Journal of proteome research

pubs.acs.org/jpr

Article

MetaLab-MAG: A Metaproteomic Data Analysis Platform for Genome-Level Characterization of Microbiomes from the Metagenome-Assembled Genomes Database

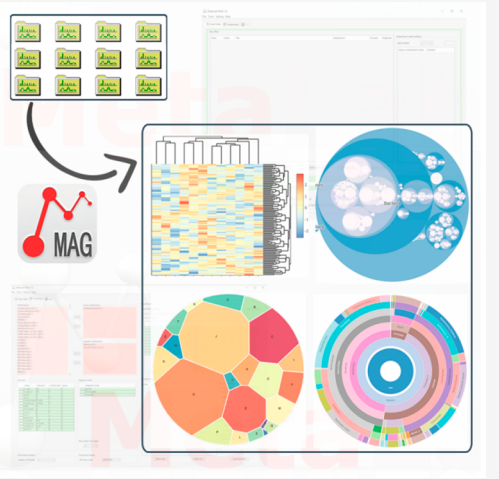
Kai Cheng, Zhibin Ning, Leyuan Li, Xu Zhang, Joeselle M. Serrana, Janice Mayne, and Daniel Figeys*

Cite This: J. Proteome Res. 2023, 22, 387−398

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The studies of microbial communities have drawn increased attention in various research fields such as agriculture, environment, and human health. Recently, metaproteomics has become a powerful tool to interpret the roles of the community members by investigating the expressed proteins of the microbes. However, analyzing the metaproteomic data sets at genome resolution is still challenging because of the lack of efficient bioinformatics tools. Here we develop MetaLab-MAG, a specially designed tool for the characterization of microbiomes from metagenome-assembled genomes databases. MetaLab-MAG was evaluated by analyzing various human gut microbiota data sets and performed comparably or better than searching the gene catalog protein database directly. MetaLab-MAG can quantify the genome-level microbiota compositions and supports both label-free and isobaric labeling-based quantification strategies. MetaLab-MAG removes the obstacles of metaproteomic data analysis and provides the researchers with in-depth and comprehensive information from the microbiomes.



KEYWORDS: *metaproteomics, human gut microbiome, isobaric labeling, label free, metagenome-assembled genomes, data analysis, software*

INTRODUCTION

The microbiome is the community of microorganisms inhabiting various environments. Research on the microbiome developed rapidly in recent years and in particular, the role of the microbiome in human health, e.g., the human gut microbiota has been associated with type 2 diabetes,[1] inflammatory bowel diseases,[2,3] cardiovascular disease,[4] and other diseases. Currently, multiomics approaches including metagenomics, metatranscriptomics, metaproteomics, and metabolomics provide comprehensive information on microbiomes.[5] Constructing metagenome-assembled genomes (MAGs) from metagenome data reveals the composition and predicted functional potential of the complex community.[6,7] In contrast, mass spectrometry (MS)-based metaproteomic strategies focus on the characterization of the expressed proteins from the microbiome. Metaproteomics provides dynamic insights into the functional, enzymatic, and pathway changes occurring at the individual microbes and at the systems microbiome level.[3,8,9]

The experimental workflows of metaproteomics have similarities to the conventional approaches in proteomics studies. Briefly, proteins are extracted from the samples and digested into peptides, then subjected to high-performance liquid chromatography-electrospray ionization tandem mass spectrometry (MS/MS) analysis. The generated MS data sets are searched against a theoretical protein sequence database to determine the peptides and proteins. Taxonomic analysis and functional annotations are performed based on the identified peptides and proteins using specialized metaproteomics postanalysis tools.[10] In general, proteomics studies focus on single species. In contrast, metaproteomic studies have to deal with complex microbiota often with hundreds of different species, each having up to a few different strains. Moreover, this is further complicated by significant differences in microbiome compositions between individuals. This complexity brings significant obstacles to the data analysis of metaproteomics studies. In particular, the microbiota protein database is derived from a gene catalog consisting of nonredundant genes, which could come in two forms: (1) metagenomics/metatranscriptomic sequencing of the sample of interest, which provides a

Special Issue: Software Tools and Resources 2023

Received: September 19, 2022
Published: December 12, 2022

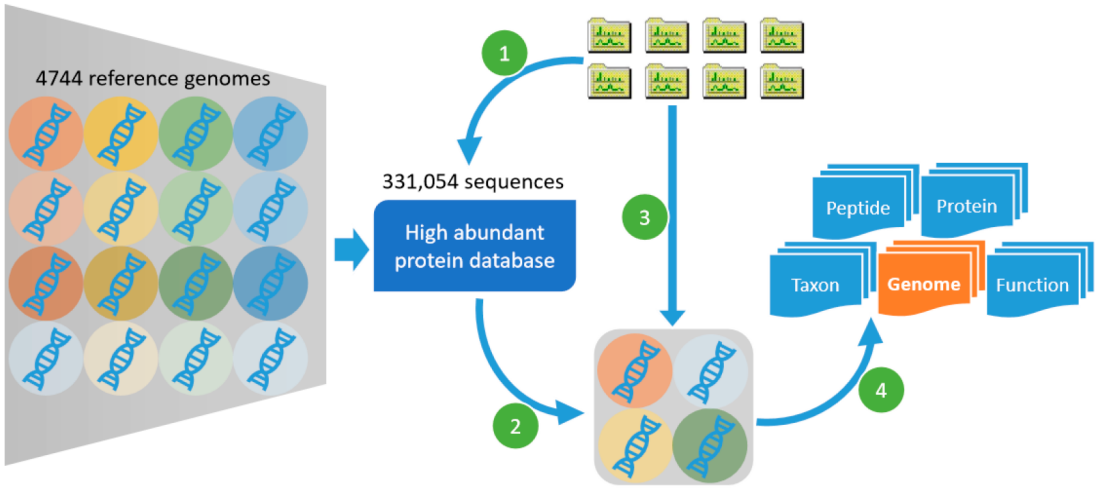


Figure 1. Workflow of MetaLab-MAG. (1) raw files were searched against the HAP database consisting of 331,054 proteins; (2) from the search result the possible components of the samples were determined, and proteins from the corresponding genomes were added to compiled to create a refined database; (3) the raws files were searched against the refined database; (4) data tables in peptide, protein, taxa, genome, and function levels were obtained.

sample-specific database, and (2) integrated public gene catalogs. The quality of individualized protein sequence databases derived from sample-specific metagenomics/metatranscriptomic sequencing affects the identification of peptides. If a gene is missed in the gene catalog database the corresponding protein cannot be detected. It can be challenging to obtain a high-quality gene catalog database for researchers who lack the resources and/or experience for deep sequencing. An alternative approach is the use of a public microbiome gene catalog database. These databases are always comprehensive but are very large. For example, a commonly used human gut microbiome database, the Integrated Gene Catalog (IGC) contains 9,879,896 sequences;[11] by contrast, the *Homo sapiens* database in Uniprot contains 20,361 sequences. A large database produces enormous search space, which will cause an extremely long data processing time and a low identification rate. We have developed the MetaLab[12,13] software to solve the problem of identifying peptides from huge gene catalog databases such as the IGC. An iterative searching strategy[14] is implemented in MetaLab, in which a first search is performed against the gene catalog database to generate a refined database. Then a second search against the refined database is performed for peptide identification. Before the first search, a spectra cluster step is performed to select the delegate spectra which will be used for the search, which can greatly decrease the processing time. We also perform an open search strategy in the second search to improve peptide identifications. Although specially designed metaproteomics data analysis software is available,[12−16] many studies still use conventional proteomic tools. A recent metaproteomic benchmark study utilized four bioinformatic workflows for data analysis and three of them were proteomic tools.[17] Their result showed that the proteomic toolset SearchGUI[18]/PeptideShaker[19] performed best, but the highest spectra identification rate was only 34.79% for fecal samples.

Another challenge of metaproteomics data analysis is the taxonomic assignment, which is usually performed based on the identified peptides/proteins using specialized metaproteomics postanalysis tools. The taxonomic information is retrieved from a peptide-to-taxon or protein-to-taxon database, which is constructed from a repository of annotated proteins such as NCBI and Uniprot. A drawback of this method is that it is a generic solution but not specific to the target microbiome. For example, a Uniprot-based peptide-to-taxon database will contain information from all the species. A peptide will likely be matched to multiple proteins from various species. For one peptide multiple taxonomic lineages will be found in the database. In this case, a commonly used algorithm named the lowest common ancestor algorithm (LCA) will be used, which results in a situation where the taxonomic information is usually obtained at a higher phylogenic level.

To address these challenges, we developed a complete metaproteomics data processing platform, MetaLab-MAG (metagenome-assembled genomes), which uses the publicly available MAGs as the resource for peptide/protein identification and genome-level taxonomic/functional annotation. Currently, four MAGs databases in MGnify[20] including Cow Rumen v1.0,[21] Unified Human Gastrointestinal Genome (UHGG) v2.0,[22] Human Oral v1.0, and Marine v1.0 were supported (https://www.ebi.ac.uk/metagenomics/browse/genomes). In this paper, we evaluated the performance of MetaLab-MAG by analyzing human gut microbiota data sets with the UHGG 2.0 database consisting of 4,744 species-level genomes. In these cases, the taxonomic identification was restricted to the MAGs of the human gut microbiome, which is more precise than acquired from NCBI or Uniprot whole databases. To accelerate and improve the data processing, a two-step database search strategy was utilized. The first search of the high-abundant protein (HAP) database is used to generate the sample-specific MAGs database,[23] and the second search against the refined MAGs databases is used to identify peptides and proteins. The results from multiple data sets demonstrated that the MS/MS identification rates were close to or even exceeded the values obtained by searching the customized or public gene catalog database from the same samples. Analyzing the same fecal data set of the above benchmark study, the average identification rate reached 52%. The reliability of the results had been confirmed by comparing the data from searching the reference or gene catalog database. MetaLab-MAG was used for qualitative and quantitative analysis and supported both label-free and isobaric labeling-based quantification. MetaLab-MAG

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

is free for academic use and can be downloaded from https://imetalab.ca/.

METHODS

The Workflow of MetaLab-MAG

The workflow of MetaLab-MAG is shown in Figure 1. In the first step, each raw file from microbiota analysis was searched against a HAP database, which was composed of all the ribosomal and elongation factor proteins from the MAGs. MetaLab-MAG utilized pFind[24] as the database search engine for the two-step search. A target-decoy strategy was adopted in both the two search steps for the assessment of the false discovery rate (FDR). Then, a peptide-spectrum-match (PSM) list with FDR < 1% was generated for each raw file. A PSM may have multiple possibilities for which genome it came from, so the Occam's razor principle was utilized here to keep a minimized genome list that can explain the sources of all the PSMs. The generated genome list was used to compose the sample-specific MAGs database by adding all the proteins from the selected genomes to the refined database. The next step was searching each raw file against the sample-specific database. The open search strategy has been demonstrated to be a valuable method not only for identifying modified peptides but also for improving the overall identification rate.[24−26] We have utilized the open search method in metaproteomics data analysis and the identification at both the peptide and taxa levels was significantly improved.[13] Therefore, the open search strategy was used in this step. The target-decoy strategy was used to evaluate the FDR, similarly. After the peptide and proteins were identified, the results from each raw file were combined. A minimized protein group list that can explain the attribution of all peptides was kept. In the peptide list, for each peptide multiple source proteins were listed. Furthermore, the protein with the highest pFind score was selected as the razor protein for this peptide. In the quantification part, the intensity of the peptide only contributed to the razor protein and the corresponding genome. Both label-free and isobaric labeling strategies were supported in this step. Taxonomic annotations based on each genome and functional annotations based on each protein were available for all of the four MGnify MAGs databases. This information could be retrieved after the quantification of peptides, proteins, and genomes was finished. Then according to the quantitative information, data tables about the peptides, proteins, genomes, taxa, and functions were exported. Visualized reports of the result were also available in the form of web pages. If the metadata was provided, multivariate statistical analyses such as a principal component analysis score plot and a hierarchical clustering heatmap were generated as well in the report.

The Construction of the MAGs Database

Currently, four MAGs databases were available in MGnify, including Cow Rumen v1.0, Unified Human Gastrointestinal Genome (UHGG) v2.0, Human Oral v1.0, and Marine v1.0. Users can select and download what they need in MetaLab-MAG. These public MAGs databases are constructed based on large-scale metagenomic sequencing data from the target microbiomes. The enormous amount of sequencing data is clustered into genomes representing the composition of the microbiomes. In other words, all the genes/proteins in the MAGs database belong to specific genomes. Proteins identified from the MAGs could be readily linked to the corresponding genomes. Let us take the UHGG for example. This database included 4,744 genomes derived from the human gut microbiome. For each genome, the protein sequence database, and taxonomic and functional information were available. Based on the UHGG database, we constructed an integrated built-in database in MetaLab-MAG for peptide/protein identification, taxonomy analysis, and functional annotation. This database contained four components: (1) the original protein sequence databases of the 4,744 genomes; (2) the functional annotation information generated by eggNOG and taxonomy information based on Genome Taxonomy Database r202 (GTDB, https://gtdb.ecogenomic.org/); (3) a HAP database including 331,054 sequences, which was created by extracting all the ribosomal and elongation factor proteins from the original database; (4) a host protein database, which was a *Homo sapiens* protein database downloaded from Uniprot (https://www.uniprot.org/). The taxonomy information provided by the UHGG project was based on GTDB. We added the corresponding NCBI taxon information for the genomes to the taxonomic database. The taxonomy and function database was packed within the MetaLab-MAG software and the other databases can be downloaded from the MetaLab-MAG software interface, no manual configuration step was required.

Sample Description and LC-MS/MS Acquisition

The data set of the bacterial strain samples was obtained from a recent work of our lab, in which we analyzed proteomic profiles of these strains cultured with or without added sugars (glucose, sucrose, and kestose) in a glucose-free Yeast Casitone Fatty Acids (YCFA) broth. Samples were analyzed on an Orbitrap Exploris 480 mass spectrometer. A 60 min gradient of 5 to 35% (v/v) buffer B at a 300 μL/min flow rate was used to separate the peptides on a tip column (75 μm inner diameter ×10 cm) packed with reverse phase beads (3 μm/120 Å ReproSil-Pur C18 resin, Dr. Maisch GmbH, Ammerbuch, Germany). Buffer A was 0.1% formic acid (v/v), and buffer B was 0.1% formic acid with 80% acetonitrile (v/v). The MS full scan ranging from 350 to 1200 m/z was recorded in profile mode with the resolution of 60,000. Data-dependent MS/MS scan was performed with the 15 most intense ions with the resolution of 15,000. Dynamic exclusion was enabled for duration of 30 s with a repeat count of one. The data set of the four intestinal aspirate samples obtained by high-pH reversed-phase fractionation and Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, USA) was obtained from our previous work, with a detailed experimental procedure described before.[27] The data set of the TMT labeling samples was obtained from our previous work with a detailed experimental precedure described before.[28]

Metagenomics Gene Catalog from Real Human Gut Microbiota Samples

Illumina paired-end reads from the microbiomes of four human gut aspirate samples, i.e., HM454, HM455, HM466, and HM503, obtained from another study[14] were used to construct a gene catalog. The raw sequence data are accessible in NCBI under the Sequence Read Archive (SRA) with the accession number SRP068619. Raw reads were filtered to remove the adapter and low-quality sequences with the trimming and quality filtering step of the MOCAT pipeline.[29] Reads with human origin were also filtered out using the SOAP2[30] package by mapping the sequences against the human genome database (hg19). The high-quality reads were assembled into contigs using MEGAHIT v1.2.9[31] with default options. The sequence data of each sample were assembled individually. The assembled contigs from the four samples were then used for gene prediction with the prodigal v2.6.3[32] software. The contigs were translated

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

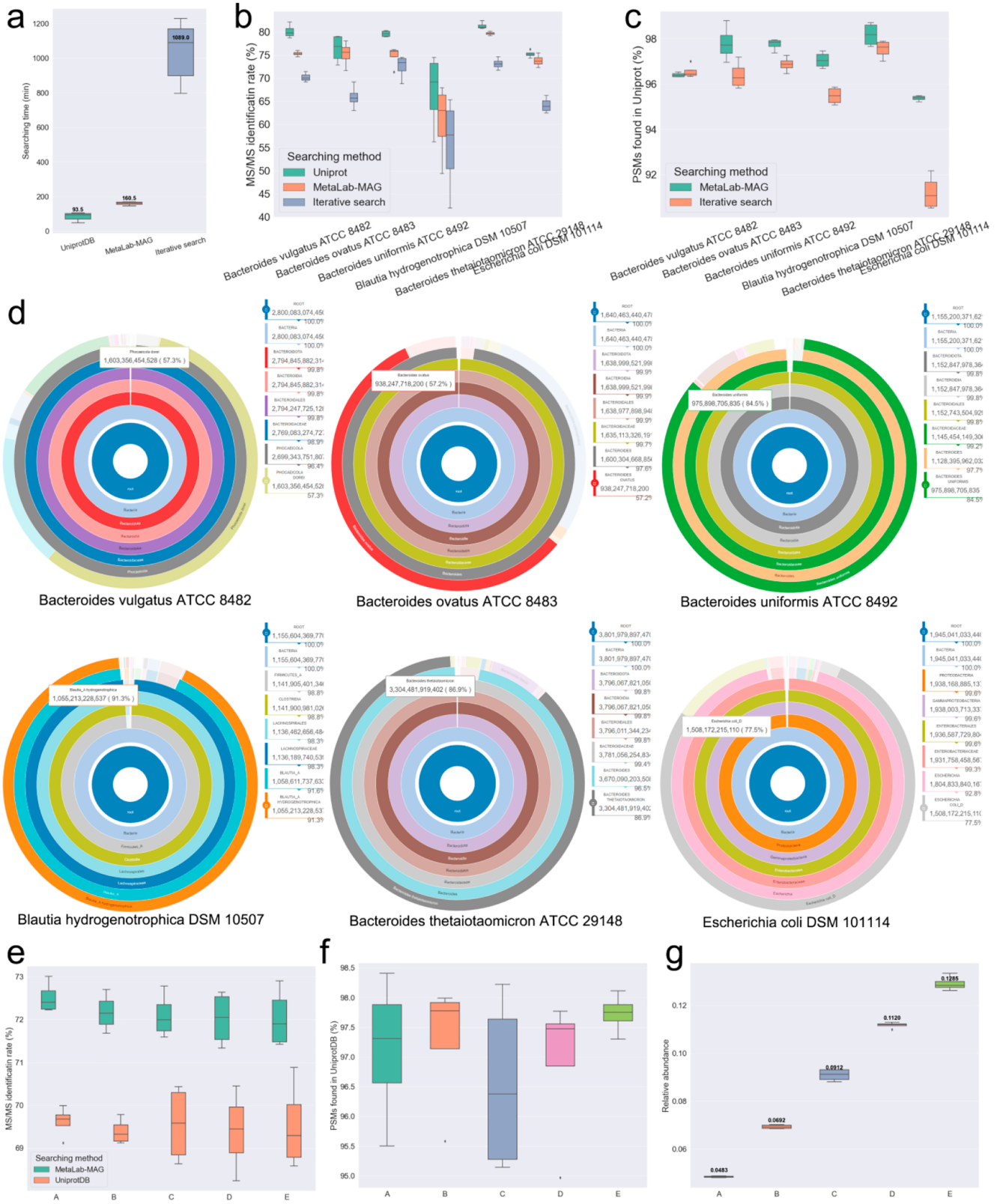


Figure 2. Performance of MetaLab-MAG in the analysis of the single species data sets. (a) The search times of the three methods including searching against the UniprotDB, MetaLab-MAG workflow, and the conventional iterative searching strategy. (b) The MS/MS identification rates of the three methods. (c) The proportions of the PSMs can be found in the UniprotDB. (d) The compositions and relative abundances of taxa in the six single species data sets. (e) The MS/MS identification rates of the human-*E. coli* samples. The axis was the different amounts of *E. coli* spiked into the samples: A (2%), B (3%), C (4%), D (5%), and E (6%). (f) The proportions of the PSMs from bacteria can be found in the *E. coli* UniprotDB. (g) the relative abundance of *E. coli* determined in the samples.

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

into amino acid sequences using the anonymous gene prediction mode (prodigal -p meta) and default parameters. The protein-coding gene sequences of the four samples were compiled into FASTA files and used as the metagenome-inferred protein database for benchmarking the MetaLab-MAG pipeline. For the taxonomic annotation, the amino acid sequences of the proteins in the catalog were searched against the UHGG database with DIAMOND v2.0.15[33] using the blastp command with default settings. To estimate the abundance of each predicted protein sequence, the high-quality reads were first aligned to the assembled contigs with minimap2 v2.24-r1122,[34] and the generated BAM files were used to create the read count matrix of the protein sequences using featureCounts v2.0.1.[35]

Metaproteomics Data Analysis

All the MetaLab-MAG workflows in this paper utilized uniform parameter settings. The protein databases were downloaded from http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0/. The host database was a reviewed human protein database from UniProtKB containing 20,387 sequences. In the pFind open search workflow, modifications were also accepted. Therefore, cysteine Carbamidomethylation was set as the fixed modification, methionine oxidation, and protein N-term acetylation were set as the variable modification. The enzyme was set as trypsin. Fully specific digestion mode was used and up to two miss cleavages were allowed. The mass tolerances of precursor and fragment ions were 10 and 20 ppm, respectively. The false discovery rate was less than 1% at the PSM level. When using the iterative searching strategy in MetaLab 2.3.0, the IGC human gut microbiome database was used.[10] Other database searches in this paper were all performed by pFind directly. The databases used for the searches were described in the corresponding sections. Other parameters were all consistent with these used in the MetaLab-MAG workflow.

RESULTS

Sensitivity and Accuracy of MetaLab-MAG in Identifying Single Bacterial Species

We first evaluated the performance of MetaLab-MAG by analyzing samples with known taxonomic compositions. Here we used a data set consisting of six sole species samples (Supplementary Table 1) to assess the sample-specific database generation, and the qualitative/quantitative identification of peptides, proteins, and taxa. As a comparison, we also analyzed this data set by searching against the Uniprot single bacteria database and the IGC database using the conventional iterative searching strategy in MetaLab's previous version (2.3.0).

We found that the processing time was significantly decreased compared with searching the IGC database in MetaLab 2.3.0 (Figure 2a). The time cost for peptide and protein identification was only about 15% of the conventional iterative searching method. The reason was although the total size of the MAGs database was big, there was no need to search the whole MAGs database directly. The first search was against a HAP database and the second search was against the selected genomes. A target-decoy search could be performed in both of the two steps which will benefit the FDR control. Then we investigated if the correct species could be selected from the MAGs database. The average MS/MS identification rate was 6.7% at FDR < 1% by searching the HAP database. We matched the identified PSMs to the corresponding Uniprot single species database and found that 94% of PSMs matched. According to these identifications, the identified genomes were selected to generate the sample-specific database. On average, 45 genomes and 117,375 proteins were contained in one database. This result showed that the correct species were identified and their proteins were selected successfully to form a proper-sized sample-specific database.

The peptide/protein identifications were obtained by searching the sample-specific MAG database, yielding MS/MS identification rates that were similar to those by searching against the Uniprot single species database (Figure 2b) and higher than those obtained from the conventional iterative search method against the IGC database. Searching the Uniprot single-species database was the ideal method for peptide identification from single-species samples. However, in the MetaLab-MAG workflow, only a slight decrease in the identification rates of about 5% was observed. Then we tested if the PSMs were identified correctly from the expected species. The Uniprot single species database was *in silico* digested, and we matched the PSMs to the theoretical peptide sequences. It was found that 97.1% of fully tryptic PSMs were matched with the corresponding Uniprot single species database (Figure 2c). This result showed that although multiple genomes were collected to create the sample-specific database, most of the obtained PSMs and peptides were from the correct species through the target-decoy search. The relative abundance of taxa identified from the samples was shown in Figure 2d. It was observed that the taxonomic identification was accurate at the genus level. For some of the samples, the relative abundance of correct species was relatively low. This was because part of the proteins was attributed to very similar species, as mentioned above, the peptide identification was correct. Overall, the quantitative results illustrated the same conclusion that the MetaLab-MAG workflow successfully identified the target species from the single species samples.

These results demonstrated that MetaLab-MAG can easily identify high-abundant bacteria. However, the human gut contains a highly diverse microbial community with a significant number of microorganisms having very low abundance. Hence, we tested whether the low abundant bacteria could be identified from a spike-in sample, i.e., a data set of human cell samples mixed with *Escherichia coli* at different concentrations (2, 3, 4, 5, or 6%).[36] MetaLab-MAG provided a workflow enabling the identification of host proteins simultaneously by concatenating a human protein database. By this strategy, we found that the MS/MS identification rate reached 72% and was higher than searching a *Homo sapiens-E. coli* combined database (Figure 2e, Supplementary Table 2). The proportion of non-*Homo sapiens* PSMs from *E. coli* was about 97% (Figure 2f). The relative quantitative results also fitted well with the expected values 1-, 1.5-, 2-, 2.5-, or 3-fold (Figure 2g). This result shows that relatively low abundant components could also be identified and quantified correctly using MetaLab-MAG.

From these results, we demonstrated that MetaLab-MAG is an efficient and reliable tool for metaproteomic data analysis. First, MetaLab-MAG outperformed the conventional iterative search strategy in sensitivity, accuracy, and searching speed. Second, compared with the ideal matched Uniprot database search, the loss of the MS/MS identification rate was less than 5%. Considering the database used in MetaLab-MAG was a generic MAGs database consisting of 4,744 genomes, identifying single species samples with high sensitivity and accuracy was encouraging.

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

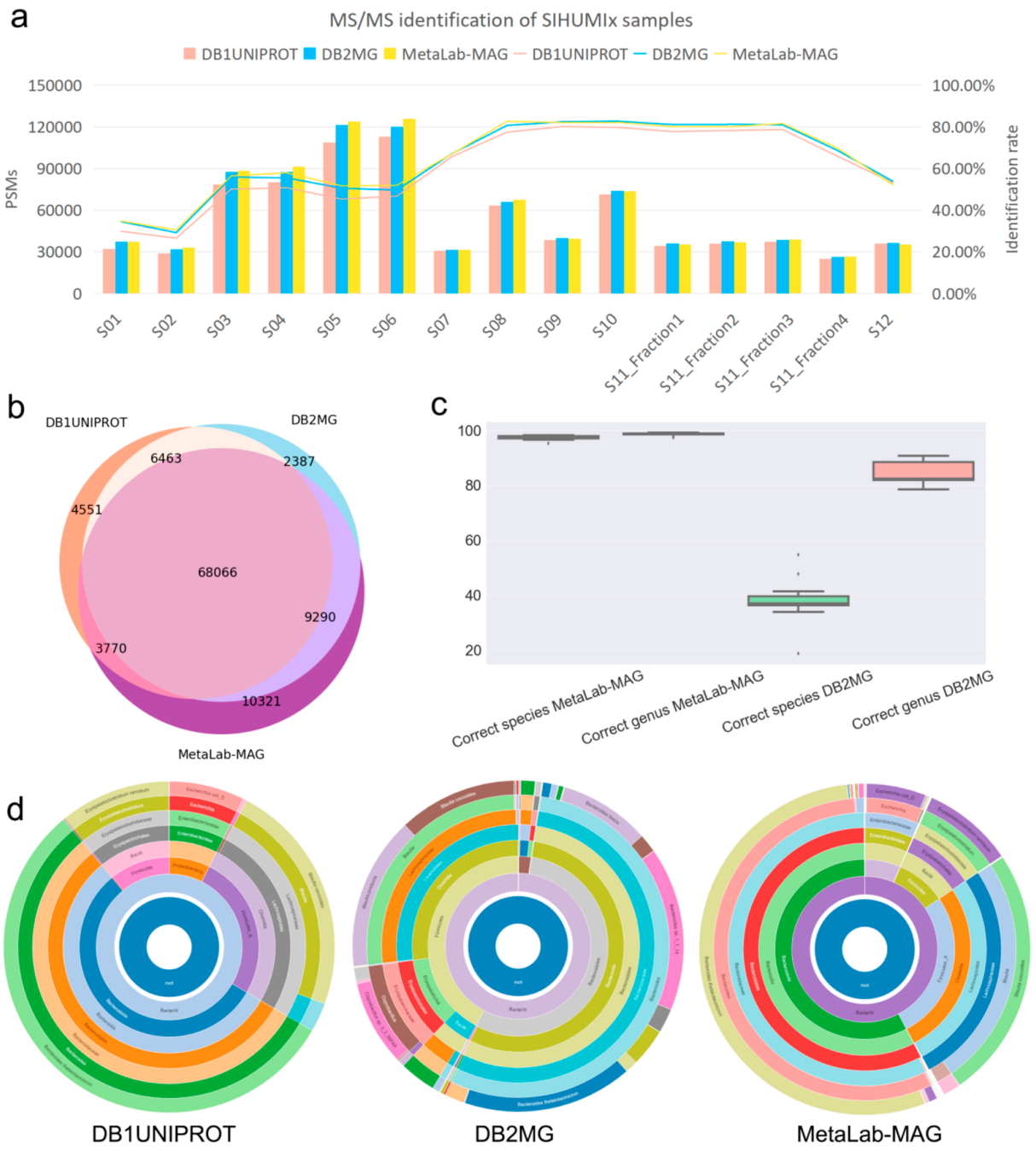


Figure 3. Use of MetaLab-MAG for the analysis of SIHUMIx samples. (a) The MS/MS identification rates and numbers of identified PSMs of the three methods including searching against the DB1UNIPROT, DB2MG, and MetaLab-MAG workflow. (b) The identified peptides shared a great common part among the three methods. (c) The proportions of the PSMs with correct species identifications and correct genus identifications. (d) The compositions and relative abundances of taxa calculated from the DB1UNIPROT, DB2MG, and MetaLab-MAG workflows.

Evaluation of MetaLab-MAG Using Synthetic Microbial Community

We then used MetaLab-MAG to analyze a data set derived from a synthetic microbial community named SIHUMIx. The SIHUMIx sample was composed of eight types of bacteria found in the human gut microbiome.[37] We determined that all eight species had corresponding genomes in the UHGG database. There were 139 genomes in UHGG that were found from the same genera of the eight microbes (Supplementary Table 3). In the following analysis, we defined the PSMs/peptides from the genomes of the correct species as "correct species identifications" and PSMs/peptides from the 139 genomes of the correct genus as "correct genus identifications". We analyzed 15 raw files by searching against the HAP database, and 67,748 PSMs were obtained at FDR < 1% with the MS/MS identification rate at 4.4%. The correct genus identifications accounted for 93.5% of the PSMs and 90.8% of the peptide. Two microbes (*Clostridium butyricum* DSMZ 10702 and *Lactobacillus plantarum* DSMZ 20174) were not found in all the 15 raw files and *Lactobacillus plantarum* DSMZ 20174 was only found in 7 raw files. Actually, these three microbes cannot be found from the metagenomics database (termed DB2MG below) based on

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

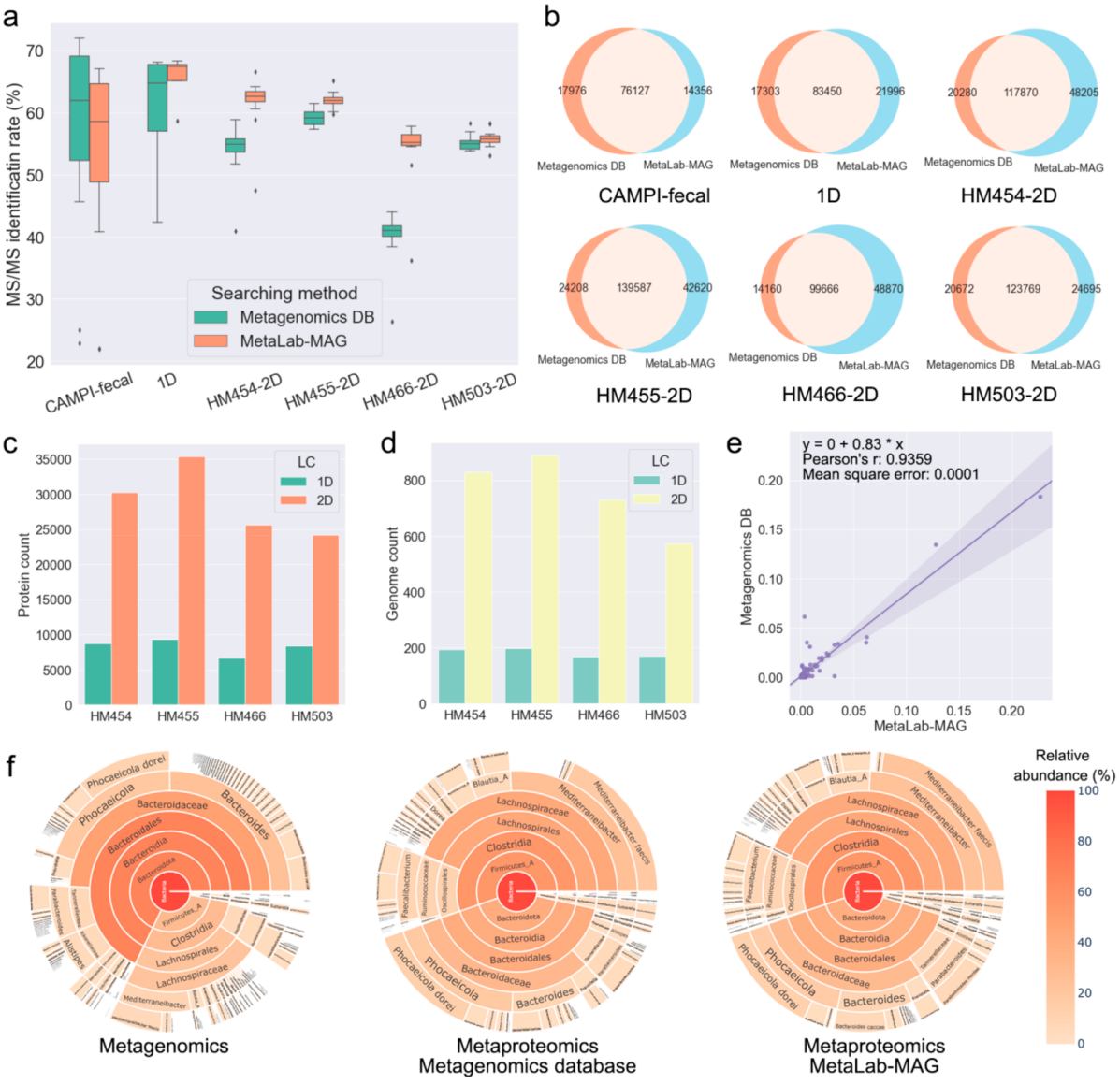


Figure 4. Analysis of real human gut microbiota samples. (a) The MS/MS identification rates of the data sets. CAMPI-fecal: fecal samples from the CAMPI projects, 15 raw files were used here; 1D: 1D-LC-MS/MS analysis of the four samples (i.e., HM454, HM455, HM466, and HM503); the 2D data sets were 2D-LC-MS/MS analysis of the above four samples. "Metagenomics DB" represents that the peptide identifications were obtained by searching against the metagenomic sequencing database directly. (b) The Venn figures showed the overlaps of the peptide identifications from the samples between MetaLab-MAG and metagenomic database searching method. (c) the protein and (d) the genome counts identified from 1D and 2D experiments of HM454, HM455, HM466, and HM503 samples. (e) The correlation of the relative abundance of genomes determined by MetaLab-MAG and metagenomic database searching method from HM454-2D. (f) The relative abundance of the taxa calculated by metagenomic read counts, metagenomic database searching method and MetaLab-MAG from HM454-2D.

the assembled contigs, either, suggesting their relative abundances were very low in the samples (Supplementary Table 3).[17] Sample-specific databases were generated based on these identifications. On average, 82 genomes and 212,644 proteins were contained in one sample-specific database. These results show that by searching the HAP database, the high-abundance components of the samples were effectively identified and could be retrieved to form the sample-specific database.

The overall identification rate using the above generated sample-specific MAG database reached 58%, which was slightly higher than searching against the DB2MG (57.1%) and the selected eight species database from Uniprot (53.4%) (Figure 3a). As well, most of the peptides identified by these three methods were the same (Figure 3b, **Data S1**, **S2**, **S3**). This suggested valid peptide identifications from searching against the three different types of databases. Theoretically, the most suitable database for metaproteomics data analysis was a reference database with proper size or a sequencing read-based database. This result proved that optimal performance was achieved using the MetaLab-MAG strategy.

Next, we investigated the results at the taxon level. It was found that 1,041,817 PSMs were identified from bacteria, at the same time 1,018,307 (97.7%) were from correct species and 1,029,947 (98.9%) were from correct genus (Figure 3c, Supplementary Table 4). By contrast, only 393,345 (38.1%) and 869,219 (84.1%) PSMs were identified from the correct species and correct genus, respectively, by searching the

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

DB2MG. The huge difference did not come from the peptide level, but the taxa level. The taxonomic annotations of the metagenomics database were against the NCBI database on the protein level, which was not perfectly matched to the theoretic compositions. For example, for the DB2MG, it was found that 382 proteins were from the species *Erysipelatoclostridium ramosum* (correct species in SHUMIx) but 1,012 proteins were from the species *Coprobacillus sp. 3_3_56FAA* (incorrect species in SHUMIx). As a result, for the DB2MG, 84,981 PSMs were from *Coprobacillus sp. 3_3_56FAA* and only 9,425 PSMs were from the correct species *Erysipelatoclostridium ramosum*. The 84,981 PSMs matched to *Coprobacillus sp. 3_3_56FAA* corresponded to 6,833 unique peptide sequences, among which 6,217 peptide sequences were also identified from the MetaLab-MAG workflow and 6,164 (99.1%) of them were from the genome "MGYG000001400" (*Erysipelatoclostridium ramosum*). This result showed a major advantage of using the MAGs database for the metaproteomics data analysis. In the MAGs database, the taxonomic annotations were generated at the genome level; moreover, these genomes were constructed from the target microbiome. As a comparison, the taxonomic annotation of the metagenomics database was performed at the protein level against the NCBI nonredundant database, which was more general and less sensitive. From the quantitative result, we also observed that the relative abundance of the taxa was more similar between the Uniprot database search results and the MetaLab-MAG results (Figure 3d).

Through the analysis of the SIHUMIx data set, we found that MetaLab-MAG fulfilled our requirement to successfully identify and quantify the components of a simplified human gut microbiome. Similar MS/MS identification rates and better performance of taxonomy analysis were achieved compared with the workflows searching against reference and metagenomic sequencing databases.

Application of Label-Free Metaproteomics of Human Gut Microbiota

We then explored the application of MetaLab-MAG for the analysis of individual human gut microbiota samples. The first data set was from fecal samples analyzed by the Critical Assessment of MetaProteome Investigation (CAMPI) project with available metagenomics and metatranscriptomics sequencing data.[17] Another human metaproteomic data set included four intestinal aspirate samples collected from pediatric individuals.[27] For the deeper measurement of these samples, in addition to the conventional liquid chromatographic separation (**1D** in Figure 4a), a two-dimensional separation was adopted. High-pH reversed-phase fractionation was used first and then each fractionation was subjected to LC-MS/MS analysis. Metagenomic sequencing data were available for this data set. The taxonomic annotation was performed by DIAMOND[38] against the UHGG 2.0 database. Here we analyzed the CAMPI and the intestinal aspirate data sets using MetaLab-MAG (**Data S4−S9**). Meanwhile, these two data sets were also searched against the corresponding multiomics databases (**Data S10**, **S11**). Figure 4a shows the MS/MS identification rates of the samples, which illustrated that for the analysis of real human gut microbiome samples, the performance of MetaLab-MAG was still similar to searching the metagenomics database. At the same time, the identification rates were 21.3% to 66.3%, which were much higher than that obtained by the SearchGUI/PeptideShaker workflow in the CAMPI study (12% to 34.8%)[17] (**Data S4**). It was worth noting that MetaLab-MAG outperformed the metagenomic database searching result from the intestinal aspirate samples. We compared the identified peptides and found that for each data set, the identified peptide sequences from the two strategies had significant overlap. (Figure 4b). MetaLab-MAG analysis of the CAMPI fecal samples identified 110,993 peptides, 17,235 protein groups, and 678 genomes. Higher MS/MS identification rates and more peptide identifications were obtained by MetaLab-MAG workflow in the analysis of the intestinal aspirate samples. The numbers of identified proteins and genomes from the four samples were shown in Figure 4c,d. Nearly 10,000 proteins and 200 genomes could be obtained from a single raw file of the 1D-separation experiments and the 2D-separation yielded about triple the number of proteins and four times genome identifications.

Since the performances of the multiomics database searching strategy and the MetaLab-MAG workflow were similar in identification, we compared the quantitative results obtained by the two methods. As a representative, Figure 4e showed a good correlation between the relative abundances of genomes quantified by the two methods from HM454-2D samples. Finally, we compared the taxa composition estimated using three different methods based on (1) the read counts of the metagenomic data; (2) the metaproteomics result searching against the metagenomic database; (3) the metaproteomics result by MetaLab-MAG (Figure 4f). We found that the relative abundances of taxa were very similar between the two metaproteomic workflows. In the previous part, different trends in the relative abundance of taxa were observed between these two types of results, mainly because the taxonomic annotation was performed based on the NCBI but not the UHGG database. The taxa information obtained by MetaLab-MAG and searching the metagenomic database were quite consistent, which demonstrated the credibility of the quantitative information provided by MetaLab-MAG and the multiomics database workflow. Obvious differences were observed compared to the metagenomics results (read counts), which suggested combining metagenomics and metaproteomics methods will help the researchers get a better profile of the microbiota samples.

Through the analysis of these data sets from the real human gut microbiota samples, we found that although the multiomics database was decent for the metaproteomics analysis, MetaLab-MAG provided an alternative solution for the interpretation of human gut microbiota samples, which showed similar or better performance. This will greatly expand the applicability of metaproteomics in the analysis of human gut microbiota samples, in particular for samples without matched multiomics data.

Application of Isobaric Quantitative Analysis of Human Gut Microbiota Samples

The most prominent advantage of isobaric labeling strategies is the multiplexing capability, enabling the relative quantification of more than 10 samples in a single MS run. This technique will greatly reduce the MS running time in high-throughput experiments such as drug screening and clinical sample analysis. The commonly used isobaric labeling methods including tandem mass tag[39] (TMT) and isobaric tags for relative and absolute quantification[40], (iTRAQ) are all supported in MetaLab-MAG. Here we tested three TMT11plex labeled human gut microbiome data sets (**Data S12−S14**). The MS/MS identification rates exceeded 50% in all three data sets

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

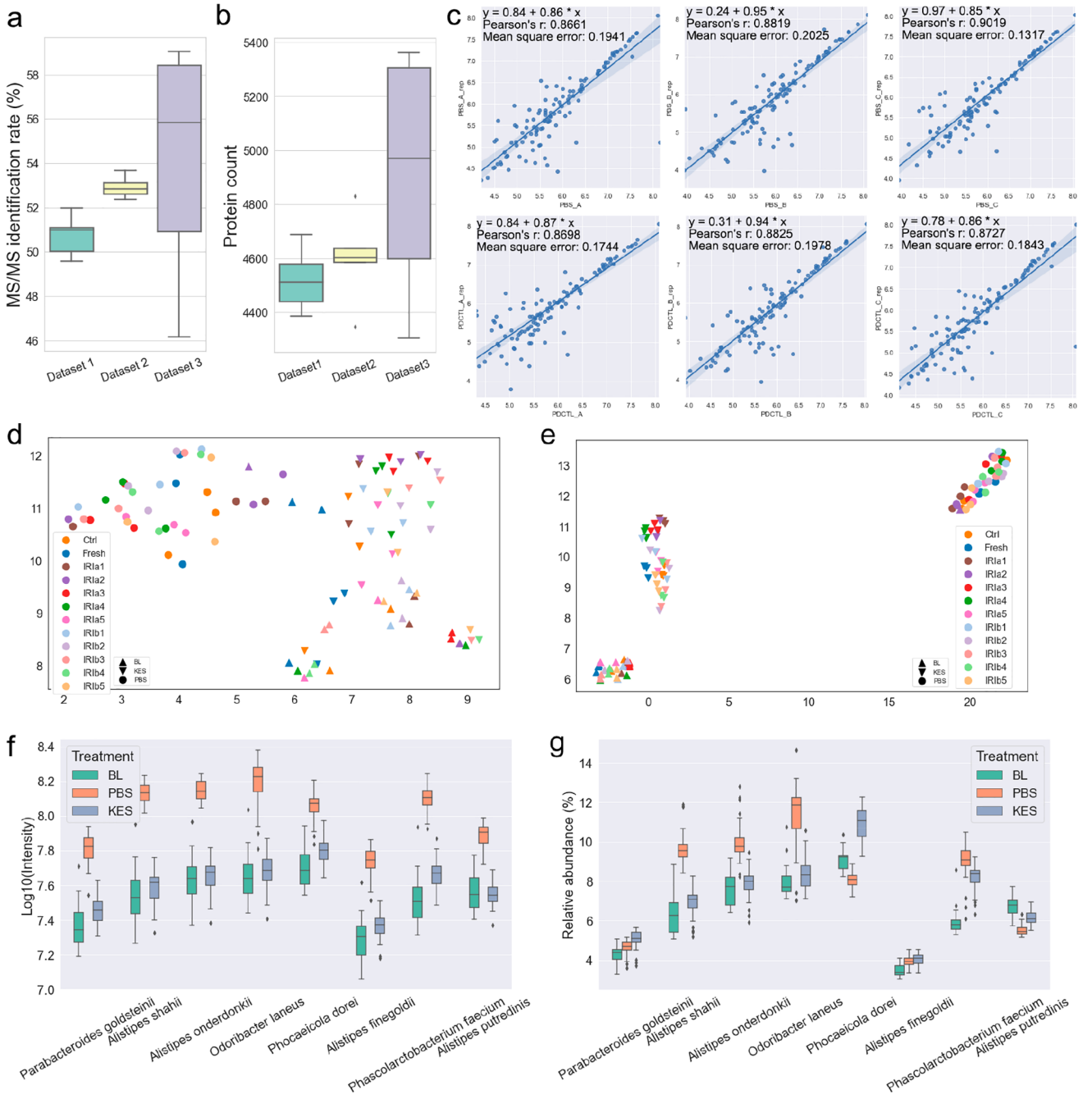


Figure 5. Qualitative and quantitative results of isobaric labeled data sets obtained by MetaLab-MAG. (a) The MS/MS identification rates of the three data sets. (b) The identified protein counts of the three data sets. (c) The correlations of the abundances of genomes between replicated samples from different experiments. The values on axis X and Y are log10(intensity) of the genomes. (d, e) UMAPs of the samples using the genome intensities (d) and protein intensities (e) show a clear trend that samples under the same treatment are clustered together. BL, KES, and PBS: the buffer solutions. (f) The abundances (represented by the log10(intensity)) and (g) relative abundances of eight high-abundance species (relative abundance >2%) in different buffer-treated samples.

(Figure 5a). Generally, over 4,500 protein groups could be identified from a single raw file (Figure 5b).

Another advantage of isobaric labeling methods was higher accuracy because samples in different channels were injected into MS analysis together, which greatly reduced the variability from different batch injections. However, if the number of samples exceeded the channels of the isobaric labeling reagent, multiple experiments were still required to quantify all the samples, which can introduce variability. To solve this problem, we took two measures. First, the MS1 intensity was utilized for the quantification and the normalization of the MS1 intensity was adopted. The intensities of the reporter ions were used to determine the relative abundances of the peptides from different channels, and the abundance of each peptide was calculated as the corresponding normalized MS1 intensity multiplied by the proportion of the reporter ion's intensity. Second, a reference channel with the same sample in every experiment was required. According to the quantitative information on the reference channel, all the channels in different experiments could be aligned and normalized. To assess the accuracy of the

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

quantitative results, we compared the calculated genome abundances between the replicate experiments from different raw files. After the normalization and alignment, good correlations between the replicate experiments were observed (Figure 5c). In particular, we noticed that the higher the abundance of the genomes, the smaller the deviation observed. This illustrated that the quantitative information on the high-abundance species was credible, even though the comparison was taken across different MS runs.

An isobaric labeling strategy was suitable for large-scale quantitative analysis. For example, in **Data 3** we tested the impacts of live microbiota frozen conditions and treatment of kestose on ex vivo cultured human gut microbiota. More details can be found in our previous study.[28] We built two UMAPs based on the quantitative information on the genomes and proteins, respectively (Figure 5d,e). It was observed that all the samples were clustered based on the culturing conditions, which was in conformity with our expectations and reflected the confidence of the quantitative results. It was also observed that the samples were better distinguished at the protein level than at the genome level. Based on the quantitative information in all the peptides, proteins, genomes/taxa, and function levels, we can interpret the data from various angles. Figure 5f showed the total intensities of eight high-abundance taxa (relative abundance above 2%) in different groups. All of them got higher abundances in PBS-treated samples. However, different trends were observed in the relative abundances (Figure 5g). The relative abundances of *Phocaeicola dorei* and *Alistipes putredinis* were lower than in BL and KES.

In this part, we analyzed multiple metaproteomic data sets with isobaric labeling quantitative methods. The performance of MetaLab-MAG was quite stable and reliable. Generally, 4,500 to 5,000 protein groups could be identified with a ~90 min MS run at the MS/MS identification rate higher than 50%. The accuracy of the quantitative results was verified by comparing the replicated samples from different labeling experiments. The information on the abundance was readily available on various levels including proteins, genomes/taxa, and functions. Investigating the responses to different conditions was straightforward based on the information provided by MetaLab-MAG. We believed that MetaLab-MAG could be a helpful tool for researchers aiming to study the changes of the human gut microbiota samples, not only on the composition level but also the active functions.

DISCUSSION

In metaproteomics data analysis, customized gene catalog databases from the metagenomics/metatranscriptomics sequencing of individual samples are generally perceived as the most reliable databases for peptide identification from the same microbiota samples. However, in the majority of projects, the sequencing information is not available, and instead searches against a generic gene catalog database would be performed. The drawback is that the generic gene catalogs contain numerous bacterial genomes, the vast majority of which are not present in a specific sample. Usually, the size of the generic database is huge, and identifying peptides against the database suffers from a long processing time and low identification rate.

MetaLab integrates the iterative search strategy and enables the characterization of the microbiome from the public gene catalog. However, the taxonomic and functional information on the gene catalog is annotated based on the gene. The relationship between different genes/proteins is nonexistent. By contrast, in this work, we develop MetaLab-MAG, a specialized metaproteomic data analysis tool, using publically available MAGs databases for peptide/protein identification. A significant advantage for metaproteomics in using the MAGs database is that the genomes are constructed with additional information such as which genes are likely from the same species. The sequencing read-based protein database from metagenomics and/or metatranscriptomics is not required. An efficient refined database generation strategy, namely high-abundance protein database search,[23] is adopted in MetaLab-MAG. Compared to the conventional iterative searching method, searching the HAP database is more efficient and accurate, yielding more identifications in less time. Moreover, the performance is comparable to or even better than searching the corresponding multiomics database. The MS/MS identification rates are similar, and the identified peptides shared significant common parts. The taxonomic information is obtained at the genome level. One current drawback of the MAGs database is the more limited taxa identification. However, with the continuous improvement of the MAGs database in the public data repository (such as MGnify), the taxonomic identification results will be more reliable.

Sample-specific metagenomics/metatranscriptomics can be used for metaproteomic data analysis; however, the reality is that they are not routinely performed on all samples and the lack of corresponding information should not hinder the application of metaproteomics. That is the motivation for the development of MetaLab-MAG. MetaLab-MAG is readily accessible to researchers with limited bioinformatics backgrounds. The researchers only need to provide their MS raw files. The MAGs databases for peptide/protein identification, taxonomy analysis, and functional annotation can be downloaded in MetaLab-MAG with a click. We anticipate that with the improvement of MGnify resources, other types of microbiomes would be supported for analysis. Rich information is generated automatically, including all data tables at the peptide, protein, genome, taxonomy, and function levels. The web-based report contains many useful charts and is usable for research manuscripts. The continuous development of MetaLab-MAG will include in the future more statistical functions to help the research community better understand their data sets. We believe that MetaLab-MAG can help researchers from various fields interested in using metaproteomics to investigate microbiomes.

ASSOCIATED CONTENT

Data Availability Statement

The mass spectrometry proteomics data from our lab and the result tables including the peptide, protein, genome, and function annotation lists (**Data S1−14**) have been deposited to the ProteomeXchange Consortium via the PRIDE[41] partner repository with the dataset identifier PXD037839. The other dataset identifiers are PXD005590 (*E. coli* spiked in human samples) and PXD023217 (SIHUMIx and fecal samples).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00554.

Supplementary Table 1: The running times, MS/MS identification rates, and the proportion of PSMs that are correctly identified in species levels of the data sets from the six sole species samples (XLSX)

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398

Supplementary Table 2: The MS/MS identification rates and the relative abundance determined by the MetaLab-MAG of the *E. coli* spike-in samples (XLSX)

Supplementary Table 3: The corresponding genomes in the UHGG MAG database from the same genera of the eight microbes in SIHUMIx samples (XLSX)

Supplementary Table 4: The MS/MS identification rates and the number of PSMs obtained by the three strategies: MetaLab-MAG, searching against the Uniprot database (DB1UNIPROT) and searching against the sequencing database (DB2MG) from the SIHUMIx samples (XLSX)

Supplementary Table 5: The MS/MS identification rates and the number of identified peptides, protein, and genomes from the isobaric labeled human gut microbiome data sets (XLSX)

AUTHOR INFORMATION

Corresponding Author

Daniel Figeys − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada;* orcid.org/0000-0002-5373-7546; Email: dfigeys@uottawa.ca

Authors

Kai Cheng − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada;* orcid.org/0000-0002-6045-0244

Zhibin Ning − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Leyuan Li − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Xu Zhang − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada;* orcid.org/0000-0003-2406-9478

Joeselle M. Serrana − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Janice Mayne − *School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00554

Author Contributions

D.F., Z.N., and K.C. designed the study. K.C. developed the software and performed the data analysis. J.M. collected patient samples. L.L. and X.Z. processed the samples and performed the experiments. J.S. processed the sequencing data. K.C., D.F., and X.Z. wrote the paper. All authors participated in the data interpretation, discussion and edits of the paper.

Notes

The authors declare the following competing financial interest(s): D.F. is a co-founder of MedBiome, a microbiome therapeutic company. The other authors declare no competing interests.

ACKNOWLEDGMENTS

This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-156 and OGI-149), the Natural Sciences and Engineering Research Council of Canada (NSERC, grant no. 210034), and the Ontario Ministry of Economic Development and Innovation (ORF-DIG-14405 and project 13440).

REFERENCES

(1) Qin, J.; Li, Y.; Cai, Z.; et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **2012**, *490*, 55−60.

(2) Mottawea, W.; Chiang, C. K.; Mühlbauer, M.; et al. Altered intestinal microbiota−host mitochondria crosstalk in new onset Crohn's disease. *Nat. Commun.* **2016**, *7*, 13419.

(3) Zhang, X.; Deeke, S. A.; Ning, Z.; et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **2018**, *9*, 2873.

(4) Witkowski, M.; Weeks, T. L.; Hazen, S. L. Gut Microbiota and Cardiovascular Disease. *Circ. Res.* **2020**, *127*, 553−570.

(5) Knight, R.; Vrbanac, A.; Taylor, B. C.; et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol* **2018**, *16*, 410−422.

(6) Almeida, A.; Mitchell, A. L.; Boland, M.; et al. A new genomic blueprint of the human gut microbiota. *Nature* **2019**, *568*, 499−504.

(7) Nayfach, S.; Shi, Z. J.; Seshadri, R.; Pollard, K. S.; Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **2019**, *568*, 505−510.

(8) Li, L.; Ning, Z.; Zhang, X.; et al. RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. *Microbiome* **2020**, *8*, 33.

(9) Zhang, X.; Ning, Z.; Mayne, J.; et al. Widespread protein lysine acetylation in gut microbiome and its alterations in patients with Crohn's disease. *Nat. Commun.* **2020**, *11*, 4120.

(10) Mesuere, B.; et al. Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. *J. Proteome Res.* **2012**, *11*, 5773−5780.

(11) Li, J.; Jia, H.; Cai, X.; et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **2014**, *32*, 834−841.

(12) Cheng, K.; Ning, Z.; Zhang, X.; et al. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **2017**, *5*, 157.

(13) Cheng, K.; et al. MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics. *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 1473−1482.

(14) Zhang, X.; Ning, Z.; Mayne, J.; et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **2016**, *4*, 31.

(15) Jagtap, P. D.; et al. Metaproteomic analysis using the galaxy framework. *Proteomics* **2015**, *15*, 3553−3565.

(16) Muth, T.; et al. The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *J. Proteome Res.* **2015**, *14*, 1557−1565.

(17) Van Den Bossche, T.; Kunath, B. J.; Schallert, K.; et al. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat. Commun.* **2021**, *12*, 7305.

(18) Vaudel, M.; et al. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11*, 996−999.

(19) Vaudel, M.; Burkhart, J.; Zahedi, R.; et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33*, 22−24.

(20) Mitchell, A. L.; et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **2019**, *48*, D570−D578.

(21) Stewart, R. D.; Auffret, M. D.; Warr, A.; et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **2019**, *37*, 953−961.

(22) Almeida, A.; Nayfach, S.; Boland, M.; et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **2021**, *39*, 105−114.

(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*, 80.
(24) Chi, H.; Liu, C.; Yang, H.; et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **2018**, *36*, 1059−1061.
(25) Chick, J.; Kolippakkam, D.; Nusinow, D.; et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33*, 743−749.
(26) Kong, A.; Leprevost, F.; Avtonomov, D.; et al. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry−based proteomics. *Nat. Methods* **2017**, *14*, 513−520.
(27) Li, L.; et al. Revealing Protein-Level Functional Redundancy in the Human Gut Microbiome using Ultra-deep Metaproteomics *bioRxiv* 2021.07.15.452564; **2021** DOI: 10.1101/2021.07.15.452564.
(28) Zhang, X.; et al. An economic and robust TMT labeling approach for high throughput proteomic and metaproteomic analysis *bioRxiv* 2022.07.30.502163; **2022** DOI: 10.1101/2022.07.30.502163.
(29) Kultima, J. R.; et al. MOCAT: a metagenomics assembly and gene prediction toolkit *PLoS One* **2012**, *7*, e47656.
(30) Li, R.; et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966−1967.
(31) 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.
(32) Hyatt, D.; et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010**, *11*, 119.
(33) Buchfink, B.; Xie, C.; Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59−60.
(34) Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094−3100.
(35) Liao, Y.; Smyth, G. K.; Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923−930.
(36) Shen, X.; et al. An IonStar Experimental Strategy for MS1 Ion Current-Based Quantification Using Ultrahigh-Field Orbitrap: Reproducible, In-Depth, and Accurate Protein Measurement in Large Cohorts. *J. Proteome Res.* **2017**, *16*, 2445−2456.
(37) Schape, S. S.; et al. The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors *Microorganisms* **2019**, *7*, 641
(38) Buchfink, B.; Reuter, K.; Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366−368.
(39) Thompson, A.; et al. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75*, 1895−1904.
(40) Ross, P. L.; et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.
(41) Perez-Riverol, Y.; Bai, J.; Bandla, C.; Hewapathirana, S.; García-Seisdedos, D.; Kamatchinathan, S.; Kundu, D.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. The PRIDE database resources in 2022: A Hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50*, D543−D552.

https://doi.org/10.1021/acs.jproteome.2c00554
J. Proteome Res. 2023, 22, 387−398