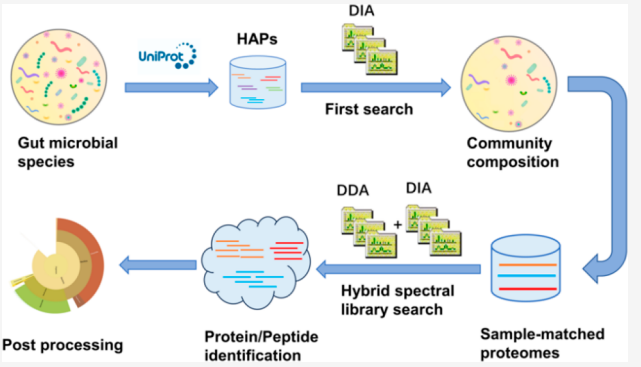
High-Abundance Protein-Guided Hybrid Spectral Library for Data-Independent Acquisition Metaproteomics

Enhui Wu, Yi Yang, Jinzhi Zhao, Jianxujie Zheng, Xiaoqing Wang, Chengpin Shen, and Liang Qiao*

Cite This: https://doi.org/10.1021/acs.analchem.3c03255 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Metaproteomics offers a direct avenue to identify microbial proteins in microbiota, enabling the compositional and functional characterization of microbiota. Due to the complexity and heterogeneity of microbial communities, in-depth and accurate metaproteomics faces tremendous limitations. One challenge in metaproteomics is the construction of a suitable protein sequence database to interpret the highly complex metaproteomic data, especially in the absence of metagenomic sequencing data. Herein, we present a high-abundance protein-guided hybrid spectral library strategy for in-depth data independent acquisition (DIA) metaproteomic analysis (HAPs-hyblibDIA). A dedicated high-abundance protein database of gut microbial species is constructed and used to mine the taxonomic information on microbiota samples. Then, a sample-specific protein sequence database is built based on the taxonomic information using Uniprot protein sequence for subsequent analysis of the DIA data using hybrid spectral library-based DIA analysis. We evaluated the accuracy and sensitivity of the method using synthetic microbial community samples and human gut microbiome samples. It was demonstrated that the strategy can successfully identify taxonomic compositions of microbiota samples and that the peptides identified by HAPs-hyblibDIA overlapped greatly with the peptides identified using a metagenomic sequencing-derived database. At the peptide and species level, our results can serve as a complement to the results obtained using a metagenomic sequencing-derived database. Furthermore, we validated the applicability of the HAPs-hyblibDIA strategy in a cohort of human gut microbiota samples of colorectal cancer patients and controls, highlighting its usability in biomedical research.



INTRODUCTION

The human digestive tract contains billions of symbiotic microorganisms, characterized by high number of microbial cells and high diversity in microbial communities.[1] Collectively referred as the gut microbiome, these microorganisms play pivotal roles in numerous physiological processes associated with human health and disease.[2,3] Dysbiosis of gut microbiota can cause a variety of diseases, such as inflammatory bowel diseases (IBD),[4] irritable bowel syndrome (IBS),[5] colorectal cancer (CRC),[6] obesity,[7] type 2 diabetes,[8] and atopic allergy.[9] Metagenomic sequencing has significantly advanced our understanding of gut microbiota and its relationship with the host.[10,11] However, while metagenomics can characterize the taxonomic profile and functional potential of gut microbiota, it falls short in revealing gene expression. Metaproteomics, based on mass spectrometry, allows for the analysis of expressed proteins within a microbial community, providing crucial functional insights into the gut microbiota.[12−15]

Currently, most metaproteomic studies use the data-dependent acquisition (DDA) mass spectrometry workflow.[16−18] However, DDA is characterized by a stochastic precursor selection, resulting in restricted identification and quantification of the protein at low abundance. To address these limitations, data-independent acquisition (DIA) methods have been proposed as an alternative.[19,20] DIA methods systematically fragment all precursor ions within defined isolation windows, offering a more comprehensive acquisition of all fragments of all precursors. DIA data analysis strategies encompass library-based and library-free methods. In library-based DIA analysis, sample-specific DDA data are searched against a protein sequence database to construct a spectral library. Subsequently, DIA data are matched against the DDA-based spectral library for peptide identification and quantification, utilizing a peptide-centric scoring algorithm.[21] Conversely, the directDIA method searches DIA data against a protein sequence database using a spectrum deconvolution algorithm.[22]

Received: July 23, 2023
Revised: December 14, 2023
Accepted: December 14, 2023

In all these data analysis strategies, protein sequence database is a key prerequisite and has a significant sway on the identification result and the accuracy of downstream analysis.[23,24] Currently, metagenomic sequencing-derived sample-specific protein sequence databases have been widely regarded as the gold standard in the field of metaproteomics.[25,26] However, metagenomic sequencing-derived databases may encounter challenges related to gene assembly and taxonomic annotation,[27,28] failing to fully represent the microbial community of a given sample.[23,26] Consequently, important proteins expressed by low abundance species may be overlooked, leading to discrepancies between the compiled gene database from metagenomic sequencing and the actual proteins in the sample. As an alternative, public proteome sequence databases are frequently employed and can provide a comprehensive coverage of possible species. However, such a huge database results in an expanded search space, which can reduce the search sensitivity and significantly increase the search time. Therefore, iterative search strategies have been developed to solve the problems.[29−31] In a recent study by Stamboulian et al.,[32] a two-step approach HAPiID was developed for DDA-based metaproteomics. In HAPiID, a database of high-abundance proteins (HAPs) serves as a guideline for constructing a target database, thereby diminishing computational time and improving peptide identification sensitivity. Nevertheless, for DIA-based metaproteomics, there is still a lack of such a two-step-based approach.

In this study, we develop a high-abundance protein-guided hybrid spectral library strategy for DIA metaproteomic analysis (HAPs-hyblibDIA). A dedicated high-abundance protein database of gut microbial species is constructed and used to mine the taxonomic information on samples by directDIA. Afterward, a target full proteome database of the selected organisms can be constructed from public proteome sequence databases, e.g., UniProt, for subsequent analysis. We evaluated the accuracy and sensitivity of the strategy using synthetic microbial community samples and human gut microbiome samples and demonstrated that the strategy can accurately identify taxonomic information from DIA-based metaproteomic data. At the peptide and species level, HAPs-hyblibDIA can serve as a complement to the results obtained with a metagenomic sequencing-derived database. It can also be used when the metagenomic sequencing data are unavailable. Furthermore, we validated the applicability of the strategy in real clinical cohorts, highlighting its usability in analyzing authentic clinical fecal samples.

EXPERIMENTAL SECTION

Protein Sequence Database Construction in the HAPs-hyblibDIA Pipeline. 647 gut microbial species were collected from HMP (https://hmpdacc.org/),[33] the Human Gastrointestinal Bacteria Culture Collection (HBC),[34] archaea,[35] and our previous study[13] (see Supplementary Data 1 for details). The ribosomal proteins and elongation factors of the 647 gut microbial species were downloaded from UniProt (https://www.uniprot.org/) to construct a HAPs database of gut microbiomes which contained 159 387 proteins from 1929 proteomes of the 647 species. In addition, HAPs from 1952 binned genomes of Uncharaterised MetaGenome Species (UMGS) were extracted based on the protein sequences and the annotation data.[32,36] We combined the HAPs of UMGS with the HAPs database of 647 microbial species, namely, the combined UMGS HAPs database, which contained 295 608 proteins. DIA data of metaproteomic sample were analyzed by directDIA using the HAPs databases to identify proteins. Then, enrichment analysis was performed on the identified proteins with the HAPs databases as background using the R package "clusterProfiler", and the p-values by hypergeometric test were adjusted by the Benjamini−Hochberg method. The organisms with a p-value of <0.05 were selected. Then, the corresponding full proteome of the selected organisms were downloaded from UniProt to construct the sample specific protein sequence database.

The UP100 database was constructed by combining the proteomes of 100 randomly selected species from UniProt (see Supplementary Data 1 for details). The 100 HAPs database contained the ribosomal proteins and elongation factors extracted from the UP100 database. The details of the UP12 and UP8 databases are presented in the Supporting Information.

Metaproteomics Data Analysis. For the directDIA analysis, raw DIA data were analyzed by Spectronaut (version 17, Biognosys AG, Schlieren, Switzerland) with default settings. Trypsin was set as the digestion enzyme. Carbamidomethyl (C) was specified as the fixed modification. Oxidation (M) was specified as the variable modification. Retention time prediction type was set to dynamic iRT. Q-value (FDR) cutoff on both precursor and protein level was 1%. For the hybrid library-based DIA analysis, the DDA and DIA raw data were searched against the corresponding protein sequence database using Spectronaut Pulsar with the default settings. Then, the generated hybrid libraries were used to analyze the DIA data by Spectronaut with the same settings aforementioned.

Taxonomic and Function Annotations and Statistical Analysis. The peptides identified using the metagenomic sequencing-derived databases were subjected to Unipept (https://unipept.ugent.be/) for taxonomic analysis employing the lowest common ancestor approach. Leucine and isoleucine were considered equal. The taxon abundance at the metagenomics level was obtained by counting the number of reads of the corresponding taxon. The taxon abundance at the metaproteomics level was determined by summing the intensities of all of the peptides corresponding to the taxon. Statistical analysis was conducted using R (version 4.1.3, https://www.r-project.org/) and python (version 3.10.10, https://www.python.org/) with packages "pandas" and "numpy". Data visualization was conducted with R packages "ggplot2" and "VennDiagram", and AntV G2 (https://github.com/antvis/g2). KEGG annotation was performed using the Ghost-KOALA Web server (https://www.kegg.jp/ghostkoala).

Protein sample preparation, LC-MS/MS analysis (LC gradients and DIA variable window settings are detailed in Table S1 and Table S2), the construction of UP12 and UP8 databases, metagenomic sequencing and data analysis, and the data sets from public resources used in this study are described in detail in the Supporting Information.

RESULTS AND DISCUSSION

Principle of the HAPs-hyblibDIA Pipeline. With the increasing focus on microbiome studies during the past years, we can obtain fruitful gut microbial taxonomic information from public resources. Recently, Stamboulian et al.[32] developed a two-step approach HAPiID for DDA-based metaproteomics and proposed a database of high-abundance proteins (HAPs). In the HPAs database by Stamboulian et

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

al.,[32] 3357 genomes from four sources were included, i.e., 612 genomes from the HMP,[33] 737 genomes from the HBC,[34] 1952 genomes from the UMGS,[36] and 56 archaea[35] genomes. The sequences of the HMP genomes and the archaea genomes were directly downloaded from RefSeq, while those of the HBC and the UMGS were translated from computationally predicted genes of the HBC contigs and the UMGS bins. In our work, we chose a similar but different strategy in the HAPs construction. We collected the species information from the HMP, HBC, and archaea databases, as well as from our previous studies, and obtained in total 647 species. Then, the corresponding high-abundance proteins (ribosomal proteins and elongation factors) were downloaded from the UniProt database. As a result, our HAPs database contained 1929 proteomes from 647 species with a size of 159 387 proteins. In addition, HAPs from 1952 binned genomes of Uncharaterized MetaGenome Species (UMGS) were combined with the HAPs database of the 647 microbial species to form the combined UMGS HAPs database. The combined UMGS HAPs database contains 295 608 proteins. As depicted in Figure 1, DIA data

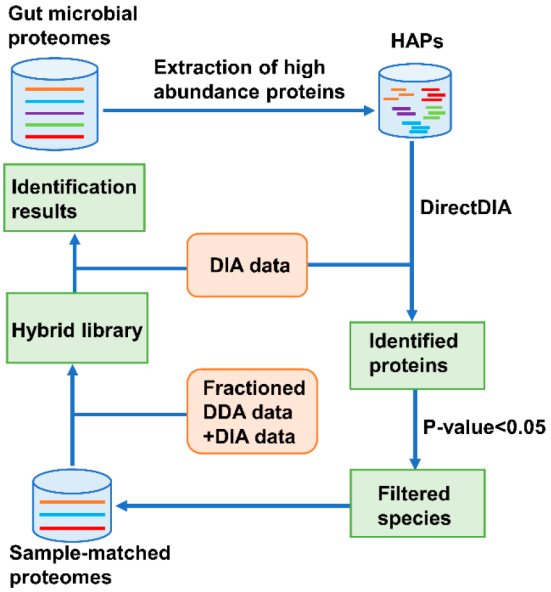


Figure 1. Workflow of the HAPs-hyblibDIA pipeline.

are first searched by directDIA against the HAPs database to select the target organisms. Then, a full proteome database of the selected organisms can be constructed as a sample-specific database. After that, spectral libraries for DIA analysis are built by using the sample-specific database from the DIA data and the corresponding DDA data for peptide-centric analysis. The database can also be used to analyze DIA data by spectrum-centric strategies, e.g., directDIA, and both identification results can be used for subsequent analysis.

Evaluation of the HAPs-hyblibDIA Pipeline Using Simulated Microbial Communities. Initially, we evaluated the performance of the HAPs-hyblibDIA pipeline by analyzing a simulated microbial community of 12 species (12mix). The data sets of the 12mix are from our previous work, and the details of the composition of the 12 species can be found in the original publication.[13] We constructed a reference database UP12 containing the proteomes of the 12 target species from UniProt. In order to evaluate the accuracy of the taxonomy identification method, we randomly selected 100 proteomes of 100 species, including the 12 target species, to construct a relatively large database UP100 (Supplementary Data 1). Then the ribosomal proteins and elongation factors in the UP100 database were extracted to construct a 100 HAPs database. By directDIA analysis of the DIA data of the 12mix against the 100 HAPs database, 666 protein groups were obtained for further taxonomy selection analysis. Figure 2A shows a plot of the adjusted p-values versus the number of organisms identified, with a clear turning point after 12 species. Under the criterion of p-value of <0.05, all of the 12 target species were identified correctly and there were no false positive identifications. The 100 HAPs-filtered database is thus identical to the reference database UP12.

We performed directDIA and hybrid library-based DIA analysis (hyblibDIA) on the DIA data of the 12mix using the 100 HAPs-filtered database (100 HAPs), the UP100 database (UP100), and the metagenomics sequencing-derived database (MG). For directDIA, three technically replicated DIA data sets were searched against the three databases. For the hyblibDIA, 12 fractioned DDA data and 3 technically replicated DIA data from the 12mix sample were combined as input files to search against the three databases by Spectronaut Pulsar. Then DIA data were searched against the hybrid spectral library to obtain the identification results. Overall, the number of proteins and peptides identified by hyblibDIA was significantly larger than directDIA as shown in Figure 2B and Figure 2C. The numbers of peptides and protein groups identified from each DIA run are shown in Table S3. The hybrid library with the 100 HAPs resulted in the identification of 11 770 protein groups and 66 457 peptides (Figure 2B and Supplementary Data 2). When comparing the different databases, it is worth noting that UP100 identified 49 more protein groups than the 100 HAPs, but the number of peptides was less than the 100 HAPs. It is attributed to the fact that the UP100 contains an overwhelming quantity of nontarget species, resulting in a mismatch during protein inference.

As for further analysis, we compared the protein groups of each species identified by the hyblibDIA. As shown in Figure 2D, the numbers of protein groups identified for *Klebsiella pneumonia* and *Escherichia coli* using the MG were significantly larger than those using the 100 HAPs, while the numbers of protein groups for *Citrobacter freundii* and *Klebsiella aerogenes* using the MG were smaller than those using the 100 HAPs. The numbers of protein groups for the others of the 12 species were similar for both databases. These differences can be attributed to the varieties in metagenomic sequencing-derived and UniProt-based databases including protein sequences and species annotations. Notably, although the total number of protein groups identified using UP100 was larger than that using 100 HAPs, the number of protein groups assigned to each target species was larger using 100 HAPs than the UP100. 747 proteins were wrongly identified from species other than the 12 target species using the UP100. Overall, these results demonstrated that the HAPs-hyblibDIA pipeline can achieve substantial proteome coverage in metaproteomic analysis.

Additionally, we applied our pipeline to analyze a publicly available data set consisting of 8 species (8mix, *Bacteroides uniformis*, *Bifidobacterium adolescentis*, *Enterococcus faecalis*, *Escherichia coli*, *Faecalibacterium prausnitzii*, *Lactobacillus acidophilus*, *Staphylococcus aureus*, and *Streptococcus pyogenes*).[14] The UP100 and the 100 HAPs database for the 8mix data set contained the proteome of the 8 target species, and the rest of

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

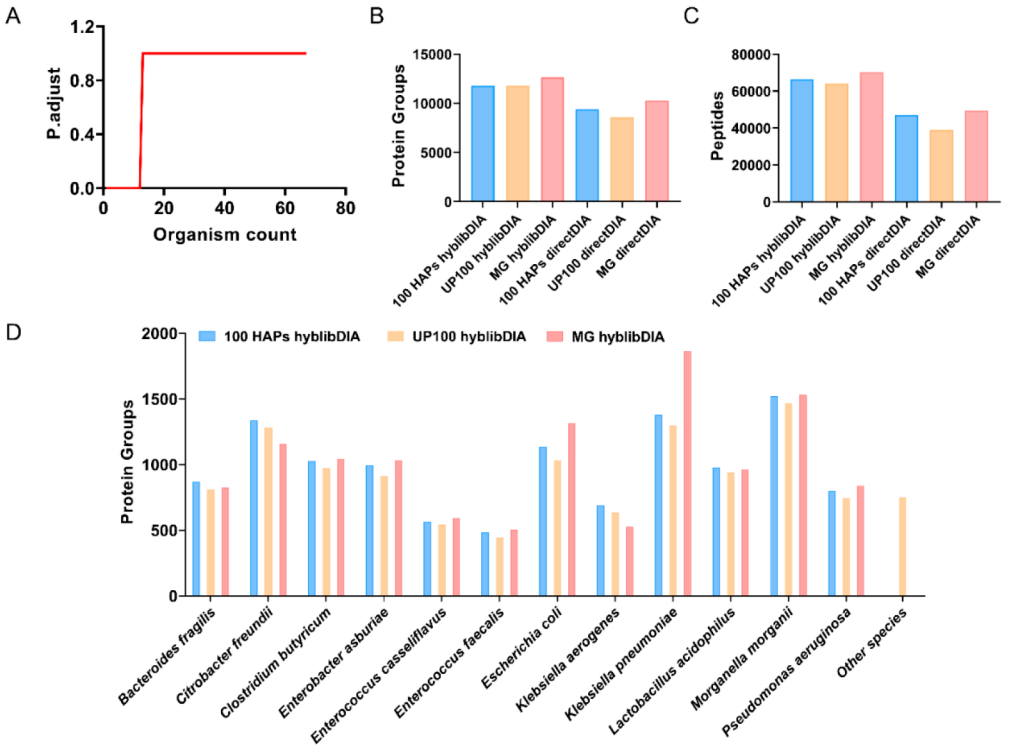


Figure 2. Performance of HAPs-hyblibDIA in the analysis of the simulated microbial community of 12 species. (A) Identification of organisms in the sample using the HAPs database by plotting the adjusted *p*-value (P.adjust) versus the number of organisms. The number of (B) protein groups and (C) peptides identified by the hyblibDIA and directDIA using a HAPs-filtered database (100 HAPs), 100 species database (UP100), and metagenomic sequencing-derived database (MG). (D) The number of protein groups for each species identified by the hyblibDIA using the different databases. The DIA data were obtained with three technique replicates, and the sums of identification results from the three replicates were reported for the comparison.

the two databases were the same as those for the 12mix data set (Supplementary Data 1). We conducted the same method to identify taxonomic information on the sample and finally identified 9 species (Figure S1). In addition to the eight target species, *Klebsiella pneumoniae* was also selected as a candidate target species by our method. *Klebsiella pneumoniae* is from the *Enterobacteriaceae* family, the same as *Enterococcus faecalis* and *Escherichia coli*. Due to the inherent challenge of searching metaproteomic data against a comprehensive and intricate database, peptides from different species with high sequence similarity can be misassigned, leading to the mis-selection of *Klebsiella pneumoniae*.

We subsequently employed the 100 HAPs-filtered 9-species UniProt database (100 HAPs), the 8 target species UniProt database (UP8), and the UP100 database for directDIA and hyblibDIA analysis. As depicted in Figure S2 and Supplementary Data 3, the identification using directDIA with the 100 HAPs gave more accurate results than those using the UP100. Fewer erroneous species were identified with the 100 HAPs, and a higher number of proteins were assigned to each target species. Figure S3 displays the identification achieved by hyblibDIA, which invariably outperformed directDIA. The numbers of peptides and protein groups identified from each DIA run are listed in Table S4. In terms of identified protein groups and peptides, the 100 HAPs and the UP8 were comparable, while the UP100 exhibited significantly lower sensitivity and higher false rate at species level. It is noted that the UP8 database led to a significantly larger number of proteins identified for *Escherichia coli* than the 100 HAPs. With the 100 HAPs, these proteins were wrongly assigned to *Klebsiella pneumoniae*, due to the high sequence similarity between peptides from *Escherichia coli* and *Klebsiella pneumoniae*.

Performance of the HAPs-hyblibDIA Pipeline in Analyzing Human Gut Microbiota. We further explored the application of the HAPs-hyblibDIA pipeline in the analysis of the human gut microbiota. One human stool sample was collected from a healthy volunteer for metaproteomic analysis and metagenomic sequencing. Both DIA and fractionated DDA data were acquired for the sample. The HAPs database of 647 microbial species and the combined UMGS HAPs database were applied to analyze the DIA data for organism selection. As shown in Figure S4, 256 organisms or 366 organisms with adjusted *p*-values of <0.05 were selected using the two databases. Then, we performed directDIA and hyblibDIA analysis with the 647 species HAPs-filtered database (HAPs), the combined UMGS HAPs-filtered database (Combined UMGS) and the metagenomic sequencing-derived database (MG) (Figure 3A,B and Supplementary Data 4). For the hyblibDIA analysis, 24 018 proteins and 114 046 peptides were identified using the HAPs, 28 585 proteins and 131 867 peptides were identified using the Combined UMGS, while 29 334 proteins and 130 595 peptides were identified using MG. For directDIA, 13 960 proteins and 55 681 peptides were identified using the HAPs, 16 586 proteins and 62 293 peptides were identified using the Combined UMGS, while 18 921 proteins and 73 086 peptides were identified using the MG. The numbers of peptides and protein groups identified from each DIA run are shown in Table S5. When the HAPs database included the UMGS, the HAPs-filtered database, and the MG

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

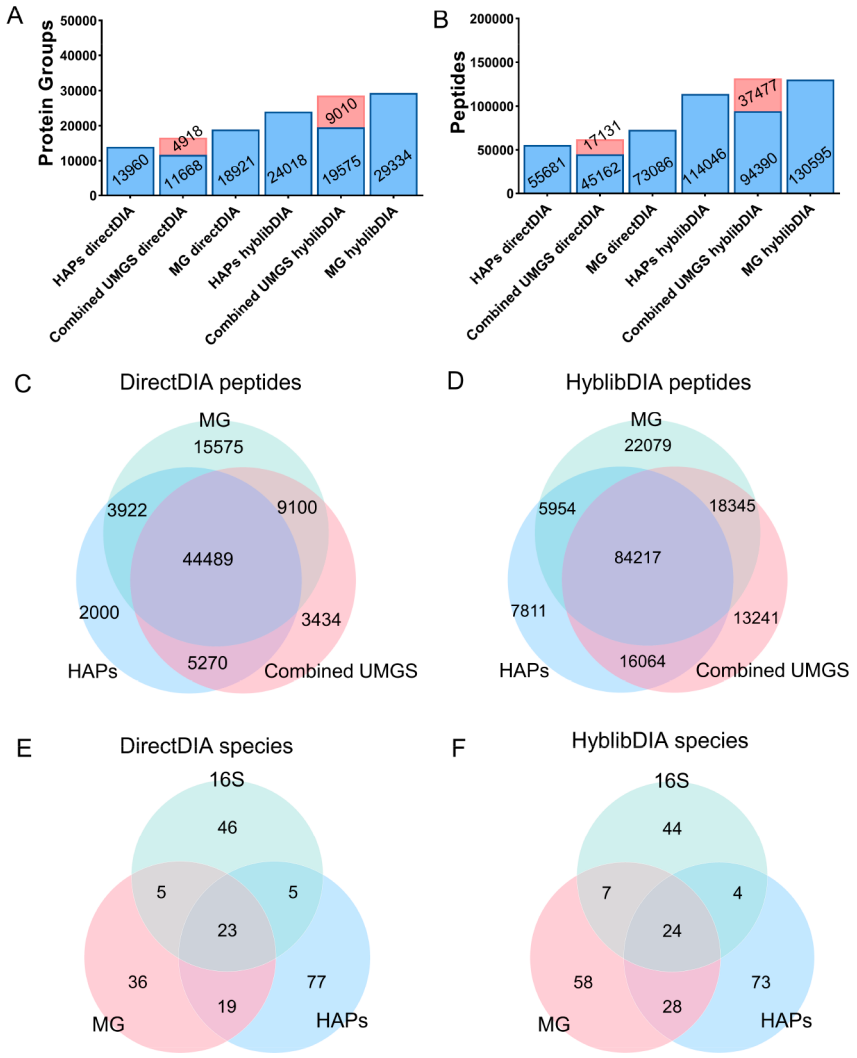


Figure 3. Analysis of stool sample. (A, B) The number of protein groups and peptides identified using different databases by directDIA and hyblibDIA analysis. (C, D) Venn diagram illustrating the overlap of peptides identified using the 647 species HAPs-filtered database (HAPs), the combined UMGS HAPs-filtered database (Combined UMGS), and the metagenomic sequencing-derived database (MG). (E-F) Overlap of species information obtained from the metaproteomics using the HAPs and the MG databases as well as from 16S rRNA sequencing results (16S). The DIA data were obtained with three technique replicates, and the sums of identification results from the three replicates were reported for the comparison.

database led to very similar numbers of protein groups and peptides identified using the hyblibDIA analysis. The results demonstrated the good performance of the HAPs-hyblibDIA pipeline in the analysis of human gut microbiota. However, most proteins of the UMGS database have a taxonomy annotation level of genera or above, and we then did not include the UMGS database for downstream taxonomy analysis and bioinformatics analysis. This result also showed that there are many human gastrointestinal bacteria without well documented information in Uniprot and without species level taxonomic information. The coverage of human gastrointestinal microbes should be improved with the development of a microbiota study in the future.

As for further comparison, we employed popular public protein sequence database for DIA data analysis, known as HMP stool_nr database (https://portal.hmpdacc.org) and JPGM,[37] both with a size of over 4.8 million protein sequences (929MB and 1.13GB). Similarly, directDIA and hyblibDIA were performed for protein identification. As shown in Figure S5, the numbers of peptides identified using the two public databases were slightly lower than those using the MG database or the combined UMGS HAPs-filtered database. However, more proteins were identified using the two public databases. This phenomenon is similar to the comparison between the HAPs-based method and the UP100 database in the analysis of the 12-mix sample, which represents false protein inference generated from the unsuitably large scale of database. In such case, protein inference becomes extremely complex where peptides can stem from homologous proteins of either closely related organisms or from well-conserved proteins in less-related organisms.[23] It should also be noted that the data analysis took a much longer time using the public databases compared to our strategy, as shown in Table S6.

To test the consistency of the results obtained with the HAPs-filtered database and the MG, we performed a comparative analysis at the peptide and taxon level. As depicted in Figure 3C,D, approximately 78% to 86% of peptides identified employing the HAPs were also detected with the MG, demonstrating the overall reliability of peptide identification by our method. The taxonomic information for

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

the results with the MG database was obtained by annotating the identified peptides using Unipept, while the taxonomic information for the results with the HAPs was directly obtained from the species information in the database. As shown in Figure S6, the consistency of taxonomic information increased when the taxonomy levels became higher. In addition, the sample was also subjected to 16s rRNA sequencing, and the species composition obtained by 16s rRNA sequencing was compared with those of the aforementioned two methods. As shown in Figure 3E,F, the consistency at species level by the three different methods was low. Considering the high consistency at the peptide level while the low consistency at species level, the difference can be from low abundance species, which had low number of peptides identified. As shown in Figure S7, boxplots show that the common species by the three methods were more abundant in view of peptide number and peptide quantity than the unique species by the HAPs. Metagenomic sequencing is more sensitive to microbes with higher cell number, while metaproteomics is more sensitive to microbes with larger protein amount. Therefore, the two methods can show different sensitivities against low abundance species.

To demonstrate the practical value of our method, a comparative study was performed to investigate the functional and taxonomic features at both the metagenomic and metaproteomic levels. Metagenomic sequencing and metaproteomic results based on the HAPs-hyblibDIA pipeline were annotated using KEGG Orthology (KO) through the GhostKOALA Web site. Figure S8A shows the comparison of the KO functional annotations derived from metagenomics and metaproteomics in terms of metabolism, genetic information processing, and environmental information processing. Notably, the KO annotation results from metagenomics and metaproteomics were generally consistent. The relative abundance of each category by metaproteomics were slightly different from those by metagenomics, which highlighted the divergence between metagenomics and metaproteomics in revealing the functional potential of the gut microbiota.

The cladogram in Figure S8B depicts the relative abundance discrepancies at the metaproteomic level and metagenomic level, where abundance distinction exists at different taxa levels. Five main phyla of bacteria, *Verrucomicrobia*, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, and *Actinobacteria*, were found in the stool sample, consistent with previous studies.[38] At the species level, there were 33 species with higher abundance by metaproteomics and 24 species with higher abundance by metagenomics. *Bilophila wadsworthia*, *Parabacteroides distasonis*, *Clostridium citroniae*, and *Lachnospiraceae bacterium*, belonging to the phyla of *Bacteroidetes* and *Firmicutes*, exhibited a significantly low metaproteomic/metagenomic ratio, while *Akkermansia muciniphila* showed a significantly high metaproteomic/metagenomic ratio. The results indicated that within the same phylum, the relative abundance of different species by metagenomics and metaproteomics can display heterogeneity. The relative abundance by metagenomics is closely related to the cell copy of a species, while the relative abundance by metaproteomics shows the total protein amount of a species. Since different bacterial species can generate significantly different amounts of proteins, it is reasonable that the relative abundances by metagenomics and metaproteomics are different. It is expected that metagenomics and metaproteomics can show different sensitivities in the analysis of various species within a microbiota community. With this consideration, our HAPs-hyblibDIA pipeline can identify proteins from species not detected by metagenomics, which are also missed by the metaproteomics based on the metagenomic sequencing-derived database. In such a case, our method can complement the result obtained using metagenomic sequencing-derived database.

HAPs-hyblibDIA Analysis of Human Gut Microbiota Sample Spiked with Known Species. To further test the accuracy and sensitivity of our method for organism identification in real samples, 6 cultured strains from 6 species were spiked into another fecal sample for metaproteomic analysis, including DIA and fractionated DDA. The 6 species were not detected from the original fecal sample before the spike-in by metagenomic sequencing. The data sets of the fecal sample spiked with 6 species are from our previous work, and the details of the composition of the 6 species can be found in the original publication.[13] The DIA data were first searched against the HAPs database of 647 microbial species, wherein the 6 added species were included in the HAPs. We screened out 370 strains (Figure S9), among which all the 6 spiked species were included. HAPs-filtered database (HAPs) was then built from the 370 strains for hyblibDIA and directDIA analysis. As shown in Figure 4 and Supplementary Data 5,

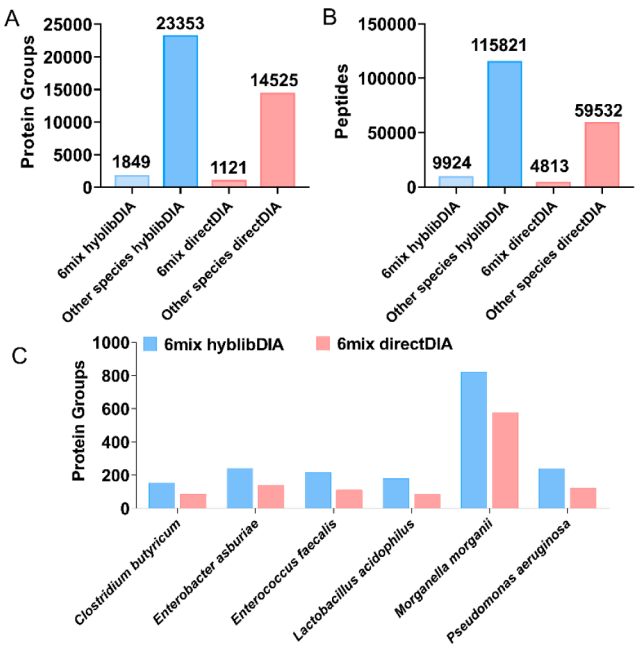


Figure 4. Analysis of the stool sample spiked with 6 species. The number of (A) protein groups and (B) peptides identified by hyblibDIA and directDIA using the HAPs-filtered database for the 6 target species and other species. (C) The number of protein groups for each of the 6 spiked species identified by hyblibDIA and directDIA using the HAPs-filtered database. The DIA data were obtained with three technique replicates, and the sums of identification results from the three replicates were reported for the comparison.

25 202 protein groups and 125 745 peptides were identified with the HAPs by hyblibDIA, of which 1849 proteins and 9924 peptides uniquely belonged to the 6 spiked species. Comparatively, 1121 protein groups and 4813 peptides were identified from the 6 spiked species with the HAPs by directDIA. The numbers of peptides and protein groups identified from each DIA run are shown in Table S7. This

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

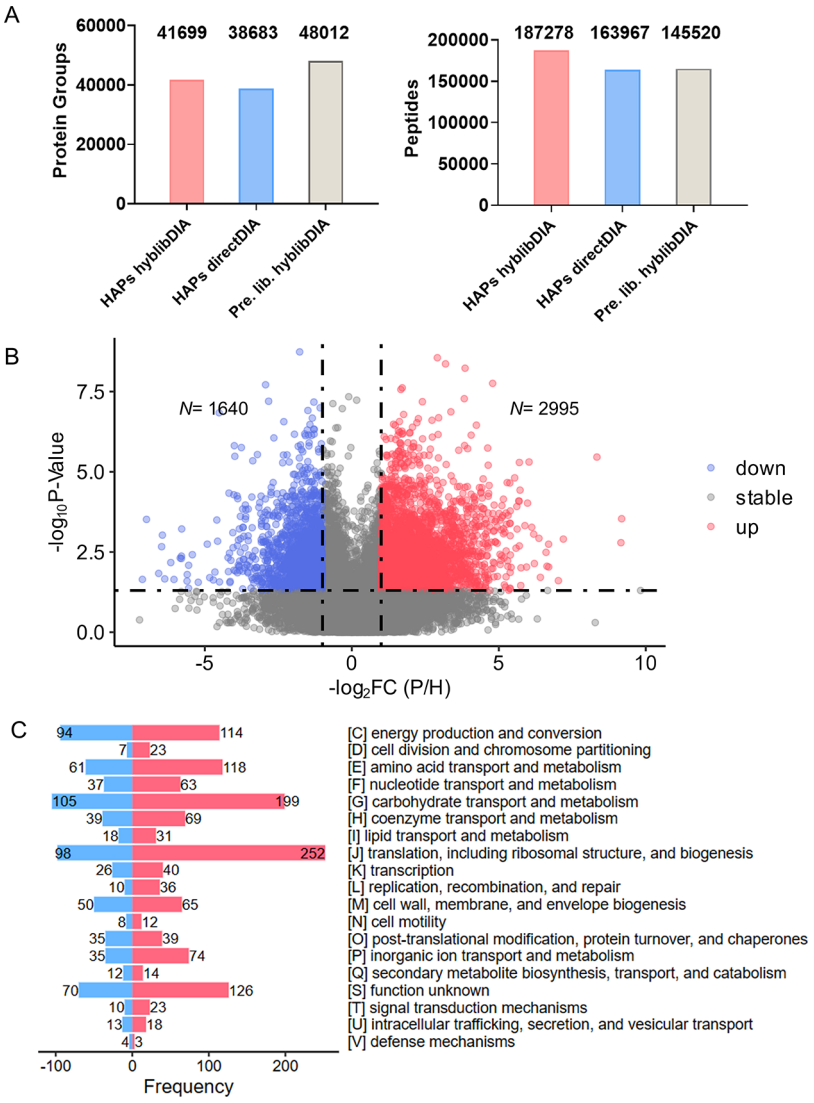


Figure 5. Analysis of a CRC gut microbial metaproteome data set. (A) The number of protein groups and peptides identified using the HAPs-filtered database and the database by previous work of Long et al.[12] (B) Volcano plot and (C) clusters of orthologous group (COG) categories of the differential proteins between the patients (P) and controls (H) obtained by hyblibDIA analysis using the HAPs-filtered database. Proteins with fold change (FC, P/H) of >2 and *p*-value of <0.05 were colored red, while those with FC < 0.5 and $p < 0.05$ were colored blue in (B). Proteins more abundant in patients were colored red, while those more abundant in controls were colored blue in (C).

result further demonstrated the sensitivity and effectiveness of our method for analyzing real gut microbial samples.

HAPs-hyblibDIA Analysis of a Cohort of Human Gut Microbiome Samples. To illustrate the utility of the HAPs-hyblibDIA pipeline in clinical microbiome research, we reanalyzed a colorectal cancer (CRC) gut microbial metaproteome data set previously published by Long et al.,[12] which includes the DIA and the fractioned DDA data of 14 CRC patients and 14 controls. With the DIA data, we screened out 405 organisms to construct a sample specific database (Figure S10). As shown in Figure 5 and Supplementary Data 6, 41 699 proteins and 187 278 peptides were obtained by hyblibDIA with the HAPs-filtered database (HAPs), while 38 683 proteins and 163 967 peptides were obtained by directDIA with the HAPs. In order to avoid discrepancies in identification numbers caused by inconsistent software versions, we reanalyzed the data with the spectral library reported by the previous study[12] using the updated Spectronaut (version 17), yielding 48 012 proteins and 145 520 peptides with the hyblibDIA. It is worth noting that although more proteins were obtained using the spectral library from the previous study, more peptides were obtained with the HAPs. This is consistent with the results of the simulated microbial community when comparing UP100 to the 100 HAPs, as well as the results of the stool sample when comparing the HAPs filtered databases to the public protein sequence databases of HMP stool_nr and JPGM.

To further test the consistency of the identification results employing the HAPs-filtered database and the database in previous work of Long et al.,[12] differential proteins between the CRC patients (P) and healthy crowds (H) were determined using the abundance FC and the statistical test. The Bonferroni method was conducted on the *p*-values given by the MS1-MS2-combined statistical test in Spectronaut for multiple testing corrections to obtain a conservative result. As shown in the volcano plot (Figure 5B and Figure S11A), 4635 and 5361 differential proteins with FC > 2 (or <0.5) and adjusted p-value of <0.05 were discovered by our method and

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX

the previously reported database, respectively (Supplementary Data 7). We then performed functional annotations of differential proteins using eggNOG.[39] The differential proteins obtained by our method were annotated into 19 clusters of orthologous group (COG) categories (Figure 5C). The differential proteins obtained using the previously reported database were annotated into 20 COG categories, wherein only one protein was annotated to the additional category, namely, chromatin structure and dynamics (category B) (Figure S11B). For most COG categories, more proteins were found with high relative abundance in the CRC patient group than in healthy control group by our method and the previously reported database. The results by our method and based on the previously reported database were in general consistent, highlighting the practical value of our method in biomedical research.

CONCLUSION

In summary, our study proposes a high-abundance protein-guided hybrid spectral library strategy for in-depth DIA metaproteomic analysis (HAPs-hyblibDIA). The accuracy and practical value of the method are demonstrated using simulated microbial communities, human fecal samples, human fecal samples spiked with known species, and clinical cohorts. We find that the method can provide a result comparable to the results with the metagenomic sequencing-derived database. This method provides a practical solution when metagenomic sequencing data are not available. Compared to other methods using the public proteome database, this method greatly reduces the calculation time and enhances the depth of peptide coverage. In addition, this study provides a dedicated HAPs database containing currently available gut microbial species, which can be used by the society of human gut microbiome research. Since our pipeline constructs HAPs-guided protein sequence databases based on public databases, it is inevitable that the method is limited to microbes with good coverage in public databases. However, we would expect the improvement of microbiota coverage in public databases with the development of microbiome, and the HAPs database can be easily extended to facilitate the application of metaproteomics in human microbiome research.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c03255.

Additional experimental methods for metaproteomic sample preparation, LC-MS/MS analysis, metaproteomic data sets, the construction of UP12 and UP8 databases, and metagenomic sequencing as well as data analysis; additional figures and tables on the analysis of the 8mix data set, a human fecal sample, a human fecal sample spiked with 6 species, and a cohort of CRC patients and controls (PDF)

Supplementary Data 1 showing details of 647 gut microbial species and the 100 species used for the HAPs, 100 HAPs, and UP100 databases (XLSX)

Supplementary Data 2 showing identification results of protein groups and peptides from the synthetic microbial community of 12 species using the 100 HAPs filtered database, the UP100, and the metagenomic sequencing-derived database (XLSX)

Supplementary Data 3 showing identification results of protein groups and peptides from the 8mix data set using the 100 HAPs filtered database, the UP100, and the 8-target species database (XLSX)

Supplementary Data 4 showing identification results of protein groups and peptides from the stool sample using the 647 species HAPs-filtered database, the combined UMGS HAPs-filtered database, the metagenomic sequencing-derived database, the JPGM database, and the HMP stool database (XLSX)

Supplementary Data 5 showing identification results of protein groups and peptides from the stool sample spiked with 6 species using the HAPs filtered database (XLSX)

Supplementary Data 6 showing identification results of protein groups and peptides from the CRC data sets using the HAPs filtered database and the database from previous work (XLSX)

Supplementary Data 7 showing the differential proteins between CRC patients and controls (XLSX)

Accession Codes

All raw MS data, spectral libraries, and search results generated in this study have been deposited to the ProteomeXchange via the iProX partner repository with accession numbers IPX0006766000 or PXD043890.

AUTHOR INFORMATION

Corresponding Author

Liang Qiao − *Department of Chemistry, and Shanghai Stomatological Hospital, Fudan University, Shanghai 200000, China;* orcid.org/0000-0002-6233-8459; Email: liang_qiao@fudan.edu.cn

Authors

Enhui Wu − *Department of Chemistry, and Shanghai Stomatological Hospital, Fudan University, Shanghai 200000, China*

Yi Yang − *Department of Chemistry, and Shanghai Stomatological Hospital, Fudan University, Shanghai 200000, China; ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 310000, China;* orcid.org/0000-0002-1330-9985

Jinzhi Zhao − *Department of Chemistry, and Shanghai Stomatological Hospital, Fudan University, Shanghai 200000, China*

Jianxujie Zheng − *Department of Chemistry, and Shanghai Stomatological Hospital, Fudan University, Shanghai 200000, China*

Xiaoqing Wang − *Shanghai Omicsolution Co., Ltd., Shanghai 200000, China*

Chengpin Shen − *Shanghai Omicsolution Co., Ltd., Shanghai 200000, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c03255

Author Contributions

E. Wu performed the majority of the experiments, analyzed the data, and wrote the first draft of the manuscript. Y. Yang assisted in the proteomic data analysis. J. Zhao assisted in the microbial and proteomic experiments. J. Zheng assisted in the proteomic data analysis and modified the manuscript. X. Wang

assisted in the proteomic data analysis. C. Shen assisted in the proteomic data analysis. L. Qiao designed and supervised all aspects of the study and finalized the manuscript.

Notes

Ethics Approval and Consent To Participate. The subjects gave their informed consent for using the biological material for research purposes. The study protocol was approved by the Ethics Committee of Fudan University and complied with all relevant laws and regulations of China.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (NSFC, Grants 22022401, 22374031, 22074022, and 21934001), Sino-German Center Mobility Programme (Grant M-0614), and the Ministry of Science and Technology of China, National Key R&D Program of China (Grants 2020YFF0304502 and 2022YFC2704300).

■ REFERENCES

(1) Malard, F.; Dore, J.; Gaugler, B.; Mohty, M. *Mucosal Immunol.* **2021**, *14*, 547−554.
(2) Bull, M. J.; Plummer, N. T. *Integr. Med. (Encinitas)* **2014**, *13*, 17−22.
(3) de Vos, W. M.; Tilg, H.; Van Hul, M.; Cani, P. D. *Gut* **2022**, *71*, 1020−1032.
(4) Lloyd-Price, J.; et al. *Nature* **2019**, *569*, 655−662.
(5) Mars, R. A.; Yang, Y.; Ward, T.; Houtti, M.; Priya, S.; Lekatz, H. R.; Tang, X.; Sun, Z.; Kalari, K. R.; Korem, T.; et al. *Cell* **2020**, *182*, 1460−1473.
(6) Tilg, H.; Adolph, T. E.; Gerner, R. R.; Moschen, A. R. *Cancer Cell* **2018**, *33*, 954−964.
(7) Zhao, L. *Nat. Rev. Microbiol.* **2013**, *11*, 639−647.
(8) Gurung, M.; Li, Z.; You, H.; Rodrigues, R.; Jump, D. B.; Morgun, A.; Shulzhenko, N. *EBioMedicine* **2020**, *51*, 102590.
(9) Fujimura, K. E.; Sitarik, A. R.; Havstad, S.; Lin, D. L.; Levan, S.; Fadrosh, D.; Panzer, A. R.; LaMere, B.; Rackaityte, E.; Lukacs, N. W.; Wegienka, G.; Boushey, H. A.; Ownby, D. R.; Zoratti, E. M.; Levin, A. M.; Johnson, C. C.; Lynch, S. V. *Nat. Med.* **2016**, *22*, 1187−1191.
(10) Wang, W. L.; Xu, S. Y.; Ren, Z. G.; Tao, L.; Jiang, J. W.; Zheng, S. S. *World J. Gastroenterol.* **2015**, *21*, 803−814.
(11) Das, P.; Babaei, P.; Nielsen, J. *BMC Genom.* **2019**, *20*, 208.
(12) Long, S.; Yang, Y.; Shen, C.; Wang, Y.; Deng, A.; Qin, Q.; Qiao, L. *npj Biofilms Microbiomes* **2020**, *6*, 14.
(13) Zhao, J.; Yang, Y.; Xu, H.; Zheng, J.; Shen, C.; Chen, T.; Wang, T.; Wang, B.; Yi, J.; Zhao, D.; Wu, E.; Qin, Q.; Xia, L.; Qiao, L. *npj Biofilms Microbiomes* **2023**, *9*, 4.
(14) Pietilä, S.; Suomi, T.; Elo, L. L. *ISME Commun.* **2022**, *2*, 51.
(15) Aakko, J.; Pietilä, S.; Suomi, T.; Mahmoudian, M.; Toivonen, R.; Kouvonen, P.; Rokka, A.; Hänninen, A.; Elo, L. L. *J. Proteome Res.* **2020**, *19*, 432−436.
(16) Tanca, A.; Palomba, A.; Pisanu, S.; Deligios, M.; Fraumene, C.; Manghina, V.; Pagnozzi, D.; Addis, M. F.; Uzzau, S. *Microbiome* **2014**, *2*, 49.
(17) Kolmeder, C. A.; Salojarvi, J.; Ritari, J.; de Been, M.; Raes, J.; Falony, G.; Vieira-Silva, S.; Kekkonen, R. A.; Corthals, G. L.; Palva, A.; Salonen, A.; de Vos, W. M. *PLoS One* **2016**, *11*, No. e0153294.
(18) Young, J. C.; Pan, C.; Adams, R. M.; Brooks, B.; Banfield, J. F.; Morowitz, M. J.; Hettich, R. L. *Proteomics* **2015**, *15*, 3463−3473.
(19) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell Proteomics* **2012**, *11*, O111.016717.
(20) Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B.; Aebersold, R. *Mol. Syst. Biol.* **2018**, *14*, No. e8126.
(21) Ting, Y. S.; Egertson, J. D.; Payne, S. H.; Kim, S.; MacLean, B.; Kall, L.; Aebersold, R.; Smith, R. D.; Noble, W. S.; MacCoss, M. J. *Mol. Cell Proteomics* **2015**, *14*, 2301−2307.
(22) Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2017**, *14*, 903−908.
(23) Schiebenhoefer, H.; Van Den Bossche, T.; Fuchs, S.; Renard, B. Y.; Muth, T.; Martens, L. *Expert Rev. Proteomics* **2019**, *16*, 375−390.
(24) Tanca, A.; Palomba, A.; Fraumene, C.; Pagnozzi, D.; Manghina, V.; Deligios, M.; Muth, T.; Rapp, E.; Martens, L.; Addis, M. F.; Uzzau, S. *Microbiome* **2016**, *4*, 51.
(25) Jouffret, V.; Miotello, G.; Culotta, K.; Ayrault, S.; Pible, O.; Armengaud, J. *Microbiome* **2021**, *9*, 195.
(26) Tanca, A.; Palomba, A.; Deligios, M.; Cubeddu, T.; Fraumene, C.; Biosa, G.; Pagnozzi, D.; Addis, M. F.; Uzzau, S. *PLoS One* **2013**, *8*, No. e82981.
(27) Muth, T.; Renard, B. Y.; Martens, L. *Expert Rev. Proteomics* **2016**, *13*, 757−769.
(28) Cantarel, B. L.; Erickson, A. R.; VerBerkmoes, N. C.; Erickson, B. K.; Carey, P. A.; Pan, C.; Shah, M.; Mongodin, E. F.; Jansson, J. K.; Fraser-Liggett, C. M.; Hettich, R. L. *PLoS One* **2011**, *6*, No. e27173.
(29) 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.
(30) Xiao, J.; Tanca, A.; Jia, B.; Yang, R.; Wang, B.; Zhang, Y.; Li, J. *J. Proteome Res.* **2018**, *17*, 1596−1605.
(31) Bassignani, A.; Plancade, S.; Berland, M.; Blein-Nicolas, M.; Guillot, A.; Chevret, D.; Moritz, C.; Huet, S.; Rizkalla, S.; Clement, K.; Dore, J.; Langella, O.; Juste, C. *J. Proteome Res.* **2021**, *20*, 1522−1534.
(32) Stamboulian, M.; Li, S. J.; Ye, Y. Z. *Microbiome* **2021**, *9*, 80.
(33) Lloyd-Price, J.; Mahurkar, A.; Rahnavard, G.; Crabtree, J.; Orvis, J.; Hall, A. B.; Brady, A.; Creasy, H. H.; McCracken, C.; Giglio, M. G.; McDonald, D.; Franzosa, E. A.; Knight, R.; White, O.; Huttenhower, C. *Nature* **2017**, *551*, 256.
(34) Forster, S. C.; Kumar, N.; Anonye, B. O.; Almeida, A.; Viciani, E.; Stares, M. D.; Dunn, M.; Mkandawire, T. T.; Zhu, A.; Shao, Y.; Pike, L. J.; Louie, T.; Browne, H. P.; Mitchell, A. L.; Neville, B. A.; Finn, R. D.; Lawley, T. D. *Nat. Biotechnol.* **2019**, *37*, 186−192.
(35) Nkamga, V. D.; Henrissat, B.; Drancourt, M. *Hum. Microbiome J.* **2017**, *3*, 1−8.
(36) Almeida, A.; Mitchell, A. L.; Boland, M.; Forster, S. C.; Gloor, G. B.; Tarkowska, A.; Lawley, T. D.; Finn, R. D. *Nature* **2019**, *568*, 499−504.
(37) Nishijima, S.; Suda, W.; Oshima, K.; Kim, S.-W.; Hirose, Y.; Morita, H.; Hattori, M. *DNA Res.* **2016**, *23*, 125−133.
(38) Thursby, E.; Juge, N. *Biochem. J.* **2017**, *474*, 1823−1836.
(39) Huerta-Cepas, J.; Szklarczyk, D.; Heller, D.; Hernández-Plaza, A.; Forslund, S. K.; Cook, H.; Mende, D. R.; Letunic, I.; Rattei, T.; Jensen, L. J.; von Mering, C.; Bork, P. *Nucleic Acids Res.* **2019**, *47*, D309−d314.

https://doi.org/10.1021/acs.analchem.3c03255
Anal. Chem. XXXX, XXX, XXX−XXX